Statistical Methods II: Foundations of general and generalized linear models

Master Mathematics

Vrije Universiteit Brussel and Universiteit Antwerpen

K. Barbé

Edition 2016/2017

Chapter 1

Introduction

The course studies the main modern research methods of parametric statistics. The linear model plays throughout data-analysis a central role under the philosophy that measurements are prone to errors. Therefore, one postulates that the observations allow a decomposition in an error and error-free contribution leading naturally to an additive model. Depending on the experimental set-up, the linear model requires dedicated estimators and hypothesis tests. This course provides a detailed overview of the different aspects of linear models where the necessary instruments are developed and studied.

1.1 General(ized) linear models

Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a random variable $Y : \Omega \to \mathcal{X}$ such that $\mathcal{X} \subseteq \mathbb{R}$. The sample space is a measure space $(\mathcal{X}, \mathcal{R}, \lambda)$ with λ typically the Lebesgue measure on $\mathbb{R}$ or the counting measure on $\mathbb{Z}, \mathbb{N}$. We assume that the random variable Y holds a probability measure $\mathbb{P}_Y$ dominated by λ.
Consider a set of random variables $U_1, U_2, \ldots, U_d$ mutually defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The following results motivates the use of linear additive models under the scenario that the random vector$(Y, U_1, U_2, \ldots, U_d)$ is jointly Gaussian.

Theorem 1.1 *Let $(Y, U_1, U_2, \ldots, U_d)$ a Gaussian random vector with expectation $\underline{\mu} = [\mu_Y, \underline{\mu_U}]'$ and covariance matrix $\Sigma = \begin{bmatrix} \sigma^2 & \underline{\rho}' \\ \underline{\rho} & \Sigma_{\underline{U}} \end{bmatrix}$ then the following decomposition holds*

$$Y = \mathbb{E}[Y|\underline{U}] + \epsilon$$

with $Cov(\mathbb{E}[Y|\underline{U}], \epsilon) = 0$ and $\mathbb{E}[Y|\underline{U}] = \mu_y + A\mu_{\underline{U}} - A\underline{U}$ where $A = -\underline{\rho}'\Sigma_{\underline{U}}^{-1}$.

Proof: We show first that the random vector $(Y, U_1, U_2, \ldots, U_d)$ allows an orthogonal decomposition of the form $Y = \mathbb{E}[Y|\underline{U}] + \epsilon$ such that $\mathrm{Cov}(Y, \epsilon) = 0$ if $\{Y, U_1, U_2, \ldots, U_d\} \subset L_2(\Omega, \mathcal{A}, \mathbb{P})$.
Let $f : \mathcal{R}^d \to \mathcal{R}$ measurable mappings, then we compute:

$$g(\underline{U}) = \underset{f(\underline{U}) \in L_2(\Omega, \mathcal{A}, \mathbb{P})}{argmin} \mathbb{E}[(Y - f(\underline{U}))^2]$$

It is straightforward to verify that

$$\mathbb{E}[(Y - f(\underline{U}))^2] = \mathbb{E}[(Y - \mathbb{E}[Y|\underline{U}])^2] + \mathbb{E}[(\mathbb{E}[Y|\underline{U}] - f(\underline{U}))^2]$$

by using the identities $\mathbb{E}\left[\mathbb{E}[W|V]\right] = \mathbb{E}[W]$ en $\mathbb{E}[Wf(V)|V] = f(V)\mathbb{E}[W|V]$. Now, we obtain that $g(\underline{U}) = \mathbb{E}[Y|\underline{U}]$. This result together with Pythagoras theorem for Hilbert spaces completes the proof of the first claim.
For the second claim, we use the property of conditional expectations $\mathbb{E}[W|V] = \mathbb{E}[W]$ if W and V are independent. Consider now the random variable $Z = Y + A\underline{U}$ such that one obtains immediately $\mathrm{Cov}(Z, \underline{U}) = \underline{0}$. Since these variables are mutually Gaussian implies that Z and U_i are independent for all $i \in \{1, 2, \ldots, d\}$. Next, we can compute: $\mathbb{E}[Y|\underline{U}] = \mathbb{E}[Z - A\underline{U}|\underline{U}] = \mathbb{E}[Z] - A\underline{U} = \mu_y + A\underline{\mu_U} - A\underline{U}$. This completes the proof. □
This result brings one to the main definition of a general linear model (GLM). Consider the matrix $\mathbf{x}$ of n-observations of the random vector $\underline{U}$ where x_{ij} is observation i of random variable U_{j-1} for $j = 2, \ldots, d+1$ and the first column $x_{i1} = 1$ for all $i = 1, 2, \ldots, n$.

Definition 1.1 *Consider a sequence of independent random variables Y_i, $i \in \{1, 2, \ldots, n\}$. We define*

$$\underline{Y} = \mathbf{x}\underline{\beta} + \underline{\epsilon}$$

a general linear model such that $\underline{\epsilon}$ follows a Gaussian distribution with expectation 0 and covariance matrix $\sigma^2 I_n$.

Definition 1.1 implies that the probability measure $\mathbb{P}_{\underline{Y}|\mathbf{x}}$ can be parametrized as $\mathcal{N}(\mathbf{x}\underline{\beta}, \sigma^2 I)$. One can consider the situation wherein the random variable Y is not following a Gaussian distribution while the random variables $U_1, U_2, \ldots, U_d$ remain a Gaussian random vector. We consider the random variables Y to be of the generalized exponential family or exponential dispersion family.

Definition 1.2 *A random variable Y with probability measure $\mathbb{P}_Y$ is a member of the exponential dispersion family with natural parameter θ if its Lebesgue density function holds the form:*

$$f_Y(y|\theta, \phi) = h(y, \phi) \exp\left(\frac{\theta y - b(\theta)}{\phi}\right) I_A(y)$$

with $A \in \mathcal{A}$ parameter independent and $h : \mathcal{X} \to \mathbb{R}^+$ an $\mathcal{A}$-measurable mapping.

For exponential dispersion families, we can compute the distribution's expected value and variance as well as its characteristic function.

Theorem 1.2 *Let Y be a member of an exponential dispersion family then the expected value and variance are respectively given by $\mathbb{E}[Y] = \frac{\partial b(\theta)}{\partial \theta}$ and $Var(Y) = \frac{\partial^2 b(\theta)}{\partial \theta^2}\phi$. Moreover its characteristic function $\psi_Y(s) = \exp(b(\theta + s\phi) - b(\theta))$.*

Proof: The characteristic function is the expectation of the Laplace transform of the random variable Y in the left complex domain. It is straightforward to verify that for $s \in \mathbb{C}$ with negative real parts:

$$\mathbb{E}[\exp(sX)] = \exp\left(\frac{b(\theta + s\phi) - b(\theta)}{\phi}\right)$$

From probability theory, we know that $\mathbb{E}[Y] = \frac{\partial}{\partial s}\mathbb{E}[\exp(sX)](0)$ and $\mathbb{E}[Y^2] = \frac{\partial^2}{\partial s^2}\mathbb{E}[\exp(sX)](0)$ which delivers the claim. □
The characteristic function also implies that the exponential dispersion family holds an interesting summation formula. Let $Y_1, Y_2, \ldots, Y_M$ be independent distributed according

to the exponential dispersion family with respective parameters (θ, ϕ_i). Then the random variable $S = \sum_{i=1}^{M} Y_i$ is also of the exponential dispersion family with parameter $\left(\theta, \sum_{i=1}^{M} \frac{1}{\phi_i}\right)$. Definition 1.1 for general members of the exponential dispersion family suggests the following *generalized* linear models.

Definition 1.3 *Consider a sequence of independent random variables Y_i, $i \in \{1, 2, \ldots, n\}$ where its distribution is of the exponential dispersion family. We define*

$$\underline{Y} = \left.\frac{\partial b(\theta)}{\partial \theta}\right|_{\theta=\mathbf{x}\underline{\beta}} + \underline{\epsilon}$$

a generalized linear model such that $\underline{\epsilon}$ follows a distribution with expectation 0 and covariance matrix $\sigma^2 I_n$ with $\sigma^2 = \phi \left.\frac{\partial^2 b(\theta)}{\partial \theta^2}\right|_{\theta=\mathbf{x}\underline{\beta}}$. One refers to the mapping $g(\theta)$ such that $g^{-1}(\theta) = \frac{\partial b(\theta)}{\partial \theta}$ as the link function.

Within the class of generalized linear models, we consider of course the Normal distribution, but also random variables Y with a Binomial distribution, Multinomial distribution, Poisson distribution, Gamma distribution, Tweedie distribution... In this course we will encounter Normal leading to Analysis of (Co)variance, Binomial leading to logistic regression and Multinomial distribution leading to multinomial regression.

Example 1.1 *The normal distribution $Y \sim \mathcal{N}(\mu, \sigma^2)$ with probability density function given by*

$$\begin{aligned} f(y|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \\ &= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2}\right) \exp\left(\frac{-y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \end{aligned}$$

The normal distribution is of the exponential dispersion family with $\theta = \mu, \phi = \sigma^2, b(\theta) = \frac{\theta^2}{2}$ and $h(y, \phi) = \exp\left(\frac{-y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)\right)$. The link function is $g(x) = x$ such that the generalized linear model boils down to the general linear model.

Example 1.2 *The Poisson distribution $Y \sim Poisson(\lambda)$ with probability density function given by*

$$\begin{aligned} f(y|\lambda) &= \frac{\lambda^y \exp(-\lambda)}{y!} \\ &= \exp\left(y\log(\lambda) - \lambda\right) \frac{1}{y!} \end{aligned}$$

The Poisson distribution is a member of the exponential dispersion family with $\theta = \log(\lambda)$, $\phi = 1$, $b(\theta) = \exp(\theta)$ and $h(y) = \frac{1}{y!}$. The link function for the generalized linear model is given by $g(x) = \log(x)$ leading to Poisson regression analysis:

$$\underline{Y} = \exp\left(\mathbf{x}\underline{\beta}\right) + \underline{\epsilon}$$

Example 1.3 *The Bernoulli distribution $Y \sim Bern(p)$ with probability density function given by*

$$\begin{aligned} f(y|p) &= p^y(1-p)^{1-y} \\ &= \exp\left(y\log\left(\frac{p}{1-p}\right) + \log(1-p))\right) \end{aligned}$$

The Bernoulli and by extension the binomial distribution is of the exponential dispersion family with $\theta = \log\left(\frac{p}{1-p}\right)$, $b(\theta) = \log(1+\exp(\theta))$ and $\phi = 1 = h(y)$. The link function for its generalized linear model is given by $g(x) = \log\left(\frac{x}{1-x}\right)$ leading to binary logistic regression:

$$\underline{Y} = \frac{\exp(\mathbf{x}\underline{\beta})}{1+\exp(\mathbf{x}\underline{\beta}x)} + \underline{\epsilon}$$

1.2 Revisiting (fixed effects) normal linear models

In this section we recall the results from earlier courses wherein Definition 1.1 plays a central role. Note that the wording "fixed effects" refers to the situation where the design matrix $\mathbf{x}$ is fixed. Even when these are realizations of possible random variables $U_1, \ldots, U_p$, the analysis is performed conditionally on the specific realizations $\mathbf{x}$ of these random variables.
In a linear model the expected value of $Y_i, i = 1, \ldots, N$ is explained linearly $\mu_i(\underline{\beta}) = \sum_{k=1}^{p+1} x_{ik}\beta_k$. Therefore, it is straightforward to see that in case $x'_{ij} = x_{ij} + \Delta$ that this implies a change in the expectation given by $\mu'_i(\underline{\beta}) = \mu_i(\underline{\beta}) + \Delta\beta_j$. This interpretation is called "ceteris paribus" or roughly translated to "others remaining equal".

Example 1.4 *Consider the random variables Y_i denoting salary in Euro for different people $i = 1, \ldots, n$. This variable's expectation can be described by the variables U_1 gender, U_2 degree with possible outcomes (i) high school, (ii) bachelor, (iii) master and (iv) phd, U_3 experience since graduation in years. Hence, this conditioned model leads to*

$$\mu_i(\beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

The parameter β_0 is the intercept receiving the interpretation of "grand average income" which is perturbed by the additional variables. The parameter β_1 receives the increment in salary between men and women with $U_1 \in \{0, 1\}$. The parameter β_2 denotes the increase in salary for an increasing degree with $U_2 \in \{1, 2, 3, 4\}$. The parameter β_3 denotes the increment of an additional year of experience.

The model in the previous example is not popular. Although it may be expected that income increases linearly as a function of their experience in year, this is not expected to progress linearly for what an increasing degree is concerned. We call the variable U_2 in the example an ordinal variable. This variable takes an increasing number of finite number of real values for which it is unclear which values these are to support this linear trend. Thus, we wish the model to be independent of its ordinal coding. Thus the result may not matter if we assign values $1, 2, 3, 4$ or $1, 10, 24, 86$ to the outcomes (i) high school, (ii) bachelor, (iii) master and (iv) phd. This idea leads to the following linear model.

Example 1.5 *Consider the random variables Y_i denoting salary in Euro for different people $i = 1, \dots, n$. This variable's expectation can be described by the variables U_1 gender, U_2 degree with possible outcomes (i) high school, (ii) bachelor, (iii) master and (iv) phd, U_3 experience since graduation in years. Hence, this conditioned model leads to*

$$\mu_i(\beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 + \beta_1^{[1]} x_{i1} + \beta_1^{[2]} x_{i2} + \beta_2^{[1]} x_{i3} + \beta_2^{[2]} x_{i4} + \beta_2^{[3]} x_{i5} + \beta_2^{[4]} x_{i6} + \beta_3 x_{i5} + \beta_4 x_{i5}^2$$

The parameter β_0 is the intercept receiving the interpretation of "grand average income" which is perturbed by the additional variables. The parameter $\beta_1^{[i]}$ with $i = 1, 2$ receives the increment w.r.t. the grand mean in salary between men and women. The parameters $\beta_2^{[i]}$ with $i = 1, 2, 3, 4$ denote the increase in salary **with respect to a high school degree** *for respective degrees bachelor, master and phd. The parameter β_3 denotes the* **linear** *increment of an additional year of experience while parameter β_4 denotes the* **quadratic** *increment of an additional year of experience. The constraint "sum-to-zero" is added to the dummy parameters such that $\beta_1^{[1]} + \beta_1^{[2]} = 0$ and $\sum_{i=1}^4 \beta_2^{[i]} = 0$.*

The first model is a purely multiple linear regression model whereas the latter is called the Analysis of Covariance or ANCOVA model.

Definition 1.4 *Consider a sequence of independent random variables Y_i, $i \in \{1, 2, \dots, n\}$. We define*

$$\underline{Y} = \mathbf{x}\underline{\beta} + \underline{\epsilon}$$

a fixed effects ANCOVA model such that $\underline{\epsilon}$ follows a Gaussian distribution with expectation 0 and covariance matrix $\sigma^2 I_n$ if categorical variables, either nominal whose outcomes have no natural order or ordinal whose outcomes have a natural order, are used as dummy variables with outcomes 0 or 1 together with the sum-to-zero constraint. The columns of the design matrix $\mathbf{x}$ associated to the categorical variables are called the ANOVA part, whereas the numerical variables are the covariates.

Remark: The sum-to-zero constraint is not the only possibility, one may apply a reference category such that $\beta_i^{[1]} = 0$. This constraint is popular in generalized linear models but less common in an ANCOVA setting.
Now we can establish the classical result that the least squares estimator is the "Uniform Minimum Variance Unbiased" (UMVU) estimator under the condition that the full design matrix $\mathbf{x}$ is of full rank.

Theorem 1.3 *Consider a sequence of random variables $Y_i, i = 1, 2, \dots, n$ satisfying Definition 1.4 with realization vector $\underline{y}$. Let the matrix $\mathbf{x} \in \mathbb{R}^{n \times p}$ be of full rank then the estimator $\hat{\underline{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\underline{Y}$ is the UMVU estimator of $\underline{\beta}$ whereas the estimator $\hat{\sigma}^2 = \frac{1}{n-p}||\underline{Y} - \mathbf{x}\hat{\underline{\beta}}||^2$ is the UMVU estimator of σ^2. Moreover $\hat{\underline{\mu}} = \mathbf{x}\hat{\underline{\beta}}$ is the orthogonal projection of $\underline{y}$ onto the subspace span$(\mathbf{x})$.*

Proof: Since $\underline{\mu} = \mathbf{x}\beta$, it follows from Example 1.1 that the statistic $\underline{T} = [\underline{Y}'\mathbf{x}, \underline{Y}'\underline{Y}]'$ is sufficient for $\underline{\theta}' = [\underline{\beta}', \sigma^2]'$ due to the factorization theorem. We first show that the sufficient statistic is complete. Recall the definition of a complete statistic $\underline{T}(\underline{Y})$:
Let $f : \mathcal{R}^p \to \mathcal{R}$ such that $f(T(\underline{Y}))$ is absolutely $\mathbb{P}_\theta$-integrable. The statistic $\underline{T}(\underline{Y})$ is complete if the following holds: $\mathbb{E}_\theta[f(\underline{T}(\underline{Y}))] = 0$ implies that $f(\underline{T}(\underline{Y})) = 0, \mathbb{P}_\theta$-almost surely with $\mathbb{P}_\theta$ following a parametrized Gaussian distribution with $\underline{\theta} = [\underline{\beta}', \sigma^2]'$ such that the expectation and covariance matrix are respectively given by $\mathbf{x}\underline{\beta}$ and $\sigma^2 I$.

First we prove that the sufficient statistic $\underline{T} = [\underline{Y}'\mathbf{x}, \underline{Y}'\underline{Y}]'$ is complete. A straightforward computation reveals that:

$$\begin{aligned}\mathbb{E}_\theta[f([\mathbf{x}'y, y'y])] &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\int_{\mathbb{R}^n} f([\mathbf{x}'y, y'y])\exp\left(\frac{-y'y}{2\sigma^2}\right)\exp\left(\frac{y'\mathbf{x}\underline{\beta}}{\sigma^2}\right)d\underline{y} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\mathcal{B}\left(f([\mathbf{x}'y, y'y])\exp\left(\frac{-y'y}{2\sigma^2}\right)\right)\left(\frac{\mathbf{x}\underline{\beta}}{\sigma^2}\right)\end{aligned}$$

where $\mathcal{B}(.)$ denotes the two-sided Laplace transform. Hence, the Laplace transform $\mathcal{B}\left(f([\mathbf{x}'y, y'y])\exp\left(\frac{-y'y}{2\sigma^2}\right)\right)(\underline{p}) = 0$ for all $p \in \text{span}(\mathbf{x})$ since $\mathbf{x}$ is of full rank. Clearly the function 0 can be analytically continued on the domain of convergence of the Laplace integral. Since analytical continuation is unique, the bilateral Laplace transform for the function $\mathcal{B}\left(f([\mathbf{x}'y, y'y])\exp\left(\frac{-y'y}{2\sigma^2}\right)\right)(\underline{p}) = 0$ for $\underline{p} \in \mathbb{C}^n$ of its domain of convergence. Since the bilateral Laplace transform is bijective, we conclude that the function itself equals 0. This establishes the completeness.
The completeness of the sufficient statistic $\underline{T} = [\underline{Y}'\mathbf{x}, \underline{Y}'\underline{Y}]'$ now provides that the UMVU-estimator for $[\underline{\beta}', \sigma^2]$ is provided by a measurable mapping $g : \mathcal{R}^{p+1} \to \mathcal{R}^{p+1}$ such that $[\hat{\underline{\beta}}', \hat{\sigma}^2]' = g(\underline{T})$. Since the UMVU estimator is $\mathbb{P}_\theta-$almost sure unique, it is sufficient to obtain a mapping $g(.)$ which maps the sufficient statistic to an unbiased estimator. We compute

$$\mathbb{E}[\underline{Y}'\mathbf{x}] = \mathbb{E}[(\underline{\beta}'\mathbf{x}' + \underline{\epsilon}')\mathbf{x}] = \underline{\beta}'\mathbf{x}'\mathbf{x}$$

It is therefore clear that unbiased estimator as a mapping of its sufficient statistic is provided through $\hat{\underline{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\underline{Y}$ which is immediately the UMVU-estimator. We further compute

$$\mathbb{E}[(\underline{Y} - \mathbf{x}\underline{\beta})'(\underline{Y} - \mathbf{x}\underline{\beta})] = n\sigma^2$$

Next it follows that

$$n\sigma^2 = \mathbb{E}[((\underline{Y} - \mathbf{x}\hat{\underline{\beta}}) + \mathbf{x}(\hat{\underline{\beta}} - \underline{\beta}))'((\underline{Y} - \mathbf{x}\hat{\underline{\beta}}) + \mathbf{x}(\hat{\underline{\beta}} - \underline{\beta}))]$$

Note that $\underline{Y} - \mathbf{x}\hat{\underline{\beta}} = (I_n - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\underline{\epsilon}$ and $\hat{\underline{\beta}} - \underline{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\underline{\epsilon}$. Thus, it is not so difficult to see that the mixed products cancel leading to the simplification:

$$n\sigma^2 = \mathbb{E}[||\underline{Y} - \mathbf{x}\hat{\underline{\beta}}||^2] + \mathbb{E}[\underline{\epsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\underline{\epsilon}]$$

To compute the second term, we apply the following trick:

$$\begin{aligned}\mathbb{E}[\underline{\epsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\underline{\epsilon}] &= \mathbb{E}[\text{trace}\left(\underline{\epsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\underline{\epsilon}\right)] \\ &= \mathbb{E}[\text{trace}\left(\underline{\epsilon}\underline{\epsilon}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\right)] \\ &= \text{trace}\left(\mathbb{E}[\underline{\epsilon}\underline{\epsilon}']\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\right) \\ &= \sigma^2\text{trace}\left(I_p\right) = \sigma^2 p\end{aligned}$$

As such, we obtain that $\hat{\sigma}^2$ is an unbiased estimator. This estimator is the UMVU estimator as it is a mapping of both sufficient statistics only. The final claim follows from noting that the inner-product between vectors $\hat{\underline{\mu}} = \mathbf{x}\hat{\underline{\beta}} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\underline{y}$ and $\underline{y} - \hat{\underline{\mu}} = (I_n - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}')\underline{y}$ equals zero. □
This result provides a method to estimate the parameters of a normal linear model but it does not allow inference. To provide a full interpretation of a normal linear model,

one requires confidence intervals on the estimates $\hat{\underline{\beta}}$ as well as the estimated mean $\hat{\underline{\mu}}$. On top of that, the validity of the model is required leading to optimal goodness-of-fit tests. Inference requires knowledge on the probability distribution of the estimators $\hat{\underline{\beta}}$ and $\hat{\sigma}^2$, the normality assumption of the data keeps the distribution of the associated estimators tractable.

Theorem 1.4 *Consider a sequence of random variables $Y_i, i = 1, 2, \ldots, n$ satisfying Definition 1.4 with realization vector $\underline{y}$. Let the matrix $\mathbf{x} \in \mathbb{R}^{n \times p}$ be of full rank and consider the UMVU estimator $\hat{\underline{\beta}}$ for $\underline{\beta}$ and $\hat{\underline{\mu}}$ for $\underline{\mu}$. The distribution of the estimators is given by:*

$$\hat{\underline{\beta}} \overset{d}{=} \mathcal{N}(\underline{\beta}, (\mathbf{x}'\mathbf{x})^{-1}\sigma^2)$$

$$\hat{\underline{\mu}} \overset{d}{=} \mathcal{N}(\underline{\mu}, \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\sigma^2)$$

This result follows directly from linear transformation of random variables. This result allows to compute $(1-\alpha)$-confidence intervals around the estimates such that we obtain:

$$\beta_i \in [\hat{\beta}_i - t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}^2 \left[(\mathbf{x}'\mathbf{x})^{-1}\right]_{ii}, \hat{\beta}_i + t_{m-p,1-\frac{\alpha}{2}}\hat{\sigma}^2 \left[(\mathbf{x}'\mathbf{x})^{-1}\right]_{ii}]$$

$$\mu_i \in [\hat{\mu}_i - t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}^2 \left[\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\right]_{ii}, \hat{\mu}_i + t_{n-p,1-\frac{\alpha}{2}}\hat{\sigma}^2 \left[\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\right]_{ii}]$$

where $[\mathbf{a}]_{ij}$ denotes the entry of matrix $\mathbf{a}$ with index (i, j) and $t_{q,\gamma}$ is the γ-quantile of the t-distribution with q degrees of freedom. The confidence intervals are an important source to provide an accurate interpretation of the estimated linear model. It offers to infer the expected effect of one parameter on the mean value of a random variable Y by increasing a covariate or factor by one in the model. On top of that, it allows predictions μ_j with dedicated confidence intervals for user-specified values of the variables $[x_{j1}, \ldots, x_{jp}]$. Additional to the interpretation of the model parameters, one wishes to assess the significance of the model or model parts. One wishes to investigate whether the model can be simplified to a lower dimensional subspace $\text{span}(\underline{x}_i | i \in I)$ of $\text{span}(\mathbf{x})$ with $I \subsetneq \{1, 2, \ldots, p\}$. Typically for dispersion exponential families the hypotheses are formulated with respect to its natural parameter $\underline{\mu}$. The tests of interest are for so called nested models.

Definition 1.5 *Let $\underline{Y}$ be a random vector of size n following a multivariate Gaussian distribution $\mathcal{N}(\underline{\mu}, \sigma^2 I_n)$. Consider the design matrix $\mathbf{x}$ and its associated vector space $V = span(\mathbf{x})$, let $V_0 \leq V$ be a vector subspace. The model given by*

$$\underline{Y} = \mathbf{x}\underline{\beta} + \underline{\epsilon}$$

with $\mathbf{x}\underline{\beta} \in V_0$ is called a nested model. One may generalize the definition to dispersion exponential distribution by adding the required link function.

We put forward the following testing problem for normal linear models where we assess the nul hypothesis $\mathcal{H}_0 : \underline{\mu} \in V_0$ versus $\mathcal{H}_1 : \underline{\mu} \in V$. Unfortunately, even when the noise parameter σ^2 is assumed known, there is no Uniform Maximum Power (UMP) test for an arbitrary subspaces V_0 versus V (Kolodziejczyk, On an important class of statistical hypotheses, Biometrica, 27(161), 1935). As such, by virtue of the Neyman-Pearson theorem, one relies on an ad-hoc approach: Likelihood Ratio (LR) tests. This class of tests has some optimality properties: the test is optimal among tests with a specific structure such that it is no longer uniformly the best test (Wolfowitz, The power of the classical tests associated with the normal distribution, Annals of mathematical statistics, 20(540), 1949), the test is however asymptotically UMP. The latter is proven under an additional regularization condition.

Definition 1.6 *Let $\underline{Y}$ be a random vector of size n following a distribution whose joint density $f(\underline{y}|\underline{\theta})$ is parametrized in $\underline{\theta} \in \Theta$ where the parameter space $\Theta \subset \mathbb{R}^m$ is compact. Consider a subset $\Theta_0 \subset \Theta$. The Likelihood Ratio test for the assessment of $\mathcal{H}_0 : \underline{\theta} \in \Theta_0$ versus $\mathcal{H}_1 : \underline{\theta} \in \Theta$ is given by*

$$L(\underline{Y}) = \frac{\max_{\theta \in \Theta} f(\underline{Y}|\theta)}{\max_{\theta \in \Theta_0} f(\underline{Y}|\theta)}.$$

Proof in favour of the alternative hypothesis is found for increasing values of the test statistic $L(\underline{Y})$.

Thus the LR test applies maximum likelihood estimators to the Neyman-Pearson lemma. This procedure is unfortunately not guaranteed to be an UMP test. The Neyman-Pearson lemma only provides an UMP test for simple hypotheses where Θ_0 and Θ_1 are sets with one element such that the test becomes dichotomous. For so called composite hypotheses the problem to obtain an UMP test is challenging but possible if Θ_0 remains a single value and the family of distributions is of the monotone likelihood ratio type. No finite sample properties can be addressed to the LR test but asymptotically the test converges to a dichotomous Neyman-Pearsona approach since ML estimators converge for a specific probability measure $\mathbb{P}_{\theta_1}$ to a single deterministic value which makes the LR approach reasonable.

Theorem 1.5 *Let $\underline{Y}$ be a random vector of size n following a multivariate Gaussian distribution with expectation $\underline{\mu}$ and covariance matrix $\sigma^2 I_n$. The LR test to assess the hypothesis $\mathcal{H}_0 : \underline{\mu} \in V_0$ versus $\mathcal{H}_1 : \underline{\mu} \in V$ such that $V = span(\mathbf{x})$ of dimension p and $V_0 \leq V$ of dimension r is given by*

$$L(\underline{Y}) = \left(\frac{||\underline{Y} - \mathbf{x}\hat{\underline{\beta}}_0||^2}{||\underline{Y} - \mathbf{x}\hat{\underline{\beta}}_1||^2} \right)^n$$

with $\hat{\underline{\beta}}_0$ and $\hat{\underline{\beta}}_1$ are the respective ML estimators w.r.t. $\{\underline{\beta}|\mathbf{x}\underline{\beta} \in V_0\}$ and $\{\underline{\beta}|\mathbf{x}\underline{\beta} \in V\}$ such that the LR test can be transformed through a monotonical mapping $f(x) = \frac{n-p}{p-r}\left(x^{\frac{1}{n}} - 1\right)$ on the domain $x \geq 1$ to

$$\frac{n-p}{p-r} \frac{||\underline{Y} - \mathbf{x}\hat{\underline{\beta}}_0||^2 - ||\underline{Y} - \mathbf{x}\hat{\underline{\beta}}_1||^2}{||\underline{Y} - \mathbf{x}\hat{\underline{\beta}}_1||^2} \overset{\mathcal{H}_0}{\sim} \mathcal{F}_{p-r,n-p}$$

where $\mathcal{F}_{p-r,n-p}$ denotes a F-distribution with $(p-r, n-p)$ degrees of freedom.

Proof: Referring to the multivariate Gaussian distribution of $\underline{Y}$ provide in Example 1.1, we obtain the likelihood:

$$f(\underline{Y}|\underline{\beta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^n \exp\left(\frac{-1}{2\sigma^2}||\underline{Y} - \mathbf{x}\underline{\beta}||^2\right)$$

It is easy to see that the ML estimators $\hat{\underline{\beta}}_i$ with $i = 0, 1$ denote the coefficients which provide $\hat{\underline{\mu}}_i$ which are orthogonal projections of $\underline{Y}$ onto the vector space V_0 and V respectively. On top of that, the ML estimators $\hat{\sigma}_i^2$ is given by $\hat{\sigma}_i^2 = \frac{1}{n}||Y - \mathbf{x}\hat{\underline{\beta}}_i||^2$. Therefore we immediately obtain that

$$L(\underline{Y}) = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^n$$

To complete the proof, we study the distribution of $T_n^{[LR]} = \frac{||\underline{Y}-\mathbf{x}\hat{\underline{\beta}}_0||^2-||\underline{Y}-\mathbf{x}\hat{\underline{\beta}}_1||^2}{||\underline{Y}-\mathbf{x}\hat{\underline{\beta}}_1||^2}$ under the hypothesis $\mathcal{H}_0$. Consider an orthogonal basis $(\underline{e}_1,\ldots,\underline{e}_n)$ of $\mathbb{R}^n$. We order the basis elements in such a way that $V_0 = \text{span}(\underline{e}_1,\ldots,\underline{e}_r)$ while $V = \text{span}(\underline{e}_1,\ldots,\underline{e}_p)$. Since $\hat{\underline{\mu}}_i$ are orthogonal projections of $\underline{Y}$ onto V_0 and V respectively, the following representations holds:

$$\underline{Y}-\mathbf{x}\hat{\underline{\beta}}_0 = \sum_{i=r+1}^{n}(\underline{Y}'\underline{e}_i)\underline{e}_i \text{ and } \underline{Y}-\mathbf{x}\hat{\underline{\beta}}_1 = \sum_{i=p+1}^{n}(\underline{Y}'\underline{e}_i)\underline{e}_i$$

The test statistic $T_n^{[LR]}$ becomes:

$$T_n^{[LR]} = \frac{\sum_{i=r+1}^{p}(\underline{Y}'\underline{e}_i)^2}{\sum_{i=p+1}^{n}(\underline{Y}'\underline{e}_i)^2}$$

Due to the fact that $Y_1, Y_2, \ldots, Y_n$ are independently distributed, the numerator and denominator of the test statistic $T_n^{[LR]}$ are independent. Next, we look at the joint distribution of the inner product $\underline{Y}'\underline{e}_i$ for $i = r+1, r+2, \ldots, p$ and $i = p+1, p+2, \ldots, n$. The inner-products clearly follow a multivariate Gaussian distribution, since the inner-products denotes a linear operator of a Gaussian random vector $\underline{Y}$. Consider the matrix $\mathbf{e}_0 = [\underline{e}_{r+1},\ldots,\underline{e}_p]$ such that the random vector of inner products becomes $\underline{Y}'\mathbf{e}_0$ with expectation $\underline{\mu}'\mathbf{e}_0$ and covariance matrix $\sigma^2 I_{p-r}$. Note however that under the hypothesis $\mathcal{H}_0$, the expectation $\underline{\mu} \in V_0$ such that $\underline{\mu}'\mathbf{e}_0 = \underline{0}$. As a consequence, the numerator $\sum_{i=r+1}^{p}(\underline{Y}'\underline{e}_i)^2 \overset{\mathcal{H}_0}{\sim} \sigma^2\chi^2_{p-r}$. Entirely the same approach leads to the conclusion $\sum_{i=p+1}^{n}(\underline{Y}'\underline{e}_i)^2 \overset{\mathcal{H}_0}{\sim} \sigma^2\chi^2_{n-p}$ which completes the proof. □

Now we can establish two specific tests to assess the significance of the design which is the goodness-of-fit test and the significance of one particular column of the design matrix which is the Wald test. Both tests are special cases of Theorem 1.5

Theorem 1.6 *(Goodness-of-fit test) Let $\underline{Y}$ be a random vector of size n following a multivariate Gaussian distribution with expectation $\underline{\mu}$ and covariance matrix $\sigma^2 I_n$. Let $\underline{\mu} \in span(\mathbf{x})$ such that $\mathbf{x} \in \mathbb{R}^{n\times p}$ of full rank. Let $\mathbf{x} = [\underline{1}, \tilde{\mathbf{x}}]$ with $\underline{1} \in \mathbb{R}^{n\times 1}$. The LR test to assess the hypothesis $\mathcal{H}_0 : \underline{\mu} = c.\underline{1}$ versus $\mathcal{H}_1 : \underline{\mu} \in span(\mathbf{x})$ is given by*

$$T_n(\underline{Y}) = \frac{MSR}{MSE} = \frac{\frac{1}{p-1}SSR}{\frac{1}{n-p}SSE}$$

with sum-of-squared-regression $SSR = ||\hat{\underline{\mu}}-\bar{Y}||^2$ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $SSE = ||\underline{Y}-\hat{\underline{\mu}}||^2$ such that

$$||\underline{Y}-\bar{Y}||^2 = ||\underline{Y}-\hat{\underline{\mu}}||^2 + ||\hat{\underline{\mu}}-\bar{Y}||^2$$

such that $\hat{\underline{\mu}}$ is the ML estimate for $\underline{\mu}$ w.r.t. $\mathcal{H}_1$.

Proof this as an exercise. One often refers to the R^2-statistic computed as $\frac{\text{SSR}}{||\underline{Y}-\bar{Y}||^2}$ which is value between 0 and 1. It specifies the fraction of the variability which is captured and described by the regression model. It is useful to compare two nested models where one prefers a model with a significantly larger R^2 value.

Theorem 1.7 *(Wald test) Let $\underline{Y}$ be a random vector of size n following a multivariate Gaussian distribution with expectation $\underline{\mu}$ and covariance matrix $\sigma^2 I_n$. Let $\underline{\mu} \in span(\mathbf{x})$ such that $\mathbf{x} \in \mathbb{R}^{n\times p}$ of full rank. Consider the statistical model with $1 \leq k \leq p$ given*

by $\underline{Y} = \tilde{\mathbf{x}}\underline{\tilde{\beta}} + \underline{x}_k\beta_k + \underline{\epsilon}$. *The LR test to assess the hypothesis* $\mathcal{H}_0 : \underline{\mu} \in span(\tilde{\mathbf{x}})$ *versus* $\mathcal{H}_1 : \underline{\mu} \in span(\mathbf{x})$ *is given by*

$$T_n(\underline{Y}) = \frac{(\underline{\hat{\beta}}_1 - \underline{\hat{\beta}}_0)'\mathbf{x}'\mathbf{x}(\underline{\hat{\beta}}_1 - \underline{\hat{\beta}}_0)}{\hat{\sigma}^2}$$

Proof is left as an exercise. This test can be simplified under the approximation that $\underline{\hat{\beta}}_1 - \underline{\hat{\beta}}_0 = [0, 0, \ldots, 0, \hat{\beta}_k, 0, \ldots, 0]'$. This implies that the test is approximately equal to:

$$T_n(\underline{Y}) \approx \frac{\hat{\beta}_k^2}{\hat{\sigma}^2[(\mathbf{x}'\mathbf{x})^{-1}]_{pp}}$$

This motivates that inspecting the confidence intervals of individual parameter estimates is an approximate significance test. One infers that if the confidence interval captures 0 that the parameter is insignificant such that the model can be simplified.

Example 1.6 *An NBA coach wishes to analyze the performance of its players by answering the research question: Is a player's performance dominated by skill or physical appearance? In an effort to answer this question, a total of 54 NBA players are investigated. The measured variables are: Height (feet), Weigth (lbs), Successful 2 point shots (%), Successful 3 point shots (%) and average number of points per match of last season. One wishes to clarify the average number of points per match as a function of the other variables. A model building step (see exercises) reveals that the general linear model's conditions are satisfied for the following model:*

$$\sqrt{Y} = \mu_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

with Y *the average number of points per match,* X_1 *height,* X_2 *Weight,* X_3 *successful 2 point shots and* X_4 *successful 3 point shots.*

The first step of the analysis is to estimate the full model and assess the significance of the model through an F-test and assess the individual parameters through a Wald test. The

ANOVA [a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9,887	4	2,472	4,328	,004 [b]
	Residual	27,984	49	,571		
	Total	37,871	53			

a. Dependent Variable: SqrtScore

b. Predictors: (Constant), successful 3 points [%], successful goals [%], Height [feet], weight [lbs]

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2,964	2,075		1,429	,159	-1,205	7,134
	Height [feet]	-,653	,415	-,354	-1,573	,122	-1,487	,181
	weight [lbs]	,001	,006	,025	,106	,916	-,012	,014
	successful goals [%]	7,422	2,194	,497	3,383	,001	3,013	11,831
	successful 3 points [%]	1,602	1,099	,190	1,458	,151	-,607	3,810

a. Dependent Variable: SqrtScore

Table 1.1: The Goodness-of-fit test see Theorem 1.6 (top table) and Wald test see Theorem 1.7 (bottom table)

table indicates that the goodness-of-fit reveals a strongly significant model (p-value<0.01)

with an F-statistic exceeding 4 implying that the MSR is four times larger than the MSE. The R^2 statistic is only 0.26 which is low such that the model may be too simple and misses variables in the design which are not taken into consideration. The Wald test only identifies one variable of significance: X_3. Due to collinearity, significance of the other parameters may be lowered as its information is spread over different variables indicating the same variable. Therefore, we inspect the correlations among the different variables X_i, X_j. The highest correlation is shared between height and weight while height is more

Correlations

		SqrtScore	Height [feet]	weight [lbs]	successful goals [%]	successful 3 points [%]
Pearson Correlation	SqrtScore	1,000	-,137	-,070	,330	,265
	Height [feet]	-,137	1,000	,834	,496	-,259
	weight [lbs]	-,070	,834	1,000	,516	-,290
	successful goals [%]	,330	,496	,516	1,000	-,019
	successful 3 points [%]	,265	-,259	-,290	-,019	1,000

Table 1.2: Empirical correlations among the different variables

correlated to $\sqrt{Y}$ than weight. Therefore it makes sense to drop the variable weight in an effort to improve the significance of the variable height. Backward elimination reduces the model by dropping stepwise the least significant variable. If this variable holds a high correlation to another regressor, the regressor remaining in the model will improve its significance. If the significance of the variables does not longer improve or there are no insignificant regressors left, the backward elimination ends. In the top table, we see the

ANOVA [a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9,887	4	2,472	4,328	,004 [b]
	Residual	27,984	49	,571		
	Total	37,871	53			
2	Regression	9,881	3	3,294	5,884	,002 [c]
	Residual	27,990	50	,560		
	Total	37,871	53			
3	Regression	8,662	2	4,331	7,562	,001 [d]
	Residual	29,209	51	,573		
	Total	37,871	53			

a. Dependent Variable: SqrtScore

b. Predictors: (Constant), successful 3 points [%], successful goals [%], Height [feet], weight [lbs]

c. Predictors: (Constant), successful 3 points [%], successful goals [%], Height [feet]

d. Predictors: (Constant), successful goals [%], Height [feet]

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	2,964	2,075		1,429	,159	-1,205	7,134
	Height [feet]	-,653	,415	-,354	-1,573	,122	-1,487	,181
	weight [lbs]	,001	,006	,025	,106	,916	-,012	,014
	successful goals [%]	7,422	2,194	,497	3,383	,001	3,013	11,831
	successful 3 points [%]	1,602	1,099	,190	1,458	,151	-,607	3,810
2	(Constant)	2,880	1,895		1,519	,135	-,927	6,687
	Height [feet]	-,620	,269	-,336	-2,301	,026	-1,160	-,079
	successful goals [%]	7,477	2,110	,500	3,543	,001	3,238	11,716
	successful 3 points [%]	1,582	1,072	,187	1,476	,146	-,571	3,735
3	(Constant)	4,624	1,499		3,084	,003	1,614	7,633
	Height [feet]	-,734	,261	-,398	-2,814	,007	-1,258	-,210
	successful goals [%]	7,885	2,116	,528	3,726	,000	3,636	12,134

a. Dependent Variable: SqrtScore

Table 1.3: The Goodness-of-fit test see Theorem 1.6 (top table) and Wald test see Theorem 1.7 (bottom table) for the backward elimination procedure

model significance as a function of the F-statistic improving from a p-value of 0.004 to 0.001 by decreasing the number of variables. The bottom table tracks which variables

are left out of the model to conclude a model with only significant variables left tested through Wald's test: X_1, X_3. Also by applying the approximate Wald test, one sees that the estimated parameters are at least 1.65 times larger than the standard errors resulting in significance at a confidence of 95% excluding 0 from its confidence intervals. Now an answer to the research question may be given. Both height and the ability of making goals are significant variables to explain the average number of goals per match of an individual player. An average NBA player makes approximately $(4.624)^2 \approx 21$ goals per match. Height lowers the dependent variable $\sqrt{Y}$ by 0.734 per feet such that on average the tallest player score less per match while skill improves the scoring per match such that the variable $\sqrt{Y}$ increases by 7.885 per %. We conclude that skill is the most dominant variable. In order to trust this interpretation, one analyzes three conditions: linearity, homoscedasticity and normality. The left plot shows a scatter diagram of the dependent

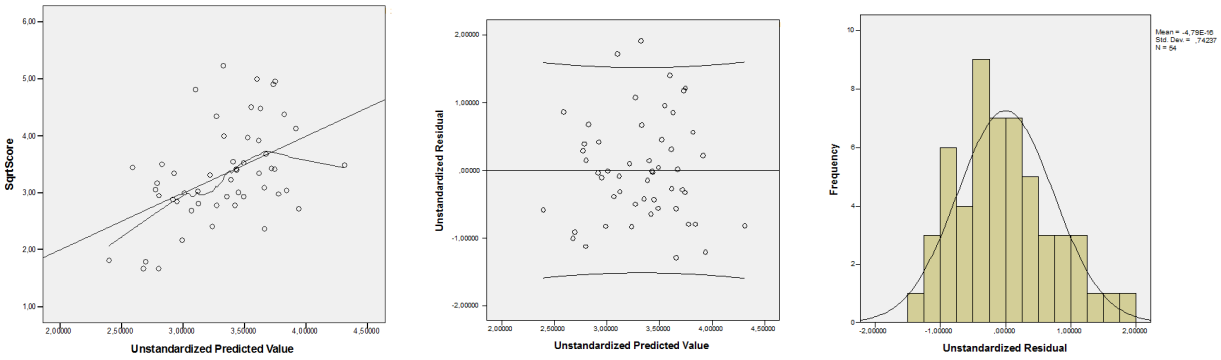

Figure 1.1: Dependent variable versus fits (left plot), Residuals versus fits plot (middle plot) and histogram with normality curve (right plot)

variable $\sqrt{Y}$ versus the fitted values $\underline{\hat{y}}$ to assess linearity. This plot allows to assess the link function $Y = g(\hat{\mu}) + \epsilon$. Therefore the identity function $g(x) = x$ is added to the plot together with a non-parametric LOWESS (see exercises). The LOWESS curve is checked to identify possible functional characteristics. The middle plot shows a scatter diagram of the residuals $\underline{\epsilon}$ versus the fitted values $\underline{\hat{y}}$ to assess homoscedasticity. One adds the mean value and the confidence bounds of the residuals. The homoscedasticity is assessed by inspecting that the confidence interval is sufficiently filled without empty areas. Finally the right plot provides a histogram of the resduals together with the density of the normal distribution with the same mean and variance as the residuals. A visual inspection is crude but sufficient to argue that the conditions are reasonably satisfied. The LOWESS curve in the first plot is close to the straight line except at the end points. Moreover the scatter diagram exhibits an elliptic shape which indicates a linear relationship between fitted and dependent variable values which argues to support linearity. The second plot seems sufficiently filled to argue that homoscedasticity is satisfied. Indeed, 95% of the scatter points should be captured within the confidence bound implying on a sample size of 54 around 2 to 3 points agreeing with the visual observation. Finally the histogram seems to follow the normal density shape approximately such that also the final conditions is acceptable.

Chapter 2

One-way analysis of variance

Consider a random vector $\underline{Y}$ of size n following a multivariate Gaussian distribution $\mathcal{N}(\underline{\mu}, \mathbf{d})$ with $\mathbf{d}$ a diagonal matrix. Assume that the random vector consists of K different multivariate Gaussian distributions such that $(Y_{1j}, Y_{2j}, \ldots, Y_{n_j,j})$ are independently and identically distributed such that Y_{ij} follows a Gaussian distribution $\mathcal{N}(\mu_j, \sigma_j^2)$. Hence, the following linear model holds:

$$Y_{ij} = \beta_0 + \beta_1^{[j]} + \epsilon_{ij}$$

with $j = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n_j$ such that $\mu_j = \beta_0 + \beta_1^{[j]}$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$. If one considers the stacked random vector $\underline{Y} = [Y_{11}, \ldots, Y_{n_1 1}, Y_{12}, \ldots, Y_{n_K K}]'$, one obtains the linear model in matrix format:

$$\underline{Y} = \mathbf{x}\underline{\beta} + \underline{\epsilon}$$

such that the design matrix equals $\mathbf{x} = \begin{bmatrix} 1_n & \underline{e}_1 & \ldots & \underline{e}_K \end{bmatrix}$ with $\underline{e}_j \in \mathbb{R}^n$ consisting of $\sum_{i=1}^{j-1} n_i$ zeros, followed by n_j ones and filled up by zeros once more until dimension n is reached. The choice to parametrize $\mu_j = \beta_0 + \beta_1^{[j]}$ allows having a linear model with an intercept term. This linear model is called the one-way ANOVA model or the K-sample problem.
In this chapter, we study three solutions to analyze the one-way ANOVA model where its main question is to estimate the parameters and their significance. In general one wishes to study that at least two groups have different expected values $\mu_i \neq \mu_j$ or equivalently $\underline{\beta}_1 \neq \underline{0}$. We discriminate the following approaches:

1. Fisher ANOVA: The matrix $\mathbf{d}$ consists of equal entries such that $\sigma_i^2 = \sigma_j^2$ for all groups i, j.

2. Welch ANOVA: The matrix $\mathbf{d}$ consists of possibly different entries over the group levels.

3. Scheffé ANOVA: The expected values μ_j vary polynomially over the group levels.

The Scheffé ANOVA is also called contrast ANOVA because different than the other ANOVA analyses, the technique is not only looking for possible differences among the group expectations but it studies a particular pattern across group levels.

2.1 Fisher's ANOVA

In this version of ANOVA we construct the F-test to test the hypotheses $\mathcal{H}_0 : \underline{\mu} = c.\underline{1}_n$ versus $\mathcal{H}_1 : \underline{\mu} \in \text{span}(\mathbf{x})$. The null-hypothesis implies that the expectation $\mu_i = \mu_j$ for $i, j \in \{1, 2, \ldots, K\}$ is tested versus the alternative where $\mu_i \neq \mu_j$ for at least one couple $i \neq j$. Note that $\dim(\text{span}(\mathbf{x})) = K - 1$ such that one must introduce a constraint to uniquely estimate the coefficients $\underline{\beta}_1 \in \mathbb{R}^K$. Two standard constraints are popular: (i) sum-to-zero or (ii) reference group. The first is popular in ANOVA while the latter is popular in general(ized) linear models. In the first constraint one assumes $\sum_{j=1}^K \beta_1^{[j]} = 0$ while in the latter one takes $\beta_1^{[K]} = 0$.
We first generalize Theorem 1.5 where the F-test is represented by orthogonal projection matrices. Consider a subspace V of $\mathbb{R}^n$ such that the matrix π_V denotes the orthogonal projection of $\mathbb{R}^n$ onto V. From functional analysis and linear algebra, orthogonal projections are well studied which this chapter revisits in the following Lemma.

Lemma 2.1 *Let $V_i \leq \mathbb{R}^n$ with $i \in \{1, 2\}$. Let $\pi_{V_i}(\underline{y}) = argmin_{\underline{x} \in V_i} ||\underline{y} - \underline{x}||^2$ the orthogonal projection operator the the following properties hold:*
1. π_{V_i} is a linear operator whose matrix representation is a symmetrical idempotent with rank equal to the dimension of V_i.
2. Two subspaces V_1, V_2 are orthogonal if and only if the projection matrices satisfy $\pi_{V_1}\pi_{V_2} = \mathbf{0}$.

Proof this result as an excerise. This result allows reformulating Theorem 1.5 with projection matrices forming the foundation of Fisher's ANOVA.

Theorem 2.1 *Let $\underline{Y}$ be a random vector of size n following a multivariate Gaussian distribution with expectation $\underline{\mu}$ and covariance matrix $\sigma^2 I_n$. The F-test to assess the hypothesis $\mathcal{H}_0 : \underline{\mu} \in V_0$ versus $\mathcal{H}_1 : \underline{\mu} \in V$ such that $V = span(\mathbf{x})$ of dimension p and $V_0 \leq V$ of dimension r is given by*

$$F = \frac{\underline{Y}'\pi_{V_0^\perp}\underline{Y}}{\underline{Y}'(I_n - \pi_V)\underline{Y}} \frac{n-p}{p-r}$$

such that $V = V_0 \oplus V_0^\perp$

Proof: Theorem 1.5 provides the characterization of the F-test in the form:

$$F = \frac{n-p}{p-r} \frac{||\underline{Y} - \mathbf{x}\underline{\hat{\beta}}_0||^2 - ||\underline{Y} - \mathbf{x}\underline{\hat{\beta}}_1||^2}{||\underline{Y} - \mathbf{x}\underline{\hat{\beta}}_1||^2}$$

with $\underline{\hat{\beta}}_0$ and $\underline{\hat{\beta}}_1$ are the respective ML estimators w.r.t. $\{\underline{\beta} | \mathbf{x}\underline{\beta} \in V_0\}$ and $\{\underline{\beta} | \mathbf{x}\underline{\beta} \in V\}$. The ML estimators coincide with the orthogonal projects such that $\pi_{V_0}\underline{Y} = \mathbf{x}\underline{\hat{\beta}}_0$ and $\pi_V\underline{Y} = \mathbf{x}\underline{\hat{\beta}}$. This leads to the equivalent representation

$$F = \frac{n-p}{p-r} \frac{\underline{Y}'(I_n - \pi_{V_0})\underline{Y} - \underline{Y}'(I_n - \pi_V)\underline{Y}}{\underline{Y}'(I_n - \pi_V)\underline{Y}} = \frac{n-p}{p-r} \frac{\underline{Y}'(\pi_V - \pi_{V_0})\underline{Y}}{\underline{Y}'(I_n - \pi_V)\underline{Y}}$$

To complete the proof, it remains to show that $\pi_V - \pi_{V_0}$ is the projection of V onto the orthogonal complement of V_0 with respect to V. First, we consider a vector $\underline{x} \in V$

such that the unique orthogonal decomposition holds: $\underline{x} = \pi_{V_0}\underline{x} + (I_n - \pi_{V_0})\underline{x}$. Consider further $\underline{x} = \pi_V \underline{y}$ for a dedicated vector $y \in \mathbb{R}^n$, then the latter decomposition equals: $\pi_V \underline{y} = \pi_{V_0}\pi_V \underline{y} + (I_n - \pi_{V_0})\pi_V \underline{y}$. However $V_0 \leq V$ a nested subspace such that the projections are transitive: $\pi_{V_0}\pi_V \underline{y} = \pi_{V_0}\underline{y}$ which proves the result. □

This theorem leads to Fisher's Analysis of Variance table where the variability measured by the Euclidean norm is partitioned over nested subspaces by virtue of Pythagoras theorem. Consider three nested subspaces: $V_\mu = \text{span}(\underline{1}_n)$, $V_{\tilde{\mathbf{x}}} = \text{span}(\tilde{\mathbf{x}})$ and $V = \text{span}(\mathbf{x})$ such that $V_\mu \leq V_{\tilde{\mathbf{x}}} \leq V$. The Fisher F-test table becomes:

$\mathcal{H}_0$	SS	df	MSS	F
$\underline{\mu} \in V_{\tilde{\mathbf{x}}}$	$\underline{Y}'\pi_{V_\mu}\underline{Y}$	1	$\underline{Y}'\pi_{V_\mu}\underline{Y}$	$\frac{\underline{Y}'\pi_{V_\mu}\underline{Y}}{\underline{Y}'(I_n-\pi_V)\underline{Y}}(n-K)$
$\underline{\mu} \in V_\mu$	$\underline{Y}'\pi_{V_{\tilde{\mathbf{x}}}}\underline{Y}$	$K-1$	$\frac{\underline{Y}'\pi_{V_{\tilde{\mathbf{x}}}}\underline{Y}}{K-1}$	$\frac{\underline{Y}'\pi_{V_{\tilde{\mathbf{x}}}}\underline{Y}}{\underline{Y}'(I_n-\pi_V)\underline{Y}}\frac{n-K}{K-1}$
$\underline{\mu} \in V$	$\underline{Y}'(I_n-\pi_V)\underline{Y}$	$n-K$	$\frac{\underline{Y}'(I_n-\pi_V)\underline{Y}}{n-K}$	

Note that the first null-hypothesis postulates that $\beta_0 = 0$ whereas the second null-hypothesis postulates that $\underline{\beta}_1 = \underline{0}$. Hence, the SS or sum-of-squares measures the loss in information if the null-hypothesis is incorrectly assumed. As a result, if this loss is too high w.r.t. the residual error, the null-hypothesis is considered false. Under the conditions set, one is able to compute the orthogonal projections on the two nested subspaces.

Theorem 2.2 *(Fisher) Consider a multivariate Gaussian random vector $\underline{Y}$ of size n given by $\mathcal{N}(\underline{\mu}, \sigma^2 I_n)$ the the following decomposition holds:*

$$Y_{ij} = \bar{Y} + (\bar{Y}_j - \bar{Y}) + (Y_{ij} - \bar{Y}_j)$$

such that $\pi_\mu(\underline{Y}), \pi_{\tilde{\mathbf{x}}}(\underline{Y})$ and $I_n - \pi_V(\underline{Y})$ are given by vectors whose entry at position $\sum_{m=1}^{j-1} n_m + i$ equals $\bar{Y} = \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n_j} Y_{ij}$, $\bar{Y}_j - \bar{Y}$ with $\bar{Y}_j = \sum_{i=1}^{n_j} Y_{ij}$ and $Y_{ij} - \bar{Y}_j$ respectively.

Proof: We first note that the statistical model $\underline{Y} = \beta_0 + \underline{\epsilon}$ provides the UMVU estimator $\hat{\beta}_0 = \bar{Y}$ such that $\pi_\mu(\underline{Y})$ is a vector whose entries are all equal to $\bar{Y}$. Completely similar the full model provides UMVU estimators such that $\hat{\beta}_0 + \hat{\beta}_1^{[j]} = \bar{Y}_j$ such that $\pi_V(\underline{Y})$ is a vector whose entry at position $\sum_{m=1}^{j-1} n_m + i$ equals $\bar{Y}_j$. Therefore we obtain that $I_n - \pi_V(\underline{Y})$ is a vector whose entry at position $\sum_{m=1}^{j-1} n_m + i$ equals $Y_{ij} - \bar{Y}_j$. The sole candidate to complete the equality results in $\bar{Y}_j - \bar{Y}$ which is the result of $\pi_V - \pi_\mu(\underline{Y}) = \pi_{\tilde{\mathbf{x}}}(\underline{Y})$ ending the proof. □

As a result, the ANOVA table can be explicitly given by:

$\mathcal{H}_0$	SS	df	MSS	F
$\underline{\mu} \in V_{\tilde{\mathbf{x}}}$	$n\bar{Y}^2$	1	$n\bar{Y}^2$	$\frac{nY^2}{\sum_{j=1}^{K}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_j)^2}(n-K)$
$\underline{\mu} \in V_\mu$	$\sum_{j=1}^{K} n_j(\bar{Y}_j - \bar{Y})^2$	$K-1$	$\frac{\sum_{j=1}^{K} n_j(\bar{Y}_j-\bar{Y})^2}{K-1}$	$\frac{\sum_{j=1}^{K} n_j(\bar{Y}_j-\bar{Y})^2}{\sum_{j=1}^{K}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_j)^2}\frac{n-K}{K-1}$
$\underline{\mu} \in V$	$\sum_{j=1}^{K}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$	$n-K$	$\frac{\sum_{j=1}^{K}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_j)^2}{n-K}$	

Example 2.1 *For the multiple sclerosis disease, investigators wish assessing differences between two experimental treatments: immunosuppresive and immunoglobine therapy. Therefore, they consider three groups: the two experimental groups and a placebo group wherein no treatment is given. Consider the patients to be in a similar stage of the MS*

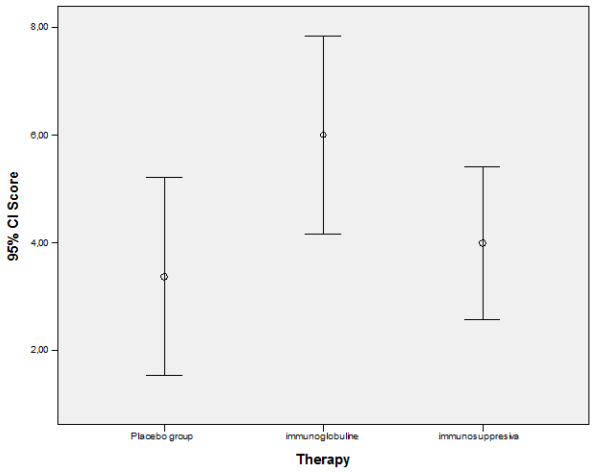

Figure 2.1: Confidence interval analysis for Multiple Sclerosis example

disease, furthermore we assume the patients' condition to be independent of each other. After two weaks of treatment, the patients are examined by a neurologist on checklist of 10 criteria. A score 0 to 10 is granted to each patient such that 0 implies severe stage of MS and 10 no symptoms of the MS disease. Consider 24 MS patients which are blindly and randomly assigned to a treatment group. We perform a one-way analysis of variance to assess differences in expected scores over the different treatment groups with a confidence of 95%.
A preliminary analysis is an approximate Wald analysis suggested by Theorem 1.7 which depicts graphically the mean scores per treatment group and its corresponding 95% confidence bounds. To analyze confidence intervals pairwise, one detects three states: the confidence intervals are disjoint suggesting a strong significant difference (expected p-value$< \alpha$), the confidence intervals intersect but the mean value of a group is not captured by the confidence interval of the other group which is called a weak effect (expected p-value$\approx \alpha$), the confidence intervals intersect and the mean value of a group is captured by the confidence interval of the other group. The confidence intervals therefore suggest:

- *Strong significant difference: no pairwise comparisons*
- *Weak significant difference: placebo vs immunoglobine and immunoglobine vs immunosuppresiva*

Next, we confirm the preliminary findings through a Fisher's ANOVA analysis. Note that the equal size of the confidence intervals over the different groups satisfy the equality of variances among the groups. The ANOVA table is revealed in rows 2,3 and 4. The

Tests of Between-Subjects Effects

Dependent Variable: Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	30,083 [a]	2	15,042	3,595	,045
Intercept	477,042	1	477,042	114,001	,000
Therapy	30,083	2	15,042	3,595	,045
Error	87,875	21	4,185		
Total	595,000	24			
Corrected Total	117,958	23			

a. R Squared = ,255 (Adjusted R Squared = ,184)

Figure 2.2: Fisher's ANOVA table for Multiple Sclerosis table

respective F-tests detect a strongly significant intercept ($p < 0.01$) and a weakly significant group effect ($p < 0.05$). This conclusion is in line with the preliminary analysis which suggested a weak effect if detected.

2.2 Welch's ANOVA

The approach of Fisher assumes equality of variances across the different groups such that the underlying linear model is called homoscedastic. It is reasonable to drop the homoscedasticity condition such that the variances may change over the different groups. Considering the same notations as previously, we obtain the linear model

$$Y_{ij} = \mu_j + \epsilon_{ij}$$

such that ϵ_{ij} are independently normally distributed with zero-mean and variances σ_j^2. Since a normal distribution is closed under linear transformations, we may transform the Welch model to the Fisher ANOVA model.

Lemma 2.2 *Consider a multivariate Gaussian vector $\underline{Y} \sim \mathcal{N}(\underline{0}, \Sigma)$ of length n such that the covariance matrix Σ is of full rank, then the random vector $\underline{Z} = \mathbf{c}\underline{Y} \sim \mathcal{N}(\underline{0}, I_n)$ if $\mathbf{c}$ is the inverse of the Cholesky factor of Σ.*

Proof: It is trivial to see that the distribution of $\underline{Z}$ is a multivariate normal distribution with expectation $\underline{0}$. Consider the Cholesky decomposition $\Sigma = \boldsymbol{\ell}\boldsymbol{\ell}'$ such that $\boldsymbol{\ell}$ is a lower triangular matrix of size $n \times n$. The matrix $\boldsymbol{\ell}$ has an inverse denoted by $\mathbf{c}$. The covariance matrix of random vector $\underline{Z}$ is given by:

$$\begin{aligned} \text{Cov}(\underline{Z}) &= \mathbb{E}[\underline{Z}\underline{Z}'] \\ &= \mathbf{c}\mathbb{E}[\underline{Y}\underline{Y}']\mathbf{c}' \\ &= \mathbf{c}\Sigma\mathbf{c}' = I_n \end{aligned}$$

□

It is easy to see that the matrix $\mathbf{c}$, the Welch ANOVA requires is given by the diagonal matrix such that the entry $[\mathbf{c}]_{mm} = \frac{1}{\sigma_j}$ if $m = \sum_{k=1}^{j-1} n_k + i$ for $1 \leq i \leq n_j$. By virtue of the matrix $\mathbf{c}$, we can consider the Welch inner-product in $\mathbb{R}^n$ given by $< \underline{x}, \underline{y} >_W = \underline{x}'\mathbf{c}'\mathbf{c}\underline{y} = \underline{x}'\mathbf{d}^{-1}\underline{y}$ and the induced Welch norm $||\underline{x}||_W = \sqrt{\underline{x}'\mathbf{d}^{-1}\underline{x}} = \sqrt{\sum_{j=1}^{K} \frac{\sum_{i=1}^{n_j} x_{ij}^2}{\sigma_j^2}}$ such that $\mathbf{x}' = [x_{11}, x_{21}, \ldots, x_{n_1 1}, x_{12}, \ldots, x_{1n_K}, \ldots, x_{n_K K}]$. The vector space $(\mathbb{R}^n, ||.||_W)$ allows describing the multivariate distribution of the random vector $\underline{Y}$:

$$f(\underline{Y}|\underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{j=1}^{K} \sigma_j^{n_j}} \exp\left(-\frac{1}{2}||\underline{Y} - \underline{\mu}||_W^2\right)$$

Welch operates initially under the assumption that the covariance matrix $\mathbf{d}$ is known such that the norm is characterized. Hence, we can generalize Theorem 2.1 in the Welch vector space. Indeed, consider the nested subspace $V_0 \leq V = \text{span}(\mathbf{x})$ with respective dimensions $r < p$. The LR-test is therefore given by

$$LR = \frac{argmin_{\underline{\mu} \in V} \exp\left(-\frac{1}{2}||\underline{Y} - \underline{\mu}||_W^2\right)}{argmin_{\underline{\mu} \in V_0} \exp\left(-\frac{1}{2}||\underline{Y} - \underline{\mu}||_W^2\right)}$$

The application of the monotonic mapping $x \mapsto \frac{2}{p-r}\log(x)$ results in a test statistic:

$$W(\underline{Y}) = \frac{1}{p-r}||(\pi_V^W - \pi_{V_0}^W)\underline{Y}||_W^2 = \frac{1}{p-r}||\pi_{V_0^\perp}^W \underline{Y}||_W^2$$

such that the operator $\pi_{V_0}^W$ denotes the orthogonal projection onto the space V accordingly to the Welch inner-product. We first compute the Welch equivalent to the 2-sample ($K = 2$) t-test or Welch t-test.

Theorem 2.3 *(Welch t-test) Consider independent Gaussian random vectors $\underline{Y}_1, \underline{Y}_2$ of sizes n_1, n_2 such that their distribution is given by $\underline{Y}_i \sim \mathcal{N}(\mu_i \underline{1}_{n_i}, \sigma_i^2 I_{n_i})$ respectively. The LR test in case the variances σ_1^2, σ_2^2 are known to test the null-hypothesis $\mu_1 = \mu_2$ versus $\mu_1 \neq \mu_2$ is given by*

$$\tilde{T}_n^W(\underline{Y}) = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Proof: We consider the subspace $V_0 = \text{span}(\underline{1}_n)$ and $V = \text{span}(\mathbf{x})$ where $\mathbf{x} = [\underline{1}, \underline{e}_1, \underline{e}_2]$. We compute $\pi_{V_0}^W(\underline{Y}) = argmin_{\mu\in\mathbb{R}}||\underline{Y} - \mu\underline{1}_n||_W^2 = \frac{\frac{n_1\bar{Y}_1}{\sigma_1^2} + \frac{n_2\bar{Y}_2}{\sigma_2^2}}{\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2}}\underline{1}_n$. Equivalently, we find that $\pi_V^W(\underline{Y}) = [\bar{Y}_1\underline{1}'_{n_1}, \bar{Y}_2\underline{1}'_{n_2}]'$. Next, we compute $(\pi_V^W - \pi_{V_0}^W)\underline{Y} = \frac{1}{\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2}}[(\bar{Y}_1 - \bar{Y}_2)\frac{n_2}{\sigma_2^2}\underline{1}'_{n_1}, (\bar{Y}_2 - \bar{Y}_1)\frac{n_1}{\sigma_1^2}\underline{1}'_{n_2}]'$. Finally, we compute the norm $||(\pi_V^W - \pi_{V_0}^W)\underline{Y}||_W^2 = (\bar{Y}_1 - \bar{Y}_2)^2\frac{\frac{n_1 n_2}{\sigma_1^2\sigma_2^2}}{\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2}}$ which completes the proof. □

The LR test of Welch depends on the unknown parameters σ_1^2 and σ_2^2. One way to circumvent this problem is replacing these remaining parameters by its UMVU estimates leading to the Welch t-test:

$$T_n^W(\underline{Y}) = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

An additional problem arises that the distribution is not a t-distribution under the null-hypothesis since the denominator is a weighted sum of independent Chi-square random variables. However Satterthwaite suggested to approximate this distribution by a t-distribution with a non-integer number of degrees of freedom denoting the efficiency loss.

Theorem 2.4 *(Satterthwaite) Under $\mathcal{H}_0$ the distribution of $T_n^W(\underline{Y})$ is approximately given by*

$$T_n^W(\underline{Y}) \overset{d}{=} \frac{\mathcal{N}(0,1)}{\sqrt{\lambda_1\chi_{n_1-1}^2 + \lambda_2\chi_{n_2-1}^2}} \approx t_\nu$$

with $\lambda_i = \frac{\sigma_i^2}{n_i(n_i-1)}\frac{1}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ and $\nu = \frac{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}}$.

Proof: The values corresponding to the weights λ_i are a straightforward calculation. The Satterthwaite approximation is now obtained as

$$\lambda_1\chi_{n_1-1}^2 + \lambda_2\chi_{n_2-1}^2 \approx r\chi_\nu^2$$

such that the first two moments are equal. Considering the first moment, we obtain that

$$\lambda_1(n_1 - 1) + \lambda_2(n_2 - 1) = r\nu$$

Thus, we obtain that $r\nu = 1$ such that $r = \frac{1}{\nu}$. Next, we consider the variance leading to

$$\lambda_1^2(n_1 - 1) + \lambda_2^2(n_2 - 1) = \frac{1}{\nu}$$

Thus, we find that $\frac{1}{\nu} = \left(\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}\right) \frac{1}{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}$ □

In practice, the degrees of freedom are computed through the UMVU estimates again. Note that this Welch t-test cannot be optimal as for equal variances $\sigma_1^2 = \sigma_2^2$, the test does not boil down to the classical two sample t-test. The lack of an optimal two sample t-test in case of different variances is known as the Behrens-Fisher problem for which only approximate solutions exists whereto Welch's test is most popular and used test. The next chapter deals with diagnostic tests wherein we study a test to verify the homogeneity of variance assumption.

Example 2.2 *A frequent complication after a heart bypass surgery is a postoperative lung dysfunction due to a reduced number of large red blood cells. Patients after massive surgery are required to take supplements of folic acid. In this example, we study the difference in folic acid concentration (μg) among patients who take supplements against patients who take a specific diet. A small dataset of 15 cases are collected after one week since surgery where 7 patients do not take supplements whereas 8 patients take supplements. To assess visually the equal variance condition, we inspect the 95%-confidence interval*

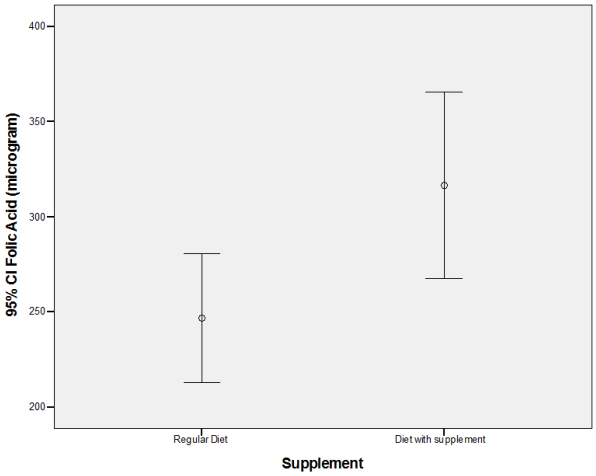

Figure 2.3: 95% Folic Acid Confidence intervals for the mean

of the mean per group. Although the mean folic acid value of the patient group taking supplements is higher than in the regular diet group, the uncertainty is also higher such that potentially differences between both groups are not significant. We observe that both intervals do not capture each others mean value although these intervals intersect. This typically suggests a p-value within the interval $[\alpha/5, \alpha]$. If both intervals are completely disjoint, one expects a p-value in the interval $[0, \alpha/5]$. The first scenario is called a weak effect, whereas the last scenario is called a strong effect. In the reported t-test table, one observes two rows of results. The first row denotes the classical t-test assuming equal variances, while the second row is the Welch correction for violations of equal variances. Levene's test (see next chapter) tests the equality of variances against violations of this equality. The reported p-value of 4.3% suggests that the violation of equality of variances

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Folic Acid (microgram)	Equal variances assumed	5,003	,043	-2,708	13	,018	-69,768	25,763	-125,425	-14,111
	Equal variances not assumed			-2,794	11,882	,016	-69,768	24,966	-124,225	-15,311

Figure 2.4: Two sample independent t-tests

is weakly significant such that Welch's test is preferred. Note that the numerator for the classical and Welch's t-test are equal such that their differences are observed for the degrees of freedom and the denominator of the test statistic reported as standard error. The p-value to assess the differences in mean value equals 1.6% which is as expected larger than the strong significance of 1%. One may expect in 97.5% of the repeated experiments that the supplement versus regular diet increased folic acid concentration by at least 15.3µg. Note that although this effect is statistically significant, an expert (in this case a surgeon) should post-assess the clinical or practical significance of this result.

The Welch two sample t-test can be generalized to K groups by application of $W(\underline{Y})$. We obtain that assess the null hypothesis $\underline{\mu} \in \text{span}(\underline{1}_n)$ versus $\underline{\mu} \in \text{span}(\mathbf{x})$. This leads to the Welch one-way ANOVA test statistic under known variances:

$$F_n^W(\underline{Y}) = \frac{\sum_{j=1}^K \frac{n_j}{\sigma_j^2} \left(\sum_{m=1}^K \frac{n_m}{\sigma_m^2} (\bar{Y}_j - \bar{Y}_m) \right)^2}{\left(\sum_{m=1}^K \frac{n_m}{\sigma_m^2} \right)^2}$$

Without explicit calculations, this test is in practice used by replacing the unknown nuisance parameters σ_j^2 by their respective UMVU estimators leading to a test statistic which is following under the null hypothesis an approximate F distribution.

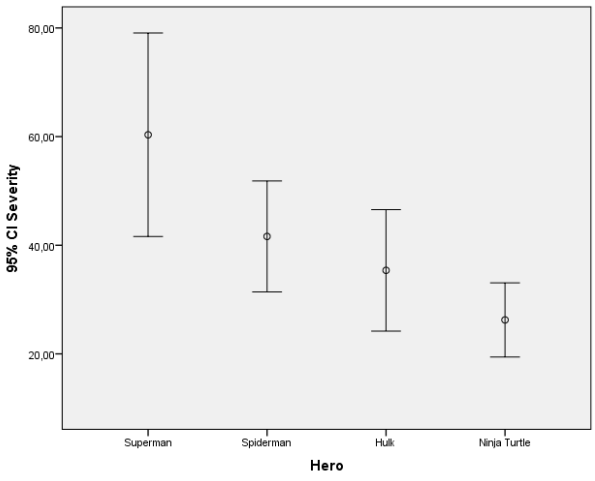

Figure 2.5: 95% Confidence intervals of the mean per hero costume

Example 2.3 *Children dressing up as super heroes often take more risks when playing through assuming the role of their hero. As a result, hospitals register more accidents with children playing if they are dressed in a super hero costume. In this example, one assess*

differences in severity among four different super hero costumes: Superman, Spiderman, the Hulk and a Ninja Turtle. The severity is measured as a discrete score ranging from 0 (not harmed) to 100 (death).
A preliminary analysis to assess possible differences of the mean severity score over the costume groups is provided through confidence intervals. These graphical representations of the approximate Wald tests indicate that no significant difference is to be expected between the Spiderman and Hulk costume group since the respective sample means are in both confidence intervals. The same is to be expected between the Hulk and Ninja turtle group. A weak effect ($\frac{\alpha}{5} < p - value < \alpha$) is expected between the Superman and Spiderman costume as the mean value of the Superman group is not reachable within the confidence bounds of the Spiderman group whereas also the mean value of the Spiderman group is at just outside of the confidence bounds of the Superman group. The effect is weak if any since both confidence bounds intersect significantly. A similar weak effect is noted between the Spiderman and Ninja Turtle group. Clearly a strong effect ($p - value < \frac{\alpha}{5}$) is observed between Superman and the Ninja Turtle group. Since at least one strong effect is concluded from the preliminary analysis, one expects the ANOVA test to detect such a difference. One may opt for the Welch ANOVA due to the increased size of the confidence interval of the Superman group compared to the other costume groups.

Robust Tests of Equality of Means

Severity

	Statistic [a]	df1	df2	Sig.
Welch	7,100	3	13,018	,005

a. Asymptotically F distributed.

Figure 2.6: Welch ANOVA F-test for differences of the mean among hero costumes

The Welch F-test confirms the preliminary analysis indicating a strong proof in favorite of the alternative hypothesis that difference among hero costumes exist ($p - value < 1\%$). Fishers ANOVA confirms the alternative hypothesis a bit liberal with a $p - value < 0.0483\%$. Welch analysis loses power due to an efficiency loss measured in degrees of freedom, it provides a stronger argument that one is not detecting a false positive due to enlarged confidence bounds. Indeed if one group holds a larger confidence bounds compared to others, Fisher's ANOVA - through applying the pooled variance - will reduce this size which may increase significance falsely.

2.3 Scheffé contrast ANOVA

Different than the approach of Fisher and Welch, Scheffé proposes an alternative analysis. Scheffé argues that one is typically not interested in detecting that at least one difference exist across the different catogrical factor levels. Rather than detecting a difference, Scheffé proposes to detect a specific trend, how the expected values varies as a function of the factor levels. Scheffé wishes to assess whether:

$$\mathcal{H}_0 : \mu_j = P_m(f_j) \text{ versus } \mathcal{H}_1 : \mu_j \neq P_m(f_j) \text{ for all levels } j$$

with f_j the categorical factor level j and $P_m(.)$ a polynomial of degree m. An immediate problem is that one does not know the specific polynomial to test. The following lemma forms the basis of the contrast hypotheses. A way to avoid this problem is through

differentation. A polynomial of degree m is reduced to zero after taking the derivative of order $m+1$. Hence the hypotheses can be rewritten in a better format:

$$\mathcal{H}_0 : \frac{\partial^{m+1}}{\partial f^{m+1}}\mu(f) = 0 \text{ versus } \mathcal{H}_1 : \frac{\partial^{m+1}}{\partial f^{m+1}}\mu(f) \neq 0$$

where $\mu(f)$ is the expectation as a function of the factor levels. This idea must be discretized as the factor levels are not a continuous scale which is done through the next lemma while replacing the derivative by finite differences or the backshift operator $z^{-1}(y_i) = y_{i-1}$.

Lemma 2.3 *Let (x_i, y_i) for $i = 0, 1, 2, \ldots, m$ with $m \geq 2$ be a sequence of points such that $x_i, i = 0, \ldots, m$ are forming an equidistant grid within a predetermined interval of $\mathbb{R}$. If the points are solutions of the polynomial equation $y = P_{k-1}(x)$ for a $k \in \{1, 2 \ldots, m\}$ then it holds that $(1 - z^{-1})^k(y_m) = 0$.*

Proof: The proof is established through induction. Let us consider the base case: $m = 2$. Consider the grid points $(x_0, y_0), (x_1, y_1)$ and (x_2, y_2). Let $k = 2$ for instance such that $y_i = b + ax_i$ with $a, b \in \mathbb{R}$. It follows that $y_i = y_{i-1} + a\Delta$ with $\Delta = x_i - x_{i-1}$. It follows that $(1 - z^{-1})(y_2) = a\Delta$ which equals 0 if $a = 0$ or if dealing with a polynomial of degree 0. Further, $(1 - z^{-1})^2(y_2) = 0$ which completes the base case.
The induction hypothesis postulates the validity of the lemma up to m. For the induction step, we consider the case $m + 1$ and $k \leq m + 1$

$$\begin{aligned}(1 - z^{-1})^m(1 - z^{-1})y_i &= (1 - z^{-1})^m(y_i - y_{i-1}) \\ &= (1 - z^{-1})^m \left[\sum_{r=0}^{k-1} c_r x_i^r - \sum_{r=0}^{k-1} c_r x_{i-1}^r\right]\end{aligned}$$

Note that $x_i^r = (x_{i-1} + \Delta)^r = \sum_{n=0}^{r} \begin{pmatrix} r \\ n \end{pmatrix} x_{i-1}^n \Delta^{r-n}$ such that one concludes that $x_i^{k-1} - x_{i-1}^{k-1} = \sum_{n=0}^{k-2} \begin{pmatrix} k-1 \\ n \end{pmatrix} x_{i-1}^n \Delta^{k-1-n}$ and hence $\left[\sum_{r=0}^{k-1} c_r x_i^r - \sum_{r=0}^{k-1} c_r x_{i-1}^r\right]$ is a polynomial in x_{i-1} of degree $k - 2$ such that the induction hypothesis ($k - 2 \leq m - 1$) implies that:

$$(1 - z^{-1})^m \left[\sum_{r=0}^{k} c_r x_i^r - \sum_{r=0}^{k} c_r x_{i-1}^r\right] = 0$$

□

The simplest situation considers two groups such that one wishes to falsify a constant trend versus a linear trend leading to the contrast (coinciding with a simple) hypothesis:

$$\mathcal{H}_0 : \mu_1 = \mu_2 \text{ versus } \mathcal{H}_1 : \mu_1 \neq \mu_2$$

These hypothesis satisfy the lemma through $\mathcal{H}_0 : \mu_1 - \mu_2 = (1 - z^{-1})\mu_2 = 0$. A slightly more difficult situation aims at rejecting a linear trend over three groups leading to

$$\mathcal{H}_0 : \mu_i = a + bi \text{ versus } \mathcal{H}_1 : \mu_i \neq a + bi \text{ for all} i \in 1, 2, 3$$

By virtue of the previous lemma, the hypothesis are equivalent to:

$$\mathcal{H}_0 : (1 - z^{-1})\mu_3 = \mu_1 - 2\mu_2 + \mu_3 = 0 \text{ versus } \mathcal{H}_1 : \mu_1 - 2\mu_2 + \mu_3 \neq 0$$

This leads to Scheffé's definition of a contrast hypothesis.

Definition 2.1 *Consider a multivariate Gaussian random vector $\underline{Y}$ satisfying the Welch ANOVA model. A contrast hypothesis is given by $\mathcal{H}_0 : \sum_{i=1}^K c_i\mu_i = 0$ versus $\mathcal{H}_1 : \sum_{i=1}^K c_i\mu_1 \neq 0$ where the vector $\underline{c}$ is called the contrast vector if $\underline{c}1_K = 0$. Note that contrast vectors are typically represented by a row vector.*

Testing for contrasts is more powerful than Fisher's ANOVA because a specific structure is analyzed. A polynomial contrast can be tested through the previous lemma but its representation is only unique if the analyzed polynomial trend one wishes to examine is of degree $K-1$.

Example 2.4 *Consider three groups such that the possible polynomial trends among the three expected values over the groups which can be detected are: parabola, straight line or constant. The null hypothesis for a straight line would hold a contrast vector $\underline{c}_1 = [1,-2,1]$. To detect that one deals with a constant over the group levels, three contrast hypotheses are suited with vectors $\underline{c}_0$: $[1,-1,0],[1,0,-1],[0,1,-1]$. In practice the contrast matrix $\mathbf{c} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix}$ is orthogonalized through Gram-Schmidt by keeping the contrast vector for the highest degree invariant. This procedure provides the polynomial contrast matrix for three groups given by $\mathbf{c} = \begin{bmatrix} 1 & 0 & -1 \\ 1 & -2 & 1 \end{bmatrix}$.*

Example 2.5 *Consider a one-way ANOVA with four groups such that the possible polynomial trends how the group expectations vary over the factorial levels is accordingly to a constant, linear, quadratic or cubic function. The most simple contrast matrix is then $\mathbf{c} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ -1 & 3 & -3 & 1 \end{bmatrix}$ which leads to the orthogonal contrast matrix $\mathbf{c} = \begin{bmatrix} -3 & -1 & 1 & 3 \\ -1 & 1 & 1 & -1 \\ -1 & 3 & -3 & 1 \end{bmatrix}$*

Now, we can generalize the t-test between two groups to a t-test for contrast hypotheses.

Theorem 2.5 *Consider a multivariate Gaussian vector $\underline{Y}$ of size n satisfying Fisher's ANOVA model. Consider the contrast hypothesis $\mathcal{H}_0 : \underline{c}\underline{\mu} = 0$ versus $\mathcal{H}_1 : \underline{c}\underline{\mu} \neq 0$ with contrast vector $\underline{c}$. The LR-test for this contrast hypothesis is equivalent to*

$$T = \frac{\sum_{j=1}^K c_j \bar{Y}_j}{\sqrt{\hat{\sigma}^2 \sum_{j=1}^K \frac{c_j^2}{n_j}}}$$

with $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{j=1}^K n_j \hat{\sigma}_j^2$.

Proof: We start by considering the subspace $V_0 = \{\underline{y} \in \mathbb{R}^n | \underline{y} = \mathbf{x}\underline{\beta} \text{ such that } \underline{c}\underline{\beta} = 0\} \leq V \leq \mathbb{R}^n$. We can apply Theorem 2.1 to obtain a dedicated F-test. We compute $\pi_{V_0}(\underline{Y})$ as a minimization problem:

$$\begin{aligned} \pi_{V_0}(\underline{Y}) &= \underset{\underline{\mu} \in \text{span}(\mathbf{x})}{argmin} ||\underline{Y} - \underline{\mu}||^2 \text{ subject to } \underline{c}\underline{\mu} = 0 \\ &= \underset{\substack{\underline{\mu} \in \text{span}(\mathbf{x}) \\ \lambda \in \mathbb{R}}}{argmin} ||\underline{Y} - \underline{\mu}||^2 + \lambda \underline{c}\underline{\mu} \end{aligned}$$

The partial derivative w.r.t. μ_j leads to $-2(\bar{Y}_j - \mu_j)n_j + \lambda c_j$ such that $\pi_{V_0}(\underline{Y})$ is a vector with entry at index $i + \sum_{m=1}^{j-1} n_m$ is given by $\underline{Y}_j - \frac{\lambda c_j}{2n_j}$. As a result, we obtain the vector $\pi_V - \pi_{V_0}(\underline{Y})$ with entry at the same index given by $\frac{\lambda c_j}{2n_j}$. We can compute the value λ since the constraint $\underline{c\mu} = 0$ leads to $\lambda = \frac{2\sum_{j=1}^{K} c_j \bar{Y}_j}{\sum_{j=1}^{K} \frac{c_j^2}{n_j}}$. As a result, we find that $\underline{Y}(\pi_V - \pi'_{V_0})\underline{Y} = \sum_{j=1}^{K} \frac{\lambda^2 c_j^2}{4n_j} = \frac{(\sum_{j=1}^{K} c_j \bar{Y}_j)^2}{\sum_{j=1}^{K} \frac{c_j^2}{n_j}}$. A straightforward application of Theorem 2.1 proceeded by taking the square root completes the proof. □

Example 2.6 *We retake Example 2.3 but instead of detecting differences in injury severity among different costume types, one want to detect specific contrasts. The first contrast is to see whether children dressed up in a flying super hero costume are harmed more seriously than children dressing up in a non-flying super hero costume. A second contrast is to analyze within the flying and non-flying super hero group differences. Figure 2.5 shows the confidence intervals for the mean of the severity score per super hero costume. This allows inspecting the second and third contrast. The second contrast requires the comparison between the two left boxplots, while the third contrast is studied in the two right boxplots. Similar to Example 2.3, the second contrast is expected to reveal a weak effect, while the third contrast is expected to be insignificant. The error bars of Figure 2.7 shows*

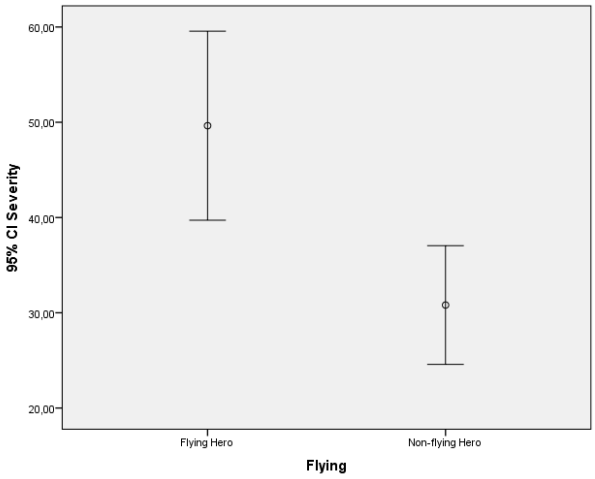

Figure 2.7: 95% Confidence intervals of the mean for flying and non-flying super hero costumes

that one may expect a strongly significant difference between the severity scores induced by wearing a costume of a flying super hero versus a non-flying super hero. Indeed, it is clear the confidence intervals are completely disjoint. As such, we discriminate three alternative contrast hypotheses:

$$\begin{aligned} \mathcal{H}_1 &: \mu_1 + \mu_2 - \mu_3 - \mu_4 \neq 0 \\ \mathcal{H}_1 &: \mu_1 - \mu_2 \neq 0 \\ \mathcal{H}_1 &: \mu_3 - \mu_4 \neq 0 \end{aligned}$$

We first analyze the polynomial trend revealed in Figure 2.5. The highest degree of the polynomial is cubic where all lower degrees are tested simultaneously. The polynomial trend is selected exhibiting the lowest p-value. The contrast ANOVA reported in Figure

ANOVA

Severity

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		4180,617	3	1393,539	8,317	,000
	Linear Term	Unweighted	4094,696	1	4094,696	24,437	,000
		Weighted	3927,138	1	3927,138	23,437	,000
		Deviation	253,479	2	126,740	,756	,479
	Quadratic Term	Unweighted	169,551	1	169,551	1,012	,324
		Weighted	160,976	1	160,976	,961	,336
		Deviation	92,503	1	92,503	,552	,464
	Cubic Term	Unweighted	92,503	1	92,503	,552	,464
		Weighted	92,503	1	92,503	,552	,464
Within Groups			4356,583	26	167,561		
Total			8537,200	29			

Figure 2.8: 95% Confidence intervals of the mean for flying and non-flying super hero costumes

2.8 reveals that the linear trend is strongly proven with a p-value ($p < 0.01$). Note that due to the fact that Example 2.3 revealed that the homogeneity of variances is violated, one preferably inspects the Weighted test which applies a Welch correction to the contrast t-test of Theorem 2.5. No significant proof is found supporting high degree polynomial patterns. Confronting this observation to the error bar diagram of Figure 2.7, we observe a linearly decreasing severity score as a function of the super hero costume. The Scheffé analysis therefore confirms that repeating the experiment will probably result in a linear trend for the severity score as a function of the super hero costumes which pinpoints that the Ninja Turtle costume is the safest, while the Superman costume is the most dangerous costume to wear. Finally, the three contrast hypotheses can be analyzed for

Contrast Coefficients

	Hero			
Contrast	Superman	Spiderman	Hulk	Ninja Turtle
1	1	1	-1	-1
2	1	-1	0	0
3	0	0	1	-1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Severity	Assume equal variances	1	40,3333	9,52692	4,234	26	,000
		2	18,7083	6,99085	2,676	26	,013
		3	9,1250	6,47227	1,410	26	,170
	Does not assume equal variances	1	40,3333	10,12200	3,985	15,097	,001
		2	18,7083	8,47064	2,209	8,388	,057
		3	9,1250	5,54104	1,647	11,568	,126

Figure 2.9: Contrast Matrix and associated contrast t-tests

their significance. Since the width of the confidence intervals as revealed in Figure **??** *are unequal, it is advisable to apply the Welch corrected contrast t-tests. As a result, we find that only the first contrast discriminating flying and non-flying super heroes is strongly significant. The other two contrasts are insignificant. Note that assuming equality of variances also identifies contrast 2 as a significant effect. However its p-value does not reflect the behavior of the confidence interval analysis in the error bars. The confidence intervals suggest a p-value which is closer to the significance threshold 5% rather than the strong significance level 1%. This further motivates the use of the Welch correct contrast tests.*

Example 2.7 *Glaucoma is an eye disease affecting the optic nerves leading to limitation of vision. The symptoms are a direct result of an increased pressure in the eye. Normal pressure in the eye is typically for 99% of the population in the range 10-21 mmHg. In this dataset a total of 150 patients suffering from Glaucoma are examined with 34 men and 116 women since the disease is 4 times as frequent among women than among men.*

Patients are divided among three treatment groups to investigate improvements to vision as a function of nutrition supplements. One group receives an increased supply of caffeine, one group receives an increased supply of vitamins and a third group is put on a diet acting as a placebo without expectations that vision would improve. The following research question is tackled: Do the treatments have an effect on the eye pressure? We

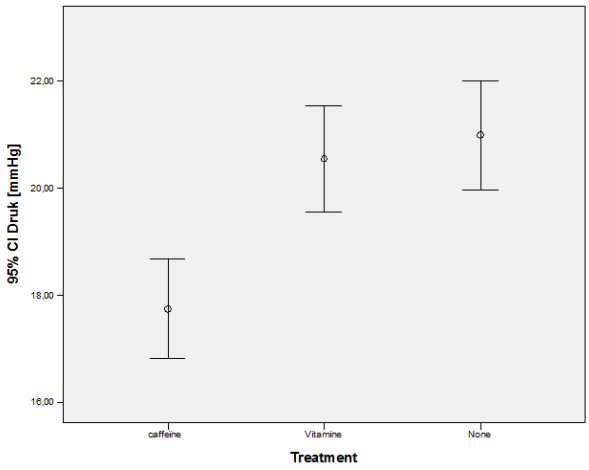

Figure 2.10: 95% Confidence intervals for the mean of eye pressure as a function of food supplements

first examine preliminarily the 95% error bars of the eye pressure as a function the supplement groups. This plot reveals that caffeine lower the pressure, compared to the placebo group significantly. This significance is not observed for what vitamin supplements are concerned. Furthermore the plot reveals a monotonically increase in eye pressure over the different treatment groups. The Scheffé contrast ANOVA shows that considering the confidence intervals, the best polynomial trend is identified to be linear. As such, we conclude that sufficient proof is found in the data that caffeine minimizes the eye pressure w.r.t. placebo and vitamin supplements.

ANOVA

Druk [mmHg]

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		308,586	2	154,293	12,944	,000
	Linear Term	Contrast	262,445	1	262,445	22,017	,000
		Deviation	46,141	1	46,141	3,871	,051
	Quadratic Term	Contrast	46,141	1	46,141	3,871	,051
Within Groups			1752,276	147	11,920		
Total			2060,863	149			

Figure 2.11: Scheffé ANOVA contrast table

Chapter 3

Multi-way Analysis of Variance

In this chapter Fisher's one-way ANOVA is generalized to a more general design where we discriminate between randomized block designs and interaction models. Finally, during the exercises we investigate analysis of covariance (ANCOVA) where also numerical instead of only categorical regressions can be used.
In the previous chapter, we dealt with contrast tests. For complicated designs, a priori hypotheses are often difficult to isolate such that one often wishes to test all possible pairwise comparisons. For this objective we introduce Tukey's t-test. Finally, we handle a residual analysis to verify the GLM conditions and propose heuristic procedures to circumvent departures from the assumptions.

3.1 Randomized block designs

A critical point where a one-way ANOVA fails is if the residual error $\underline{\epsilon} = (I_n - \pi_{\mathbf{x}})\underline{Y}$ is dominant in the sense that it captures a large MSS. This is potentially the case if latent or unobserved categorical variables are missing in the model and hence present in the residual error. A way to reduce the residual error is to consider block designs.

Definition 3.1 *Consider a Gaussian random vector $\underline{Y}$ of size n such that its covariance matrix satisfies $\sigma^2 I_n$ Consider M categorical variables F_m with $m \in \{1, 2, \ldots, M\}$ each with K_m levels respectively. The categorical variables are called factors. The random vector $\underline{Y}$ satisfies a block design if each observation y_i belongs to a unique set of levels, one per factor. If the assignment of the individual factor levels is random, the block design is called randomized.*

Example 3.1 *A grocery store claims to be the cheapest within a competition radius of 5 km. Within that area 3 competitors are found: Colruyt, Delhaize, Match and Carrefour. To test the claim 7 joint products are randomly selected to assess their price. The linear model therefore becomes:*

$$Y_{ij} = \mu + F_i^{[1]} + F_j^{[2]} + \epsilon_{ij}$$

assuming that $\sum_i F_i^{[j]} = 0$ for all j. Although one is only interested in $\underline{F}^{[1]}$ the price depends on the type of product. By disregarding the type of product the residual's variance would blow up. Hence product is a confounder which stabilizes the residuals.

Definition 3.2 *Consider a block design then the matrix $\tilde{\mathbf{x}}_m$ denotes the columns of the regression matrix associated to factor m. The full regression matrix is given by $\mathbf{x} =$*

$[\underline{1}, \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_M]$ *with parameter vector* $\underline{F} = [\mu, F_1^{[1]}, \ldots, F_{K_1}^{[1]}, \ldots, F_1^{[M]}, \ldots, F_{K_M}^{[M]}]'$. *The block design is a M-way ANOVA if the matrix* $\mathbf{x}$ *is of rank* $n - M$.

Next we can generalize Fisher's one-way ANOVA to the multi-way case. Consider the vector space $V_{\tilde{\mathbf{x}}_m} = \text{span}(\tilde{\mathbf{x}}_m)$. This automatically leads to the multi-way ANOVA table by virtue of Theorem 2.1:

$\mathcal{H}_0$	SS	df	MSS	F
$\underline{\mu} \in V_\mu^\perp$	$\underline{Y}'\pi_{V_\mu}\underline{Y}$	1	$\underline{Y}'\pi_{V_\mu}\underline{Y}$	$\frac{\underline{Y}'\pi_{V_\mu}\underline{Y}\left(n-\sum_{m=1}^M (K_m-1)-1\right)}{\underline{Y}'(I_n-\pi_V)\underline{Y}}$
$\underline{\mu} \in V_{\tilde{\mathbf{x}}_m}^\perp$	$\underline{Y}'\pi_{V_{\tilde{\mathbf{x}}_m}}\underline{Y}$	$K_m - 1$	$\frac{\underline{Y}'\pi_{V_{\tilde{\mathbf{x}}_m}}\underline{Y}}{K_m-1}$	$\frac{\underline{Y}'\pi_{V_{\tilde{\mathbf{x}}_m}}\underline{Y}}{\underline{Y}'(I_n-\pi_V)\underline{Y}} \frac{\left(n-\sum_{m=1}^M (K_m-1)-1\right)}{K_m-1}$
$\underline{\mu} \in V$	$\underline{Y}'(I_n - \pi_V)\underline{Y}$	$\left(n-\sum_{m=1}^M (K_m-1)-1\right)$	$\frac{\underline{Y}'(I_n-\pi_V)\underline{Y}}{\left(n-\sum_{m=1}^M (K_m-1)-1\right)}$	

This table can be filled out explicitly through generalizing Theorem 2.2 to the M-way ANOVA case. It turns out that a block design M-way ANOVA boils down to different one-way ANOVAs. Consider the following notation

$$Y_{i_1,i_2,\ldots,i_M,k} = \mu + \sum_{m=1}^M F_{i_m}^{[m]} + \epsilon_{i_1,i_2,\ldots,i_M,k}$$

denoting the linear model for observation k in group levels $i_1, \ldots, i_M$ of the respective M factors.

Theorem 3.1 *Consider a Gaussian random vector* $\underline{Y}$ *of size* n *such that its covariance matrix satisfies* $\sigma^2 I_n$. *Let the matrix* $\mathbf{x}$ *satisfy a M-way ANOVA block design, then the following decomposition holds:*

$$Y_{i_1,i_2,\ldots,i_M,k} = \bar{Y} + \sum_{m=1}^M (\bar{Y}_{..i_m\ldots} - \bar{Y}) + (Y_{i_1,i_2,\ldots,i_M,k} - \sum_{m=1}^M (\bar{Y}_{..i_m\ldots} + (M-1)\bar{Y})$$

where $\pi_{\tilde{\mathbf{x}}_m}(\underline{Y})$ *gives a vector with components* $\bar{Y}_{..i_m\ldots} - \bar{Y}$ *and* $\pi_\mu(\underline{Y})$ *equals a vector with entries equalling* $\bar{Y}$.

Proof: Since the matrix $\mathbf{x}$ is of rank $n - M$, the M-constraints $\sum_{i=1}^{K_m} F_{i_m}^{[m]} = 0$ with $m \in \{1, \ldots, M\}$ implies a unique solution $\hat{\underline{\mu}} = \pi_{V_\mathbf{x}}(\underline{Y})$ of the minimization problem:

$$\underset{\underline{F}\in\mathbb{R}^{1+\sum_{m=1}^M K_m}}{\text{argmin}} \sum_{i_1,\ldots,i_M,k} (Y_{i_1,i_2,\ldots,i_M,k} - \mu - \sum_{m=1}^M F_{i_m}^{[m]})^2 \text{ subject to } \sum_{i=1}^{K_m} F_{i_m}^{[m]} = 0 \forall m$$

The constraints provide further that the solution is written by virtue of the direct sum as the decomposition:

$$\hat{\underline{\mu}} = \sum_{m=0}^M \pi_{V_{\tilde{x}_m}}(\underline{Y})$$

Now it is clear that $\pi_{V_{\tilde{x}_m}}(\underline{Y})$ is a one-way ANOVA w.r.t. factor m leading to the entries $\bar{Y}_{..i_m\ldots} - \bar{Y}$ which completes the proof. □

The M-way ANOVA table can now be explicitly written by

$\mathcal{H}_0$	SS	df	MSS
$\underline{\mu} \in V_\mu^\perp$	$n\bar{Y}^2$	1	$n\bar{Y}^2$
$\underline{\mu} \in V_{\tilde{\mathbf{x}}_m}^\perp$	$\sum_{i_m=1}^{K_m} n_{i_m}(\bar{Y}_{..i_m\ldots}-\bar{Y})^2$	K_m-1	$\frac{\sum_{i_m=1}^{K_m} n_{i_m}(\bar{Y}_{..i_m\ldots}-\bar{Y})^2}{K_m-1}$
$\underline{\mu} \in V$	$\sum_{i_1,\ldots,i_M}\sum_k (Y_{i_1\ldots i_M k} - \bar{Y})^2$	$\left(n-\sum_{m=1}^M (K_m-1)-1\right)$	$\frac{\sum_{i_1,\ldots,i_M}\sum_k (Y_{i_1\ldots i_M k}-Y)^2}{\left(n-\sum_{m=1}^M (K_m-1)-1\right)}$

3.2 Post-hoc analysis of main effects

In case pre-determined contrast tests are not available all pairwise comparisons tests per main effect is performed. This battery of tests is called the post-hoc analysis where the next definition is a more formal formulation.

Definition 3.3 *Consider a Gaussian random vector $\underline{Y}$ of size n such that its covariance matrix satisfies the homoscedasticity assumption $\sigma^2 I_n$. Let the matrix $\mathbf{x}$ satisisfy a M-way ANOVA block design then for factor $m \in \{1, 2, \ldots, M\}$ the post-hoc tests assess the hypotheses:*

$$\mathcal{H}_0 : F_{i_m}^{[m]} = F_{j_m}^{[m]} \text{ versus } \mathcal{H}_1 : F_{i_m}^{[m]} \neq F_{j_m}^{[m]}$$

for $i_m \neq j_m$ such that $i_m, j_m \in \{1, 2, \ldots, K_m\}$

It is clear that a post-hoc analysis for a factor with K levels allows $\begin{pmatrix} K \\ 2 \end{pmatrix} = \frac{K(K-1)}{2}$ hypotheses such that every one is tested at a risk of making a type-I error (i.e. false positive). We must control the overall or family-wise type-I error. Therefore, we consider 4 types of post-hoc tests which are simultaneously interesting so one should not make a choice but use its results jointly: Least Significant Difference (LSD or Fisher's t-test) test, Bonferroni's correction, Sidak's correction and Tukey's t-test. Before studying the different tests, we formally define the familywise type I-error and significance.

Definition 3.4 *Consider a post-hoc analysis for an ANOVA main effect with K-levels. Consider the test statistics T_k to test pairwise comparison idex $k \in \{1, 2, \ldots, \frac{K(K-1)}{2}\}$. The familywise significance α_F is the probability that at least one type I-error is committed over the different pairwise tests: $\alpha_F = \mathbb{P}\left(\cup_{k=1}^{\frac{K(K-1)}{2}} \{|T_k| > c\} \,|\mathcal{H}_0\right)$ with $c \in \mathbb{R}$ the user-defined critical value.*

Let α denote the user-defined significance which is desired to be achieved by the tests used, then the post-hoc analysis offers the following possibilities:

1. The test is called liberal if $\alpha_F > \alpha$
2. The test is called conservative if $\alpha_F < \alpha$
3. The test is called honest if $\alpha_F = \alpha$

Typically liberal tests are used cautiously as it may detect an excess of false positives while a conservative test overcompensates increasing false negatives. Without compensation, a sequence of t-tests would be conservative as the familywise significance is larger than the desired significance. As a result it is possible that the ANOVA F-test fails to detect significant differences while a battery of t-tests detects significantly different pairs. That would immediate identify a false positive. This is the reason why post-hoc tests are never used without an initial ANOVA F-test to identify possible false positives. Next result describes the LSD test or Fisher's t-test which is a specific contrast test.

Theorem 3.2 *Consider a Gaussian random vector $\underline{Y}$ of size n with covariance matrix $\sigma^2 I_n$. Let the matrix $\mathbf{x}$ satisfy a M-way ANOVA block design then the LR-test to assess $\mathcal{H}_0 : F_{i_m}^{[m]} = F_{j_m}^{[m]}$ versus $\mathcal{H}_1 : F_{i_m}^{[m]} \neq F_{j_m}^{[m]}$ is given by*

$$T_{i_m, j_m} = \frac{\bar{Y}_{..i_m...} - \bar{Y}_{..j_m...}}{\sqrt{MSE\left(\frac{1}{n_{i_m}} + \frac{1}{n_{j_m}}\right)}}$$

with $MSE = \frac{\sum_{i_1,\ldots,i_M}\sum_k (Y_{i_1\ldots i_M k}-\bar{Y})^2}{\left(n-\sum_{m=1}^M (K_m-1)-1\right)}$

This follows very straightforward from the contrast t-tests in Theorem 2.5.
Since the previous test is liberal due to a lack of correcting the familywise type I-error, we can introduce two corrections for Fisher's t-test. Both overcompensate in some sense and are therefore conservative.

Theorem 3.3 *Consider the LSD t-test under the conditions set in the previous theorem. Bonferroni's correction applies each LSD test with a significance* $\alpha_B = \frac{2\alpha}{K(K-1)}$ *while Sidak's correction applies for each test* $\alpha_S = 1-(1-\alpha)^{\frac{2}{K(K-1)}}$. *The following properties hold:*

1. $\alpha_F < \alpha$ *if Bonferroni's correction is applied*
2. $\alpha_F = \alpha$ *if Sidak's correrction is applied to orthogonal contrast tests*

Proof: For claim (1), we compute

$$\mathbb{P}\left(\cup_{k=1}^{\frac{K(K-1)}{2}} \{|T_k| > c\} \,|\mathcal{H}_0\right) \leq \frac{K(K-1)}{2}\alpha_B = \alpha$$

For the second claim, we compute

$$\begin{aligned}\mathbb{P}\left(\cup_{k=1}^{\frac{K(K-1)}{2}} \{|T_k| > c\} \,|\mathcal{H}_0\right) &= 1-\mathbb{P}\left(\cap_{k=1}^{\frac{K(K-1)}{2}} \{|T_k| \leq c\} \,|\mathcal{H}_0\right)\\ &= 1-\left(\mathbb{P}\left(|T_k| \leq c|\mathcal{H}_0\right)\right)^{\frac{K(K-1)}{2}}\\ &= 1-(1-\alpha_S)^{\frac{K(K-1)}{2}} = \alpha\end{aligned}$$

□

As a result, Bonferroni overcompensates turning the test conservative. Sidak's correction is a type of honest test but only for othogonal pairwise comparisons. Unfortunately for all pairwise comparisons Sidak's correction does not lead to a honest test. To motivate this, we can use a factor with 4 levels such that all possible pairwise comparisons lead to the contrast matrix

$$\mathbf{c} = \begin{bmatrix} 1 & -1 & 0 & 0\\ 1 & 0 & -1 & 0\\ 1 & 0 & 0 & -1\\ 0 & 1 & -1 & 0\\ 0 & 1 & 0 & -1\\ 0 & 0 & 1 & -1 \end{bmatrix}$$

which does not provide orthogonal contrasts. Tukey introduces a modification of the LSD test which does not require correction of the significance although leading to an honest test.

Definition 3.5 *Consider a Gaussian random vector* $\underline{Y}$ *of size* n *with covariance matrix* $\sigma^2 I_n$. *Let the matrix* $\mathbf{x}$ *satisfy the M-way ANOVA block design then Tukey's test to assess all pairwise comparisons* $\mathcal{H}_0 : F_{i_m}^{[m]} = F_{j_m}^{[m]}$ *versus* $\mathcal{H}_0 : F_{i_m}^{[m]} \neq F_{j_m}^{[m]}$ *is given by the LSD test but its p-value is computed by the studentized range distribution*

$$\mathcal{Q} = \frac{\max_{i_m} \bar{Y}_{..i_m\ldots} - \min_{i_m} \bar{Y}_{..i_m\ldots}}{\sqrt{\frac{2MSE}{n_m}}}$$

Tukey's tests abandons the use of the t-distribution but switches to the studentized range distribution $\mathcal{Q}$ which boils down to the t-distribution in the case of only two groups. As such the distribution takes into consideration that multiple hypotheses are being tested accordingly to a t-statistic in this specific case the LSD t-statistic. The next theorem proves that the studentized distribution leads to an honest test.

Theorem 3.4 *(Tukey) Consider Tukey's test under the conditions set previously. If the design is balanced with respect to factor m, then $\alpha_F = \alpha$*

Proof: Consider the familywise type-I error and the $1-\alpha$-quantile of the Q-distribution denoted by q such that we compute

$$\begin{aligned}
\mathbb{P}\left(\cup_{k=1}^{\frac{K(K-1)}{2}} \{|T_k| > q\} \,|\mathcal{H}_0\right) &= \mathbb{P}\left(\max_k |T_k| > q|\mathcal{H}_0\right) \\
&= \mathbb{P}\left(\frac{\max_{i_m} \bar{Y}_{..i_m...} - \min_{i_m} \bar{Y}_{..i_m...}}{\sqrt{\frac{2MSE}{n_m}}} > q|\mathcal{H}_0\right) \\
&= \mathbb{P}\left(\mathcal{Q} > q|\mathcal{H}_0\right) = \alpha
\end{aligned}$$

□

In order to compute the critical value q, we characterize the distribution $\mathcal{Q}$. The following theorem motivates the name Tukey's t-distribution. The difference with Welch's t-distribution is that in Tukey's case the numerator is altered while in Welch's case the denominator is modified which introduces a departure from the actual t-distribution.

Theorem 3.5 *Consider $X_1, X_2, \ldots, X_n$ independently but identically distributed Gaussian random variables with parameters μ and σ^2, then it holds that:*

$$\mathcal{Q}^*_{n,n-1} = \frac{\max_i X_i - \min_i X_i}{\hat{\sigma}_n} \overset{d}{=} \frac{R(n)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

with cumulative distribution of the range R given by

$$F_R(t) = n \int_{\mathbb{R}} \Phi^n(t+s)\Phi^{n-1}(-s)\phi(s)ds$$

with $\Phi(t)$ the cumulative and $\phi(t)$ the density functions of a standard normal random variable.

Proof: Consider random variables $Z_i = \frac{X_i - \mu}{\sigma}$ the standardized versions of random variabels X_i, then we compute

$$\begin{aligned}
R(n) &= \frac{\max_i X_i - \min_i X_i}{\sigma} \\
&= \max_i \left(Z_i + \frac{\mu}{\sigma}\right) - \min_i \left(Z_i + \frac{\mu}{\sigma}\right) \\
&= \max_i Z_i - \min_i Z_i
\end{aligned}$$

Next, we compute the cumulative distribution of $R(n)$ through the law of total probability for absolutely continuous random variables. We adopt the notation $f_{\min_i Z_i}(t)$ the density

function of the random variable $\min_i Z_i$ and $F_{\max_i Z_i}(t)$ the cumulative distribution of the random variable $\max_i Z_i$.

$$\begin{aligned}\mathbb{P}(R(n) \leq t) &= \mathbb{P}(\max_i Z_i - \min_i Z_i \leq t) \\ &= \int_{\mathbb{R}} \mathbb{P}(\max_i Z_i \leq t+s) f_{\min_i Z_i}(s) ds \\ &= \int_{\mathbb{R}} F_{\max_i Z_i}(t+s) f_{\min_i Z_i}(s) ds\end{aligned}$$

We compute $F_{\max_i Z_i}(t+s)$ and $f_{\min_i Z_i}(s)$ explicitly:

$$\begin{aligned}F_{\max_i Z_i}(t+s) &= \mathbb{P}(\max_i(Z_i) \leq t+s) \\ &= \mathbb{P}(\cap_i\{Z_i \leq t+s\}) \\ &= (\mathbb{P}(Z_i \leq t+s))^n \\ &= \Phi^n(t+s)\end{aligned}$$

and similarly we obtain

$$\begin{aligned}F_{\min_i Z_i}(s) &= \mathbb{P}(\min_i Z_i \leq s) \\ &= 1 - \mathbb{P}(\min_i(Z_i) > s) \\ &= 1 - \mathbb{P}(\cap_i\{Z_i > s\}) \\ &= 1 - (1-\Phi(s))^n\end{aligned}$$

such that $f_{\min_i Z_i}(s) = n(1-\Phi(s))^{n-1}\phi(s)$ which characterizes the distribution of $R(n)$, completing the proof. □

Based on this theorem we define a general studentized range distribution with (n, m) degrees of freedom.

Definition 3.6 *A random variable $\mathcal{Q}^*_{n,m}$ holds a studentized range distribution with (n, m) degrees of freedom if for two independent random variables $R(n)$ and χ^2_m the following holds:*

$$\mathcal{Q}^*_{n,m} \stackrel{d}{=} \frac{R(n)}{\sqrt{\frac{\chi^2_m}{m}}}$$

As a result, Tukey's t-test follows under the null-hypothesis the distribution $\mathcal{Q}^*_{n,n-\sum_m(K_m-1)-1}$.

Example 3.2 *A supermarket brand claims to be the cheapest among their competitors. To statistically verify this claim, a specific supermarket of the brand is selected together with its retail competitors in a 10km radius leading to 3 competitors. A random selection among 7 of its products is drawn. This leads to a two-way ANOVA where the dependent variable price is analyzed as a function of two categorical variables: items and supermarkets. Note that items is a random effect such that the cheapest claim only holds for the items considered.*

The first preliminary analysis is a one-way errorbar analysis of price over the different supermarkets and items separately. An inspection of the error-bars reveals that one cannot discriminate significant difference among supermarkets due to the large variability within each supermarket. The right plot shows that the variability within a supermarket can be further explained by the significant differences among items which explains the large

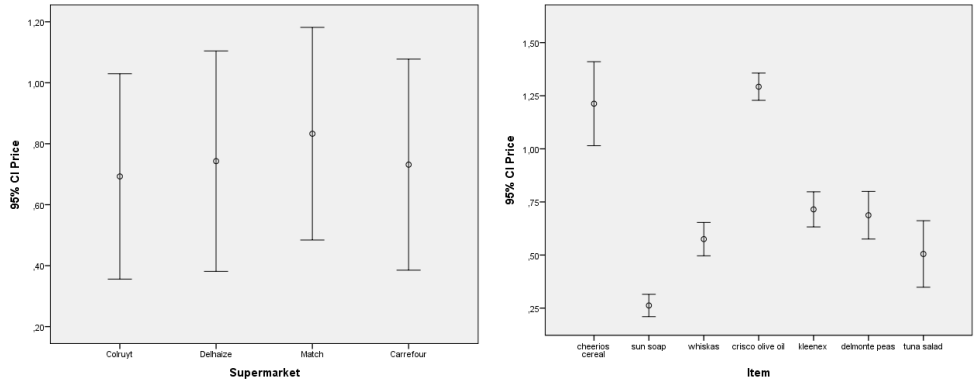

Figure 3.1: 95% Confidence intervals for the mean of price as a function of supermarkets (left) and items (right)

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	,074 [a]	3	,025	,173	,914
Intercept	15,750	1	15,750	110,957	,000
Supermarket	,074	3	,025	,173	,914
Error	3,407	24	,142		
Total	19,230	28			
Corrected Total	3,480	27			

a. R Squared = ,021 (Adjusted R Squared = -,101)

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3,367 [a]	6	,561	103,542	,000
Intercept	15,750	1	15,750	2906,415	,000
Item	3,367	6	,561	103,542	,000
Error	,114	21	,005		
Total	19,230	28			
Corrected Total	3,480	27			

a. R Squared = ,967 (Adjusted R Squared = ,958)

Table 3.1: One-Way ANOVA for the price as a function of supermarkets (left) and items (right)

variability of price within a supermarket. As such, the categorical variable item is a confounder disturbing the analysis. Inspection of the respective one-way ANOVA in Table 3.1 reveals that the one-way ANOVA for supermarkets leads to an insignificant model which only explains 2.1% *of the variability. However, the one-way ANOVA with respect to items reveals that the price variability can be explained by discriminating among items for* 96.7% *indicating a dominant confounder. Theorem 3.1 shows that a multi-way ANOVA consists of a sum of one-way ANOVAs such that we can correct the preliminary errorbar analysis. We re-analyze the error bars of the residuals of the one-way ANOVA performed on the confounding variable. The confidence intervals shown in Figure 3.2 identifies differences among the grocery stores. Colruyt seems the cheapest as it reveals a weak effect w.r.t. both Delhaize and Carrefour, while a strong effect w.r.t. Match which is also the strongest difference. Therefore, we expect the two-way ANOVA who identifies the strongest pair to lead to a p-value lower than* 1%*. The two-way table confirms that both item and supermarket are strongly significant with p-values below* 1%*. The model obtains a high* R^2*-statistic where nearly* 99% *of the variability is explained through the two categorical variables. The post-hoc analysis identifies Match as the most expensive supermarket but fails to identify a cheapest grocery store. Colruyt is positioned within the group to which also Carrefour and Delhaize belong. The weak differences observed from the preliminary analysis has not been confirmed. This may suggest a problem, although a visual analysis is preliminary and less accurate. Nevertheless critically revisiting Figure 3.2 shows a possible violation of the homogeneity of variance assumption. One may wish to perform a Welch correction on the studentized range distribution for the post-hoc analysis known as the Games-Howell test. Hence, Games-Howell's test is a Welch corrected Tukey test. Since Games-Howell test has been designed for one-way ANOVAs only, the Games-*

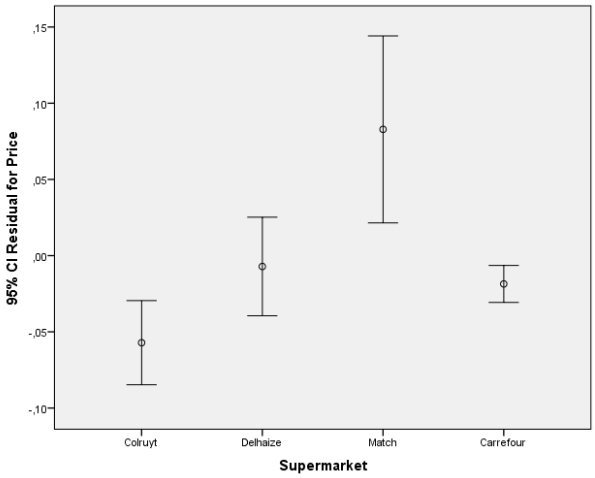

Figure 3.2: 95% Confidence intervals for residuals of the one-way ANOVA performed on the categorical variable item

Howell post-hoc is performed on the residuals of the one-way ANOVA w.r.t. the factor item. Although the Games-Howell post-hoc reveals the same conclusions, its descriptive statistics are more in line with the preliminary assessment. We first note that the standard errors of the pairwise differences are not equal, although the numerators representing the differences remain equal. This is in line with a Welch correction which only modifies the denominators of the respective t-tests. The p-values drop for differences between Colruyt and both Delhaize as well as Carrefour. Although these are still higher than the significance level set to 5%, one may expect a reduction in costs per item up to 10 Eurocents w.r.t Delhaize and up to 8 Eurocents w.r.t. Carrefour. This reduction per item achieves even 23 Eurocents when compared to Match.

Tests of Between-Subjects Effects

Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3,440 [a]	9	,382	171,524	,000
Intercept	15,750	1	15,750	7067,308	,000
Item	3,367	6	,561	251,776	,000
Supermarket	,074	3	,025	11,021	,000
Error	,040	18	,002		
Total	19,230	28			
Corrected Total	3,480	27			

a. R Squared = ,988 (Adjusted R Squared = ,983)

Table 3.2: Two-Way ANOVA for price as a function of both supermarket and item

3.3 Interaction models

Previously, it was assumed that each factor contributes separately. One deviates from this assumption by inclusion of cross terms.

Definition 3.7 *Consider an M-way ANOVA model with factors* $\underline{F}^{[m]}$*. An interaction of order L is given by the categorical variable consisting of the product of L factors. Typically L is restricted to* $L = 2$ *or* $L = 3$*.*

Dependent Variable: Price

	(I) Supermarket	(J) Supermarket	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Colruyt	Delhaize	-,0500	,02523	,231	-,1213	,0213
		Match	-,1400*	,02523	,000	-,2113	-,0687
		Carrefour	-,0386	,02523	,442	-,1099	,0327
	Delhaize	Colruyt	,0500	,02523	,231	-,0213	,1213
		Match	-,0900*	,02523	,011	-,1613	-,0187
		Carrefour	,0114	,02523	,968	-,0599	,0827
	Match	Colruyt	,1400*	,02523	,000	,0687	,2113
		Delhaize	,0900*	,02523	,011	,0187	,1613
		Carrefour	,1014*	,02523	,004	,0301	,1727
	Carrefour	Colruyt	,0386	,02523	,442	-,0327	,1099
		Delhaize	-,0114	,02523	,968	-,0827	,0599
		Match	-,1014*	,02523	,004	-,1727	-,0301

	(I) Supermarket	(J) Supermarket	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	Colruyt	Delhaize	-,0500	,02523	,063	-,1030	,0030
		Match	-,1400*	,02523	,000	-,1930	-,0870
		Carrefour	-,0386	,02523	,144	-,0916	,0144
	Delhaize	Colruyt	,0500	,02523	,063	-,0030	,1030
		Match	-,0900*	,02523	,002	-,1430	-,0370
		Carrefour	,0114	,02523	,656	-,0416	,0644
	Match	Colruyt	,1400*	,02523	,000	,0870	,1930
		Delhaize	,0900*	,02523	,002	,0370	,1430
		Carrefour	,1014*	,02523	,001	,0484	,1544
	Carrefour	Colruyt	,0386	,02523	,144	-,0144	,0916
		Delhaize	-,0114	,02523	,656	-,0644	,0416
		Match	-,1014*	,02523	,001	-,1544	-,0484

	(I) Supermarket	(J) Supermarket	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	Colruyt	Delhaize	-,0500	,02523	,378	-,1248	,0248
		Match	-,1400*	,02523	,000	-,2148	-,0652
		Carrefour	-,0386	,02523	,863	-,1133	,0362
	Delhaize	Colruyt	,0500	,02523	,378	-,0248	,1248
		Match	-,0900*	,02523	,013	-,1648	-,0152
		Carrefour	,0114	,02523	1,000	-,0633	,0862
	Match	Colruyt	,1400*	,02523	,000	,0652	,2148
		Delhaize	,0900*	,02523	,013	,0152	,1648
		Carrefour	,1014*	,02523	,005	,0267	,1762
	Carrefour	Colruyt	,0386	,02523	,863	-,0362	,1133
		Delhaize	-,0114	,02523	1,000	-,0862	,0633
		Match	-,1014*	,02523	,005	-,1762	-,0267

Based on observed means.
The error term is Mean Square(Error) = ,002.
*. The mean difference is significant at the ,05 level.

Table 3.3: Post-hoc analysis price w.r.t supermarket: Tukey (top left), LSD (top right) and Bonferroni (bottom)

Multiple Comparisons

Dependent Variable: Residual for Price

	(I) Supermarket	(J) Supermarket	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Games-Howell	Colruyt	Delhaize	-,0500	,01738	,060	-,1018	,0018
		Match	-,1400*	,02749	,004	-,2272	-,0528
		Carrefour	-,0386	,01232	,054	-,0778	,0006
	Delhaize	Colruyt	,0500	,01738	,060	-,0018	,1018
		Match	-,0900*	,02834	,046	-,1783	-,0017
		Carrefour	,0114	,01412	,848	-,0343	,0571
	Match	Colruyt	,1400*	,02749	,004	,0528	,2272
		Delhaize	,0900*	,02834	,046	,0017	,1783
		Carrefour	,1014*	,02556	,025	,0150	,1879
	Carrefour	Colruyt	,0386	,01232	,054	-,0006	,0778
		Delhaize	-,0114	,01412	,848	-,0571	,0343
		Match	-,1014*	,02556	,025	-,1879	-,0150

Based on observed means.
The error term is Mean Square(Error) = ,002.
*. The mean difference is significant at the 0,05 level.

Table 3.4: Post-hoc analysis price w.r.t supermarket - Games-Howell test

Example 3.3 *Consider a two way ANOVA model with 4 and 6 levels respectively. The second order interaction is the categorical variable* $F^{[1]} \times F^{[2]}$ *with factor levels* $(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(2,6),\ldots,(4,1),\ldots,(4,6)$ *holding 24 levels. The statistical model equals:*

$$Y_{ijk} = \mu + F_i^{[1]} + F_j^{[2]} + (F_i^{[1]} \times F_j^{[2]}) + \epsilon_{ijk}$$

Hence, the model holds a total of $1+3+5+15 = 24$ *estimable parameters due to the three sum-to-zero constraints.*

The estimation of the parameters has been sequential in the sense that first μ is estimated as $\bar{Y}_{...}$, secondly main effects can be estimated which in turn are again sample means groupwise on the residual data $Y_{ijk} - \bar{Y}_{...}$ leading to $\hat{F}_i^{[1]} = \bar{Y}_{i..} - \bar{Y}$ and $\hat{F}_j^{[2]} = \bar{Y}_{.j.} - \bar{Y}$. Consider the averaging operator $A_{i_m}^{[m]}(\underline{Y}) = \bar{Y}_{..i_m..}$ with multiplication rule $A_{i_{m_1}}^{[m_1]} A_{i_{m_2}}^{[m_2]}(\underline{Y}) = \bar{Y}_{..i_{m_1}.i_{m_2}.}$. The operator tells which index to remove from the averaging process. The operation $A^{[0]}$ includes all indices in the averaging.

Theorem 3.6 *Consider a Gaussian random vector* $\underline{Y}$ *of size* n *such that its covariance matrix equals* $\sigma^2 I_n$*. We obtain the UMVU-estimator for the interaction term* $F_{i_1}^{[1]} \times F_{i_2}^{[2]} \times$

$\cdots \times F_{i_l}^{[l]}$ *is given by:*

$$\prod_{k=1}^{l} \left(A^{[0]} - A_{i_k}^{[k]} \right) \underline{Y}$$

The theorem can be proven inductively but is omitted as it is particularly lengthy. Hence, we can explicitly compute the two way interaction $F_i^{[1]} \times F_j^{[2]}$ which is a vector with entries:

$$\begin{aligned} \left(A^{[0]} - A_i^{[1]} \right) \left(A^{[0]} - A_j^{[2]} \right) \underline{Y} &= \left(A_i^{[1]} A_j^{[2]} - A_i^{[1]} - A_j^{[2]} + A^{[0]} \right) \underline{Y} \\ &= \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...} \\ &= \bar{Y}_{ij.} - \bar{Y}_{...} - \hat{F}_i^{[1]} - \hat{F}_j^{[2]} \end{aligned}$$

The final equation illustrates the sequential approach. Hence, the ANOVA table for a 2-way design with interaction receives an additional row. For a balanced design with l samples per cell the 2-way table becomes:

$\mathcal{H}_0$	SS	df	MSS
$\mu = 0$	$nY_{...}^2$	1	$nY_{...}^2$
$\underline{F}^{[1]} = \underline{0}$	$lK_2 \sum_{i=1}^{K_1} (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$K_1 - 1$	$\frac{lK_2 \sum_{i=1}^{K_1} (\bar{Y}_{i..} - \bar{Y}_{...})^2}{K_1 - 1}$
$\underline{F}^{[2]} = \underline{0}$	$lK_1 \sum_{j=1}^{K_2} (\bar{Y}_{.j.} - \bar{Y}_{...})^2$	$K_2 - 1$	$\frac{lK_1 \sum_{j=1}^{K_2} (\bar{Y}_{.j.} - \bar{Y}_{...})^2}{K_2 - 1}$
$\underline{F}^{[1]} \times \underline{F}^{[2]} = \underline{0}$	$l \sum_i^{K_1} \sum_{j=1}^{K_2} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$	$(K_1 - 1)(K_2 - 1)$	$\frac{l \sum_i^{K_1} \sum_{j=1}^{K_2} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2}{(K_1 - 1)(K_2 - 1)}$

Example 3.4 *The beer-goggle effect is the hypothesis that higher alcohol intoxications imply an overrated physical attraction. To test the hypothesis, a cohort of 48 students in the age category of 18-22 years old was randomly selected to participate in the test consisting of 24 men and 24 women. Each gender group was randomly divided in 3 groups of 8 people which could only consume either (i) alcohol-free beverages, (ii) light alcoholic beverages like pilser or (iii) heavier alcoholic beverages like stronger beers or wines. The participants must consume 5 beverages. Every participant was asked to chat or dance with a physically attractive date. At the end of the night, a picture was taken of the couple where the participant's choice was rated for his or hers physical attraction independently by 10 people on a 100 points scale. The research question is two-folded: (i) is the physical attraction of a person's choice depending on the amount of alcohol consumed? (ii) Is the dependence different among men or women?*

We start the analysis by performing a preliminary analysis based on the error-bars, in which we discriminate between a one-way approach where the attraction scores are explored as a function of gender and alcohol group. Additionally within each gender group, the attraction scores are explored as a function of alcohol level which is called an interaction plot. The errorbar analysis w.r.t. gender reveal no significant effects, although the variability within the man group is significantly higher. The errorbars w.r.t. the alcohol group show that one can expect that for higher alcohol levels the average selected data is rated less attractive. This effect is strongly significant as the error-bars of the first two alcohol groups do not intersect with the third alcohol group. Note that the effect of alcohol is as such only significant for higher intoxications. The profile plot shows that the effect of alcohol is only present for the man group (dashed line) whereas for the women no significant of alcohol on the attraction scores is observed as all confidence bounds intersect where its individual means are situated in the other groups' errorbar interval.

To confirm the visual analysis of the errorbars, we perform a two-way ANOVA with interaction or a full factorial model. The two-way ANOVA confirms the visual analysis, where gender is considered insignificant whereas alcohol intoxication is strongly signifcant although its effect is strongly dependent due to the presence of the interaction of gender. We conclude that the highest intoxication typically reduces the attraction score of the selected date by approximately 30 points for men while this is only approximately 5 points for women.

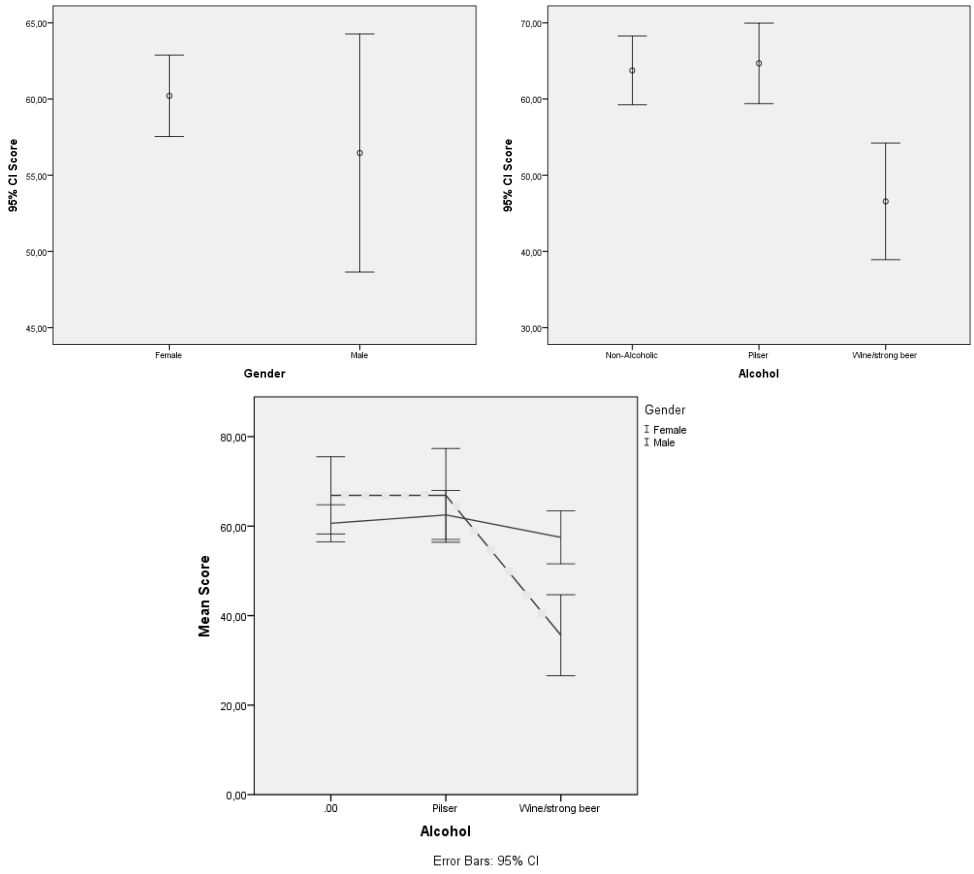

Figure 3.3: 95% Confidence intervals for the attraction score as a function of gender (left), alcohol level (right) and profile between gender and alcohol level (bottom)

Tests of Between-Subjects Effects

Dependent Variable: Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5479,167 [a]	5	1095,833	13,197	,000
Intercept	163333,333	1	163333,333	1967,025	,000
Gender	168,750	1	168,750	2,032	,161
Alcohol	3332,292	2	1666,146	20,065	,000
Gender * Alcohol	1978,125	2	989,062	11,911	,000
Error	3487,500	42	83,036		
Total	172300,000	48			
Corrected Total	8966,667	47			

a. R Squared = ,611 (Adjusted R Squared = ,565)

Table 3.5: Two-Way ANOVA for attraction score as a function of gender and alcohol level

3.4 Residual Analysis

Two important assumptions are tested: (i) normality of the residuals and (ii) homogeneity of variances over the group levels. Since the residual analysis is typically conducted per main effect, the tests are developed by considering a one-way design only. By virtue of the central limit theorem we may expect some robustness to departures of the normality assumption. This condition can be tested graphically through a qq-plot which serves as a preliminary analysis. If the qq-plot exhibits a scatter diagram whose points are scattered around a straight line, the normality assumption can be defended, otherwise the normality assumption is violated.

Definition 3.8 *Let $\underline{\hat{\epsilon}}$ be the observed residual vector given by $\underline{Y}-\underline{\hat{Y}}$. The vector is ordered in ascending order denoted by $\underline{\tilde{\epsilon}}$. The empirical cumulative distribution function associated to $\underline{\tilde{\epsilon}}$ is given by the vector $\left(\frac{n-i}{n}\right)_i$ with i the vector index. If the residuals follow a Gaussian distribution, the theoretical quantiles $\underline{\tilde{\epsilon}}_0$ are given element-wise by $\tilde{\epsilon}(i) = \Phi^{-1}\left(\frac{n-i}{n}\right)\hat{\sigma}$ such that $\Phi(.)$ denotes the cumulative distribution of the standard normal distribution, with $\hat{\sigma}^2$ the UMVU-estimator of the variance. The qq-plot is the scatter diagram of $\underline{\tilde{\epsilon}}$ as a function of $\underline{\tilde{\epsilon}}_0$.*

A hypothesis test formalizing the qq-plot is the Shapiro-Wilk test which assesses the hypotheses $\mathcal{H}_0 : F_\epsilon = \mathcal{N}(0,\sigma^2)$ versus $\mathcal{H}_1 : F_\epsilon \neq \mathcal{N}(0,\sigma^2)$. The test compares the Gaussian-based UMVU estimator $\hat{\sigma}^2$ for σ^2 against a rank based estimator $\tilde{\sigma}^2 = \left(\sum_{i=1}^n a_i\tilde{\epsilon}_0(i)\right)^2$ unbiased for any type of distribution. The Shapiro-Wilk test assesses the test statistic:

$$SW = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2}$$

Theorem 3.7 *Consider an observed residual vector $\underline{\hat{\epsilon}}$ and $\underline{\tilde{\epsilon}}$ the ordered residual vector in ascending order. If $\tilde{\epsilon}(i) = \Phi^{-1}\left(\frac{n-i}{n}\right)\sigma+\delta(i)$ such that $\delta(i)$ is independently and identically distributed with mean zero, then the least squares estimator for σ is given by*

$$\tilde{\sigma} = \sum_{i=1}^{n} a_i\tilde{\epsilon}(i)$$

with $a_i = \frac{\Phi^{-1}\left(\frac{n-i}{n}\right)}{\sum_{i=1}^n\left(\Phi^{-1}\left(\frac{n-i}{n}\right)\right)^2}$

The proof is an easy computation. Note that in the literature one reports the estimator under the relaxed assumption that the random variable $\delta(i)$ is not necessarily identically distributed, although this is currently for the application at hand not needed. The assumption that $\delta(i)$ is zero-mean implies that the qq-plot exhibits a scatter diagram around a straight line which makes the SW test reasonable and a formalization of the qq-plot. Similarly one can formalize a test based on the pp-plot which assesses the alignment of the empirical cumulative distribution $\frac{n-i}{n}$ to the theoretical one given by $\Phi\left(\frac{\tilde{\epsilon}_0(i)}{\hat{\sigma}}\right)$. The test underlying the pp-plot is known as the Kolmogorov-Smirnov test.
The next assumption to check is the homoscedasticity condition. This condition is graphically assessed through a scatterdiagram between the predicted and observed model residuals. The confidence interval of the scatter diagram should be sufficiently filled. In case a particular trend is forming the model can be nonlinearly transformed to stabilize the variance by elimination of the model error. The model error is often coupled to the model

which is in violation of the additivity assumption of linear models. In case of a one-way ANOVA, a formal LR test can be derived to test the hypothesis of homoscedasticity known as Bartlett's test (1938).

Theorem 3.8 *Consider a random vector $\underline{Y}$ of size n satisfying the Welch ANOVA design. The LR-test to assess the hypotheses $\mathcal{H}_0 : \sigma_i^2 = \sigma_j^2$ for all $i, j \in \{1, 2, \ldots, K\}$ versus $\mathcal{H}_1 : \sigma_i^2 \neq \sigma_j^2$ for a pair $i \neq j \in \{1, 2, \ldots, K\}$ can be monotonically transformed through mapping $x \mapsto 2\log(x)$ to*

$$F = \sum_{i=1}^{K} n_i \left(\log(\hat{\sigma}^2) - \log(\hat{\sigma}_i^2)\right)$$

where $\hat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$ and $\hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^{K}(n_i - 1)\hat{\sigma}_i^2$. Asymptotically ($n \to \infty$) the test statistic F converges under the null-hypothesis in distribution to χ_{K-1}^2 random variable.

Note that to improve the finite sample properties of the χ_{K-1}^2 approximation, Bartlett gave a finite sample correction leading to the F-statistic:

$$F = \frac{\sum_{i=1}^{K}(n_i - 1)\left(\log(\hat{\sigma}^2) - \log(\hat{\sigma}_i^2)\right)}{1 + \frac{3}{K-1}\left(\left(\sum_{i=1}^{K}\frac{1}{n_i - 1}\right) - \frac{1}{n-K}\right)}$$

Proof: We consider one-Way ANOVA with densities:

$$f(\underline{y}|\underline{\mu}, \underline{\sigma}^2) = \begin{cases} \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2\right) & \text{under } \mathcal{H}_0 \\ \frac{1}{(2\pi)^{\frac{n}{2}}\prod_{i=1}^{K}\sigma_i^{n_i}}\exp\left(\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{(y_{ij} - \mu_i)^2}{\sigma_i^2}\right) & \text{under } \mathcal{H}_1 \end{cases}$$

Repeating an earlier exercise reveals the UMVU estimators under the null-hypothesis:

$$\hat{\mu}_i = \bar{Y}_i \text{ and } \hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

while under the alternative hypothesis these UMVU estimator become:

$$\hat{\mu}_i = \bar{Y}_i \text{ and } \hat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

The LR test easily reveals a test statistic of the form:

$$LR = \frac{\hat{\sigma}^n}{\prod_{i=1}^{K}\hat{\sigma}_i^{n_i}}$$

Hence, we compute

$$F = 2\log(LR) = \sum_{i=1}^{K} n_i \left(\log(\hat{\sigma}^2) - \log(\hat{\sigma}_i^2)\right)$$

Instead of computing the distribution of the statistic F, we will study the asymptotic distribution of statistic $\tilde{F} = \sum_{i=1}^{K}(n_i - 1)\left(\log(\hat{\sigma}^2) - \log(\hat{\sigma}_i^2)\right)$. Note that $|F - \tilde{F}| = \sum_{i=1}^{K}\left(\log(\hat{\sigma}^2) - \log(\hat{\sigma}_i^2)\right) \leq \frac{1}{\min(n_i - 1 | i \in \{1,2,\ldots,K\})}\tilde{F} = \frac{K}{n}\tilde{F}$ where the last equality holds for

balanced designs. As such, we obtain that $\lim_{n\to\infty} F - \tilde{F} = 0$ in probability at a rate of n^{-1} where Slutsky's lemma guarantees that both statistics F and $\tilde{F}$ hold the same asymptotic distribution.
To compute the asymptotic distribution, we apply a second order Taylor approximation $\log(x) \approx \log(x_0) + \frac{1}{x_0}(x - x_0) - \frac{1}{2x_0^2}(x - x_0)^2$. Due to the consistency of the sample variances this approximation holds the same asymptotic distribution as the original statistic $\tilde{F}$ again by virtue of Slutsky's Lemma. We compute through the Taylor-approximation

$$\begin{aligned}
\tilde{F} &\approx \frac{1}{2}\sum_{i=1}^{K}(n_i - 1)\left[\left(\frac{\hat{\sigma}_i^2}{\sigma^2} - 1\right)^2 - \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right)^2\right] \\
&\stackrel{d}{=} \frac{1}{2}\sum_{i=1}^{K}(n_i - 1)\left[\left(\frac{1}{n_i - 1}\chi^2_{n_i-1} - 1\right)^2 - \left(\frac{1}{n-K}\left(\sum_{m=1}^{K}\chi^2_{n_m-1}\right) - 1\right)^2\right]
\end{aligned}$$

Next, we can apply the central limit theorem to the individual $\chi^2_{n_i-1}$ distributions wherein it is assumed that if $n \to \infty$ each group size $n_i \to \infty$. This is not necessarily true except for balanced designs. At this point, we continue for balanced designs although this is not strictly needed if one assumed all group sizes tending towards infinity. We adopt the notation $\left(\frac{1}{(n_i-1)}\chi^2_{n_i-1} - 1\right) \stackrel{d}{\to} \sqrt{\frac{2}{n_i-1}}Z_i$ such that Z_i are independent and identically Gaussian random variables with zero-mean and unit variance. We continue calculations:

$$\begin{aligned}
F &\stackrel{d}{=} \sum_{i=1}^{K} Z_i^2 - \frac{1}{n-K}\left(\sum_{m=1}^{K}\sqrt{(n_m - 1)}Z_m\right)^2 \\
&\stackrel{n_i=\frac{n}{K}}{=} \sum_{i=1}^{K} Z_i^2 - \frac{1}{K}\left(\sum_{m=1}^{K} Z_m\right)^2 \\
&= \sum_{i=1}^{K}\left(Z_i - \bar{Z}\right)^2 \stackrel{d}{=} \chi^2_{K-1}
\end{aligned}$$

which completes the proof of Bartlett's theorem. □
Unfortunately, statistician George Box[1] showed in 1953 that the power of Bartlett's test depends on the kurtosis of the residuals given by $\gamma = \frac{\mu_4}{\mu_2^2}$ with μ_k the k-th central moment of the data. Hence, departures from normality may lead to a dramatic drop in power rendering the test useless. In an effort to compensate for this effect, Box provided a modified Bartlett test with test statistic:

$$F_{Box} = \frac{2F}{\gamma - 1}$$

The test coincides with Bartlett's test if the Kurtosis equals 3 which is its value for a Gaussian distribution. This test only partially alleviates the robustness issue against departures of the normal assumption. The kurtosis of the residuals is typically unknown which requires to be estimated from the data. In case the sample size is small the kurtosis is not accurately estimated which has a strong negative effect on the power of the test once more. In 1960 Levene proposed a heuristic test based on the random variables

[1]George Box married the second daughter of Ronald Fisher. He studied the performance of the test while working for Imperial Chemical Industries. In 1960 he took a carreer switch towards academia.

$Z_{ij} = |Y_{ij} - \bar{Y}_i|$. He noted that the distribution of the random variables Z_{ij} holds a mean value which depends on the standard deviation $\hat{\sigma}_i$ of the group for which he suggested the use of a One-Way ANOVA on Z_{ij} to assess homoscedasticity. The test is a heuristic since the LR approach on Z_{ij} leads similarly to the Bartlett test. Levene's suggestion is based on the half-normal distribution.

Theorem 3.9 *Consider a random vector $\underline{Y}$ of size n satisfying the conditions of a Welch ANOVA design. The random vector $\underline{Z}$ with elements $Z_{ij} = |Y_{ij} - \bar{Y}_i|$ follows a half-normal distribution*[2] *with multivariate density*

$$f_{\underline{Z}}(\underline{z}) = \left(\frac{2}{\pi}\right)^{\frac{n}{2}} \frac{1}{\prod_{i=1}^{K} \sigma_i^{n_i}} \exp\left(-\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{z_{ij}^2}{\sigma_i^2}\right)$$

with $\mathbb{E}[Z_{ij}] = \sigma_i\sqrt{\frac{2}{\pi}}$ and variance $Var(Z_{ij}) = \sigma_i^2\left(1 - \frac{2}{\pi}\right)$.

Proof: We compute the cumulative distribution of Z_{ij}:

$$\begin{aligned} F_{Z_{ij}}(t) &= \mathbb{P}(Z_{ij} \leq t) \\ &= \mathbb{P}(Y_{ij} - \bar{Y}_i \leq t) + \mathbb{P}(Y_{ij} - \bar{Y}_i \geq -t) \\ &= \Phi\left(\frac{t}{\sigma_i}\right) + 1 - \Phi\left(\frac{-t}{\sigma_i}\right) \\ &= 2\Phi\left(\frac{t}{\sigma_i}\right) \end{aligned}$$

The density becomes: $f_{Z_{ij}}(t) = \sqrt{\frac{2}{\pi\sigma_i^2}} \exp\left(-\frac{t^2}{2\sigma_i^2}\right)$. The expectation can be computed by evaluation of the integral where the substitution $u = \frac{-t^2}{2\sigma_i^2}$ is used. One can verify as an exercise that the expectation equals $\sqrt{\frac{2}{\pi}}\sigma_i$. Since the second moment is trivially equal to σ_i^2 implying a variance of $\sigma_i^2\left(1 - \frac{2}{\pi}\right)$. This completes the proof. □

Levene's test is the one-way ANOVA F-statistic on the data $\underline{Z}$ given by:

$$F_{\text{lev}} = \frac{n-K}{K-1} \frac{\sum_{i=1}^{K} n_i(\bar{Z}_i - \bar{Z})^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_i)^2}$$

Note that Levene's test does not apply the Welch ANOVA although the previous theorem clearly shows that the data used Z_{ij} has a group dependent variance which is in violation of Fisher's homogeneity of variance assumption. Alternatively one may apply a transformation to stabilize the variance through a log-transformation and as such operate the one-way ANOVA to the data $\log(Z_{ij})$. Both alternatives have not been reported in the literature.

Example 3.5 *We conduct a residual analysis for the supermarket example 3.2. Based on the model residuals of the two-way ANOVA with main effects supermarket and items whereas no interaction is considered, we start with assessing possible departures from the normality condition.*
Assessment of the tables reveal a Shapiro-Wilk test which does not detect significant departures from normality. This is further emphasized through the table of descriptive statistics

[2]The distribution is part of the family of folded-normal distributions.

Tests of Normality

	Kolmogorov-Smimov [a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for Price	,128	28	,200*	,955	28	,259

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Descriptives

			Statistic	Std. Error
Residual for Price	Mean		,0000	,00728
	95% Confidence Interval for Mean	Lower Bound	-,0149	
		Upper Bound	,0149	
	5% Trimmed Mean		,0003	
	Median		-,0034	
	Variance		,001	
	Std. Deviation		,03854	
	Minimum		-,11	
	Maximum		,09	
	Range		,20	
	Interquartile Range		,04	
	Skewness		-,003	,441
	Kurtosis		1,918	,858

Table 3.6: statistical tables for price residuals: Normality tests (top), descriptive statistics (bottom)

in which one observes a median and mean closer than one standard error of the mean which indicates a sufficiently symmetrical residual distribution. Furthermore, we observe that the range equals 0.2 which is approximately equal to 6 standard deviations supporting good descriptions of confidence bounds, this is in agreement with the inter-quartile range of 0.04 such that the standard deviation equals approximately 0.75 times the inter-quartile range. Finally, the skewness is closer to zero than one standard error. The only excess is seemingly the kurtosis which is outside of one standard error around the reference value of 3 indicating a significant smaller kurtosis or a flat residual distribution. Nevertheless, it is deemed insignificant accordingly to the SW test but it may explain why the p-value is not higher than approximately 26%. Alternatively, one may consult the histrogram and qq-diagram. The sample size makes the use of the histogram not appropriate, the qq-plot allows inspection that the reference line is followed except for what the distribution tails are concerned.

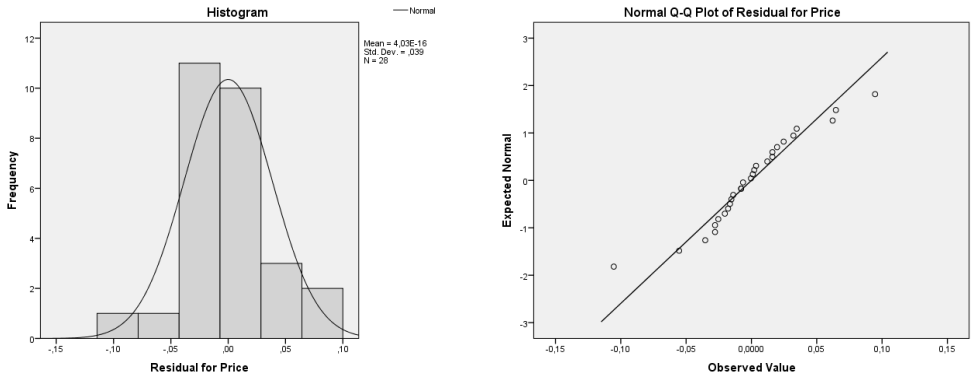

Figure 3.4: diagrams for price residuals: histogram with normality fit (left), quartile-quartile plot (right)

In the next step of the residual analysis, one assesses the homoscedasticity condition. We inspect the condition through a scatter diagram of the model residuals as a function of the predicted values as well as the levels of the main factor of interest (i.e. supermarket).

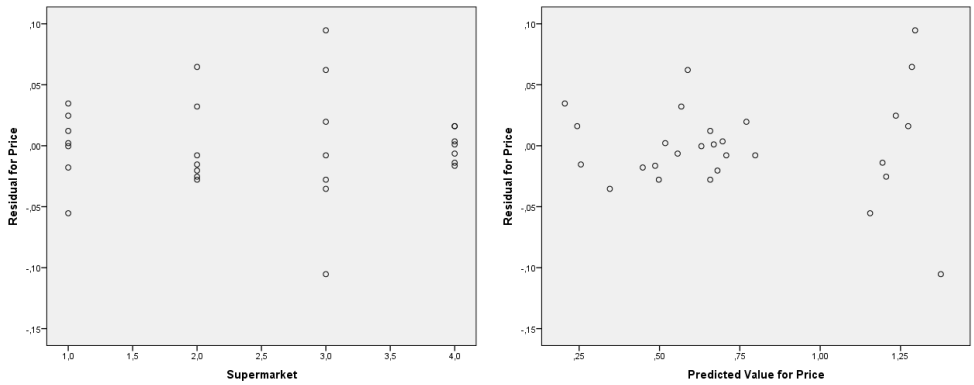

Figure 3.5: scatter diagram of price residuals with respect to group levels of supermarket (left) and predictions (right)

It is clear from the scatter diagrams that the range of the price residuals is not equal across the supermarkets which is also true by considering the scatter diagram as a function of the model predictions. The larger price predictions impose an increased uncertainty where equi-variance is better supported for the smaller and moderate price predictions. This can be explained by the presence of supermarket Match exhibiting the more expensive supermarket as well as the larger error bars for its prices per item. Levene's test confirms the violation of homoscedasticity to be significant which supports the use of the Games-Howell test as a Welch corrected Tukey test in the post-hoc analysis over supermarkets.

Test of Homogeneity of Variances

Residual for Price

Levene Statistic	df1	df2	Sig.
3,654	3	24	,027

Table 3.7: Levene's test for homoscedasticity of the residuals over levels of supermarket

Chapter 4

Sample size calculations

By virtue of the weak law of large numbers, any deviation from the null hypothesis will be detected if the sample size is sufficiently large. This does not mean that any departure is particularly relevant. As a result, one wishes to ensure that practical relevance and statistical significance coincides. Therefore, a sample size calculation strives bringing these together.
To introduce the required ingredients to compute the required sample size, we start to study Wald tests or z-tests for one parameter value.

Definition 4.1 *Consider a random vector of identically and independently distributed random variables* $\underline{Y}$*. Let the probability density function be parametrized in* $\theta \in \mathbb{R}$ *such that* $\hat{\theta}_{ML}$ *is the ML-estimator of* θ *w.r.t. its joint density. The Wald test to assess hypotheses* $\mathcal{H}_0 : \theta = \theta_0$ *versus* $\mathcal{H}_1 : \theta \neq \theta_0$ *is given by*

$$T = \frac{\hat{\theta}_{ML} - \theta_0}{\sigma_{\hat{\theta}_{ML}}} \text{ with } \sigma^2_{\hat{\theta}_{ML}} = Var\left(\hat{\theta}_{ML}\right)$$

The ML-estimator is asymptotically distributed as a Gaussian random variable $\hat{\theta}_{\mathrm{ML}} \overset{\mathcal{H}_0}{\sim} \mathcal{N}(\theta_0, \sigma^2_{\hat{\theta}_{\mathrm{ML}}})$ or $\hat{\theta}_{\mathrm{ML}} \overset{\mathcal{H}_1}{\sim} \mathcal{N}(\theta_1, \sigma^2_{\hat{\theta}_{\mathrm{ML}}})$ which is used to compute the required sample size. The following ingredients are needed for the computation:

1. Specify the minimal departure w.r.t. the null-hypothesis (i.e. for Wald tests $\Delta(\theta) = |\theta_0 - \theta_1|$) to be detected in order to return practical relevant conclusions.

2. Specify the desired confidence level $1 - \alpha$ and the desired power $1 - \beta$. Typically, one selects powers $\alpha < \beta \leq 4\alpha$.

3. The variance of the ML-estimator follows the Cramer-Rao lower bound. The Cramer-Rao bound must be computed and quantified.

Theorem 4.1 *A Wald test for a difference* $\Delta(\theta) = |\theta_1 - \theta_0|$ *with significance* α *and power* $1 - \beta$ *requires a sample size exceeding* $n \geq \left|\frac{z_{1-\beta} + z_{1-\frac{\alpha}{2}}}{e(\theta)}\right|$ *with effect size* $e(\theta) = \frac{\Delta(\theta)}{\sigma_{\hat{\theta}_{ML}}\sqrt{n}}$.

Proof: We compute the asymptotic power π under alternative θ_1, for notational simplicity

the subscript ML will be omitted:

$$\begin{aligned}
\pi &= \mathbb{P}_1\left(|T| > z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}_1\left(T > z_{1-\frac{\alpha}{2}}\right) + \mathbb{P}_1\left(T < z_{\frac{\alpha}{2}}\right) \\
&= \mathbb{P}_1\left(\frac{\hat{\theta}-\theta_1}{\sigma_{\hat{\theta}}} > z_{1-\frac{\alpha}{2}} + \frac{\theta_0-\theta_1}{\sigma_{\hat{\theta}}}\right) + \mathbb{P}_1\left(\frac{\hat{\theta}-\theta_1}{\sigma_{\hat{\theta}}} < z_{\frac{\alpha}{2}} + \frac{\theta_0-\theta_1}{\sigma_{\hat{\theta}}}\right) \\
&= \Phi\left(z_{\frac{\alpha}{2}} - \frac{\theta_0-\theta_1}{\sigma_{\hat{\theta}}}\right) + \Phi\left(z_{\frac{\alpha}{2}} + \frac{\theta_0-\theta_1}{\sigma_{\hat{\theta}}}\right) \\
&\geq \Phi\left(z_{\frac{\alpha}{2}} + \frac{\Delta(\theta)}{\sigma_{\hat{\theta}}}\right)
\end{aligned}$$

The power is minimally $1-\beta$ if $z_{\frac{\alpha}{2}} + \frac{\Delta(\theta)}{\sigma_{\hat{\theta}}} \geq z_{1-\beta}$ such that we obtain:

$$\sqrt{n} \geq \frac{z_{1-\beta} + z_{1-\frac{\alpha}{2}}}{e(\theta)}$$

which establishes the result. □

Note that if one sided Wald tests are used, one should apply the same formula for 2α. Another way to define the effect size is through the non-centrality parameter. The non-centrality parameter of a test statistic is given by the difference $\delta(\theta_1) = |T_0 - T_1|$ such that one obtains for the Wald test $\delta(\theta_1) = \frac{\Delta(\theta)}{\sigma_{\hat{\theta}_{\text{ML}}}}$. Thus the effect size is a normalized non-centrality parameter where the normalization is provided w.r.t. the WLLN. This allows a formal definition of effect sizes accordingly to Cohen, known as Cohen effect sizes.

Definition 4.2 *Consider a hypothesis testing problem to discriminate $\mathcal{H}_0 : \underline{\theta} = \underline{\theta}_0$ versus $\mathcal{H}_1 : \underline{\theta} \neq \underline{\theta}_0$ with test statistic T. Let test statistic T_1 be appropriate to evaluate hypotheses $\mathcal{H}_0 : \underline{\theta} = \underline{\theta}_1$ versus $\mathcal{H}_1 : \underline{\theta} \neq \underline{\theta}_1$. The non-centrality parameter is given by $\delta(\underline{\theta}_1) = |T_1 - T_0|$ and Cohen's effect size for the test statistic T is defined as $e(\underline{\theta}_1) = \frac{\delta(\underline{\theta}_1)}{\sqrt{n}}$ with n the total sample size.*

Theorem 4.2 *The Cohen effect size for the one and two sample t-tests are respectively:*

1. $e_1(\mu_1) = \frac{|\mu_1-\mu_0|}{\sigma}$
2. $e_2(\mu_1,\mu_2) = \frac{|\mu_1-\mu_2|}{\sqrt{\sigma_1^2+\sigma_2^2}\sqrt{2}}$ *if balanced* $n_1 = n_2$

Proof: We will only proof the result for the two-sample t-test. Consider the test statistic:

$$\begin{aligned}
T &= \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{(n_1-1)\hat{\sigma}_1^2+(n_2-1)\hat{\sigma}_1^2}{n_1+n_2-2}}} \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \\
&= \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\hat{\sigma}_1^2+\hat{\sigma}_2^2}} \sqrt{\frac{n}{2}}
\end{aligned}$$

Furthermore we obtain the statistic T_1 given by

$$T_1 = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}_1^2+\hat{\sigma}_2^2}} \sqrt{\frac{n}{2}}$$

with noncentrality parameter

$$\delta(\mu_1, \mu_2) = \frac{|\mu_1 - \mu_2|}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} \sqrt{\frac{n}{2}}$$

Hence, the effect size is computed by:

$$e(\mu_1, \mu_2) = a.s. \lim_{n \to \infty} \frac{\delta(\mu_1, \mu_2)}{\sqrt{n}} = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \sqrt{\frac{1}{2}}$$

where $a.s. \lim_{n \to \infty}$ denotes the stochastic limit almost surely. Verify as an exercise that the Welch t-test implies the same effect size □

Note that the effect size of Welch's and student's t-tests are equal which is an effect of the asymptotic definition of effect sizes. However, the degrees of freedom of the two tests are not equal such that Welch's t-test undergoes an efficiency loss. By virtue of Sattertwaithe theorem, the efficiency loss can be computed to improve the sample size calculation for Welch's t-test.

Theorem 4.3 *To compensate the loss in degrees of freedom of the Welch t-test w.r.t. the two-sample t-test, the sample size n for the two sample t-test is increased to:*

$$n_{Welch} = 2n \frac{\sigma_1^4 + \sigma_2^4}{(\sigma_1^2 + \sigma_2^2)^2} \geq n$$

where the equality holds if and only if $\sigma_1^2 = \sigma_2^2$.

Proof: The degrees of freedom of Welch's t-test is given by

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}} \overset{n_1=n_2=\frac{n}{2}}{=} \frac{(\sigma_1^2 + \sigma_2^2)^2}{2(\sigma_1^4 + \sigma_2^4)}(n-2)$$

In order to obtain the same degrees of freedom as for the unpaired Student's t-test, one should modify the sample size by the factor $2\frac{\sigma_1^4+\sigma_2^4}{(\sigma_1^2+\sigma_2^2)^2}$. It is straightforward to see that:

$$\begin{aligned} \Leftrightarrow 0 &\leq (\sigma_1^2 - \sigma_2^2)^2 \\ \Leftrightarrow 0 &\leq \sigma_1^4 + \sigma_2^4 - 2\sigma_1^2\sigma_2^2 \\ \Leftrightarrow (\sigma_1^2 + \sigma_2^2)^2 &\leq 2(\sigma_1^4 + \sigma_2^4) \\ \Leftrightarrow 1 &\leq 2\frac{\sigma_1^4 + \sigma_2^4}{(\sigma_1^2 + \sigma_2^2)^2} \end{aligned}$$

where equality only holds if and only if $\sigma_1^2 = \sigma_2^2$. □

In order to compute a reasonable effect size for t-tests used, one must specify the hypothetical difference $|\mu_1 - \mu_2|$ and the group standard deviations. The hypothetical difference proceeds the particular research question and can therefore be easily quantified. The uncertainty is more difficult. Typical values are extracted from available scientific literature. In case literature is lacking, the following rule-of-thumb can be applied.

Theorem 4.4 *Assume a random vector* $\underline{Y}$ *of size n with idd components. Assume the random variables to follow approximately a normal distribution. Consider the standard deviation of Y unknown but the range of the data equals r. Then one considers effect sizes which satisfy:*

1. $e_1(\mu_1) > \frac{6|\mu_1-\mu_0|}{r}$
2. $e_2(\mu_1,\mu_2) > \frac{3|\mu_1-\mu_2|}{r}$ *if balanced* $n_1 = n_2$

The proof follows immediately from $r > 6\sigma$.

Example 4.1 *We take the folic acid example described in Example 2.2. Assume we wish to compute the required sample size with a power of* 80%*, a confidence of* 95% *wherein the minimal difference to be detected equals* $70\mu g$*. Consider standard deviations for the groups regular diet and supplements which equal respectively 36 and 58* μg*. The effect size equals* $e_2 = 0.725$ *with a minimal sample size of* $n \geq 15$ *leading to a balanced design with approximately 8 sample per group.*
One can compensate the loss in efficiency due to the differences in standard deviations. The Satterthwaith formula induces an efficiency loss leading to increase the total sample size by 1.1969 such that groups sizes of 9 samples is required.

Example 4.2 *Let us consider the glaucoma application in Example 2.7. Assume one wishes to detect a* 20% *decrease in eye pressure w.r.t. the untreated group revealing an average pressure of 21 mmHg. Consider that the pressure range is 17 mmHg for each of the treatment groups. The effect size is minimally given by* $e_2(\mu_1,\mu_2) > \frac{3|\mu_1-\mu_2|}{r} = \frac{12.6}{17} \approx 0.74$*. Hence, a confidence of* 95% *and power of* 80%*, we obtain a sample size which is minimally 15 or balanced 8 samples per group.*
Of course one wishes to compare two treatments to the control group, leading to two pairwise comparisons. A Bonferroni correction can be applied which results in a confidence of 97.5% *or a minimal sample size of 18 or 9 samples per group.*

Example 4.3 *Determining differences in Blood-Activity of radiolabeled Nanobodies. The different prosthetic groups and chelators involved in labeling the nanobody, together with the radioisotopes (F-18/Ga- 68) employed could affect the blood residence time of the nanobody. Hence, the blood radioactivity will be evaluated in C57BL/6 (WT) mice.*
The following experimental groups are considered:

- *anti-MMR 3.49 nanobody with aminooxy conjugated via lysines labeled with* ^{18}F*-FDR.*
- *anti-MMR 3.49 nanobody with site-specific aminooxy conjugation labeled with* $^{1}8F$*-FDR.*
- *anti-MMR 3.49 nanobody with site-specific RESCA conjugation labeled with* $Al^{18}F$*.*
- *anti-MMR 3.49 nanobody with site-specific NOTA conjugation labeled with* ^{68}Ga*.*

The baseline is set at 100% *indicating the starting radioactivity of injected activity per gram of tissue (%IA/g)). The alternative set at* 80% *to detect indicates the radioactivity in blood measured at different time points (%IA/g), given by Gallium-68 and Fluorine-18 half life. A power of 80% is requested with a confidence of 90%.*
First we compute the minimal effect size where we consider the range of 100% *leading to* $e_2(\mu_1,\mu_2) > \frac{3|\mu_1-\mu_2|}{r} = \frac{60}{100}$*. A total of 4 experimental groups leads to 6 pairwise tests such that accordingly to a Bonferroni correction a confidence of* $1 - \frac{\alpha}{6} = 98,33\%$ *is applied. Hence, we use* $z_{1-\frac{\alpha}{12}} = z_{0.9917} = 2.395$ *while the power issues* $z_{1-\beta} = z_{0.80} = 0.84$*. We obtain a sample size of minimally* $n \geq 30$ *or per group 15 samples. The total sample size equals 60 mice.*

Example 4.4 *Mice will be injected subcutaneously with 3 × 105 B16-F10/HER2 tumor cells (in 50 µl PBS), while the mice are sedated with 2.5% isoflurane. Tumor growth will be followed on a daily basis. When tumors reach a volume of* $200mm^3$*, half of the mice will be treated with apyrase, as apyrase degrades ATP released upon ICD. The other half are left untreated. Next the mice of the apyrase treated and non-treated group will be subjected to a high, intermediate or low dose of radioactive chemotherapy consisting of:*

1. *classical RT*
2. *β-particle TRNT using 177Lu-Nbs*
3. *α-particle TRNT using 211At-Nbs*
4. *non radioactive labeled Nbs (control 1)*
5. *non-bound 177Lu (control 2)*
6. *non-bound 211At (control 3)*

Tumor growth will be followed on a daily basis. After 20 days, the groups in experimental conditions (4-6) both in the pretreated apyrase and non-pretreated group is expected to hold tumors reaching 1500 mm^3*, whereas for conditions (1-3) the pretreated apyrase group is expected to hold a tumor growth reaching 1250* mm^3 *and 750* mm^3 *in the non-pretreated group. The standard deviation in the tumor growth can be considered equalling 250* mm^3
For every experimental condition pretreatment with apyrase is compared to no pretreatment with a power of 90% *and confidence of* 95%*.*
To perform the sample size calculations, we observe that no effect is expected in conditions (4-6) serving as control groups, while in groups (1-3) one expects a difference in tumor size of 500 mm^3*. Hence, the effect size equals:* $e_2(\mu_1, \mu_2) = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}\sqrt{2}} = 1$*. A Bonferroni correction requiring 6 pairwise comparisons leads to a confidence level of* $1 - \frac{\alpha}{6} = 99.17\%$ *leading to* $z_{1-\frac{\alpha}{12}} = 2.395$*. The power of* 90% *implies* $z_{1-\beta} = z_{0.80} = 1.28$*. As a result, the minimal sample size is given by* $n \geq 14$ *mice or balanced groups of 7 mice. In total one requires 84 mice.*

LITERATURE

The course follows the main ideas of [A. Agresti, Foundations of linear and generalized linear models, Wiley, 2015] where additional details are added. The mathematical details are in part coming from [A. Stuart, K. Ord and S. Arnold, Kendall's advanced theory of statistics - Vol. 2A, Wiley 2004] and own notes and calculations.