3 Models

3.1 Vector Random Variables

a **Random Variables and Samples.** In (univariate) statistics, we begin with observations that, up to "fluctuations" should give the same result: The length of a sepal leaves of an Iris setosa, measured on 50 plants, gives 50 "in principle equal" but in detail different values x_i. To describe this situation, we turn to a probability model. The individual values are "realizations" of a **random vairable** X with a distribution. The length of the sepal leaves is the random variable, which has a certain value for each plant. The expected value (or another location parameter) of the distribution of the random variable is the "basic value" from which the measured values "randomly deviate". To get further – into inferential statistics – we have to assign each observed entity (plant) i its own random variable X_i. In the simplest case of a random sample, all X_i have the same distribution – the "distribution of X" and are stochastically independent. With such a model we can then determine the distribution of estimators and test statistics.

b **Random Vector.** In the preceding chapter we have considered n observations of multiple variables $X^{(j)}$, $j = 1, 2, ..., m$, and written them as vectors $\underline{x}_i$. Instead of an individual random variable X, as just recalled, we now need the model of a **random vector**

$$\underline{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(m)} \end{bmatrix} ,$$

characterized by a distribution, namely the **joint distribution of the random variables** $X^{(1)}, X^{(2)}, ..., X^{(m)}$. In the next step we consider a **sample of random vectors** $\underline{X}_i$, which all display the same (joint) distribution and are independent of each other. Finally, we can summarize all of the data from the sample as a random **data matrix**

$$\boldsymbol{X} = \begin{bmatrix} \underline{X}_1^T \\ \ldots \\ \underline{X}_n^T \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & X_1^{(2)} & \ldots & X_1^{(m)} \\ X_2^{(1)} & X_2^{(2)} & \ldots & X_2^{(m)} \\ \vdots & \vdots & & \vdots \\ X_n^{(1)} & X_n^{(2)} & \ldots & X_n^{(m)} \end{bmatrix} .$$

Careful! The random vectors $\underline{X}_i$ are written as column vectors (like the $\underline{x}_i$), even though each corresponds to a row of the data matrix $\boldsymbol{X}$!

c **Expected Value and Variance.** We have already treated the mean vector and covariance matrix for a matrix of data values (2.5.d, 2.5.f). The transition to the corresponding theoretical parameters occurs exactly like in univariate statistics: All means $\frac{1}{n}\sum\ldots$ and "near-means" $\frac{1}{n-1}\sum\ldots$ are replaced by the expected value.

Needless to say: The expected value of a random vector (or a random matrix) is simply the vector of the expected values of the elements, which are usual random variables,

$$\underline{\mu} = \mathcal{E}\langle \underline{X}\rangle \ = \begin{bmatrix} \mathcal{E}\langle X^{(1)}\rangle \\ \mathcal{E}\langle X^{(2)}\rangle \\ \vdots \\ \mathcal{E}\langle X^{(m)}\rangle \end{bmatrix} .$$

From the empirical variances $\widehat{\text{var}}$ and covariances $\widehat{\text{cov}}$, we get the "theoretical" values var und cov,

$$\boldsymbol{\Sigma} = \text{var}\langle \underline{X}\rangle = \begin{bmatrix} \text{var}\langle X^{(1)}\rangle & \text{cov}\langle X^{(1)}, X^{(2)}\rangle & \ldots & \text{cov}\langle X^{(1)}, X^{(m)}\rangle \\ \text{cov}\langle X^{(2)}, X^{(1)}\rangle & \text{var}\langle X^{(2)}\rangle & \ldots & \text{cov}\langle X^{(2)}, X^{(m)}\rangle \\ \vdots & \vdots & \ldots & \vdots \\ \text{cov}\langle X^{(m)}, X^{(1)}\rangle & \text{cov}\langle X^{(m)}, X^{(2)}\rangle & \ldots & \text{var}\langle X^{(m)}\rangle \end{bmatrix} .$$

d **Covariance Matrix as Expected Value.** For simple random variables, $\text{var}\langle X\rangle = \mathcal{E}\left\langle (X-\mu)^2\right\rangle = \mathcal{E}\langle X^2\rangle - \mu^2$. For a random vector, there is a corresponding result:

$$\boldsymbol{\Sigma} = \text{var}\langle \underline{X}\rangle = \mathcal{E}\left\langle \left(\underline{X}-\underline{\mu}\right)\left(\underline{X}-\underline{\mu}\right)^T\right\rangle = \mathcal{E}\langle \underline{X}\underline{X}^T\rangle - \underline{\mu}\underline{\mu}^T .$$

$\underline{X}-\underline{\mu}$ is a column vector of length m and therefore $(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^T$ is an $m \times m$ matrix! For the proof of this formula, we simply choose an arbitrary element $\boldsymbol{\Sigma}_{jk}$ and determine that the corresponding result equals what is known from the theory for simple random variables.

e **Linear Transformations.** All of the formulas that were derived in the last chapter for means, empirical variances, and covariance matrices of linear combinations, projections, and linearly transformed vectors of variables, are also true for the corresponding theoretical values. The most important are

$$\mathcal{E}\langle \underline{a} + \boldsymbol{B}\underline{X}\rangle = \underline{a} + \boldsymbol{B}\,\mathcal{E}\langle \underline{X}\rangle \ , \qquad \text{var}\langle \underline{a} + \boldsymbol{B}\underline{X}\rangle = \boldsymbol{B}\,\text{var}\langle \underline{X}\rangle\,\boldsymbol{B}^T .$$

f **Sums of Independent Random Vectors.** If we add two independent random vectors $\underline{X}_1$ and $\underline{X}_2$, then the expected value and variance matrix add as in univariate statistics,

$$\mathcal{E}\langle \underline{X}_1 + \underline{X}_2\rangle = \mathcal{E}\langle \underline{X}_1\rangle + \mathcal{E}\langle \underline{X}_2\rangle \ , \quad \text{var}\langle \underline{X}_1 + \underline{X}_2\rangle = \text{var}\langle \underline{X}_1\rangle + \text{var}\langle \underline{X}_2\rangle .$$

* For proof, we can combine the two vectors to a single one, $\underline{X}_{12}$, twice as long, and apply the rules of linear transformation on $\boldsymbol{B}\underline{X}_{12}$ with $\boldsymbol{B} = [\boldsymbol{I} \quad \boldsymbol{I}]$.

3.2 The Multidimensional Normal Distribution

a **Multidimensional distribution.** The distribution of a random variable is determined by the cumulative **distribution function** $F\langle x\rangle = P\langle X \leq x\rangle$. For a multidimensional distribution the distribution function is analogously defined; $\underline{X} \leq \underline{x}$ should mean that for all variables, $X^{(j)} \leq x^{(j)}$. So,

$$F\langle \underline{x}\rangle = P\Big\langle X^{(1)} \leq x^{(1)}, X^{(2)} \leq x^{(2)}, ..., X^{(m)} \leq x^{(m)}\Big\rangle \; .$$

As in the univariate case, continuous distributions also have a **density** $f\langle \underline{x}\rangle$, which represents the derivative of the cumulative distributionfunction with respect to $\underline{x}$ – in the sense of the partial derivatives with respect to all components. From this, we can get the probabilities for events via integration.

An event which is determined via the value of the random vector, is a set $\mathcal{A}$ in the m dimensional space; the event "occurs" if $\underline{X} \in \mathcal{A}$ is true or, in words, if the realization of the random vector gives a point in $\mathcal{A}$. The probability of the event is obtained from the density as $P\langle \mathcal{A}\rangle = \int_{\underline{u}\in\mathcal{A}} f\langle \underline{u}\rangle \, du^{(1)}...du^{(m)}$. In particular, for the distribution function

$$\begin{aligned} F\langle \underline{x}\rangle &= P\langle X^{(1)} \leq x^{(1)}, ..., X^{(m)} \leq x^{(m)}\rangle \\ &= \int_{u^{(1)}\leq x^{(1)},...,u^{(m)}\leq x^{(m)}} f\langle \underline{u}\rangle du^{(1)}...du^{(m)} \; . \end{aligned}$$

b **Multidimensional Standard normal distribution.** This all seems rather abstract.

One of the simplest multidimensional distributions is the m dimensional standard normal distribution, which we denote as Φ_m. It is characterized by the fact that the components are independent and standard normal distributed,

$$\underline{Z} \sim \Phi_m \quad \Longleftrightarrow \quad Z^{(j)} \sim \Phi_1 \; , \quad \text{independent} \; .$$

The density of this distribution can be calculated using the fact that the densities of independent random variables multiply to the density of the joint distribution;

$$f\langle \underline{z}\rangle = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\Big\langle -z^{(j)2}/2\Big\rangle = (2\pi)^{-m/2} \exp\Big\langle -\|z\|^2/2\Big\rangle = \widetilde{f}\Big\langle \|z\|^2\Big\rangle \; .$$

There density is therefore only a function of the length $\|z\|$ of the vector $\underline{z}$. Its contours for $m=2$ form circles, in higher dimensions they form (hyper) spherical surfaces (Figur 3.2.g).

Naturally, this distribution is seldom a reasonable model for real data; it allows no dependency between variables and therefore describes uninteresting situations.

c **Distribution of Linear Combinations.** We need a model that can describe correlated variables. First a preliminary consideration: From the theory of the usual normal distribution it is known that every linear combination $b_1 Z^{(1)} + b_2 Z^{(2)} + ... + b_m Z^{(m)}$ of normally distributed random variables n $Z^{(j)}$ is again normally distributed. From earlier, we know the expected value and variance (3.1.e or 2.6.b and 3.1.f or 2.6.c): $b_1\, \mathcal{E}\langle Z^{(1)}\rangle + b_2\, \mathcal{E}\langle Z^{(2)}\rangle + ... + b_m\, \mathcal{E}\langle Z^{(m)}\rangle$ and $b_1^2\, \mathrm{var}\langle Z^{(1)}\rangle + b_2^2\, \mathrm{var}\langle Z^{(2)}\rangle + ... + b_m^2\, \mathrm{var}\langle Z^{(m)}\rangle$. It is also true for $X = \underline{b}^T \underline{Z}$

$$X = \underline{b}^T \underline{Z} \sim \mathcal{N}\langle 0, \textstyle\sum_j b_j^2\rangle = \mathcal{N}\langle 0, \|\underline{b}\|^2\rangle \; .$$

In 2.6.e we have considered multiple linear combinations jointly and introduced the concept of the linear transformation. We now consider a linear transformation of a standard normally distributed random vector $\underline{Z}$, $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$! We have

$$\mathcal{E}\langle \underline{X}\rangle = \underline{\mu}\;, \qquad \mathrm{var}\langle \underline{X}\rangle = \boldsymbol{B}\,\boldsymbol{B}^T \; .$$

d **Multivariate normal distribution.** The family of the m dimensional normal distributions is the family of the distributions of all random vectors $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$, where $\underline{Z}$ is m dimensional standard normally distributed and $\boldsymbol{B}$ is square (and $\underline{\mu}$ some m dimensional vector).

It is straightforward to interpret the values $\underline{\mu}$ and $\boldsymbol{B}$ as parameters of this distribution family. However, it turns out that $\boldsymbol{B}$ is not appropriate for this: If $\boldsymbol{B}$ is orthogonal – for example a rotation in the two dimensional case – then $\underline{Z}$ and $\boldsymbol{B}\underline{Z}$ have the same distribution. We can not use different parameter values ($\boldsymbol{B}$ and $\boldsymbol{I}$) that identify the same distribution, each distribution of a family should have a unique set of parameters. This requirement is called **identifiability**.

Besides the expected value $\underline{\mu}$, the appropriate parameter is the covariance matrix $\boldsymbol{\Sigma}$. If two matrices $\boldsymbol{B}$ and $\boldsymbol{B}'$ lead to the same $\boldsymbol{\Sigma}$, in other words if $\boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{B}'\,\boldsymbol{B}'^T$, then (and only then) are the distributions of the corresponding random vectors $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$ and $\underline{X} = \underline{\mu} + \boldsymbol{B}'\underline{Z}$ the same.

e* For the case of a regular covariance matrix $\boldsymbol{\Sigma}$ this is not difficult to show: We let $\boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{B}'\boldsymbol{B}'^T = \boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}$ is invertible, $\boldsymbol{B}$ must also be (as is probed in linear algebra). We multiply the last equation with $\boldsymbol{B}^{-1}$ from the left and $(\boldsymbol{B}^{-1})^T$ from the right and get $\boldsymbol{I} = \boldsymbol{B}^{-1}\boldsymbol{B}'\boldsymbol{B}'^T(\boldsymbol{B}^{-1})^T = (\boldsymbol{B}^{-1}\boldsymbol{B}')(\boldsymbol{B}^{-1}\boldsymbol{B}')^T$. So, $\boldsymbol{Q} = \boldsymbol{B}^{-1}\boldsymbol{B}'$ is an orthogonal matrix. It is true that $\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{Q}$. We look at $\underline{X}' = \underline{\mu} + \boldsymbol{B}'\underline{Z} = \underline{\mu} + \boldsymbol{B}\,\boldsymbol{Q}\underline{Z}$. Since the distribution of $\underline{Z}$ and $\underline{Z}' = \boldsymbol{Q}\underline{Z}$ is the same – namely the standard normal distribution – $\underline{X}' = \underline{\mu} + \boldsymbol{B}'\underline{Z}$ has the same distribution as $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$.

f The **multivariate normal distribution** with expected value $\underline{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, written as $\mathcal{N}_m\langle \underline{\mu}, \boldsymbol{\Sigma}\rangle$, is the distribution of $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$, where $\underline{Z} \sim \Phi_m$ and $\boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{\Sigma}$.

As parameter values for $\boldsymbol{\Sigma}$, all $m \times m$ matrices that are symmetric and “positive semi-definite” are suitable (2.6.m). For any such matrix, martrices $\boldsymbol{B}$ can be found for which $\boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{\Sigma}$ (2.6.m).

g **Density.** If $\boldsymbol{\Sigma}$ is not singular, the multivariate normal distribution has the density

$$f\langle \underline{x} \rangle = \tfrac{1}{c} \cdot \exp\Big\langle -(\underline{x} - \underline{\mu})^T \boldsymbol{\Sigma}^{-1} (\underline{x} - \underline{\mu})/2 \Big\rangle$$

with the normalization constant $c = (2\pi)^{m/2} \det\langle \boldsymbol{\Sigma} \rangle^{1/2}$. (det stands for "determinant.") If $\boldsymbol{\Sigma}$ is singular, then no density exists; the distribution is then concentrated on a subspace.

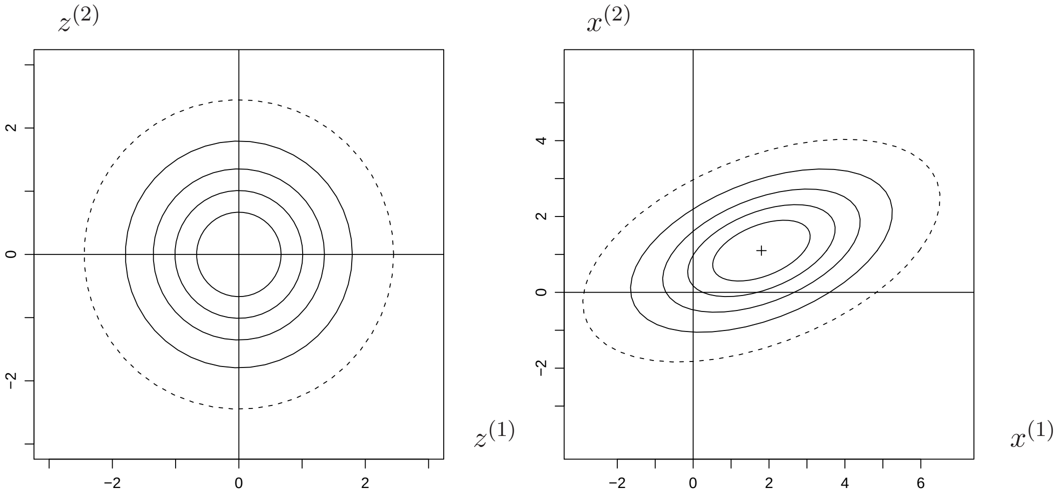

Figure 3.2.g: "Contour lines" of equal density for the standard normal distribution and a general normal distribution

The density is constant for $(\underline{x}-\underline{\mu})^T \boldsymbol{\Sigma}^{-1} (\underline{x}-\underline{\mu})$ (see Fig. 3.2.g). In two dimensions, this is the equation of an ellipse; in higher dimensions we speak of **ellipsoids**. Concentric ellipses (ellipsoids) arise for varying constants; the center is always the expected value vector $\underline{\mu}$.

h **Estimation of the Parameters.** Models should describe data. To show this with the iris flower example, we like to determine the normal distribution that is the best fit. In other words, we must estimate the parameters $\underline{\mu}$ and $\boldsymbol{\Sigma}$. It is obvious to use the mean vector $\widehat{\underline{\mu}} = \overline{\underline{X}}$ and the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}$ for these. Fig. 3.2.h shows the data for the first two variables for the Iris setosa flowers, together with the ellipses that represent the fitted normal distribution (see 3.2.m). The next chapter deals with the properties of these estimators.

i **Linear Transformation.** If a multivariate normally distributed random vector $\underline{X} \sim \mathcal{N}_m\langle \underline{\mu}, \boldsymbol{\Sigma} \rangle$ is linearly transformed to $\underline{Y} = \underline{a} + \boldsymbol{B}\underline{X}$, then the transformed vector is also multivariate normally distributed,

$$\underline{Y} \sim \mathcal{N}_m\langle \underline{a} + \boldsymbol{B}\underline{\mu}, \boldsymbol{B}\,\boldsymbol{\Sigma}\,\boldsymbol{B}^T \rangle \ .$$

Proof: $\underline{X}$ arises through a linear transformation from $\underline{Z}$. If we compose the two linear transformations, then it is clear that $\underline{Y}$ is another linearly transformed standard normally distributed random vector. If $\boldsymbol{B}$ is square, $\underline{X}$ corresponds to the definition of a multivariate normal distribution.

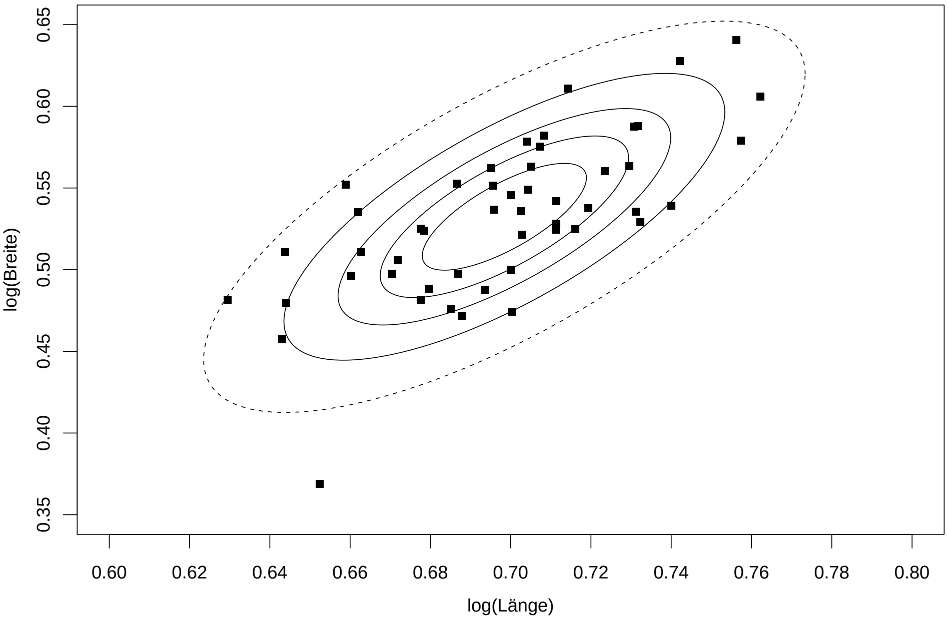

Figure 3.2.h: Data and fitted normal distribution in the iris flower example for the species Iris setosa

* If $\boldsymbol{B}$ has fewer rows than columns, the result $\underline{Y}$ is shorter than the input vector $\underline{X}$. We can then arbitrarily expand $\boldsymbol{B}$ to a square matrix and get an expanded result vector $\widetilde{\underline{Y}}$. The desired distribution is the "marginal distribution" of the first components and the result follows from 3.2.p. Finally, if $\boldsymbol{B}$ has fewer columns than rows, we supplement it with zeros to form a square matrix and get the result as for a square $\boldsymbol{B}$.

We already know the formulas for determining the parameters of the normal distribution of $\underline{Y}$ from those of $\underline{X}$ (2.6.i, 3.1.e).

j* **Characterization.** The following property "characterizes" the multivariate normal distribution:

If every linear combination $\underline{b}^T\underline{X}$, $\underline{b} \in \mathbb{R}^m$, is normally distributed (or possibly degenerate), then $\underline{X}$ is multivariate normally distributed.

k **Standardized Random Vector.** Using a linear transformation on a random vector $\underline{Z}$ with expected value $\underline{0}$ and covariance matrix $\boldsymbol{I}$, we have produced a vector $\underline{X}$, which has an expected value $\underline{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. Conversely, we will now produce a standardized random vector from $\underline{X}$, as we have already done for samples (2.6.m). For this, in the formulas mentioned there we have to substitute the mean vector $\underline{\overline{x}}$ by the expected value $\underline{\mu}$ and the the empirical $\widehat{\boldsymbol{\Sigma}}$ with the theoretical covariance matrix $\boldsymbol{\Sigma}$. We get

$$\underline{Z} = \boldsymbol{B}^{-1}(\underline{X} - \underline{\mu}) \quad \text{mit} \quad \boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{\Sigma}\ .$$

This random vector has expected value $\underline{0}$ and covariance matrix $\boldsymbol{I}$, even if $\underline{X}$ was not normally distributed. If the normal distribution is assumed, then clearly, $\underline{Z}$ is standard normally distributed.

l **Chi-square Distribution.** The χ^2 distribution with m degrees of freedom is *defined* as the distribution of the sum of m independent, squared standard normally distributed $Z^{(j)}$, $U = \sum_{j=1}^m Z^{(j)2}$, $Z^{(j)} \sim \mathcal{N}\langle 0, 1\rangle$, or

$$U = \sum\nolimits_{j=1}^{m} Z^{(j)2} = \|\underline{Z}\|^2 , \qquad \underline{Z} \sim \Phi_m .$$

It is thus the distribution of the squared length of a standard normally distributed random vector.

The χ^2_m distribution has the density

$$f_m\langle u\rangle = \frac{1}{2^{m/2}\Gamma\langle m/2\rangle} \cdot u^{m/2-1} e^{-u/2} .$$

(For the normalization constant, the so-called gamma function Γ is required.)

m **Mahalanobis Distance.** We combine the last two thoughts: Beginning with a multivariate random vector $\underline{X}$ we form the squared length of the corresponding *standardized* vector $\underline{Z}$,

$$d^2\langle \underline{X}, \underline{\mu}; \boldsymbol{\Sigma}\rangle = \|\underline{Z}\|^2 = \underline{Z}^T \underline{Z} = (\underline{X} - \underline{\mu})^T \boldsymbol{C}^T \boldsymbol{C} (\underline{X} - \underline{\mu}) = (\underline{X} - \underline{\mu})^T \boldsymbol{\Sigma}^{-1} (\underline{X} - \underline{\mu}) .$$

This value is named after an Indian statistician and called the squared "Mahalanobis distance" of $\underline{X}$ to its expected vector $\underline{\mu}$.

If $\underline{X}$ is normally distributed then d^2 has, as stated, a χ^2 distribution with m degrees of freedom.

n* **Drawing Contours of the Density.** The distance determines the probability density of the normal distribution, as we can see from 3.2.g; the vectors $\underline{x}$ that fulfill $d^2\langle \underline{x}, \underline{\mu}; \boldsymbol{\Sigma}\rangle$ = constant form a contour of equal density. To be able to determine them concretely for a drawing like in Figur 3.2.h, we must find for $\widehat{\boldsymbol{\Sigma}}$ an associated $\boldsymbol{B}$ matrix $\widehat{\boldsymbol{B}}$ – for example, with the Cholesky decomposition –, transform the points $\underline{z}_k = c \cdot [\cos\langle k\Delta\rangle, \sin\langle k\Delta\rangle]^T$ of a circle to $\underline{x}_k = \underline{\widehat{\mu}} + \underline{z}_k \widehat{\boldsymbol{B}}$ and draw them into the display. We determine the factor c so that the circle includes a desired probability π; so c^2 must be the π quantile of the χ^2 distribution with m degrees of freedom.

o **Q-Q Diagram.** If we want to check assumptions about the distribution for a simple random sample, we can compare a histogram with a fitted density curve or look at a quantile-quantile diagram (for example, see section 11.2 in Stahel (2002)). The joint distribution of two variables can best be evaluated by the simultaneous representation of the data and the contours of the model distribution, like in Figur 3.2.h. For higher dimensions, such graphical methods are not possible.

At least we can check one aspect of the joint distribution; the distribution of the Mahalanobis distances $d^2\langle \underline{X}_i, \underline{\widehat{\mu}}; \widehat{\boldsymbol{\Sigma}}\rangle$ should approximately be a χ^2 distribution with m degrees of freedom. We compare the values d, the roots of the d^2 values, with roots of the quantiles of the χ^2 distribution rather than comparing d^2 values with the χ^2 distribution. This has the advantage that the figure (for a small dimension number m) is not too badly dominated by the large values.

Fig. 3.2.o shows a good agreement of the data with the theoretical picture for the Iris setosa sepal leaves, except for an outlier, which was already visible in Figur 3.2.h.

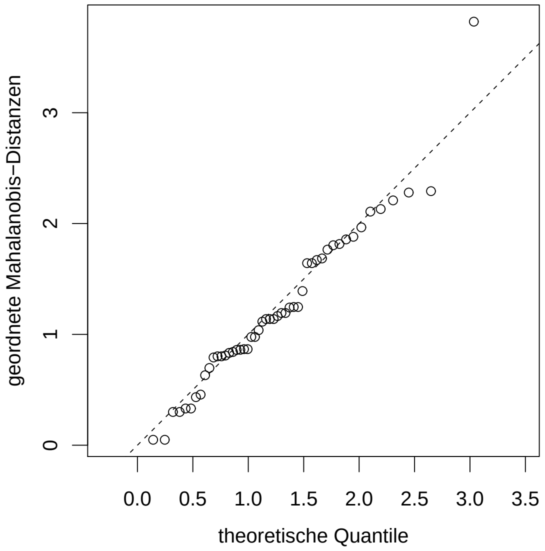

Figure 3.2.o: Q-Q diagram of the Mahalanobis distances for the log lengths and widths of the sepal leaves of the 50 Iris setosa plants

p **Marginal Distributions.** If the joint distribution of multiple random variables is given, then the distribution of the individual random variables, considered in isolation, is also determined. These distributions are called **marginal distributions**. (The expression is clear for discrete variables: The joint distribution of two discrete random variables can be given in a table. The "marginal sums" in such a table reflect the distributions of the two variables.)

For more than two random variables, we can look at "multidimensional margins": For three variables, we can be interested in the (joint) distribution of the first two, without considering the value of the third. This is the marginal distribution of $[X^{(1)}, X^{(2)}]^T$.

For simplicity, we assume that we are interested in the first p variables. (The general case is only more difficult in terms of notation.) We write the first p components of $\underline{X}$ as $\underline{X}^{[1:p]}$. We need (and will occasionally require later) a notation for **partitioned vectors and matrices:**

$$\underline{a} = \left[\begin{array}{c} \underline{a}^{[1:p]} \\ \underline{a}^{[(p+1):m]} \end{array} \right] \quad = \quad \left[\begin{array}{c} a^{(1)} \\ \dots \\ a^{(p)} \\ \hline a^{(p+1)} \\ \dots \\ a^{(m)} \end{array} \right]$$

$$\boldsymbol{C} = \begin{bmatrix} C^{[1:p]}_{[1:p]} & C^{[1:p]}_{[(p+1):m]} \\ C^{[(p+1):m]}_{[1:p]} & C^{[(p+1):m]}_{[(p+1):m]} \end{bmatrix} = \left[\begin{array}{ccc|ccc} C_{11} & ... & C_{1p} & C_{1,p+1} & ... & C_{1m} \\ & ... & & & ... & \\ C_{p1} & ... & C_{pp} & C_{p,p+1} & ... & C_{pm} \\ \hline C_{p+1,1} & ... & C_{p+1,p} & C_{p+1,p+1} & ... & C_{p+1,m} \\ & ... & & & ... & \\ C_{m1} & ... & C_{mp} & C_{m,p+1} & ... & C_{mm} \end{array}\right]$$

If the distribution of $\underline{X}$ has a density, then formally we get the density of the marginal distribution of $\underline{X}^{[1:p]}$ by integration over the "superfluous" variables,

$$f^{[1:p]}\left\langle \underline{x}^{[1:p]} \right\rangle = \int_{u^{(m)}} ... \int_{u^{(p+1)}} f\left\langle x^{(1)}, ..., x^{(p)}, u^{(p+1)}, ..., u^{(m)} \right\rangle du^{(p+1)}...du^{(m)} \ .$$

q The multidimensional normal distribution behaves as we might hope: Each marginal distribution is itself a normal distribution. The parameters are given by the corresponding element(s) of the expected value vector μ and the corresponding row(s) and columns(s) of $\boldsymbol{\Sigma}$. So,

$$\underline{X}^{[1:p]} \sim \mathcal{N}_p\left\langle \underline{\mu}^{[1:p]}, \boldsymbol{\Sigma}^{[1:p]}_{[1:p]} \right\rangle \ .$$

r* **Correlation and Independence.** Two uncorrelated random variables do not have to be independent. There are simple examples of this (cf. Stahel (2002), 3.2.i). One of them is a quadratic regression and consists of one random variable $X^{(1)}$ symmetric about 0, and a second random variable $X^{(2)} = (X^{(2)})^2 + E$ with a variable E, independent of $X^{(1)}$. For reasons of symmetry, we can easily see that the correlation of $X^{(1)}$ and $X^{(2)}$ is zero.

We might hope that zero correlation at least implies independence if both variables are normally distributed. Unfortunately, this is not true, as the following example (Th.3.2.8 in Flury (1997)) shows: Let $X_1 \sim \mathcal{N}\langle 0, 1\rangle$ and $X_2 = X_1$ with probability 0.5, $X_2 = -X_1$ otherwise. Then, $\text{cov}\langle X_1, X_2\rangle = 0$, but the joint distribution lies on the two "diagonals" – the lines through the zero point with slope 1 respektive -1 –, so X_1 and X_2 are not independent.

If the joint distribution of X_1 and X_2 is a two dimensional normal distribution then it is reasonable to conclude that: Zero correlation can only occor for independent random variables. We show this as follows: If $\boldsymbol{\Sigma}_{12} = 0$, then $\boldsymbol{\Sigma}^{-1} = \text{diag}\langle 1/\boldsymbol{\Sigma}_{11}, 1/\boldsymbol{\Sigma}_{22}\rangle$. If we replace this in the formula for the density, we see that this can be "factored", or written as the product $f\langle x_1, x_2\rangle = f_1\langle x_1\rangle \cdot f_2\langle x_2\rangle$. So, the two random variables are independent.

s **Other Distributions.** In univariate statistics, the normal distribution plays a crucial role, but for the distribution of data there are a number of other commonly used models: Log-normal, exponential, Weibull, and gamma distributions, and others.

In multivariate statistics, on one hand, there are many more possibilities for defining models, but on the other, there are barely any usable alternatives to the multivariate normal distribution. This is due to two difficulties: First, we are much more dependent on the simplifications that result from the properties of the normal distribution since in higher dimensions, the possibility of solving problems with "brute force" diminishes rapidly. Second, the many possibilities for joint distributions rapidly lead to an arbitrariness that makes it hard for any one model to prevail.

Still, one more distribution is worth mentioning: The **multivariate log-normal distribution** is defined, analogous to the univariate case, so that the vector of the log components has a normal distribution.

t* **Elliptical distributions.** Somewhat well-known are the so-called elliptical distributions. They arise in the way we introduced the multivariate normal distribution. In 3.2.f we simple replace the standard normal distribution with another distribution, whose density only depends on $\|\underline{z}\|$.

The big disadvantage of this model is that they can't describe independent variables; for the "standard distribution" with $\boldsymbol{\Sigma} = \boldsymbol{I}$ the variables are already not independent - even if uncorrelated!

3.3 Theoretical Results for Other Areas

a **Distribution of the Estimated Coefficients in Multiple Linear Regression.** In matrix notation, the model for multiple linear regression is

$$\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{E} \; .$$

With the terms known now, the assumptions $E_i \sim \mathcal{N}\langle 0, \sigma^2 \rangle$, independent, can be written as

$$\underline{E} \sim \mathcal{N}_n\langle \underline{0}, \sigma^2 \boldsymbol{I} \rangle \; .$$

The least squares estimator of the coefficient vector $\underline{\beta}$ was

$$\widehat{\underline{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \underline{Y} \; .$$

Now we use the model! This produces

$$\begin{aligned} \widehat{\underline{\beta}} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T(\boldsymbol{X}\underline{\beta} + \underline{E}) = \beta + \boldsymbol{C}\underline{E} \;\; \text{mit} \\ \boldsymbol{C} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \; . \end{aligned}$$

This is a linear transformation of the random vector $\underline{E}$. So, $\widehat{\underline{\beta}}$ is multivariate normally distributed. The expected value is $= \underline{\beta}$, da $\mathcal{E}\langle \underline{E} \rangle = \underline{0}$. The covariance matrix is

$$\begin{aligned} \text{var}\langle \widehat{\underline{\beta}} \rangle &= \boldsymbol{C}(\sigma^2 \boldsymbol{I})\boldsymbol{C}^T = \sigma^2 \cdot \boldsymbol{C}\boldsymbol{C}^T \\ &= \sigma^2 \cdot \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \cdot \boldsymbol{X} \left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)^T \\ &= \sigma^2 \cdot \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \; . \end{aligned}$$

An important result that can only be derived very tediously without random vectors and covariance matrices!

b* **Independence of Mean and Empirical Standard Deviation of a Random Variable.** In univariate statistics, the fact that the mean and the empirical standard deviation are stochastically independent if the data is normally distributed is useful. This can be easily proven with the previously introduced results.

It is true that $\underline{X} = [X_1, X_2, ..., X_n]^T \sim \mathcal{N}\langle \mu\underline{1}, \sigma^2 \boldsymbol{I} \rangle$. The mean equals $\underline{1}^T \underline{X}/n$ and therefore up to a factor $\sqrt{n}$ equals $\underline{q}_1^T \underline{X}$ with $\underline{q}_1 = \underline{1}/\sqrt{n}$. We write this in such a complicated way so that with $\underline{q}_1$ we get a vector of length 1.

Now we supplement the vector $\underline{q}_1^T$ with $n-1$ additional rows to form an orthogonal matrix $\boldsymbol{Q}$. The vector $\underline{Y} = \boldsymbol{Q}(\underline{X} - \mu\underline{1})$ has distribution $\underline{Y} \sim \mathcal{N}\langle \underline{0}, \sigma^2 \boldsymbol{I} \rangle$, since $\mathrm{var}\langle \underline{Y} \rangle = \boldsymbol{Q} \sigma^2 \boldsymbol{I} \boldsymbol{Q}^T = \sigma^2 \boldsymbol{I}$. So, the Y_k are independent of each other. Additionally, $\|\underline{Y}\| = \|\underline{X}\|$ and $Y_1 = \sqrt{n}\overline{X}$.

The empirical variance, multiplied with $n-1$, can be written as

$$\sum_i X_i^2 - n\overline{X}^2 = \|\underline{X}\|^2 - n\overline{X}^2 = \|\underline{Y}\|^2 - Y_1^2 = \sum_{k=2}^n Y_k^2$$

and therefore depends only on $Y_2, ..., Y_n$. It is therefore independent of $Y_1 = \sqrt{n}\overline{X}$, and therefore also of $\overline{X}$.