4 Statistics of Normally Distributed Data

4.1 One Sample

a **The Three Basic Questions of Inferential Statistics.** Inferential statistics form the bridge between the probability models that structure our thinking and the reality that we try to capture with data. If parametric models are used for describing the reality, then the three basic questions of inferential statistics go as follows:

1. For a (each) parameter, which value is **most plausible?** The answer leads to the **estimate** of the parameter.
2. Is **a certain value** for the parameter plausible? This question is answered with a statistical **test**.
3. **Which values** of the parameter are plausible at all? The set of all parameter values that are plausible (according to a certain test) usually forms, if only one parameter is considered, an interval, the **confidence interval**. In the case of multiple parameters a more general set arises, which is usually cohesive and is called a **confidence interval**.

b **Estimations** of the parameters $\underline{\mu}$ and $\mathbf{\Sigma}$ of the normal distribution were already introduced in descriptive statistics: the mean vector $\overline{\underline{X}}$ and the empirical covariance matrix $\boldsymbol{S}$. As in the univariate case, these are the most common estimates and, for multivariate normally distributed data, the optimal.

As in that case, they are still not robust. Robust estimations exist, but here is not the place to treat them.

So, let's get to the tests!

c **Test for the Expected Value.** In univariate statistics, there is a simple generic question, of whether a treatment causes a change, for example whether a sleep aid actually lengthens sleep.

If we now consider two or more variables of interest, for example blood pressure and pulse, then we can again ask whether a medicine causes a statistically detectable change. The change $\underline{X}$ is now a two dimensional variable, for which we can test the hypothesis "no change" as $\mathcal{E}\langle\underline{X}\rangle = \underline{0}$.

If we provide that the data has the normal distribution, it is true that $\underline{X}_i \sim \mathcal{N}_m\langle\underline{\mu}, \mathbf{\Sigma}\rangle$, and the **null hypothesis** $\underline{\mu} = \underline{0}$ should be tested.

d The obvious test statistic to start with is the estimator of the expected value $\mathcal{E}\langle \underline{X} \rangle$, so the **mean vector** $\overline{\underline{X}}$. In order to be able to judge whether it is "too large" and thus the null hypothesis should be rejected, we must know it's **distribution** under the null hypothesis. This is not difficult: If, as before, we write $\mathcal{E}\langle \underline{X}_i \rangle = \underline{\mu}$ and $\mathrm{var}\langle \underline{X}_i \rangle = \mathfrak{\Sigma}$, according to 3.1.f

$$\begin{aligned} \mathcal{E}\langle \overline{\underline{X}} \rangle &= \tfrac{1}{n} \sum_{i=1}^{n} \mathcal{E}\langle \underline{X}_i \rangle = \tfrac{1}{n}\, n\underline{\mu} = \underline{\mu} \\ \mathrm{var}\langle \overline{\underline{X}} \rangle &= (\tfrac{1}{n})^2 \sum_{i=1}^{n} \mathrm{var}\langle \underline{X}_i \rangle = \tfrac{1}{n^2}\, n\mathfrak{\Sigma} = \tfrac{1}{n}\mathfrak{\Sigma} \,. \end{aligned}$$

If we still require the normal distribution for the $\underline{X}_i$, then

$$\overline{\underline{X}} \sim \mathcal{N}_m\langle \underline{\mu}, \tfrac{1}{n}\mathfrak{\Sigma} \rangle \;,$$

in analogy to the univariate case.

e **Standardization of the test statistic.** The analogy also suggests to make the test statistic independent of the parameters via standardization. According to 3.1.e and 2.6.m we determine a matrix $\boldsymbol{B}$ with $\boldsymbol{B}\boldsymbol{B}^T = \mathfrak{\Sigma}$, with which the individual observations as well as their mean can be standardized,

$$\underline{Z}_i = \boldsymbol{B}^{-1}(\underline{X}_i - \underline{\mu}) \,, \qquad \overline{\underline{Z}} = \boldsymbol{B}^{-1}(\overline{\underline{X}} - \underline{\mu}) \,.$$

Then, $\sqrt{n}\,\overline{\underline{Z}}$ is multivariate standard normally distributed.

Now the question of which values of $\overline{\underline{Z}}$ should be considered "too large" has a natural answer: The acceptance region for $m = 2$ is a circle, for larger dimensions a (hyper)sphere. It is given by $\sqrt{n}\|\overline{\underline{Z}}\| \le c$, where c is chosen so that the acceptance region under the null hypothesis has probability 95%. According to 3.2.m this is true if c^2 is the 95% quantile of the chi-square distribution with m degrees of freedom.

f Let's wrap up! We are testing the null hypothesis $\underline{\mu} = \underline{0}$ or, more generally, a null hypothesis of the form $\underline{\mu} = \underline{\mu}_0$. We initially assume that the covariance matrix $\mathfrak{\Sigma}$ is known! Then, given the observations $\underline{X}_1, ..., \underline{X}_n$, we can calculate the associated $\overline{\underline{Z}}$ (with $\underline{\mu} = \underline{\mu}_0$) and get the acceptance region $\sqrt{n}\|\overline{\underline{Z}}\| \le c$.

For clarity, it is also useful to express this region in terms of $\overline{\underline{X}}$. This is

$$\begin{aligned} \|\overline{\underline{Z}}\|^2 &= \overline{\underline{Z}}^T\overline{\underline{Z}} = (\overline{\underline{X}} - \underline{\mu})^T(\boldsymbol{B}^{-1})^T\boldsymbol{B}^{-1}(\overline{\underline{X}} - \underline{\mu}) \\ &= (\overline{\underline{X}} - \underline{\mu})^T \left(\boldsymbol{B}\boldsymbol{B}^T\right)^{-1} (\overline{\underline{X}} - \underline{\mu}) = (\overline{\underline{X}} - \underline{\mu})^T\mathfrak{\Sigma}^{-1}(\overline{\underline{X}} - \underline{\mu}) \\ &= d^2\langle \overline{\underline{X}}, \underline{\mu}; \mathfrak{\Sigma} \rangle \le c^2/n \,. \end{aligned}$$

The mean vector $\overline{\underline{X}}$ should thus, measured with the Mahalanobis distance d, lie near the expected value $\underline{\mu}_0$ to be tested.

g **Studentization.** Since we usually don't know $\boldsymbol{\Sigma}$, we have to estimate it from the data. From doing this in the univariate case, we wind up with a Student's t-distribution from the normal distribution. In the multivariate case the test statistic

$$T^2 = n\, d^2\langle \overline{\underline{X}}, \underline{\mu}; \boldsymbol{S}\rangle = n\, (\overline{\underline{X}} - \underline{\mu})^T\, \boldsymbol{S}^{-1} (\overline{\underline{X}} - \underline{\mu})$$

is used and called **Hotelling's** T^2, since this man also found it's distribution: A multiple of T^2 shows an F-distribution,

$$\frac{n(n-m)}{(n-1)m}\, d^2\langle \overline{\underline{X}}, \underline{\mu}; \boldsymbol{S}\rangle \sim \mathcal{F}\langle m, n-m\rangle\ .$$

h ▷ As an **example** we again take up the four specimens of Iris setosa flowers (2.5.b). The null hypothesis shall be $\underline{\mu} = [5,\ 2.5]^T$. The Mahalanobis distance is

$$d^2 = [-0.175,\ 0.7] \begin{bmatrix} 0.0492 & 0.0300 \\ 0.0300 & 0.0467 \end{bmatrix}^{-1} \begin{bmatrix} -0.175 \\ 0.7 \end{bmatrix} = 23.6\ .$$

We compare $4(4-2)/((4-1)2)\cdot 23.6 = 31.4$ with an $\mathcal{F}\langle 2,2\rangle$-distribution, and thus get a p-value of 3.1%, so barely signficant. The two univariate t-tests give no significant deviations. (Admittedly, the null hypothesis was chosen so that this effect occurred, and it is understood that only in the most extreme emergency would we test such a null hypothesis with only four observations.) ◁

i **Confidence Region.** After we know how the test should be done, we can determine the set of all parameter vectors $\underline{\mu}^*$ for which the test does not rejected the null hypothesis $\underline{\mu} = \underline{\mu}^*$. This case occurs if

$$d^2\langle \overline{\underline{X}}, \underline{\mu}^*; \boldsymbol{S}\rangle \le \frac{m(n-1)}{n(n-m)}\, q$$

where q is the 95% quantile of the F-distribution with m and $n-m$ degrees of freedom. The vectors $\underline{\mu}^*$ that fulfill this inequality fill an **ellipse** for $m = 2$ (Fig. 4.1.i), in higher dimensions an "ellipsoid".

4.2 Statistics of the Covariance Matrix

a **Test for lack of correlations.** A hypothesis that is often interesting in practice is that of the independence of two variables. Under this hypothesis, the covariance and thus also the correlations are equal to zero, so $H_0:\ \boldsymbol{\Sigma}_{jk} = 0$ (mit $j \neq k$) or $H_0:\ \rho\langle X^{(j)}, X^{(k)}\rangle = 0$.

The estimated correlation serves as test statistic $\widehat{\rho}_{jk} = S_{jk} \Big/ \sqrt{S_{jj}\, S_{kk}}$ (siehe 2.4.b und 2.5.g). It can be shown that $T = \widehat{\rho}_{jk}\sqrt{n-2} \Big/ \sqrt{1-\widehat{\rho}_{jk}^2}$ is t-distributed with $n-2$ degrees of freedom if the two variables follow a bivariate normal distribution.

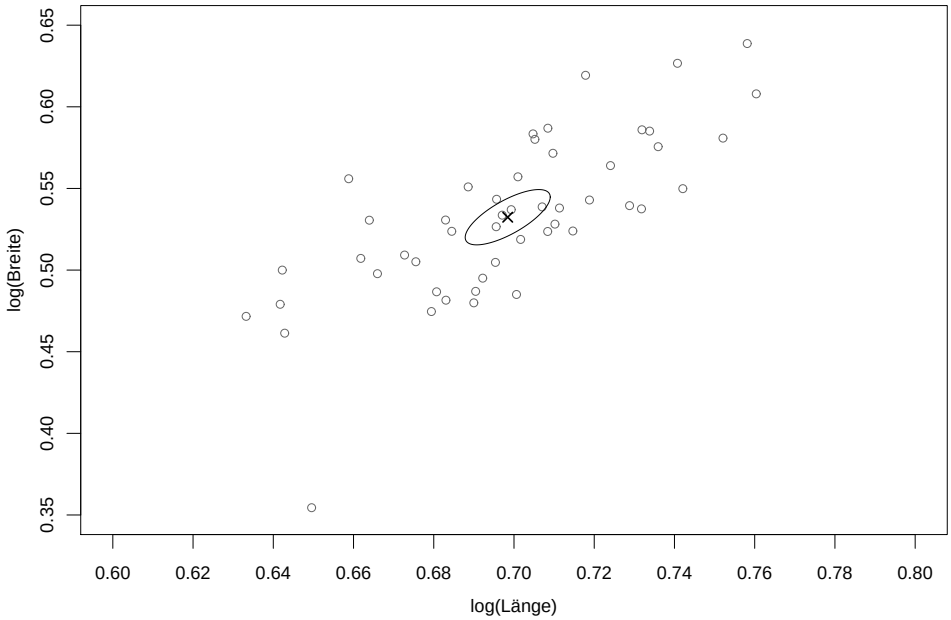

Figure 4.1.i: Confidence region for the expected value of the log length and width of the sepal leaves of Iris setosa plants

b **Confidence Interval for ρ.** The distribution of the estimated correlations is also known for correlations $\rho_{jk} \neq 0$. It can be stated more simply if we apply the "z-transformation" : We let

$$\zeta = \frac{1}{2} \log_e \left\langle \frac{1+\rho}{1-\rho} \right\rangle \ .$$

The corresponding transformed estimated variable is approximately normally distributed with constant variance,

$$\widehat{\zeta} \approx\sim \mathcal{N} \left\langle \zeta, 1/(n-3) \right\rangle \ .$$

This leads to the confidence interval $[\, \widehat{\zeta} - 1.96/\sqrt{n-3},\ \widehat{\zeta} + 1.96/\sqrt{n-3}\,]$. We get the interval for ρ through a back transformation of these boundaries by means of the formula $\rho = (\eta - 1)/(\eta + 1)$ with $\eta = \exp\langle 2\zeta \rangle$. It is often distinctly asymmetric with respect to $\widehat{\rho}$.

c ▷ In the **example** of the **Iris flowers** with species setosa, for the first two variables we have $S_{11} = 0.1242,\ S_{22} = 0.1437,\ S_{12} = 0.0992$ and from this $\widehat{\rho} = 0.0992/\sqrt{0.1242 \cdot 0.1437} = 0.743$ and $\widehat{\zeta} = 0.956$. The confidence interval for ζ is $[0.670,\ 1.24]$. From the corresponding η values 3.82 and 11.99 the correlation between length and width of the flower leaves has limits 0.585 and 0.846, so an asymmetric interval with respect to $\widehat{\rho} = 0.743$ is obtained. From this we conclude that the correlation can not be zero, since zero does not lie in this interval. The same conclusion results from the direct test: We get $T = 0.743 \cdot \sqrt{48}/\sqrt{1 - 0.743^2} = 7.68$ with p-value 0.000.

However, all these values should be interpreted cautiously, since due to the outliers apparent in Figure 2.1.a, the observations do not seem to follow a normal distribution. ◁

d **Tests for the Covariance Matrix.** We can formulate general null hypotheses for the covariance matrix $\mathfrak{Sigma}$ or parts of it. To derive the corresponding tests we need the distribution of the estimated covariance matrix. If we assume the normal distribution for the observations, then this distribution is commonly known by the name **Wishart distribution**. It depends, as we easily see, only on the true covariance matrix $\mathfrak{Sigma}$ and the sample size n. Instead of n we use the number of degrees of freedom $n-1$ as parameter and write

$$\boldsymbol{S} \sim \mathcal{W}\langle \mathfrak{Sigma}, n-1 \rangle \ .$$

e* The derivation of this distribution can be simplified if we first consider the case $\mathfrak{Sigma} = \boldsymbol{I}$. Then the $X^{(j)}$ are independent and standard normally distributed. Then the estimated variances S_{jj} are independent and each $(n-1)S_{jj}$ is chi-square distributed with $n-1$ degrees of freedom. This results in the standard Wishart distribution $\mathcal{W}\langle \boldsymbol{I}, n-1 \rangle$.

The distribution for a general $\mathfrak{Sigma}$ is obtained by interpreting $\boldsymbol{S}$ as a linear function of a matrix $\boldsymbol{S}^{(0)}$, which has this standard Wishart distribution .

We won't go into depth about the distribution of the estimated covariances S_{jk} here. (Perhaps this will change in a later version.)

4.3 Two Samples

a In the **Iris flowers example** the species setosa is clearly different from the other two in the measured variables. Are there also differences between versicolor and virginica? To simplify the presentation, we will limit ourselves to length and width of the sepal leaves. From Figures 1.2.b (i) and (ii) we can suspect that the two distributions are significantly different for the length, while this is not apparent for the width. However, instead of addressing the question of finding a difference for each of the two variables separately, here we ask whether the **joint distribution** of length and width is different for the two species.

b **Model.** For the joint distribution of the variables within each group (species) h, the simplest model again proves to be the normal distribution,

$$\underline{X}_{hi} \sim \mathcal{N}_m\langle \underline{\mu}_h, \mathfrak{Sigma}_h \rangle \ , \quad h = 1, 2 \ , \quad i = 1, 2, ..., n_h \ ,$$

and all *observations* should be stochastically independent – the *variables* are related as expressed by the covariance matrix $\mathfrak{Sigma}_h$.

We collect the differences of the expected values in the vector $\underline{\Delta}$,

$$\underline{\Delta} = \underline{\mu}_2 - \underline{\mu}_1$$

c **Estimation of the Differences.** Naturally, $\underline{\Delta}$ is estimated via the difference of the means $\widehat{\underline{\Delta}} = \overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.}$. How is this estimator distributed?

The expected value equals $\underline{\Delta}$, since the group means $\overline{\underline{X}}_{h.}$ are unbiased for the expected values $\underline{\mu}_h$. Since the $\overline{\underline{X}}_{h.}$ are independent, they add their covariance matrices

$$\mathrm{var}\langle\widehat{\underline{\Delta}}\rangle = \tfrac{1}{n_1}\,\mathfrak{Σ}_1 + \tfrac{1}{n_2}\,\mathfrak{Σ}_2 \ .$$

The exact distribution is then easy to derive and describe if we assume that the observations are normally distributed and the covariance matrices are the same between groups, $\mathfrak{Σ}_1 = \mathfrak{Σ}_2 =: \mathfrak{Σ}$. Then we have

$$\widehat{\underline{\Delta}} \sim \mathcal{N}_m\Big\langle \underline{\Delta}\ ,\ \Big(\tfrac{1}{n_1} + \tfrac{1}{n_2}\Big)\,\mathfrak{Σ}\Big\rangle \ .$$

This idea gives us the basis for the following test.

d **Test.** The null hypothesis shall be $\underline{\Delta} = \underline{\mu}_2 - \underline{\mu}_1 = \underline{0}$. The procedure is analogous to the univarate t-test on the one hand, and to the one-sample multivariate test on the other hand. The obvious test statistic to start with is the estimate $\widehat{\underline{\Delta}}$. However, it is multidimensional and thus is not a very suitable test statistic. As in the one-sample case, the length of the mean difference of standardized observations serves as a one dimensional test statistic. It can be written as

$$d^2\langle\overline{\underline{X}}_{1.},\overline{\underline{X}}_{2.};\,\mathfrak{Σ}\rangle = (\overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.})^T\,\mathfrak{Σ}^{-1}(\overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.})$$

e **Estimation of $\mathfrak{Σ}$.** Again, the covariance matrix $\mathfrak{Σ}$ is usually unknown and must be estimated from the observations. This is done via the mean of the two estimated covariance matrices for the groups. If the groups are of different sizes, it is best to weight with the number of degrees of freedom $n_h - 1$, and we thus get

$$\boldsymbol{S} = \tfrac{1}{n-2}\Big(\sum\nolimits_{i=1}^{n_1}(\underline{X}_{1i} - \overline{\underline{X}}_{1.})(\underline{X}_{1i} - \overline{\underline{X}}_{1.})^T + \sum\nolimits_{i=1}^{n_2}(\underline{X}_{2i} - \overline{\underline{X}}_{2.})(\underline{X}_{2i} - \overline{\underline{X}}_{2.})^T\Big) \ .$$

Substituting $\mathfrak{Σ}$ in the last version of the test statistic leads to the "studentized" difference

$$d^2\langle\overline{\underline{X}}_{1.},\overline{\underline{X}}_{2.};\,\boldsymbol{S}\rangle = (\overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.})^T\,\boldsymbol{S}^{-1}(\overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.})$$

as a reasonable test statistic. It measures the distance of the means "in the metric of the joint covariance matrix". Because of its underlying meaning, this statistic is aso called the **standard distance** between the two samples. In summary, we obtain the following testing procedure.

f **Hotelling's T^2 Test for Two Samples.** The test statistic, analogous to the univariate t-test, is
$T^2 = d^2\langle\overline{\underline{X}}_{2.},\overline{\underline{X}}_{1.};\,\boldsymbol{S}\rangle/(1/n_1 + 1/n_2)$. The distribution can be given in terms of the well-known F distribution for a multiple of T^2,

$$\frac{(n-m-1)}{m(n-2)(1/n_1 + 1/n_2)}\ d^2\langle\overline{\underline{X}}_{2.},\overline{\underline{X}}_{1.};\,\boldsymbol{S}\rangle \sim \mathcal{F}\langle m, n-m-1\rangle \ ,$$

if the null hypothesis $\underline{\mu}_1 = \underline{\mu}_2$ is fulfilled – and if the observations follow the normal distribution with equal covariance matrices.

g ▷ **T**he two species virginica and versicolor in the **iris flowers example** are very similar in the measured variables. Are differences present? We get the following different means:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
versicolor	5.94	2.77	4.26	1.33
virginica	6.59	2.97	5.55	2.03

The differences turn out to be beyond doubt, with a p-value given as $< 2 \cdot 10^{-16}$. ◁

4.S S-Functions

a The multivariate **two sample problem** can, like the univariate, be considered as a special case of a regression, with a binary variable that reflects group membership as an input variable.

b **Regression and analysis of variance** with multiple variables of interest can be performed with the same functions as for a univariate variable of interest, where on the left of the tilde sign $\sim$ a matrix is given. For the two sample problem, this appears as follows:

```
> t.y  <- as.matrix(iris[,1:4])
> t.r  <- lm(t.y~Species, data=iris, subset=Species!="setosa")
```

However, the function `summary(t.r)` shows for a `t.r` produced with `lm` only the results of the separate regressions of the individual variables of interest.

c The multivariate Hotelling test is obtained from the function `manova`

```
> t.r  <- manova(t.y Species, data=iris, subset=Species!="setosa")
> summary(t.r)
```

(The function `manova` essentially calls `aov` and sets as the class of the result `manova`, so that `summary` then summarizes the appropriate results.)