Escuela Universitaria de
Ingeniería
Técnica Industrial
Universidad Zaragoza

DETERMINATION OF ASPARTAME IN SOFT DRINKS USING HPLC

BACHELOR OF CHEMICAL ENGINEERING

FACULTY OF ENGINEERING - UNIVERSITY OF SOUTHERN DENMARK

SPRING 2012

Mohammad Adnan Hajjaj

Morten Gildsig Poulsen

Morten Hjort Løvendahl Nielsen

Juan Viamonte Dominguez

TABLE OF CONTENTS

FIGURE LIST

TABLE LIST

Resume

Dette projekt handler om at bestemme indholdet af aspartam i forskellige colaprodukter ved hjælp af High Pressure Liquid Chromatography (HPLC). Der undersøges fire forskellige mærker; Coca-Cola Zero, Coca-Cola Light, Pepsi Max og Harboe cola Minus.

Der er skrevet teori om aspartam, miljø og lovgivning for stoffet. Derudover er der lavet et teoretisk afsnit omkring HPLC og statistik. I laboratoriet fremstilles en metode til undersøgelse af aspartam i colaprodukter. Denne metode benyttes til at indsamle data for de forskellige fabrikater, som undersøges ved hjælp af forskellige statistiske parametre.

I metodeudviklingen bliver det fundet, at aspartam kan adskilles fuldt fra de øvrige indholdsstoffer i colaprodukterne. Aspartamindholdet er bekræftet ved hjælp af et certificeret standardstof. I laboratoriet benyttes en Agilent HPLC med en C-18 kolonne, 150 mm længde med 5 µm partikeldiameter. Metoden benytter et gradientprogram med acetonitril og 22 mM fosfatbuffer pH 2,5 [31, s. 2].

Metoden bekræftes af statistiske beregninger, der viser at data for den samme cola er både repeterbart og reproducerbart. Gennem tests af nyere colaer af samme mærker, vises det at koncentrationen af aspartam for det samme mærke varierer signifikant fra flaske til flaske. Genfindingen på metoden vises at være tæt på 100 %, hvilket er yderst essentielt for metodens validitet.

Abstract

The aim of this project is to determine the levels of aspartame in various cola products using High Pressure Liquid Chromatography (HPLC). We examine four different brands, Coca-Cola Zero, Coca-Cola Light, Pepsi Max and Harboe Cola Minus.

There is written a section of theory about aspartame, the environment and legislation for the substance. Additionally, there is made a theoretical section on HPLC and statistics. In the laboratory, a method for the examination of the aspartame in the coke product is created. This method is used to collect data for the various fabricates which are examined using various statistical parameters.

In method development, it is found that aspartame can be separated completely from the other ingredients of the cola products. Aspartame content is verified by means of a certified standard substance. In the laboratory an Agilent HPLC is used with a C-18 column, 150 mm long and a particle diameter of 5 micrometers. The method uses a gradient program of acetonitrile and 22 mM phosphate buffer pH 2.5 [31, p. 2].

The method is confirmed by statistical calculations that show that the data for the same cola is both repeatable and reproducible. Through tests of newer colas of the same brands, it is shown that the concentration of aspartame in the same brand vary significantly from bottle to bottle. The recovery of the method is close to 100 %, which is extremely essential for the validation of the method.

Preface

This report describes the results of the project on the 4th semester at the Institute of Chemistry, Biotechnology and Environmental Technology (KBM) at the Technical Faculty of the University of Southern Denmark. The report is a written account of the practically performed laboratory work and the statistical tests.

The report is formulated and created by the project group. The written statement should provide a basis for the project examination in June 2012th.

The aim of the project is to determine the levels of aspartame in cola products. During the project, the basic theory behind this issue and any problems that might arise during the practical work are examined.

To describe this, the course "instrumental analytical chemistry and applied statistics" is used.

This report is conducted by group 1 in the laboratories of KBM. Therefore, the group would like to thank the laboratory staff for their expertise and help.

In addition, the group will also show its gratitude to the supervisors: Ole Thygesen and Victoria Blanes-Vidal.

Morten Gildsig Poulsen

Morten Hjort Løvendahl Nielsen

Mohammad Adnan Hajjaj

Juan Viamonte Dominguez

1. Introduction

1.1. Problem Statement

Project title: Determination of aspartame in soft drinks using HPLC.

1.1.1. Introduction

In the modern days, use of sweeteners in soft drinks and other light products are widespread. Before sweeteners were introduced in everyday products, sugar was the most common compound used to sweeten food. As a consequence, a high calorie intake was inevitable for those who wanted sweet food. This could cause problems like obesity, which could in turn cause heart disease and other lifestyle illnesses [30, p. 2579].

The use of sweeteners has an influence on the health of the human body. Research has shown that sweeteners like Aspartame can decompose into harmful products. A consequence of this could for example be cancer. The sweeteners decompose in the human body, meaning that they are converted via a chemical reaction to dangerous chemicals like methanol [30, p. 2589].

1.1.2. The aim of the project

- To investigate the amount of aspartame in soft drinks
- Compare amounts of aspartame in different types of soft drinks
- Check if the amount of aspartame in a specific soft drink is consistent
- Compare the amount of aspartame to what is legal

To obtain the information we seek, a High Pressure Liquid Chromatography apparatus will be used. Standard curves will be created for aspartame, so the amounts in the soft drinks can be determined.

1.1.3. Hypothesis

We would like to investigate the use of aspartame in soft drinks. To do this, we use HPLC to find the concentration of aspartame in different soft drinks. We would also like to look at a single product, and find the variance in the concentration of aspartame per drink.

Besides finding what we previously described, we would like to validate, and perhaps improve on a method for analyzing aspartame on an HPLC apparatus.

1.1.4. Reason for using HPLC

We have chosen HPLC, because there are methods available for almost any kind of analysis. The biggest reason for using HPLC is that all literature found on the subject uses HPLC for the analysis. We know that an analysis of aspartame via HPLC is possible, and we can compare our research to other sources.

1.1.5. The analysis

We use Reversed-Phase Chromatography for the analysis of aspartame. The detection device would be Ultra Violet-PDA (Photodiode Array) or Ultra Violet-DAD (Diode Array Detector). For the stationary phase, a C-18 column is used.

A standard curve will be made, so that it is possible to determine the concentrations of the samples. Two different methods are found:

1.1.6. Method 1

The sample is prepared by degassing (removing the CO_2 gasses by ultrasonic bath) the soft drinks. The drinks are then mixed with 1 mL of two solutions known as Carrez solutions. The mixture is then diluted with water to 25 mL, and centrifuged. The liquid phase is then diluted (1:1) in phosphate buffer solution containing 34 % acetonitrile and 4 % methanol, centrifuged and filtered.

A volume of 20 µL of the sample is injected. The mobile phase consists of 81 % phosphate buffer, 2 % methanol and 17 % acetonitrile. The pH is 4.3. A gradient program is used: 0-1 min: 0.7 mL/min. 1-2 min: 1 mL/min. 2-8 min: 1 mL/min [3, p. 164].

1.1.7. Method 2

The sample is prepared by diluting 1:5 with water, and filtering with a 0.45 µm filter. A sample volume of 100 µL is injected [3, p. 460-462].

1.2. Background

1.2.1. History of Aspartame

Food additives are strongly used and required by the modern food technology production, fulfilling consumer demands and necessities for convenience, variety and choice. Sugar was the most known compound utilized to sweeten food before artificial sweeteners were introduced in household products. Due to this reason, an elevated calorie intake was inevitable for consumers who wanted sweet aliments. This was an important cause of problems like obesity, which eventually causes heart diseases and other illnesses.

Artificial sweeteners are used in food to enhance its keeping quality, maintain its nutritive quality and making it attractive or to aid in its storage, packaging or processing. Among them, the use of sweeteners as aspartame in soft drinks and other light products is widely used around the world [1].

Aspartame (N-L-α-aspartyl-L-phenylalanine methyl ester) is a low-calorie artificial sweetener used to reduce calorie beverages and foods and to sweeten a large variety of low-calorie products, including low-calorie soft drinks [2].

During digestion, aspartame is separated in three components: methanol and the amino acids aspartic acid and phenylalanine, which are then absorbed by the blood and utilized in common body processes. Aspartame or its components are not accumulated in the body. These three components are utilized in the body in the same way as when they come from common aliment.

Despite the fact that aspartame can safely be used by a healthy individual, it is recognised for a long time that disproportionate intake of phenylalanine, can pose a hazard to people affected from phenylketonuria, an inherited metabolic disorder. Due to this reason, all alimentary products containing aspartame must indicate the presence of phenylalanine on the label [1].

In 1996, J.W. Olney developed an article suggesting the existence of a connection between the marketing of aspartame and the increase of the incidence of brain tumours in the EEUU, which again started the debate on the risks of aspartame to human health by its consumption. The debate has been followed up by the media, with noticeable importance on the Internet In 2007, as public concern about aspartame still continued despite the risk assessments and studies that had been undertaken. The Advisory Forum of EFSA (the European Food Safety Authority), consisting of the main national food safety authorities, decided to hold a series of meetings with national experts with relevant scientific knowledge related with aspartame, nominated by their Member States. All the published literature was reviewed and all the additional

evidences and literature and data that EFSA had gathered until 2008 was taken into consideration. In 2010, a report compiled from these meetings was presented along with observations from stakeholders received presented as a public consultation. Experts concluded that no new evidences were identified to suggest reconsideration of the previous opinions of SCF (Scientific Committee on Food) and the EFSA, but also perceived that the public concern relating to aspartame remains high [3].

Nowadays, aspartame is used by over 200 million consumers throughout the world and is utilized in more than 6.000 alimentary products including carbonated and powder soft drinks, tabletop sweeteners, chewing gum, frozen desserts, confections, yogurt and even in some pharmaceuticals such as sugar-free cough drops and vitamins (European Parliament and Council Directive 94/35/EC 2011). The Joint Expert Committee on Food Additives (JECFA) of the World Health Organization, the Food and Drug Administration, the Scientific Committee for Food of the European Community and several regulatory agencies present in more than 100 countries have already reviewed aspartame and found this sweetener safe for use [1].

1.2.2. Characteristics and properties of aspartame

Aspartame (N-L-aspartyl-L-phenylalanine-1-methylester) is a white, odourless, crystalline powder which is obtained by synthesis from two amino acids, L-phenylalanine and L-aspartic acid. Around 16.000 tons/year of aspartame are produced for worldwide consumption. Aspartame is approximately 180 times sweeter than common sugar in standard concentrations. In Europe, E 951 is the E number of aspartame. It was marketed for the first time by NutraSweet AG and later by Holland Sweetener Company and Ajinomoto.

As a peptide, aspartame has a caloric value around 4 kcal (= 17 kJ) per gram. The amount of aspartame needed to produce sweet taste is so small that its caloric contribution is almost negligible.

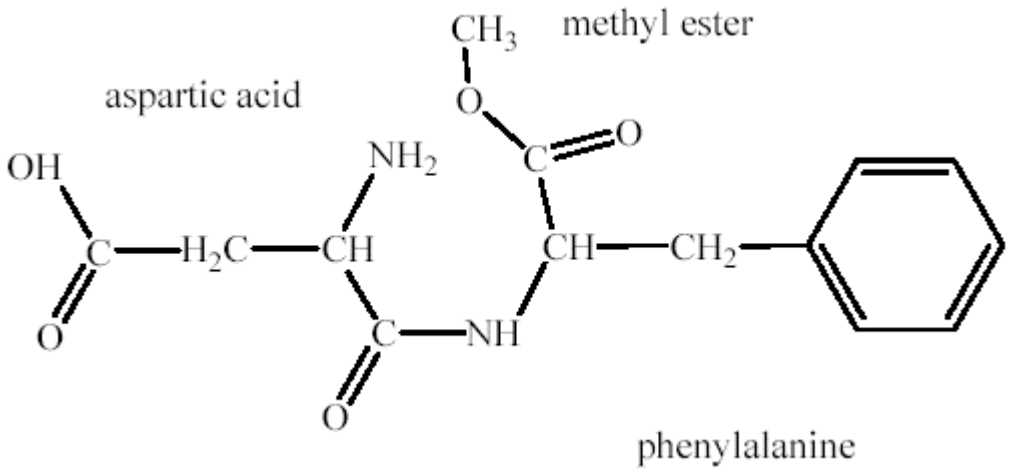


Figure 1.1: Structure of aspartame and its three main components [3].

Although relatively stable in its dry form, pH, temperature, and time are very important factors affecting its stability in solution. It is very stable in dry form: at 105°C a loss of 5% is observed after 100 hours under this condition. At 120°C and after 80 hours of treatment, a 50% loss is obtained. Under pH=3, aspartame is unstable and it hydrolyzes to produce aspartyl phenylalanine and above pH=6, it cyclizes to form its main impurity, 5-benzyl-3,6-dioxo-2-piperazine acetic acid (diketopiperazine). Both forms result in a considerable loss of sweetness.

Its molecular weight is 294.3 Daltons.

The minimum solubility of aspartame is reached at pH 5.2 (13.5 mg/ml at 25°C) and the maximum solubility at pH 2.2 (20 mg/ml at 25°C) [4].

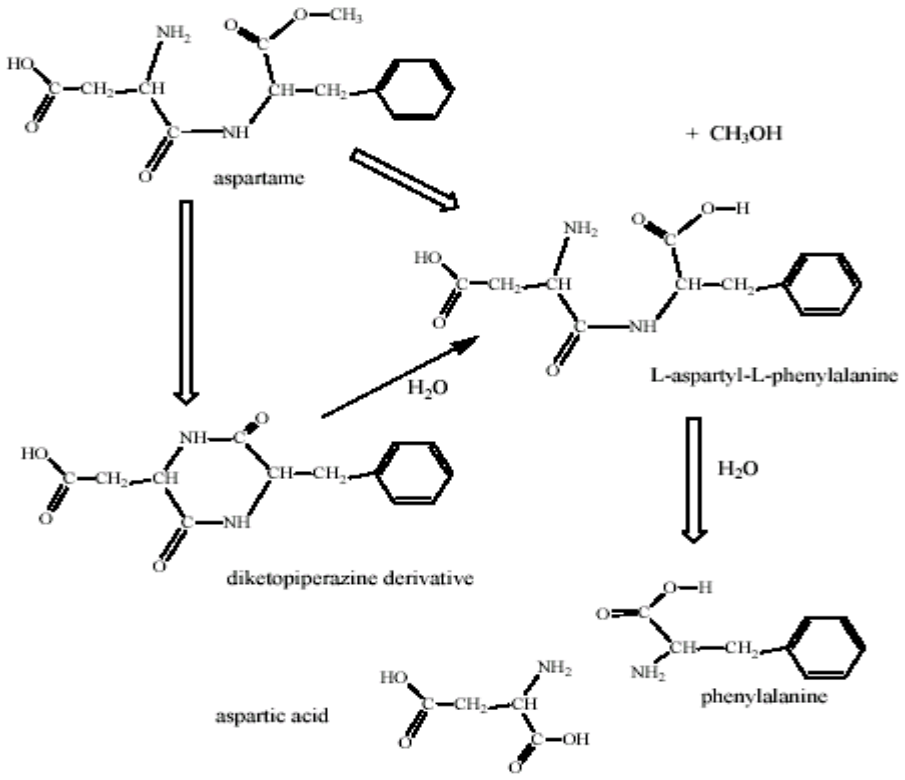


Figure 1.2: Decomposition of aspartame into L-aspartic acid and L-phenylalanine [3].

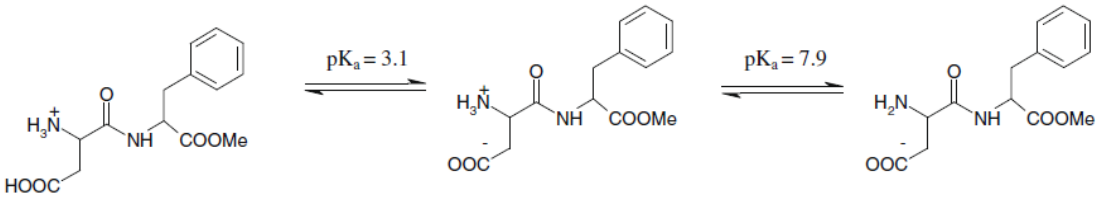


Figure 1.3: Molecular structure and pka values of aspartame [2].

1.3. Objectives of study

We would like to investigate the use of aspartame in soft drinks. To do this, we use HPLC to find the concentration of aspartame in different soft drinks. We would also like to look at a single product, and find the variance in the concentration of aspartame per drink. These tests would be done over several days.

Besides finding what we previously described, we would like to validate, and perhaps improve, on a method for the HPLC for analyzing aspartame.

1.4. Legislation

Aspartame has been authorised for use in soft drinks, foods and as a table-top sweetener by some Member States since the 1980s. The European legislation arranged its use in food production in 1994 following several safety evaluations carried on by the Scientific Committee on Food (SCF) in the years 1984 and 1988. Continuous reviews of the compound data were developed in 1997 and 2002 by the SCF. No danger concerns regarding possible developmental and reproductive toxicity, carcinogenicity or genotoxicity were found. Furthermore, in 2006 and in 2009 the Food Additives and Nutrient Sources added to food (ANS) and the Scientific Panels on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food (AFC) evaluated two long-term carcinogenicity tests in multiple rats orally exposed to aspartame. This study was developed by the European Ramazzini Foundation (ERF) and in both investigations a quantitative dose-related growth of malignant tumours in female and male rats was reported. The ANS and AFC Panels determined that testing all the evidence available, there was no manifestation of any carcinogenic or genotoxic potential of the substance and there was no reason to update the Acceptable Daily Intake (ADI) for aspartame of 40mg/kg person body weight previously established. In a study published in 2010, the EU Member States asked National Experts to review the scientific data and literature about aspartame since 2002. They manifested that there was no necessity to reevaluate the previous opinions on the substance published by the SCF and the European Food Safety Authority (EFSA) Panels [5].

In 2010, two scientific articles were published, describing important injurious for health consequence of sweetener ingestion. The paper by Halldorsson [6] suggests a relation between a proliferated risk of preterm delivery and use of aspartame sweetened soft drinks. The publication by Soffritti [7] describes that aspartame is a risky carcinogenic agent in mice.

The EFSA informed the ANS that on February 2011, the European Commission requested EFSA for scientific assistance (according to Article 31 of Regulation (EC) No 178/2002) to start technical evaluation to check if the two previous mentioned publications should prompt a revision of the existing assumptions of EFSA related to the risk of artificial food additive sweeteners [8].

In May 2011, the European Commission asked EFSA to overtake the complete re-evaluation of the risks of aspartame to 2012. Already arranged for completion by 2020, the analysis of this sweetener is part of the efficient re-evaluation of all artificial food additives authorised by the EU organisms earlier to 20 January 2009, as assumed under Regulation EU 257/2010 [5].

Although all the controversy, the current legislation is: "Directive 94/35/EC - sweeteners for the use in foodstuffs" [9] and it is still using the following legal limits in soft drinks:

94/35/EC

EC No	Name	Foodstuffs	Maximum usable dose
E 951	Aspartame	**Non-alcoholic drinks**	
		— Water-based flavoured drinks, energy-reduced or with no added sugar	600 mg/l

Table 1.1: "Directive 94/35/EC - sweeteners for the use in foodstuffs"

In Denmark the Ministry of Food, Agriculture and Fisheries (FVM) established the "positivliste" (eng: positivelist) which contains the limits of chemical substances that the food companies are allowed to use in their production. The legal limit of amount or concentration of the aspartame that is added to the food is set to be 600 mg/L [10] as in the rest of the European countries. So the food companies have to undergo these conditions and rules, so that the production of a product containing aspartame does have a concentration below 600 mg of aspartame pr. Liter product [11].

2. HPLC theory

HPLC, or High Pressure Liquid Chromatography, is a separations technique that is used in modern laboratory practice. Like the name, this method is a kind of a chromatography, where a sample is separated by using an eluent and a reference. But HPLC is a much more advanced chromatographic technique, which gives a more precise result, than the normal TLC plates.

HPLC is not a very old technique, the first sign of chromatography was discovered by the Russian botanist Mikhail Tswett in the 1903 [12, p.3]. He separated plant pigments on chalk packed in glass columns. Since then the technique has been developed during time, and in 1952 the first chromatography machine was invented and was called the GC (Gas Chromatography).

The discovery of this machine and its theoretical background was essential for the development of the LC (liquid Chromatography) technique. About 10 years later, the first High Performance Liquid chromatography machine was made. And during time improvement has made the machine more precise and better to use.

2.1. Advantages and disadvantages

Using advanced machines like HPLC have advantages and limitation like every machine around the globe. It has an incredible precision and versatility, and these two factors make it special. HPLC can measure almost everything that can absorb UV-light or be ionized by mass spectrometric detection. And the measurement can be done with very good precision. But if the sample contains chemical substances that doesn't absorb UV-light or cannot be ionized, then it's problematic for the HPLC to make the measurement. HPLC has a very good detection limit, which can detect up to nano, pico and femtogram levels. It can also make measurements of up to 80 % of all existing chemical compounds, compared to the GC that can analyze around 15 %.

2.2. HPLC device

The HPLC machine is divided into a column, pump, degasser and a valve.

The column is made of metal or glass, and is used to press the mobile phase and the sample through it, with very high pressure.

The degasser is used to remove gasses that are in the sample. Sometimes ultrasound is used if the machine doesn't have a degasser. The valve is used to purge the machine, to remove chemicals.

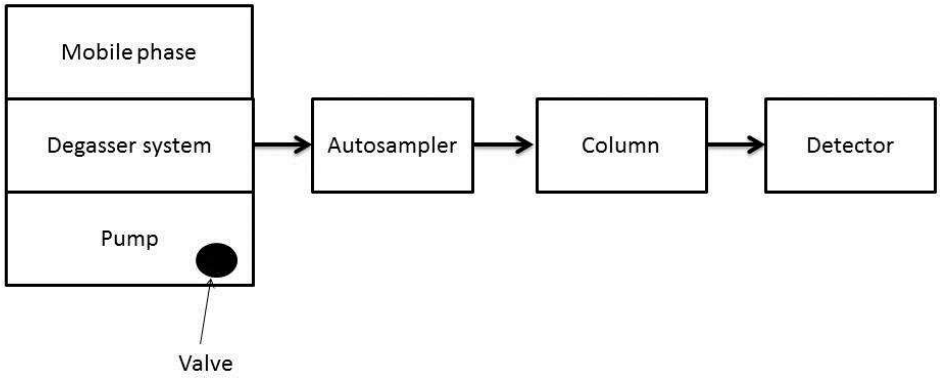


Figure 2.1: HPLC main components.

The sample the machine needs to analyze is separated in the machine by a distribution of its analytes in two different phases, and known as the mobile and stationary phase.

These two phases have two different polarities, and attracts different analytes that are in the sample.

The mobile phase is a liquid, and the stationary phase is sorbents packed inside the column. There are many different organic solvents that could be used as the mobile phase, such like hexane that is a very unpolar organic solvent. The stationary phase could for example be porous silica particles packed inside the column.

The separation simply happens in the column (packed with sorbents) where the mobile phase is pumped inside the column at high pressure. The analytes will then interact with the phase with the same polarity as their own.

The following picture explains how this works:

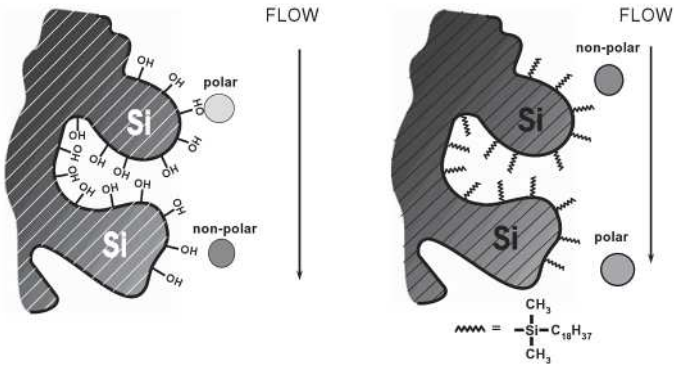


Figure 2.2: Schematic diagrams depicting separation modes of (a) normal-phase chromatography (NPC) and (b) reversed-phase chromatography (RPC). [12,p.6]

The figure shows two different flows of the same sample, which contains polar and non-polar components. The figure to the left shows the interaction between the polar “silanol” inside the column, and the polar component. The non-polar component will just pass on with the flow without any interaction.

The figure to the right is like the one to the left, but the interaction here happens with the non-polar phase and non-polar components. The polar components pass through the column with the flow.

2.3. Separations Modes

There are different kinds of separations modes for the HPLC. Each separation has its own name, and is named after the purpose it is used for.

The modes are:

- NPC (Normal Phase Chromatography)
- RPC (Reversed Phase Chromatography)
- IEC (Ion Exchange Chromatography)
- SEC (Size Exclusion Chromatography)

The most common modes that are used when running HPLC, are NPC and RPC.

NPC is a separation mode based on the adsorption and desorption of the analytes in the chemical sample, with the polar stationary phase. The mobile phase is here a nonpolar organic compound.

This type of chromatography is used when the analytes are nonpolar, because they elute first with the mobile phase, and then the polar analytes elutes slowly after, because of their interaction with the polar phase.

The RPC is the reversed or opposite mode of the NPC. The mobile phase is a polar compound and the stationary phase is an organic nonpolar compound. The mobile phase could be water, methanol or acetonitrile (ACN). The stationary phase is solid particles that are covered with a long chained organic and nonpolar compound.

This kind of separation mode is used for polar analytes and is the most popular mode that is used in more than 80 % of all HPLC analyses [12, p.7].

IEC is a separation mode based on the exchange of ions. The stationary phase is typically cationic, where the mobile phase is anionic [12, p.7].

SEC is also a kind of HPLC analysis, which is based on the molecular size of the particle, where the large particles migrate quickly, while small particles penetrate through the pores and migrates slowly through the column [12, p.9].

2.4. Retention time & Void time

The result of a HPLC analysis is a chromatogram. The chromatograms have on the x-axis the time it takes for the analytes to be detected and the absorbance on the y-axis.

When a sample has been injected, the time between the injection and the top of a peak is called the retention time(t_R) .

The "dead time" (retention time in the mobile phase), or the void time (t_M), is the first peak, or also called the first baseline disturbance by the sample [12, p.17].

The adjusted retention time (t'_R) is calculated by substract the retention time from the void time:

$$t'_R = t_R - t_M$$

The adjusted retention time is the time, the sample remains in the stationary phase.

A chromatogram from a HPLC analysis illustrates how to find these times:

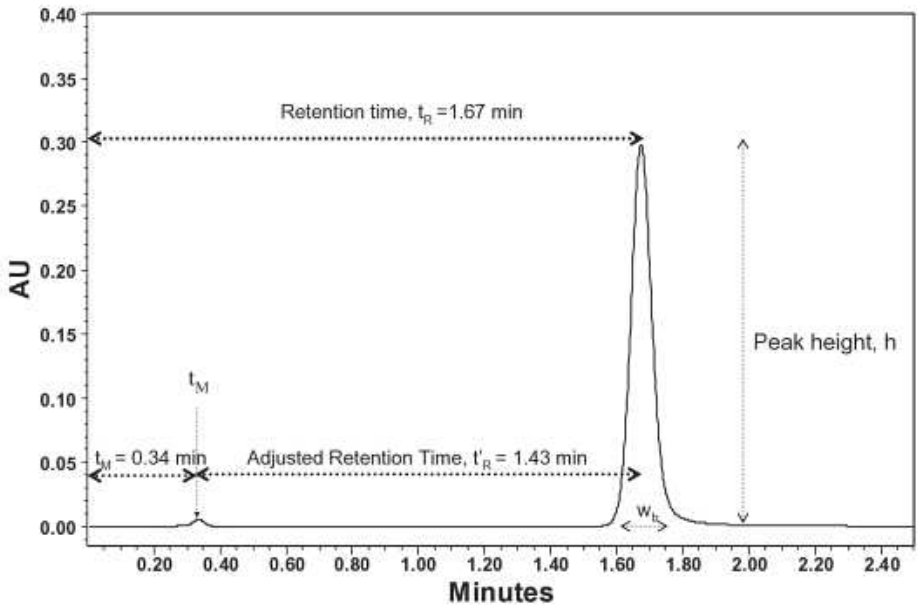


Figure 2.3: A chromatogram showing retention time (tR), void time (tM), peak width (wb), and peak height (h) [12, p.17].

Where W_b is the width of the peak and h is the peak height.

2.5. Volumes

There are different kinds of volumes, during analysis of chromatograms; retention volume (V_R), Void volume (V_M) and the peak volume.

The retention volume, describes the volume required of the mobile phase to elute the sample analyte, at a particular flowrate (F):

$$V_R = t_R * F$$

The void volume is the total amount of the mobile phase contained in the column. The void volume can be estimated by 3 different equation:

$$V_M = t_M * F$$
$$V_M = 0{,}65 * V_c$$
$$V_M = 0{,}65 * \pi * r^2 * L$$

Where:

V_c = Volume of empty column

r = inner radius of column

L = length of the column

The peak volume is the volume of mobile phase containing the eluted peak:

$$Peak\ Volume = w_b * F$$

The peak volume can also be calculated by using the equation that contains the number of theoretical plates (N) and retention factor (k).

$$Peak\ Volume = \frac{4V_R}{\sqrt{N}} = \frac{4V_M(1+k)}{\sqrt{N}}$$

The definitions of theoretical plate number and retention factor will be explained further during this section.

2.6. Retention factor (k) & partition coefficient (K)

The degree of retention of the sample in the column is called the retention factor. It is defined as k, and is the time difference between the adjusted retention time (t'_R) and the retention time (t_M):

$$k = \frac{t'_R}{t_M} = \frac{t_R - t_M}{t_M}$$

The k value decides if the component is retained or unretained in the stationary phase.

k=0 it is unretained

k>20 it is retained

Usually the k value in most analyses is between 1 and 20.

The distribution, or the different between the concentration of the analytes in the stationary and mobile phase, is described by the partition coefficient K, and is estimated by dividing these two concentrations:

$$K = \frac{[X]_s}{[X]_m}$$

Where

$[X_s]$ = concentration of analytes in the stationary phase

$[X_m]$ = concentration of analytes in the mobile phase

2.7. Selectivity (α) & Solvent strength

The selectivity or the separation factor is the ratio or difference between two retention factors.

$$\alpha = \frac{k_2}{k_1}$$

For a good peak separation, the selectivity must be >1.

A change in the stationary phase, and the composition of the mobile phase, affect the value of the partition coefficient K.

Variations affect the selectivity too, because a change in the phases means a change of the retention time and the retention factors.

The solvent strength refers to the capability of a chemical substance to elute analytes through a column. The strength of chemicals used in HPLC was defined by Hilderbrand and are listed in a scale called the hilderbrands elution strength scale (E^0):

Solvent	Solvent strength (E°)	bp (°C)	Viscosity (cP) at 20°C	UV cut-off (nm)	Refractive index
n-Hexane	0.01	69	0.31	190	1.37
Toluene	0.29	78	0.59	285	1.49
Methylene chloride	0.42	40	0.44	233	1.42
Tetrahydro-furan	0.45	66	0.55	212	1.41
Acetonitrile	0.55–0.65	82	0.37	190	1.34
2-Propanol	0.82	82	2.30	205	1.38
Methanol	0.95	65	0.54	205	1.33
Water	Large	100	1.00	<190	1.33

E° (solvent elution strength as defined by Hildebrand on alumina). Data extracted from reference 2 and other sources.

Table 2.1: Common HPLC Solvents and Their Properties [12, p.27]

Solvent strength is associated with the polarity of the solvents. In NPC (normal phase chromatography) the nonpolar compound hexane is a weak solvent, because NPC have a polar stationary phase. The opposite is true in RPC, because the stationary phase is a nonpolar compound. An increasing of the solvent strength will decrease the retention time (t_R), the retention factor (k), the selectivity (α) and the resolution (R_s).

2.8. Buffers

In some analyses the modification of the pH of the mobile phase is required, if the analytes will elute through the column. Ionized form of analytes doesn't partition very well with the nonpolar stationary phase in RPLC, and it has therefor a lower k-value, which means a lower retention time and bad separation of the peaks.

An example is given from the book [12, p. 31] to explain the effect of the pH on the separation:

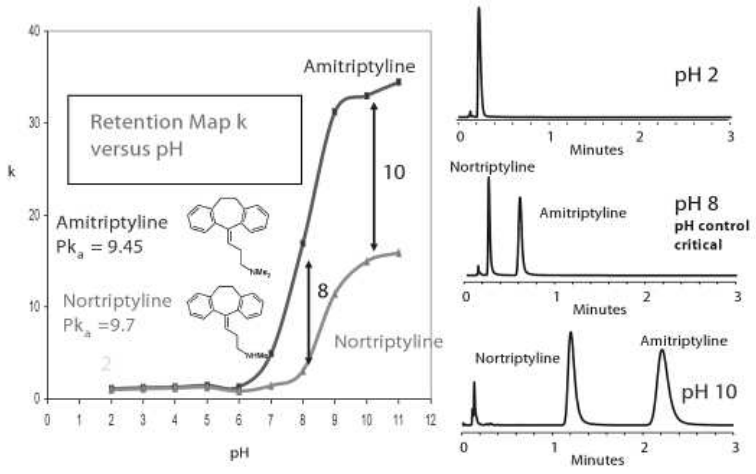


Figure 2.4: Retention map and chromatograms of two basic antidepressants using mobile phases at various pH with percentage organic modifier being kept constant. The diagram illustrates the importance of pH in the separation of basic analytes [12, p.31].

These two drugs ionize at two different pH-values. At pH = 2 there is no separation and therefor one peak is provided. A more basic condition about pH=8 gives a slightly good separation, but at pH=10 it's a perfect separation. So the purpose with this example is to illustrate the importance of the pH on the separation of the analytes.

2.9. Column efficiency & number of plates

The column efficiency depends on the number of plates in the column. A column with many plates is a very efficient column. An efficient column produces perfectly good and sharp peaks. The separation of the samples is also much better with increasing efficiency. The number of theoretical number plates (N) is defined by this equation:

$$N = \left(\frac{t_R}{\sigma}\right)^2 = \left(\frac{4t_R}{w_b}\right)^2 = 16\left(\frac{t_R}{w_b}\right)^2$$

Where:

σ = standard deviation of the peak

2.10. HETP (Height Equivalent to a Theoretical Plate)

The distillation process from the industry was the first with the concept of a column with plate. A longer column would have an increasing number of plates and a good separations technique, to separates materials to many fractions of distillates. An HPLC column doesn't really have plates, but it's the same concept. The correlation between the column height and number of plates is described by this equation:

$$HETP = \frac{L}{N}$$

Where

L = length of column

N = Number of plates

2.11. Resolution (R_s) & Peak Symmetry

Resolution (R_s) is the degree of separation of one or more analytes from other chemical compounds in a sample. The resolution is mathematically describe by dividing the difference of the retention time with the average of the width for two peaks:

$$R_s = \frac{t_{R2} - t_{R1}}{\frac{w_{b2} - w_{b1}}{2}}$$

R_s = 0 there is no separation, as bigger the R_s value get, a better separation is achieved. This figure illustrates graphically the R_s value:

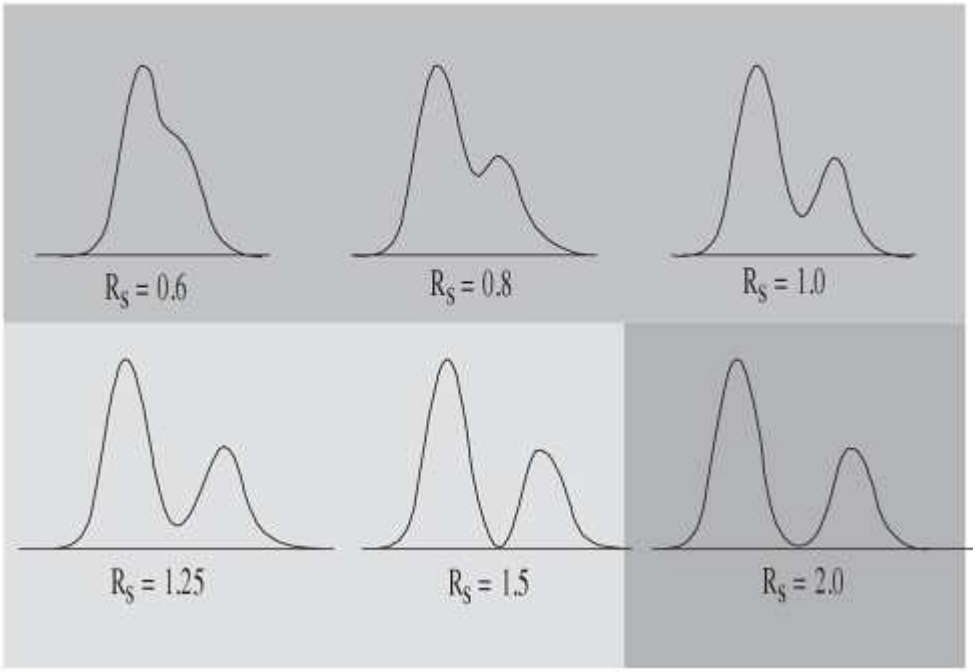


Figure 2.5: Diagrams showing two closely eluting peaks at various resolution values from 0.6 to 2.0 [12, p.25].

The resolution can also be found by using the resolution equation. This particular equation is dependent of the two thermodynamical factors; Retention (k) and Selectivity (α). But it's also dependent on the kinetic factor, column efficiency (N):

$$R_s = \left(\frac{\sqrt{N}}{4}\right)\left(\frac{k}{k+1}\right)\left(\frac{\alpha-1}{\alpha}\right)$$

Efficiency Retention Selectivity

$$R = \frac{\sqrt{N}}{4} \quad \frac{k}{k+1} \quad \frac{\alpha-1}{\alpha}$$

To maximize the resolution a large k-value is required. A smaller k-value will only result in lower resolution and therefore imperfect separation. So if k=0, then R_s = 0 (no separation).

The selectivity should be between 1,01 and 1,50. The selectivity describes the chemistry of the solvent, used as a mobile phase. If the solvent is changed for example from acetonitrile to methanol, it will affect the selectivity.

The column efficiency should be maximized by making the column longer, then the column is more efficient because of its many plates, which gives better separation and therefore a good resolution.

This figure shows graphically how these factors affect the resolution:

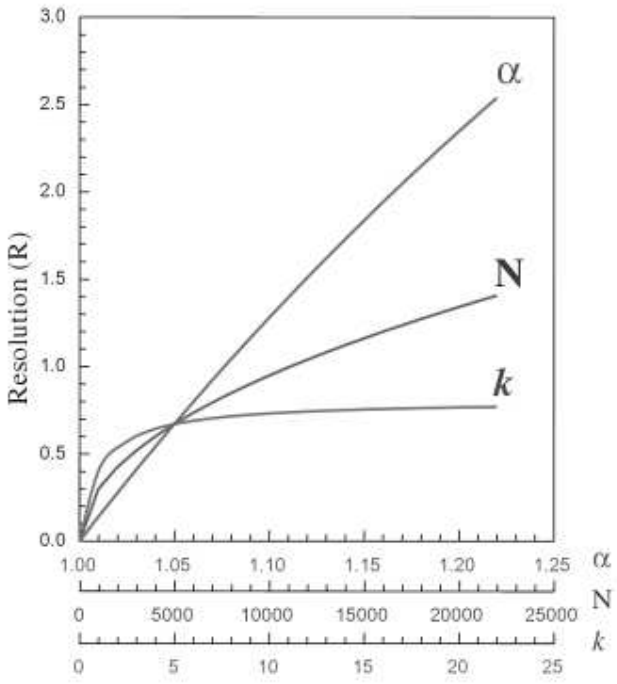


Figure 2.6: Graph illustrating the effects of α, k, and N on resolution [12, p.35].

The best and ideal shape of peaks is the Gaussian peak shape that is symmetrical. It's not all peaks that are completely symmetrical, some of them can be tailing or fronting. A factor called the asymmetry factor (A_s)

illustrates the degree of peak symmetry, and is defined as 10 % of the total peak height ($W_{0.1}$). The tailing factor (T_f) is obligatory in the pharmaceutical industry, because it indicates how symmetrical the peak is and therefor how perfect the method is. A perfectly symmetrical peak has a $T_f = 1$. The asymmetry factor (A_s) have almost the same value as T_f, but sometimes it's larger than T_f. These two factors are described by this figure:

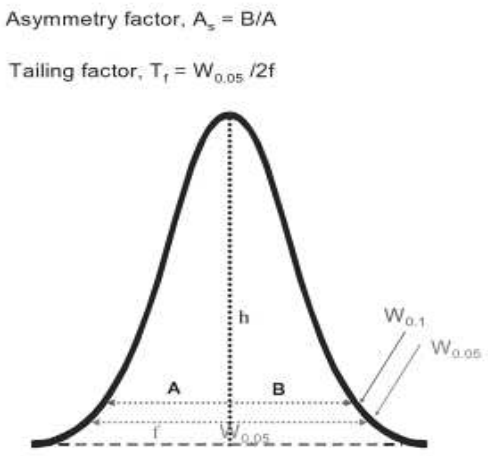


Figure 2.7: A diagram showing the calculation of peak asymmetry (A_s) and tailing factor (T_f) from peak width at 5% height ($W_{0.05}$) according to the USP. Inset diagrams show fronting and tailing peaks. [12, p.25].

2.12. Isocratic & Gradient flow

An HPLC analysis can be done in two different ways or conditions. These conditions are called the isocratic or gradient. Most of analyses are performed with isocratic condition, were no changes in the mobile phase are made.

The isocratic condition is good to use when the mixture or sample is a simple solution, while gradient analysis is good for the more complex solutions. The gradient analysis applies a change in the solvent strength of the mobile phase. Using of the gradient flow will give a better resolution, sensitivity and a higher peak capacity. The disadvantage of using that kind of flow is that the system needs to be optimized. The method development, implementation and transfer are more difficult because of the changes in the strength of the mobile phase. A contamination could also be a result, if the system isn't cleaned well.

2.13. Peak capacity

The equation for the column efficiency, described before, can't be used for the gradient flow, because it's developed for the isocratic condition. Therefore a new equation has to be developed, and that's done by using the calculation for the peak capacity (n).

The peak capacity is the maximum number of peaks that can fit in the chromatogram with a resolution (R_s) value of one. Peak capacity is used to describe the performance of the column under a gradient condition.

The peaks shape isn't the same under isocratic condition, they have different broadness, but the width using the gradient condition is the same for every peak.

$$n = \frac{t_G}{w_b}$$

Where:

t_G = gradient time

Gradient analyses are more difficult to optimize or develop because there are many additional parameters that control the separation process, like for example the beginning and ending solvent strength, the flow rate (F), and gradient time (t_G). Even the retention factor is more difficult in gradient analysis, because there are many different k values, therefor an average k-value has to be presented (k*):

$$k^* = \frac{t_G F}{1.155 \Delta\varphi V_M}$$

Where:

S= constant (close to 5)

F=flowrate

Δϕ= change in volume fraction

V_M = void volume

To know when a gradient analysis or isocratic is favorable, as usual there is a thumb rule called the 0.25 Δt_G rule. This rule is used to check if the sample runs best on gradient or isocractic conditions.

3. Statistical Theory

3.1. Introduction

Statistical parameters are used to validate experimental results in order to properly treat data obtained from an analysis. By means of statistical parameters, it is possible to conclude on the results, to see whether they are credible or not.

3.2. The normal distribution

A data distribution can be described in different ways. One of those ways is by using the normal distribution to describe the mean and the variation of the mean by its probability distribution.

Data follow the normal distribution if plots of the values assume a bell-shaped curve, evenly distributed around the mean. The normal distribution, also known as the normal probability density is the most widespread distribution, and follows the formula [13, p. 125]:

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

Where f(x) denotes the probability distribution of x, while μ is the mean of the population and σ is standard deviation of the population. The graph of the normal distribution will look like this:

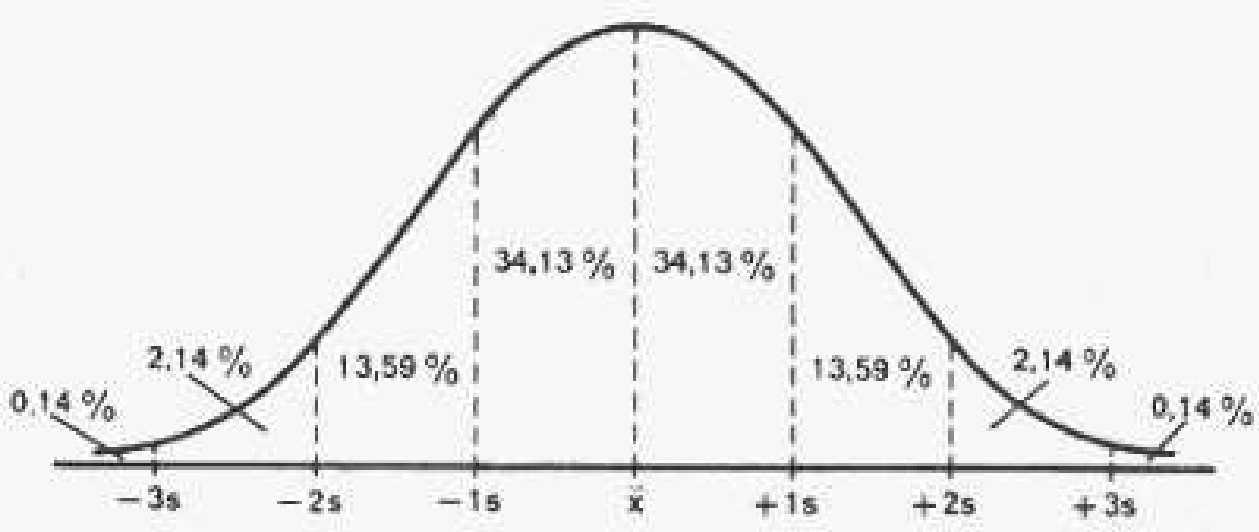


Figure 3.1: The typical graph of the normal distribution [14].

The graph center point is the mean of the data sample. A tall and narrow graph indicates that the standard deviations between the individual values are low, while a low and wide graph indicates that the standard deviations between the individual values are high.

As can be seen from the graph, the graph is divided into a number of intervals. It is indicated how large a percentage of the measurements are found within the given range. On the graph, $\overline{x}$ is mean of all the values, while s indicates the standard deviation of the sample. The intervals are divided into stages on the basis of the mean value, to which s is added or subtracted for each step.

The objective is to show that the chance to obtain a value away from the mean value becomes smaller, and finally infinitesimally small because the total deviation becomes larger. 34.13% of the measurements are located in the middle interval. This can be demonstrated by integrating the function of the normal distribution as given above (in this case for a mean of 0 and a standard deviation of 1):

$$\int_0^1 \frac{1}{\sqrt{2\pi} * 1} e^{\frac{-(x-0)^2}{2*1^2}} = 0.34135 = 34.13\,\%$$

13.59 % of the measurements are located in the interval ($\overline{x}$+s ; $\overline{x}$+2s). This can be calculated like above:

$$\int_1^2 \frac{1}{\sqrt{2\pi} * 1} e^{\frac{-(x-0)^2}{2*1^2}} = 0.13591 = 13.59\,\%$$

If the entire curve is integrated, the total area under the curve will be equal to 1.

The used mean values and standard deviations can be of any value; a normal distribution with a mean of 0 and a standard deviation of 1 is known as a standard normal distribution [13, p. 126].

Calculation of statistical parameters is done using the formulas:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

And for the standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}}$$

Where x_i is each individual data statistic of the given number series, while n is the total amount of numbers [13, p. 25+27].

To calculate a probability using the normal distribution, special tables are used that contain values of a variable z; this value follows the modified version of the equation of the probability distribution, which is:

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt = P(Z \leq z)$$

The last part of the equation shows that F(z) is a cumulative probability [13, p. 126]. This means that a given value of z corresponds to a probability P that only increases as z increases. This equation pertains to a specific table that uses the standard normal distribution.

In order to determine whether an amount of data is normally distributed, a standard quantile plot is used [13, p. 163]. This plot is a special graph, which effectively displays if all / some of the data differs from a normal distribution. For maximum assurance whether the data is normally distributed or not, it is best with at least 15-20 samples. To determine whether the data is normally distributed, all the measurements must be within the dotted lines. An example of a normal quantile plot is:

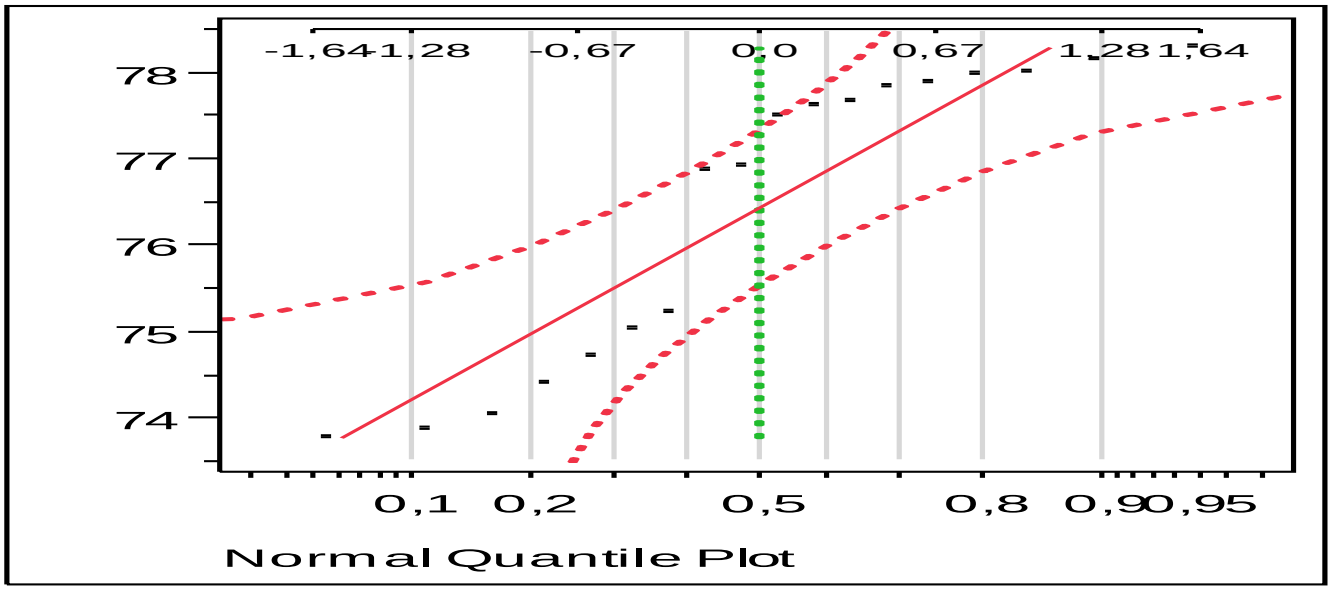


Figure 3.2: Normal quantile plot [15].

3.3. t-distribution

The t-distribution is very similar to normal distribution, inasmuch as the graphs of the two distributions are very much alike:

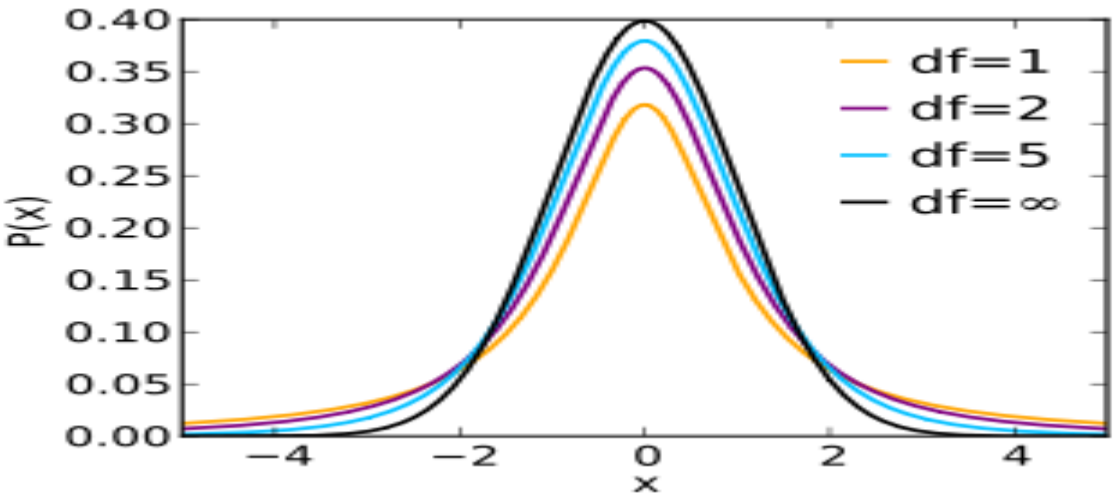


Figure 3.3: t-distribution compared to the normal distribution [16].

The t-distribution, or student's t-distribution, is symmetrical around the mean and bell-shaped like the normal distribution, but Figure 3.3 displays a function that is not found in the normal distribution; df, or degrees of freedom. The degrees of freedom, v, are calculated via the formula [13, p. 187]:

$$v = n - 1$$

As can be seen on the graph, the lower the number of degrees of freedom, the wider the graph is. The higher the number of degrees of freedom, the more the t-distribution approaches the normal distribution. Hence, the t-distribution is a version of the normal distribution which allows for more variance when the number of samples, n, is low. It requires 30 samples or more for the t-distribution to become a good approximation to the normal distribution [13, p. 188].

Furthermore, the t-distribution allows for the use of the sample standard deviation, s, where the normal distribution does not. For a sample with a mean of $\bar{x}$ and a standard deviation s, the random variable:

$$t = \frac{\bar{\bar{x}} - \mu}{\frac{s}{\sqrt{n}}}$$

Where μ is the population mean, follows the t-distribution. A table of data with values of probabilities for different values of t can be used to estimate the probability of a given sample mean in comparison to a

given population mean and sample standard deviation [13, p. 188]. This comes in handy when comparing different samples to each other, to see whether they share a similar mean or not.

3.4. F-distribution

It is often assumed, when testing two different samples of size n_1 and size n_2, that the variances equal each other.

A problem could be to find the difference in these sample variances and to check whether they are similar to each other or not. This is an important factor to consider, when testing to see if two populations have the same variance.

For two variances s_1^2 and s_2^2 with populations n_1 and n_2 respectively, the F-distribution has the random variable F [13, p. 190]:

$$F = \frac{s_1^2}{s_2^2}$$

With the parameters, the degrees of freedom:

$$\nu_1 = n_1 - 1$$

$$\nu_2 = n_2 - 1$$

To test whether two sample variances are different or similar, the F-value is calculated. This F-value pertains to a value on a graph:

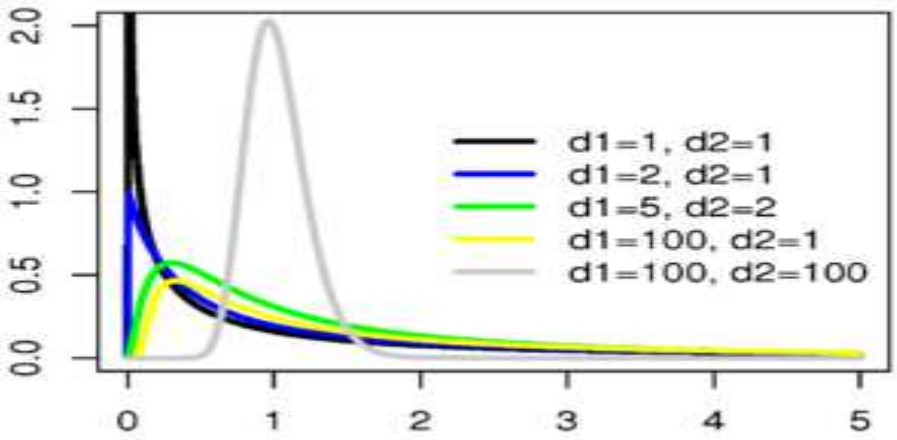


Figure 3.4: An F-distribution, showing how the degrees of freedom change to appearance of the graph [17].

The F-value corresponds to an area to the right on the graph which equals the probability of the two variances being the same. An F-value can be found in a special table, from which it can be pointed out by

using the degrees of freedom and the α-value (the confidence level). In this way, two hypotheses are tested:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

The F-value from the table can be compared to the F-value that is calculated. If the calculated value is bigger than the value obtained from the table, we can reject H_0, also known as the null hypothesis.

3.5. Confidence interval

The confidence interval is a statistic tool used to estimate the area where a specific population mean μ is located. The interval uses the sample mean $\overline{x}$ and gives a lower and upper boundary for the real position of the population mean, with only little error.

There are different values of the level of significance, and these indicate the probability that the true mean lies within the given boundary. The values that are typically used for the levels of significance are:

- 95 %
- 97.5 %
- 99 %
- 99.9 %

The higher the level of significance chosen, the higher the chance that the real mean will be situated in the given area. Statistically, it can be formulated like:

$$-z_{\frac{\alpha}{2}} < \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}$$

This can be rewritten to:

$$\overline{x} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

Where the sample mean forms the basis for the interval of the real population mean, by creating an upper and lower limit [13, p. 209]. The value of z_α is usually chosen amongst the values given above. This formula only applies for a sample amount of 30 and over, and uses the z-values from the normal distribution. It is also the population standard deviation that is in use.

It is possible to make another approximation and use the sample standard deviation s instead of the population standard deviation σ though, if the sample size is large. This leads to the formula [13, p. 210]:

$$\bar{x} - z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

For a small sample of a normal population, the t-distribution can be used instead of the normal distribution, to provide more precise approximations:

$$\bar{x} - t_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

It is important to know that, for a level of significance of α, there is a (1-α) chance that the calculated interval misses the real value of the population mean. The interval is centered at $\bar{x}$ and increases proportionally with the sample standard deviation [13, p. 211].

3.6. Hypothesis Testing

As mentioned in the in the section about the F-distribution, we use hypothesis testing to show whether there is a difference in a statistical value or not. It is important to have an objective method of looking at statistical values, such as means and variances, to determine whether there is a difference or not. The first step is to set up a hypothesis, consisting of a null hypothesis H_0 and an alternative hypothesis H_1 [13, p. 227].

The null hypothesis is the basis of the test; it is the hypothesis that the two parameters you are testing are not different from each other. The alternative hypothesis is the hypothesis that the two parameters are different; this can be expressed by a one-sided or a two-sided test. One-sided tests test whether the value of the parameter is either higher or lower than the parameter you are comparing it with. In the two-sided test, it is tested if the parameter is higher or lower than the parameter you are comparing it with, going either way.

If you are testing to see, whether a sample mean is equal to a true population mean, then the hypotheses could be [13, p. 230]:

Null hypothesis:	**Alternative hypothesis:**	**Reject null hypothesis if:**
H_0: $\mu = \mu_0$	H_1: $\mu < \mu_0$	$Z < -z_\alpha$
	H_1: $\mu > \mu_0$	$Z > z_\alpha$
	H_1: $\mu \neq \mu_0$	$Z < -z_{\alpha/2}$ Or $Z > z_{\alpha/2}$

The test is performed by calculating a specific value of Z, which corresponds to an area on the graph of the standard normal distribution. This area gives the probability for the null hypothesis to be true. The formula is [13, p. 229]:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Once the Z-value is calculated, it can be compared to the value in the tables for the standard normal distribution. Like with the confidence intervals, a level of significance is chosen. This normally falls on the same arbitrarily chosen values as the confidence intervals, like $\alpha = 0.05$ or $\alpha = 0.01$. Once you compare the two values, the one you calculated to the one you find from the level of significance, then you can decide whether to accept your null hypothesis or not. As can be seen on the table above, rejecting the null hypothesis or not depends of your choice of alternative hypothesis, which depends on the case.

Like in most cases, if the population standard deviation σ is unknown, it can be substituted with the sample standard deviation s [13, p. 232]:

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

This is for a sample size that is large though, ($n > 30$). If the sample size is small, and σ is unknown, the Z-value can be replaced by the t-value, assuming the population is normal [13, p. 233]:

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

This parameter follows the t-distribution, but the principle is the same as in the table shown before:

Null hypothesis:	**Alternative hypothesis:**	**Reject null hypothesis if:**
H_0: $\mu = \mu_0$	H_1: $\mu < \mu_0$	$t < -t_\alpha$
	H_1: $\mu > \mu_0$	$t > t_\alpha$
	H_1: $\mu \neq \mu_0$	$t < -t_{\alpha/2}$ Or $t > t_{\alpha/2}$

Since the hypothesis test is based on a level of confidence of α, there is a probability of falsely rejecting the hypothesis, even though it is true. This probability corresponds to α. There are two types of errors [13, p. 227]:

- Type I error: rejecting H_0 when H_0 is true
- Type II error: Not rejecting H_0 when H_1 is true

While α is the chance of committing a type I error, the chance to commit a type II error is defined by the letter β. The chance of committing a type II error is higher the lower the value of α is. For this reason, one should take care when choosing the value of α.

It's possible to calculate the chance that the sample mean is of the same or a higher value than the one already observed. This is called the P-value. The P-value is calculated by using the above formulas, and checking for the probability given at the Z- or t-value obtained from that. This would give an idea, whether or not the value of the sample mean is correct, or if there is a big probability that it would actually be of a different value.

Comparing two different samples with each other to test whether the means are the same or not is hypothesis testing with two samples. For these two samples, some assumptions have to be made; both the samples need to be independent of each other, and have the means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2 respectively. Also, the samples will be of size n_1 and n_2. For a large sample, the statistic Z is approximately normal, and can be calculated by using the formula [13, p. 247]:

$$Z = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where δ is the difference in the means $\mu_1 - \mu_2$. A confidence interval can be calculated for the difference, in which the fixed value of the difference will be located with a probability of 1 – α [13, p. 247]:

$$\overline{x} - \overline{y} \pm z_{\frac{\alpha}{2}} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

When doing a hypothesis test on two means from two independent samples, the null hypothesis is formulated as a difference in the means, as seen above. The difference can be set to a specific value, depending on how big a difference in the means one wants to test. It can also be set to zero, meaning that the test would be done to conclude if there is any difference at all. In general, the null hypothesis is [13, p. 248]:

$$H_0: \mu_1 - \mu_2 = \delta_0$$

Where the difference is a specific value. The alternative hypothesis is similar to the alternative hypotheses of testing one mean; it can be both one-sided and two-sided, and is the hypothesis that the difference in the two means is either greater than, lower than or not equal to the specific value. For a large sample, the Z-value is calculated by the formula:

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In the table below, the hypotheses are shown [13, p. 249]:

Null hypothesis:	**Alternative hypothesis:**	**Reject null hypothesis if:**
H_0: $\mu_1 - \mu_2 = \delta_0$	H_1: $\mu_1 - \mu_2 < \delta_0$	$Z < -z_\alpha$
	H_1: $\mu_1 - \mu_2 > \delta_0$	$Z > z_\alpha$
	H_1: $\mu_1 - \mu_2 \neq \delta_0$	$Z < -z_{\alpha/2}$ Or $Z > z_{\alpha/2}$

For small samples, the t-distribution can once again be used. This requires that more assumptions be made; both the populations being tested must be normal, and that the standard deviations must have a common value [13, p. 251]. The formula for calculating the t-value is then [13, p. 252]:

$$t = \frac{\bar{X} - \bar{Y} - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where the variance is estimated by using a pooled estimator; this is done by pooling the sums of the squared deviations of the two different samples. The estimator s_p is thus calculated by:

$$s_p^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}$$

It is important to note that the degrees of freedom now are calculated by:

$$n_1 + n_2 - 2$$

Here, the hypotheses are [13, p. 253]:

Null hypothesis:	**Alternative hypothesis:**	**Reject null hypothesis if:**
H_0: $\mu_1 - \mu_2 = \delta_0$	H_1: $\mu_1 - \mu_2 < \delta_0$	$t < -t_\alpha$
	H_1: $\mu_1 - \mu_2 > \delta_0$	$t > t_\alpha$
	H_1: $\mu_1 - \mu_2 \neq \delta_0$	$t < -t_{\alpha/2}$ Or $t > t_{\alpha/2}$

The confidence interval for the difference can also be calculated. This is also done using the pooled estimator:

$$\bar{x} - \bar{y} \pm t_{\frac{\alpha}{2}} * \sqrt{s_p} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

3.7. Linear Regression

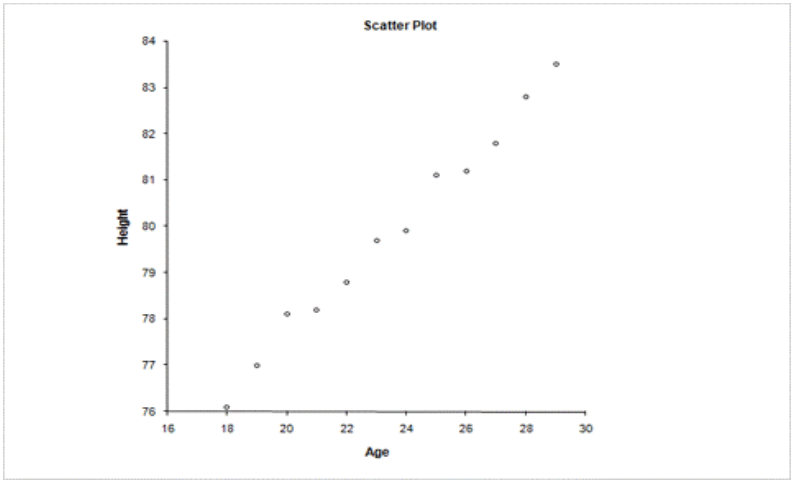
The purpose of looking at linear regression from a statistical point of view is to be able to state objectively whether there is a correlation between a set of paired data. To do this, the method of least squares is used to find the best regression between paired data, which can then be analyzed by various means.

The regression curve of a linear relationship is given as:

$$Y = \alpha + \beta x + \varepsilon$$

Where Y is a random variable, said to be dependent on x, Y being the dependent variable, x being the independent variable. The two constants α and β denote the intersection and the slope respectively, while the random variable ε accounts for any possible error, or other unknown factors besides α and β that may affect Y [13, p. 302].

In reality, a set of paired data would yield a graph, from which a linear regression curve would be constructed:



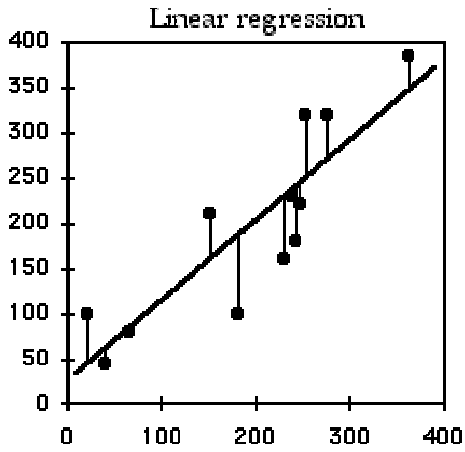


Figure 3.5 (left): An example of a scatter plot of data showing a linear correlation. [18]. Figure 3.6 (right): an example of a linear regression plot showing the vertical deviations [19].

While Figure 3.5 shows a scatter plot of the data, Figure 3.6 gives a linear regression curve based on the method of least squares. This linear regression curve would go by the formula:

$$\hat{y} = a + bx$$

Where the hat sign (^) shows that $\hat{y}_i$ is an estimate of the real value y_i from the data set, for a specific value of x_i. The constants "a" and "b" are also estimates of α and β respectively. The error for each statistic in the data set is:

$$e_i = y_i - \hat{y}_i$$

The errors are also known as the residuals, and the purpose of the least squares method is then to reduce the residuals so that they are as small as possible. This is done by making sure the estimators a and b make the equation:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

As numerically minimal as possible [13, p. 303]. To calculate a and b from a set of paired data (x,y) with n observations, the sum of squares and sum of cross products are calculated like:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})$$

From which “a” and “b” are calculated like:

$$b = \frac{S_{xy}}{S_{xx}} \qquad and \qquad a = \bar{y} - b * \bar{x}$$

Now that the regression curve is obtained, the sum of squares equations above can be used to calculate other parameters as well. The residual sum of squares (SSE) is calculated by the formula [13, p. 304]:

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

If it is assumed that the random variable Y from the very first equation is normally distributed with the mean $\alpha+\beta x_i$ and the variance σ^2, then an estimate of the variance can be obtained by using the formula:

$$s_e^2 = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2} = \frac{SSE}{n-2}$$

The reason for using n-2 instead of the more common n-1 in the previous formula is to make the estimate of σ^2 unbiased. It is worth noting that the sum of squares S_{xx} and S_{yy} are closely related to their respective sample variances, as:

$$s_x^2 = \frac{S_{xx}}{n-1}$$

And the same goes for S_{yy} [13, p. 310]. Perhaps one of the most important calculations to make when considering the correlation between a set of paired data, is the sample correlation coefficient [4, p. 336]:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} * S_{yy}}}$$

The correlation coefficient gives a value between -1 and 1, -1 for perfect linear graphs going from upper left to bottom right, and 1 one for graphs going from bottom left to upper right. The coefficient is 0 if there is no linear relationship:

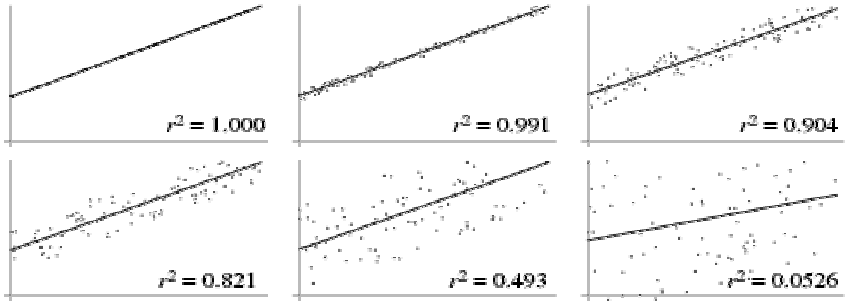


Figure 3.7: the interpretation of the correlation coefficient [20].

On figure 3.7, the value used is r^2, which gives the proportion of variability in y which is explained by the linear relation. Thus, if $r^2 < 1$, then there is some unexplained variability in the value of y which is not caused by the linear regression curve [13, p. 338-339].

Another way to determine whether there is a difference from linearity is by using a residual plot. A residual plot is a graph of the estimated value of each data point of the dependent value $\hat{y}$, versus the value of the residuals of each point. If the data points of the graph show a straight line of points, then the correlation between the two data sets are acceptable:

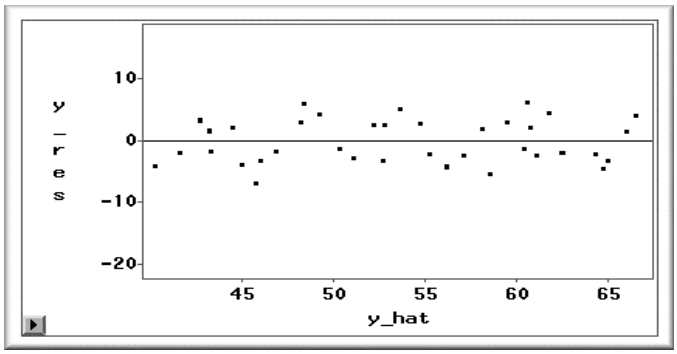


Figure 3.8: A residual plot showing the ideal situation [21].

A confidence interval can also be calculated for the estimators “a” and “b” to show the fixed value of the parameters α and β with a level of significance of α (not the same α as the regression coefficient). With n-2 degrees of freedom, the formulas are [13, p. 311]:

$$\alpha: \quad a \pm t_{\frac{\alpha}{2}} * s_e * \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$$\beta: \quad b \pm t_{\frac{\alpha}{2}} * s_e * \frac{1}{\sqrt{S_{xx}}}$$

3.8. Boxplots

Boxplots are used to identify possible outliers in a specific sample. A boxplot is a graphic that displays the sample points in a way, so that outliers are easily seen. The boxplot uses the median, and the quartiles:

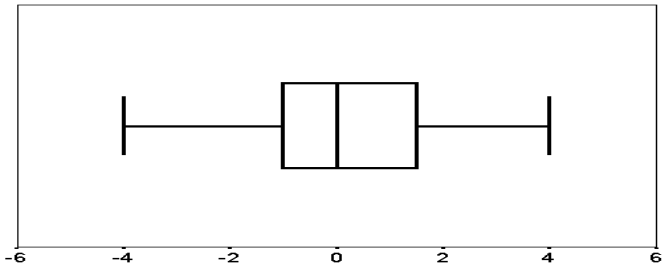


Figure 3.9: A boxplot showing the median and quartiles [22].

In the box, the middle line is the median, and the two outer lines are the 25 % and 75 % quartiles. The thin lines extend to the most extreme values; but only if they are not too far away. In this way, outliers can be seen as dots for themselves on the plot [13, p. 31-32].

3.9. Q-test

Another way to check for outliers in a sample is to use a Q-test [23, p. 63-64]. The Q-test (also known as Dixons Q-test) is a simple analytical tool that uses the difference between the outlier and the nearest point to determine whether it can be rejected or not, divided by the difference between the highest and lowest point of the data:

$$Q = \frac{x_0 - x_{nearest}}{x_{max} - x_{min}}$$

Where x_0 ia the outlier, and $x_{nearest}$ is the point nearest to the outlier numerically.

The obtained Q-value is compared to a table value. If the calculated value of Q exceeds the table value, then the outlier can be removed from the data with a confidence of α. For a confidence level of α=0.05, the table is:

Sample size	3	4	5	6	7	8	9	10
Q-value	0.970	0.831	0.717	0.621	0.570	0.524	0.492	0.464

3.10. Analysis of Variance (ANOVA)

ANOVA, or Analysis of Variance, is used when comparing three or more means to each other. In this case, the t-distribution would be insufficient, as it can only be used when comparing two means to each other at most, as shown in the section about hypothesis testing. The ANOVA uses the F-distribution to determine, whether there is a difference in the means or not. Thus, the null hypothesis is [13, p. 361]:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$

If it is chosen to reject the null hypothesis, it would not be possible *at first* to determine which mean/which means that are deviating. For this reason, the alternative hypothesis is:

$$H_1: \mu_x \neq \mu_y$$

To decide which mean/which means are different, it is possible to use confidence intervals. The confidence intervals will be calculated for each set of difference in means:

$$\bar{y}_x - \bar{y}_y \pm t_{\frac{\alpha}{2}} * \sqrt{s^2 * \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

Where the variance used as an estimator is the mean square error, which will be explained later. While this method is good for a low amount of means, it would require a lot of work for a test with many means. When all the intervals have been calculated, it is possible to see which mean(s) is different, by looking at which interval(s) is different [13, p. 366].

To make the necessary calculations to find the F-value that is needed, to reject the null hypothesis or not, the first thing to do is to find the total population, which corresponds to the sum of the number of observations, n_i from each sample k:

$$N = \sum_{i=1}^{k} n_i$$

The next step is to find the total sum of all observations y_{ij}, from each sample k [13, p. 357]:

$$T_{\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

From these two, a new value, the correction term for the mean is calculated:

$$C = \frac{T_{\cdot}^{2}}{N} = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^{2}}{\sum_{i=1}^{k} n_i}$$

From this, it is now possible to calculate the total sum of squares by the formula:

$$Total\ Sum\ of\ Squares = SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^{2} - C$$

And the sum of the squares of the treatments:

$$Treatment\ Sum\ of\ Squares = SS(Tr) = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} y_{ij}\right)^{2}}{n_i} - C$$

From these, the sum of squares error can be calculated by simple subtraction [13, p. 363]:

$$Error\ Sum\ of\ Squares = SSE = SST - SS(Tr)$$

Once these values have been calculated, an ANOVA-table [13, p. 362] can be set up, in which the remaining values can be calculated, that are needed to find the F-value:

Analysis of Variance table				
Source of variation	**Degrees of freedom**	**Sum of squares**	**Mean square**	**F-value**
Treatments	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F = \frac{MS(Tr)}{MSE}$
Error	$N-k$	SSE	$MSE = \frac{SSE}{N-k}$	
Total	$N-1$	SST		

Once the F-value has been calculated, it can be compared to an F_{α}-value from a table. The α-value is again the level of significance, and is chosen in the same way as in other hypothesis tests. As explained in the section about the F-distribution, the calculated F-value has to be to the left of the found value of F_{α}, to confirm the null hypothesis.

When the comparison is done, the steps explained in the start of this section can be used to see which mean/which means are different.

4. Method Validation theory

"Validation of an analytical method is the process by which it is established by laboratory studies, that the performance characteristics of the method meet the requirements for the intended analytical application [24].”

Validation of the method is needed for any new or revised method to guarantee that it is proper of giving reliable and reproducible execution, when used by different laboratory operators using the same products and equipment in different or in the same laboratories. The kind of validation sequence required depends completely on the specific method and its suggested applications [25].

4.1. Specificity

Specificity for an experiment guarantees that the signal determined comes from the analyte of interest, and that there are no interferences from traces and/or impurities and/or degradation products.

4.2. Range

Range of a method is the gap between the lower and upper levels of a substance that have been analyzed with satisfactory linearity, accuracy and precision. Range is determined on either a nonlinear or linear response graphic curve and is usually presented in identical units as the final test results.

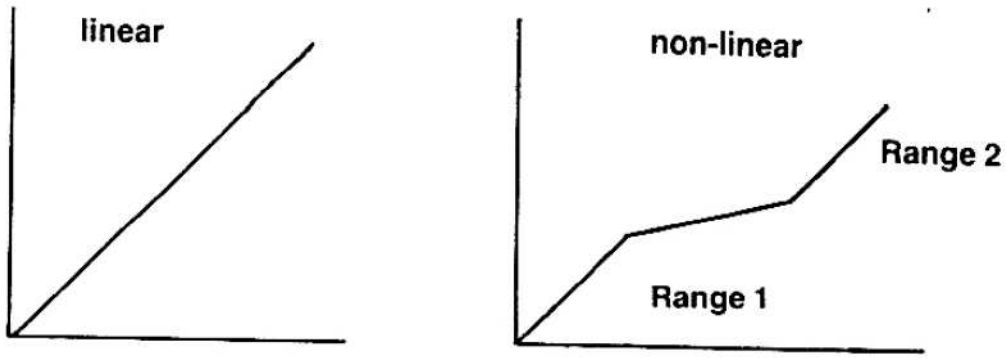


Figure 4.1: Differences between linearity and non-linearity ranges [24].

4.3. Linearity

Linearity is the ability of the method to obtain results which are directly proportional to the concentration of the substance within a specified range. This validation tool is obtained by determining the analyte concentration vs. the regression line (with some mathematical treatment of the final results as least mean squares).

4.4. Precision

Precision is the determination of the reproducibility of the full method (including analysis and sample preparation) under regular operating variables. Precision is calculated by utilizing the method to evaluate a sample for an enough number of times to get statistically correct results. Precision is then designated as the relative standard deviation (CV%):

$$\%RSD = \frac{\text{std dev x } 100\%}{\text{mean}}$$

4.5. Accuracy

Accuracy expresses the deviation between the true value and the mean value found. It is calculated by implementing the method to samples containing known quantities of analyte. The samples have to be analysed against blank and standard solutions to guarantee the elimination of interferences. Accuracy is then determined from the test values as a percentage of the amount of analyte retrieved by the measurement.

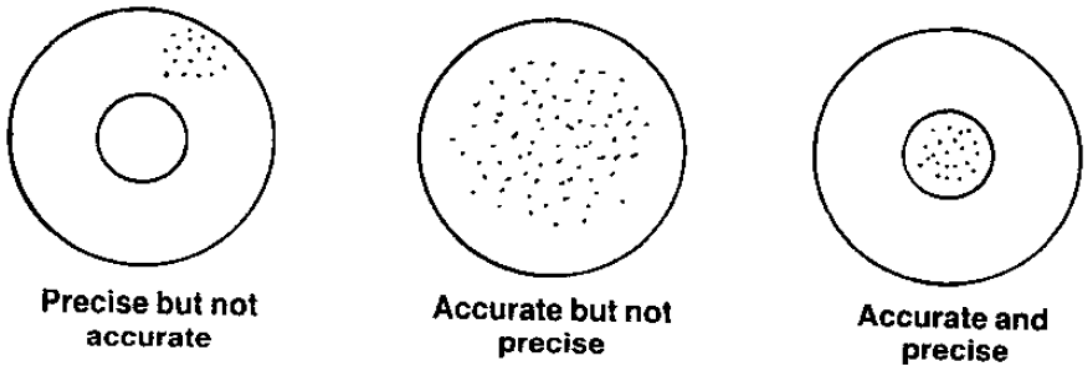


Figure 4.2: Differences between precision and accuracy [24].

Errors in determinations can be divided into two main categories: random errors and systematic errors.

Systematic errors appear from traceable sources due to the operator, the instrument or the methodology, and affect both the precision and the accuracy of the determination. A random error affects only the precision, and is complicated to remove, because this error is the result of random variations in the obtained signal, due to noise and different factors.

Random errors are equivalent to the root of the summation of the squares of each individual contribution. The imprecision of a method is often governed in the most imprecise step by the random errors.

4.6. Limit of detection (LOD)

LOD Is the lowest concentration of a sample that is detectable under the experimental conditions. This limit is important for the evaluation of dosages containing low analyte levels and for impurity tests.

It is normally related as the concentration producing a signal-to-noise ratio of 2:1 and is then ratified by analyzing a determinate number of samples near this relation with the following equation. Signal-to-noise ratio is calculated by:

$$s = H/h$$

Where:

H = height of the component peak.

h = absolute value of the largest noise variation from the chromatogram baseline of a blank solution.

4.7. Limit of quantification (LOQ)

LOQ is the lowest concentration value of analyte in a sample that can be calculated with satisfactory accuracy and precision. It is related as the concentration producing a signal-to-noise ratio of 10: 1 and is ratified by analyzing several samples near this relation.

4.8. Ruggedness

Ruggedness is the grade of reproducibility of the results acquired by the analysis of exactly the same sample under different normal experiment conditions ie different laboratories, instruments, analysts, assay temperatures, reagents, different days, small variations in mobile phase, etc [26].

5. Method Optimization

The 3 essential components for any HPLC method are: base sample preparation, analysis of HPLC results and standardization (calculations). All these components have to been investigated during the development of the project in order to obtain the final method optimization [27].

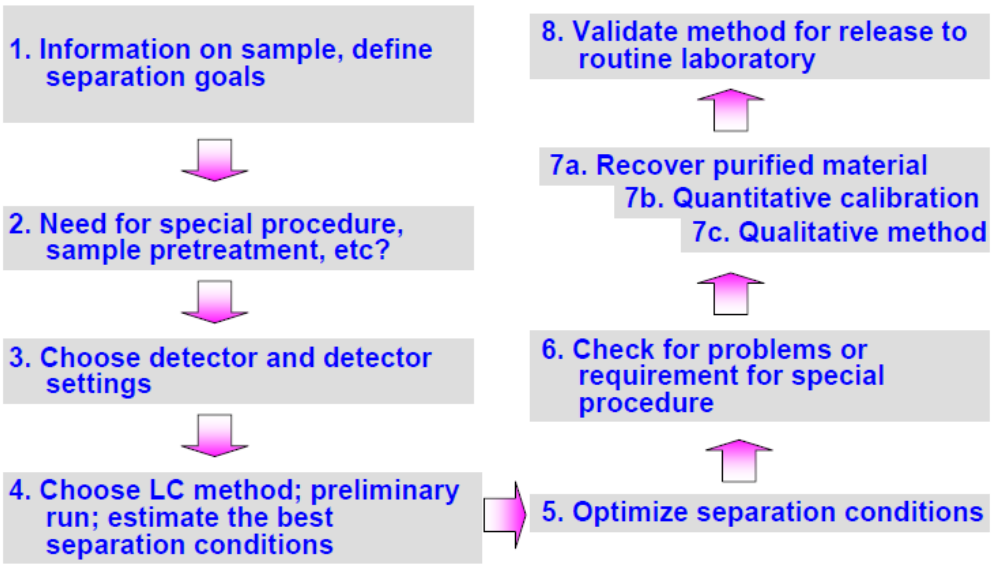


Figure 5.1: Steps for HPLC method development [28].

5.1. Initial Method

This report was initially based in the article “Determination of aspartame and phenylalanine in diet soft drinks by high-performance liquid chromatography with direct spectroflourimetric detection”, from Wróbel, K., Wróbel, K., and accepted in 1996 in the Journal of Chromatography A 773. According to this method a standard curve was made, so it is possible to determine the concentrations of the samples [29, p.2].

The sample was prepared by degassing (removing the CO_2 gasses by ultrasonic bath) the soft drinks. The drinks were mixed 1:1 (1 mL to one mL) with a solution known as a Carrez solution. The mixture was then diluted with water to 25 mL, and centrifugated. The liquid phase was then diluted (1:1) in phosphate buffer solution containing 34 % acetonitrile and 4 % methanol, the solution was again centrifuged and then filtered. These samples were prepared with different types of soft drinks.

A volume of 20 µL of the sample is injected. The mobile phase consists of 81 % phosphate buffer, 2 % methanol and 17 % acetonitrile. The pH is 4.3. A gradient program is used: 0-1 min: 0.7 mL/min. 1-2 min: 1 mL/min. 2-8 min: 1 mL/min. [30, p.2579, p.2589].

5.2. New Method

Due to the long time required to prepare the samples using the previous method, a new and optimized method for the analysis of Aspartame with HPLC has been found during the research of this project. With shorter sample time preparation, the preparation errors are decreased and the statistical variables improved. This method optimization is based on the article "Direct HPLC-UV determination of cyclamate, saccharine and aspartame from soft drinks" from M.D. Croitoru, I. Fülöp, M. Kincses Ajtay, C. Balogh and M.T. Dogaru accepted in September 2011 [31, p.459-465]. In this article the following materials and methods are used:

Equipment and reagents:

- Merck HPLC system consisting of: quaternary pump L-7100, auto sampler L-7200, column thermostat L-7360, Diode Array Detector (DAD) L-7455, interface L-7000, solvent degasser L-7612 and HMS manager software;
- LiChroCART 250-4 LiChrospher 100 RP-18 (5 µm) Merck column;
- gradient grade acetonitrile (Merck);
- aspartame p.a.;
- purified water HPLC grade;
- phosphoric acid, disodium phosphate, sodium hydroxide p.a. (Merck).

HPLC Method:
The mobile phase gradient and composition are shown in Table 5.1:

Time (min)	22 mM phosphate buffer pH 2.5	Acetonitrile	Flow rate (ml min^{-1})
0	93	7	1.400
8	93	7	1.400
8.1	85	15	1.400
18	85	15	1.400
18.1	93	7	1.400

Table 5.1: HPLC pump setup [31].

In the article, total analysis time was 24 min and sample volume was 100 µl. With a wavelenght of 196 nm the best chromatogram was extracted.

5.3. Optimized method

Optimization was necessary because the laboratory provided a different column than the one from the article since that one was being used to perform other analysis. The one provided is a Phenomenex Luna C18 5µ 100A 150x4,60 mm mentioned before. This column is 15cm, shorter than the one used in the article with 18cm and it will allow obtaining a shorter retention time.

Following the method explained in the "Method" section, using the mobile phase gradient and the HPLC configuration from the article and utilizing the column provided by the laboratory, a high and defined peak for the Aspartame was obtained. Due to these initial good results and the change in the column, a simple modification was required to perform an acceptable optimization. In order to improve the signal absorbance, the wavelength of the UV detector was changed from 196nm to 210nm, and the results were significantly better.

The experimental conditions for the optimized HPLC aspartame separation are shown in the Table 5.2:

Separation Variable	**Initial choice**
COLUMN	
Model	Luna C18 (00F-4252-E0)
Dimensions (length, I.d.)	150x4.60 mm
Particle size	5 µm
Stationary phase	C-18
MOBILE PHASE	
Solvents A/B	Buffer - ACN
%-B	Table 5.1
Buffer (pH, concentration)	2.5, 22 mM
Flow rate	1.4 ml/min
TEMPERATURE	25ºC
pH	2.5
SAMPLE SIZE	
Volume	20 µl

Table 5.2: Experimental conditions for the HPLC separation

5.4. Optimization results

Due to the change in the length of the column, less retention time was obtained. Consequently, the necessary change in the wavelength of the UV detector allowed to obtaining a better absorbance.

The method is robust in routine operation and usable by all the laboratories due to the rigour of the method standards and the statistics results. These data results are shown in the “Calculations of analytical parameters on HPLC chromatogram” and in the “Statistical data analysis” sections of the report.

The repeatability of the chromatogram is confirmed, and there is enough time elapsed between samples for the column to reach the equilibrium with the new condition of the mobile phase.

6. Description of experimental work

6.1. Chromatograms

The purpose of this experimental work in the laboratory is done to determine the amount of aspartame in light products. These products are soft drinks and we have focused on the most favorable cola products; Coca Cola Zero, Coca cola Light, Pepsi Max and Harboe Cola Minus.

The method we use in the experimental work is from a scientific article found on Scifinder database. The specific manual or method can be found in the appendix.

Safety, during an experimental work in a laboratory, is a very important issue. Therefore, the experimental work in the laboratory needs to be precise and carefully planned to obtain good results.

To begin with we made some preliminary test of a diluted Coca cola Zero sample, to have an idea of the different peaks and retention times. By analyzing the sample we optain the following chromatogram:

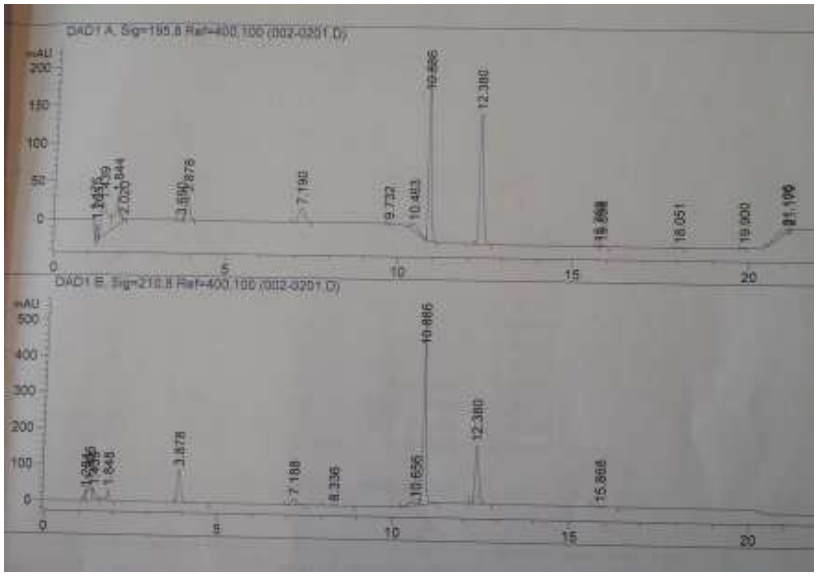


Figure 6.1: Chromatogram of the preliminary test.

As we can see there are two major peaks with retention time 10 and 12 min, and there is a difference in the absorbance of the two measured wavelength, therefore we select the wavelength which gives the largest absorbance. A larger absorbance gives a better sensitivity, therefore we use the data,measured at 210 nm, in the following analysis.

To determine between the two major peaks, we run a sample solution of pure aspartame. The standard solution has a concentration near 500 mg/L. We use the area of the first run to estimate the concentration

of the cola sample. We use the estimated concentration to be sure that the calibration curve covers the range of aspartame concentration in the cola samples. The first run of the standard solution of aspartame showed that aspartames retention time is around 12 minutes. Therefor we can define the retention time for aspartame to be near 12 minutes.

The chromatogram of the first aspartame standard run:

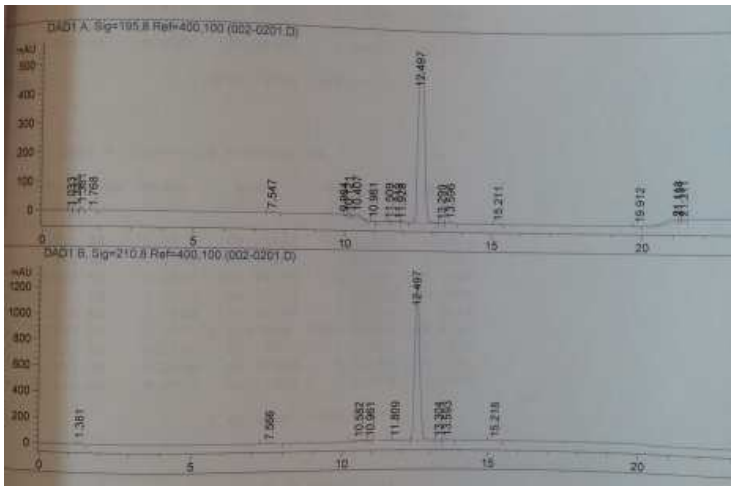


Figure 6.2: Chromatogram of the standard solution.

6.2. Materials

HPLC HP Agilent 1100 series is used with Reverse Phase Chromatography (RPC) for the analysis of Aspartame. The detection device would be UV-PDA or UV-DAD. For the stationary phase, the column utilized is a Phenomenex Luna C18 5µ 100A 150x4,60 mm.

Aspartame was supplied by Superco Analytical and the buffer solution compounds by the SDU laboratories. For the practical work in the laboratory, we used the following chemicals:

Acetonitrile

HiPerSolv CHROMANORM

For HPLC Supergradient

Batch:12B027806

Company: VWR

Hydrochloric Acid fuming 37 %

For analysis

Batch: 1723711

Company: Merck KGaA

Ortho-phosphoricacid 85 %

GR for analysis

Batch: 1715164

Natriumdihydrogenphosphat monohydrat

Pro analysis

Batch: A674146

Company: Merck KGaA

SodiumHydroxide 32%

AnalaR NORMAPUR

Batch: 11K170502

Company VWR

Aspartame 99 %

Production date: February 2012

Packed from: R474775

Lot nr. : LB64940

Highly filtrated Water

Figure 6.3 (left): Aspartame supplied by Superco Analytical. Figure 6.4 (right): HPLC Agilent 1100 series.

6.3. Method description

The method we use is based on measuring the concentration of aspartame in a range of 0-500 mg /L. The analysis is performed by using the HPLC with a reversed phase chromatography. The mobile phase that is used is a mixture of two chemical substances, which are Acetonitrile and phosphate buffer, in a gradient program.

The program is illustrated schematically in the Table 6.1:

Time (min)	22 mM phosphate buffer pH 2.5	Acetonitrile	Flow rate (mL/min)
0	93	7	1.400
8	93	7	1.400
8.1	85	15	1.400
18	85	15	1.400
18.1	93	7	1.400

Table 6.1: Mobile phase method program.

Aspartame is detected by using a UV-detector (Diode Array Detector), and in the method from the scientific article they used a wavelength on 196 nm. But this wavelength seemed to be low, because acetonitrile can absorb with this specific wavelength.

Therefor we chose to measure at two different wavelengths, one at 196 nm and 210 nm. We achieved the best absorbance at 210 nm, therefor we chose that wavelength in all further analyses.

The injection volume used in the scientific article is 100 µL, but we had chosen to inject with only 20 µL, to avoid overloading the system.

In the article the temperature of the column is not mentioned, but we had chosen a constant temperature at 25 °C to obtain the same conditions for the entire run.

6.3.1. Samples

We establish a list of samples that we need to run with our optimized method to be able to validate it. Other than this, we also run cola samples to test for the contents once we have validated the method. The sample list is given in the table below:

	Number of samples
Standard curve:	6 x 3 samples
Repeatability:	3 x 3 samples
Reproducibility:	3 x 3 samples
Precision:	Same 3 x 3 samples as above
Spiking:	3 x 5 samples
Cola samples:	3 x 3 3 x 4

The standard curve is created from 6 different concentrations (one of those concentrations is zero), with 3 measurements per concentration.

The repeatability is 9 measurements of the same cola (Coca Cola Zero) with 3 different sample preparations.

The reproducibility is 9 measurements of the same cola with 3 different sample preparations, carried out on a different day than the repeatability tests.

The precision is calculations done on the samples from the repeatability and reproducibility.

The spiking is 5 samples of 3 different concentrations (of which one is just normal Coca Cola Zero).

While all the samples above are carried out on a specific type of cola, we also test 3 other brands (Coca Cola Light, Pepsi Max and Harboe Minus). We run 3 samples of each of the 3 cola brands. We also run 3 samples of all of the 4 different brands, this time from a new can, to check the consistency of the contents of aspartame.

6.3.2. Reagents

To the sample preparation we use highly filtrated water to ensure that we don't have a contamination in the samples, with the exception of the standards that are diluted with dilution water, a mixture of 93% phosphate buffer and 7% acetonitrile. All samples are filtrated with a nylon filters (0,45 µm) before the injection on the HPLC.

6.3.3. Dilution Water

The dilution water is made in a 400 mL bottle by diluting 372 mL phosphate buffer with 28 mL acetonitrile.

6.3.4. Stock Solution

The stock solution is made from 99.99 % pure aspartame, which is diluted with eluent. The concentration of the stock solution was set to 1 g/L. This is done by mixing 100 mg of pure aspartame in a 100 mL volumetric flask with dilution water. The weight was done on a 4 significant digits weighing scale. The stock solution is used to make the standard solutions.

Figure 6.5: Analytic scale.

6.3.5. Standards

The standards are made in these concentrations: 500 mg, 300 mg/L, 200 mg/L, 100 mg/L, and 50 mg/L.

The standards are made from the stock solution by using a bulb with a suction ball. On the next page, a table describing the standards is set up:

The standards are made from the table below:

		V2 [mL]	C2 [mg/L]	C1 [mg/L]	V1 [mL]	Producing procedure
1	Blind	50	0	1000	0	0 mL stock solution into a 50 mL Volumetric flask
2	50 mg/L	50	50	1000	2,5	2,5 mL stock solution into a 50 mL Volumetric flask
3	100 mg/L	50	100	1000	5	5 mL stock solution into a 50 mL Volumetric flask
4	200 mg/L	50	200	1000	10	10 mL stock solution into a 50 mL Volumetric flask
5	300 mg/L	50	300	1000	15	15 mL stock solution into a 50 mL Volumetric flask
6	500 mg/L	50	500	1000	25	25 mL stock solution into a 50 mL Volumetric flask

Table 6.2: Description of the standard solutions.

6.3.6. Cola Samples

The cola products are degassed by ultrasonic equipment in 2 hours, to be sure that all kind of gasses that are present in the cola, are removed. Thereafter the cola are diluted with highly filtrated water; the dilution factor is 1:5. This is done by using a suction ball and bulb, where 20 mL of cola is added into a 100 mL volumetric flask and filled with highly filtrated water. The diluted cola mixture is then filtrated by a nylon filter (0,45 µm) and added into the vials. This procedure is done for all 4 different cola products.

6.3.7. Spiking

Prepare a spiked solution for a run is very simple. An amount of the stock solution is added to the diluted cola. This mixture are filtered with a nylon filter (0,45 µm) and added into vials. We should, after a run, be able to recover the added concentration from the stock solution (1000 mg/L) that is added to the diluted cola.

We spike with two different volumes of the stock solution, to obtain two spiking levels.

The spiked sample level 1: 3 ml stock solution to 30 ml diluted cola.

The spiked sample level 2: 6 ml stock solution to 30 ml diluted cola.

6.3.8. Buffer

The buffer is prepared by phosphoric acid (H_3PO_4) and its corresponding base salt ($H_2PO_4^-$). The chemical equation shows how this triprotic acid deprotonates:

$$H_3PO_4 \leftrightharpoons H_2PO_4^- \leftrightharpoons HPO_4^{2-} \leftrightharpoons PO_4^{3-}$$

We know the concentration (22 mM) and the pH (=2.5). Using the buffer equation we can find the required amount of the acid and its corresponding base.

$$pH = pk_a + \log\left(\frac{n_B}{n_A}\right) = 2.15 + \log(x) = 2.5$$

$$\Rightarrow 2.399 = \frac{[Base]}{[Acid]} \Rightarrow [Base] = 2.399 * [Acid]$$

$$0{,}022\ M = [Base] + [Acid] \Rightarrow [Base] = 0.022M - [Acid]$$

$$2.399 * [Acid] = 0.022 - [Acid]$$

$$\Rightarrow [Acid] = 0.00647 \frac{mol}{Liter}$$

$$[Base] = 0.022\ M - 0.00647M = 0.01553 \frac{mol}{Liter}$$

By multiplying the concentration with the molar weight of the base (NaH_2PO_4,H_2O), M = 137,99 g/mol, we get how much base we should add in a 1 liter volumetric flask:

$$m = 137.99 \frac{g}{mol} * 0.01553\ mol = 2{,}1426\ gram\ NaH_2PO_4 \cdot H_2O$$

Now we find the amount of the phosphoric acid, this calculation is also done to made 1 liter buffer:

Since the phosphoric acid is a liquid and not a solid, we cannot use the same method of calculation as above.

$$v = \frac{n}{c} = \frac{0{,}00647\ mol}{14{,}6495\ M} = 0{,}000442\ L = 442\ \mu L$$

The concentration used in the equation is for the 85 % phosphoric acid.

The determined amount of acid, 442 µL is added to the same 1 liter volumetric flask as the base and filled up with highly filtrated water. The pH is adjusted to 2.5 with either a base (NaOH) or acid (HCl).

7. Calculations of analytical parameters on HPLC chromatogram

7.1. Retention time (t_R), Void Time (t_M) & Retention factor (k)

A random chromatogram is chosen to illustrate how the parameters can be calculated or found.

The retention time (t_R) for aspartame from the chosen chromatogram is 12.066 minutes. The retention times vary a little in general because of the buffer used.

The void time (t_M) is determined to be 1.166 minutes. The adjusted retention time can now be calculated:

$$t_R' = t_R - t_M$$

$$12.066 - 1.166 = 10.9\ minutes$$

The retention factor (k) can also be found:

$$k = \frac{t_R - t_M}{t_M}$$

$$\frac{12.066 - 1.166}{1.166} = 9.3482$$

7.2. Volumes

The volumes which can be found are the Retention volume (V_R), Void volume (V_M) and the Peak volume.

$$V_R = t_R * F$$

$$12.066\ min * 1.400 \frac{mL}{\text{min}} = 16.8924\ mL$$

$$V_M = t_M * F$$

$$1.166\ min * 1.400 \frac{mL}{min} = 1.6324\ mL$$

The amount of eluent used up to the retention time is 16.89 mL, and the void time 1.63 mL.

The width of the peak (t_R=12.066) is measured to be 2 mm = 0.2857 min (35 mm = 5 min, 1 min = 0.143 mm):

$$Peak\ Volume = w_b * F$$

$$0.2857\ min * 1.400 \frac{mL}{min} = 0.3999\ mL$$

The volume used to get this specific peak with the retention time 12.066 minutes is 0.4 mL.

7.3. Selectivity (α)

To determine the selectivity we need to calculate the retention factor for another peak. The new retention time (t_R) for the other peak is 10.886 minutes.

$$k_1 = 9.3482$$

$$k_2 = \frac{10.886 - 1.166}{1.166} = 8.3361$$

The selectivity can now be found, as the difference between the two retention factors:

$$\alpha = \frac{k_2}{k_1}$$

$$\alpha = \frac{9.3482}{8.3361} = 0.8917$$

The selectivity has a value higher than 1, which indicates that the peaks are completely separated.

7.4. Column efficiency (N)

We will now find out how efficient the column is, which is done by this formula:

$$N = 16\left(\frac{t_R}{w_b}\right)^2$$

$$N = 16\left(\frac{12.066}{0.2857}\right)^2 = 28535.3748 \approx 28535$$

So we can conclude that the column is very efficient and has 28535 theoretical numbers of plates.

7.5. HETP

The height equivalent to a theoretical plate (HETP) is determined by dividing the length of the column with number of plates (N):

$$HETP = \frac{L}{N}$$

The length of the column (L) we used is 15 cm = 150 mm

$$HETP = \frac{150000\ \mu m}{28535.3748\ plates} = 5.2566$$

The value indicates the height of the theoretical plate in the column, and in this case the plate height is 5.2566 µm.

7.6. Resolution (R_s)

The resolution is determined to see how good a separation between the peaks is. This is done by using the formula for resolution:

$$R_s = \frac{t_{R2} - t_{R1}}{\frac{w_{b2} - w_{b1}}{2}}$$

$$R_s = \frac{12.066 - 10.886}{\frac{0.2857 - 0.2}{2}} = 27.5379$$

So that indicates a very good resolution, and the two peaks are completely separated.

7.7. Asymmetry (A_s) & tailing factor(T_f)

To calculate the tailing factor and asymmetry factor, we need to know the value of A, B, $W_{0.05}$ and f. These values are measured to be:

A = 0.8 mm

B =0.8 mm

$W_{0.05}$ = 1.85 mm

f=2

$$T_f = \frac{W_{0.05}}{2f}$$

$$T_f = \frac{0.264286}{2 * 2} = 0.0661$$

The T_f value is under 1.0, so the peak doesn't tailing or fronting.

$$A_s = \frac{B}{A}$$

$$A_s = \frac{0.8}{0.8} = 1$$

That indicates no asymmetry, so the peaks have a Gaussian peak shape with good symmetry.

These calculated parameters have shown us that the method has a good separation of the different peaks. The resolution value is bigger than 2, so therefore we can conclude that the peaks are completely separated from each other.

The efficiency of the column is very high because of the large number of theoretical plates in the column. This means that the method and the column easily can separate the components in the samples from each other.

The peaks symmetry seems to be very good because the asymmetry factor is close to 1, and the tailing factor is close to 0. Which indicates that the peak shape is following the Gaussian peak shape, with perfect symmetry.

8. Statistical analysis of results

8.1. Standard curve

For each standard solution created, three measurements of peak areas are made. Including three measurements at zero to prove that there is no measurement of aspartame at a concentration of zero, the results are:

Concentration [mg/L]	Area [mAU*s]		Concentration [mg/L]	Area [mAU*s]
0	0		200	6060,07
0	0		200	6072,01
0	0		200	6099,92
50	1515,64		300	8935,52
50	1527,02		300	8802,61
50	1513,77		300	8970,02
100	3048,2		500	14348,6
100	3036,82		500	14341,9
100	3049,83		500	14151,9
		Mean	**191,67**	**5637,44**

Where the mean of the two columns are given in the end of the table. Note that the table (which is from Microsoft Excel) uses comma instead of dots to separate the integers from the decimals. The means are calculated using:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The concentration is the independent variable, whereas the peak area is the dependent variable. We can calculate the values of S_{xx}, S_{yy} and S_{xy}, from which all the statistical parameters can be calculated:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{18}(0-191.67)^2 + (0-191.67)^2 + \ldots + (500-191.67)^2 = 5.16 * 10^5$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{18}(0-5637.44)^2 + (0-5637.44)^2 + \ldots + (14151.9-5637.44)^2 = 4.23 * 10^8$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x}) * (y_i - \overline{y}) = \sum_{i=1}^{18}(0-191.67) * (0-5637.44) + \ldots + (500-191.67) * (14151.9-5637.44) = 1.48 * 10^7$$

From these, the values of a and b are found:

$$b = \frac{S_{xy}}{S_{xx}} \qquad and \qquad a = \bar{y} - b * \bar{x}$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{1.48 * 10^7}{5.16 * 10^5} = 28.61$$

$$a = \bar{y} - b * \bar{x} = 5637.44 - 28.61 * 191.67 = 153.00$$

The error of sum of squares is also calculated:

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 4.23 * 10^8 - \frac{(1.48 * 10^7)^2}{5.16 * 10^5} = 4.28 * 10^5$$

It is now possible to estimate the variance σ^2:

$$s_e^2 = \frac{SSE}{n-2} = \frac{4.28 * 10^5}{18-2} = 2.67 * 10^4$$

From this variance, the confidence intervals of α and β can be found:

$$\alpha: \quad a \pm t_{\frac{\alpha}{2}} * s_e * \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} = 153.00 \pm 2.12 * \sqrt{2.67 * 10^4} * \sqrt{\frac{1}{18} + \frac{191.67^2}{5.16 * 10^5}} = 153.00 \pm 123.38$$

$$\beta: \quad b \pm t_{\frac{\alpha}{2}} * s_e * \frac{1}{\sqrt{S_{xx}}} = 28.61 \pm 2.12 * \sqrt{2.67 * 10^4} * \frac{1}{\sqrt{5.16 * 10^5}} = 28.61 \pm 0.48$$

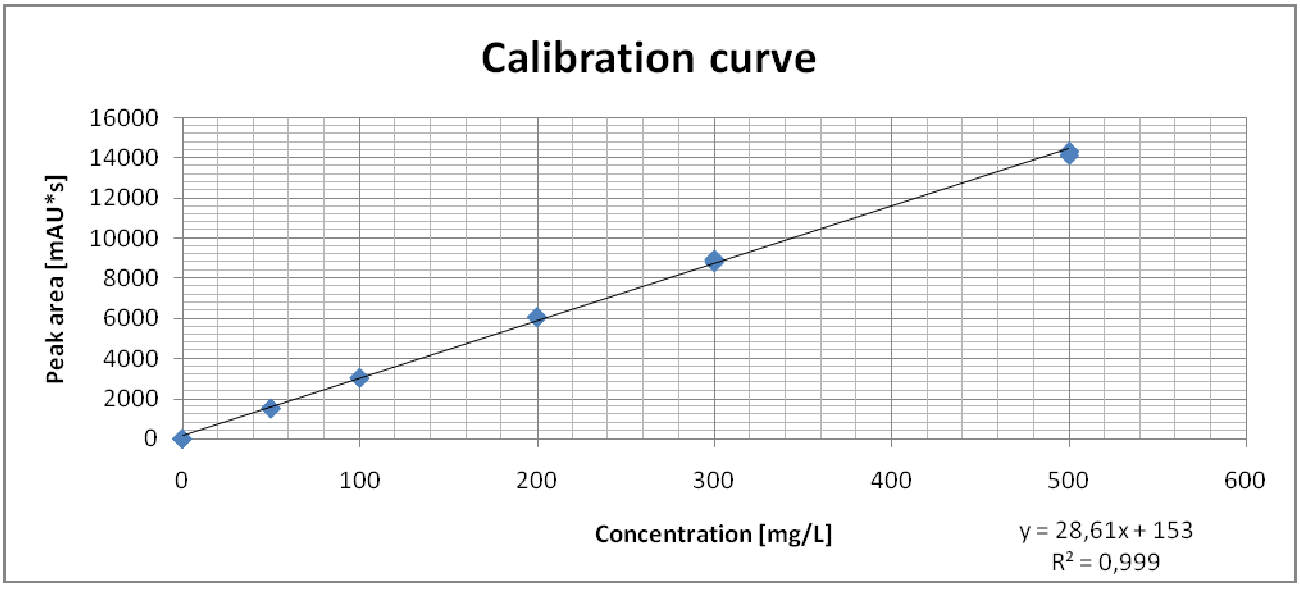
Where $t_{\alpha/2}$ is for n-2 degrees of freedom, from [13, p. 516]. It is also possible to find the coefficient of correlation:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} * S_{yy}}} = \frac{1.48 * 10^7}{\sqrt{5.16 * 10^5 * 4.23 * 10^8}} = 0.999$$

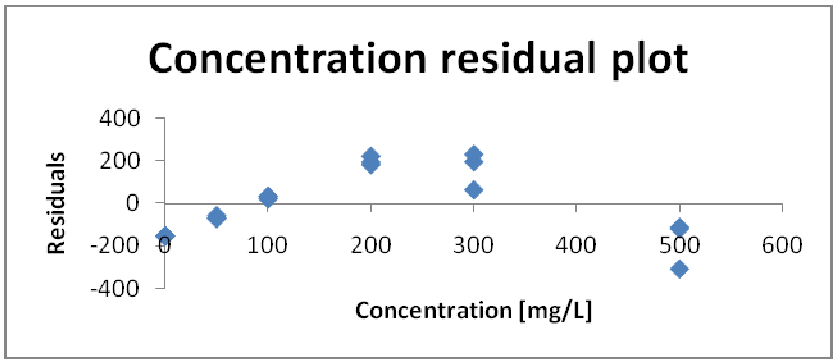
The value of R^2 can also be found:

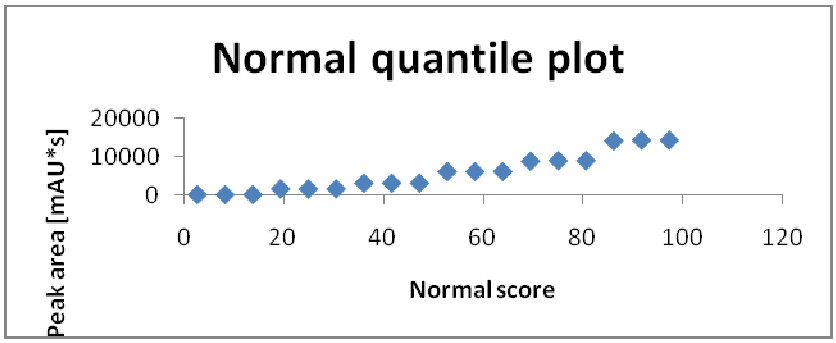
$$R^2 = 0.999495^2 = 0.998989$$

Excel gives the following graph of the standard curve:



With the residuals and normal quantile plot:





The residuals do not look random. This is most likely because the standard curve is not entirely linear in the tested area – there is a tendency for the curve to straighten out at higher concentrations. This is also what can be seen from the residual plot. The normal quantile plot appears to be linear.

8.2. Repeatability & reproducibility

We check the repeatability and reproducibility by sampling a specific Coca Cola Zero nine times a day over two days. Samples are made so that we, in total, obtain 18 samples from six different sample preparations (not different in the way that the methods are different). The results are shown in the table below, along with the calculated concentrations of the samples:

Zero tests, first day	Area [mAU*s]	Concentration [mg/L]	Concentration in cola [mg/L]
1,1	2271,98	74,05	370,26
1,2	2267,20	73,88	369,42
1,3	2264,23	73,78	368,91
2,1	2282,38	74,42	372,08
2,2	2353,22	76,89	384,45
2,3	2354,42	76,93	384,66
3,1	2374,42	77,63	388,16
3,2	2375,90	77,68	388,42
3,3	2370,88	77,51	387,54

Zero tests, second day	Area [mAU*s]	Concentration [mg/L]	Concentration in cola [mg/L]
1,1	2381,67	77,89	389,43
1,2	2393,47	78,30	391,49
1,3	2385,58	78,02	390,11
2,1	2389,31	78,15	390,76
2,2	2380,72	77,85	389,26
2,3	2384,88	78,00	389,99
3,1	2291,11	74,72	373,60
3,2	2305,66	75,23	376,14
3,3	2300,1	75,03	375,17
Total mean (both days)	**2340,40**	**76,44**	**382,21**

Here, the different sample preparations are given as 1, 2 and 3 for each day, while each sample from each of the different preparations have been numbered 1, 2 and 3 as well. The concentration is calculated by using the formula obtained from the standard curve:

$$y[mAU*s] = 28.61*x\left[\frac{mg}{L}\right] + 153 \Rightarrow x\left[\frac{mg}{L}\right] = 0.035*y[mAU*s] - 5.35$$

The first sample of the first day:

$$x\left[\frac{mg}{L}\right] = 0.035*y[mAU*s] - 5.35 = 0.035*2271.98\ mAU*s - 5.35 = 74.05$$

As the samples have been diluted five times, a column with the actual concentration in the cola has been added to the table:

$$x\left[\frac{mg}{L}\right] * dilution\ factor = actual\ cola\ concentration\ \left[\frac{mg}{L}\right]$$

$$74.05\frac{mg}{L} * 5 = 370.26\frac{mg}{L}$$

We can use a t-test to determine whether the means measured over the two days are similar to each other. That is, we can test to see whether they come from the same population. We have two sets of data (n=9 in both cases), all coming from different sample preparations as well. For each, there is a mean and a standard deviation:

	First day, cola concentration [mg/L]	Second day, cola concentration [mg/L]
	389,43	370,26
	391,49	369,42
	390,11	368,91
	390,76	372,08
	389,26	384,45
	389,99	384,66
	373,60	388,16
	376,14	388,42
	375,17	387,54
Mean:	**385,11**	**379,32**
Standard deviation:	7,655	8,834

We set up the following hypotheses for the test:

$$H_0: \mu_1 - \mu_2 = \delta_0 = 0$$

$$H_1: \mu_1 - \mu_2 \neq \delta_0$$

The t-value is then calculated by first finding the pooled standard deviation:

$$s_p^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2} = \frac{(9-1) * 7.655^2 + (9-1) * 8.834^2}{9+9-2} = 68.32$$

$$t = \frac{\bar{X} - \bar{Y} - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{385.11 - 379.32 - 0}{\sqrt{68.32} * \sqrt{\frac{1}{9} + \frac{1}{9}}} = 1.486$$

For the degrees of freedom:

$$n_1 + n_2 - 2 = 9 + 9 - 2 = 16$$

We find a t-value of:

$$t_{table}(16) = 2.120$$

For a level of significance of α=0.05, and a two sided test, we fail to reject the null hypothesis, meaning that the two sets of data have the same means. This means that we can say with 95 % confidence that we can reproduce data so that it has the same mean. As the two data sets come from the same cola, this of course only goes for the same cola. As stated earlier, we will also test new colas.

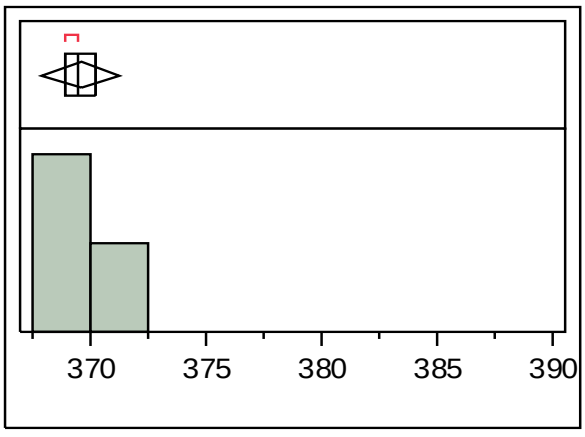
We will now look at the deviations that come from the different factors tested. As mentioned, we made 9 samples one day with three different sample preparations, and 9 samples the next day, also with three different sample preparations. This gives us different variances that can be considered:

- Variance between repetitions, created by the machine
- Variance between sample preparations, created by human error
- Variance between different days, created by difference in the laboratory conditions (person, weather, chemicals etc.)

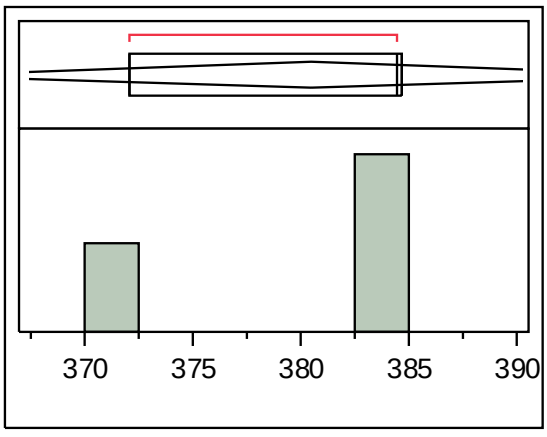
Using SAS JMP, we test the variance of each of these different parameters. First, we test the machine variance for each day, and for each of the three sample preparations each day:

First day

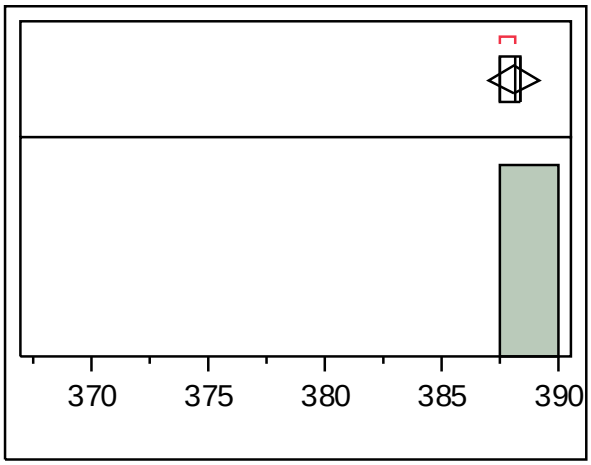
Sample preparation 1



Sample preparation 2

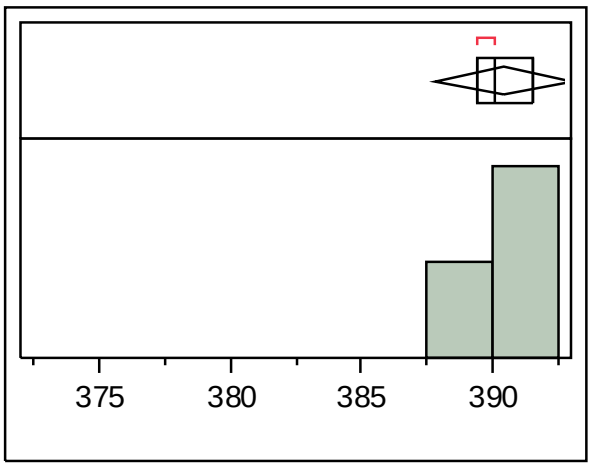


Sample preparation 3

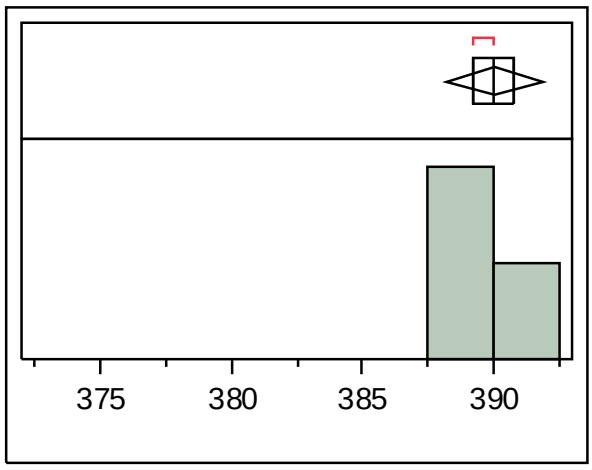


Second day

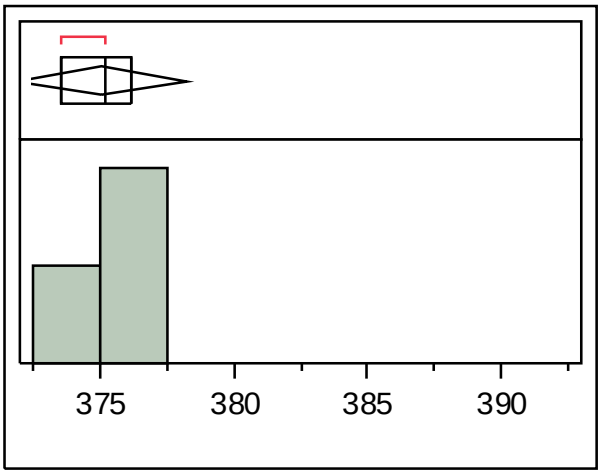
Sample preparation 1



Sample preparation 2



Sample preparation 3



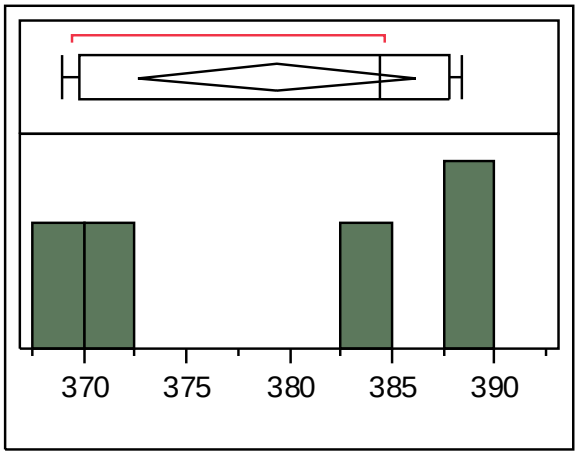
The means and standard deviations can be seen in the table below:

	First day			Second day		
	Sample prep. 1	Sample prep. 2	Sample prep. 3	Sample prep. 1	Sample prep. 2	Sample prep. 3
Mean	369.53	380.40	388.04	390.34	390.00	374.97
Standard deviation	0.6817	7.203	0.4521	1.050	0.7501	1.282

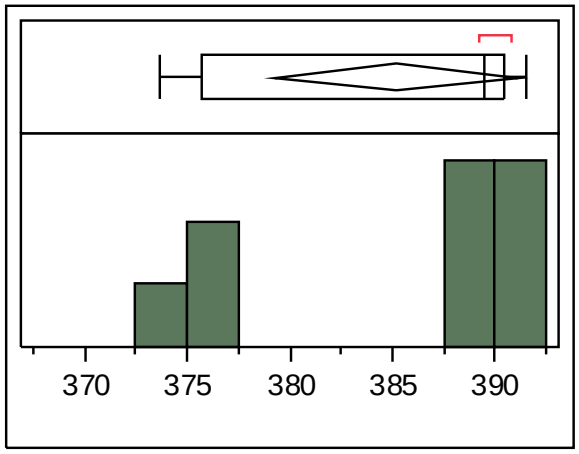
It can be seen that the different sample preparations differ a lot in the obtained means, even though they are supposed to give the exact same means. Specifically, the 2nd sample preparation of the first day has some error to it – the standard deviation is way out of scale, and should not be considered as normal. However, the table gives a good idea of the standard deviation of the machine. Differences in the mean are most likely caused by differences in the method preparation.

We can also look at the standard deviation of the entire day, including both the deviation from the machine and the deviation from the different sample preparations. The data for each day is again put into SAS JMP, which gives the standard deviations:

First day



Second day



The data for the graphs are given in the table below:

	First day	Second day
Mean	379.32	385.11
Standard deviation	8.833	7.657

If we eliminate the variation from the apparatus by taking the means of the three different sample preparations of each day, then we get the standard deviation of the sample preparations:

	First day	Second day
Sample preparation 1	369,53	390,34
Sample preparation 2	380,40	390,00
Sample preparation 3	388,04	374,97
Standard deviation	**9,302**	**8,777**

It can be seen that these standard deviations are higher. This is because the low deviation of the apparatus is no longer counted as a part of the total standard deviation. Between the days, the standard deviation can be calculated as the standard deviation of the means of the two days as given in the table at the top of the page:

First day	379.32
Second day	385.11
Standard deviation	4.089

Here it can be seen, that the standard deviation between the two days are lower than between the different sample preparations.

On the next page, we show all the standard deviations as CV% in a table. The CV% is calculated using:

$$CV\% = \frac{s}{\bar{x}} * 100\%$$

This table shows the calculated CV% that give information about the variations in the measurements of the method.

CV% for a single sample preparation					
First day			Second day		
1	2	3	1	2	3
0,18	1,89	0,12	0,27	0,19	0,34
CV% for entire day					
First day			Second day		
2,33			1,96		
CV% for different sample preparations					
First day			Second day		
2,45			2,28		
CV% for different sample preparations between 2 days					
1,07					

The values of the CV% of the single sample preparations give the deviations of the apparatus, which is seen to be very low in relation to the other values of the CV%. For the different sample preparations, the CV% is much higher, which is caused by human error in the laboratory. A combination of these two is the CV% of an entire day.

Also given is the CV% between two days. As this value is low, there is good reproducibility of the method. It has to be remembered that all the tests done so far have been on one specific Coca Cola Zero, so the variations coming from different colas has yet to be examined.

8.3. Precision

The aim, as stated earlier, with the repeatability and reproducibility is to check whether the test samples all have the same variance, and the same mean. To test this, we use one-way analysis of variance (ANOVA). The data for the first day is lined up in columns, with the real concentration of the colas arbitrarily chosen as the factor to be examined (the peak area, sample concentration and actual concentration are all proportional):

First day		
Column 1	Column 2	Column 3
370,26	372,08	388,16
369,42	384,45	388,42
368,91	384,66	387,54

The colas from day one (1,1 – 1,3) are seen to be in the first column, and so forth. The first data point of the second column is seen to be way more different from the other data points, in relation to the deviations in the other columns. This can be checked by using a Q-test:

$$Q = \frac{|x_0 - x_{nearest}|}{x_{max} - x_{min}} = \frac{|372.08 - 384.45|}{384.66 - 372.08} = 0.983$$

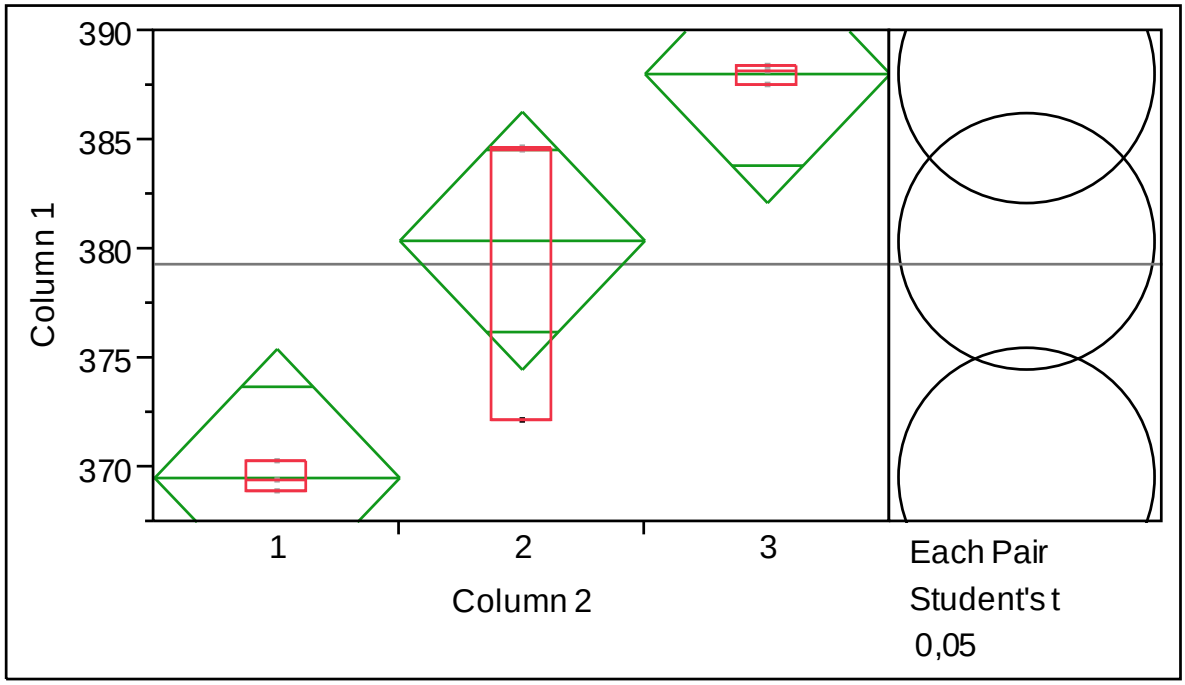
The Q-value at a confidence level of α=0.05 for three samples is 0.970. The outlier can therefore be rejected. We choose to keep it for the ANOVA, to have a higher amount of samples.

The ANOVA can now be calculated from the columns above. We use SAS JMP, and enter the data into the program. SAS JMP calculates the data for the ANOVA, which is then set up in the table below:

Analysis of Variance table					
Source of variation	**Degrees of freedom**	**Sum of squares**	**Mean square**	**F-ratio**	**Probability > F**
Treatments	**2**	**519.13**	**259.56**	**14.817**	0.0048
Error	**6**	**105.11**	**17.518**		
Total	**8**	**624.24**			

The parameter “Probability > F” gives the probability for F to reach that value. When the probability is that low, it is safe to say that the null hypothesis of the ANOVA can be rejected. This means that there is a difference in the means of the sample preparations.

A visual representation of the data, also given by SAS JMP is shown on the graphic below:



Here, the value of the means can be seen on the y-axis, with the different sample preparations on the x-axis. The outlier can clearly be seen in column 2. From this we can conclude, that the deviation in the different sample preparations is relatively large.

This could be because of an error made during the laboratory work, or an error in the apparatus, but it is most likely caused by a human error, since the apparatus works with very low variance as we showed earlier, and the fact that the sample preparation contains several steps in which it is possible to make errors.

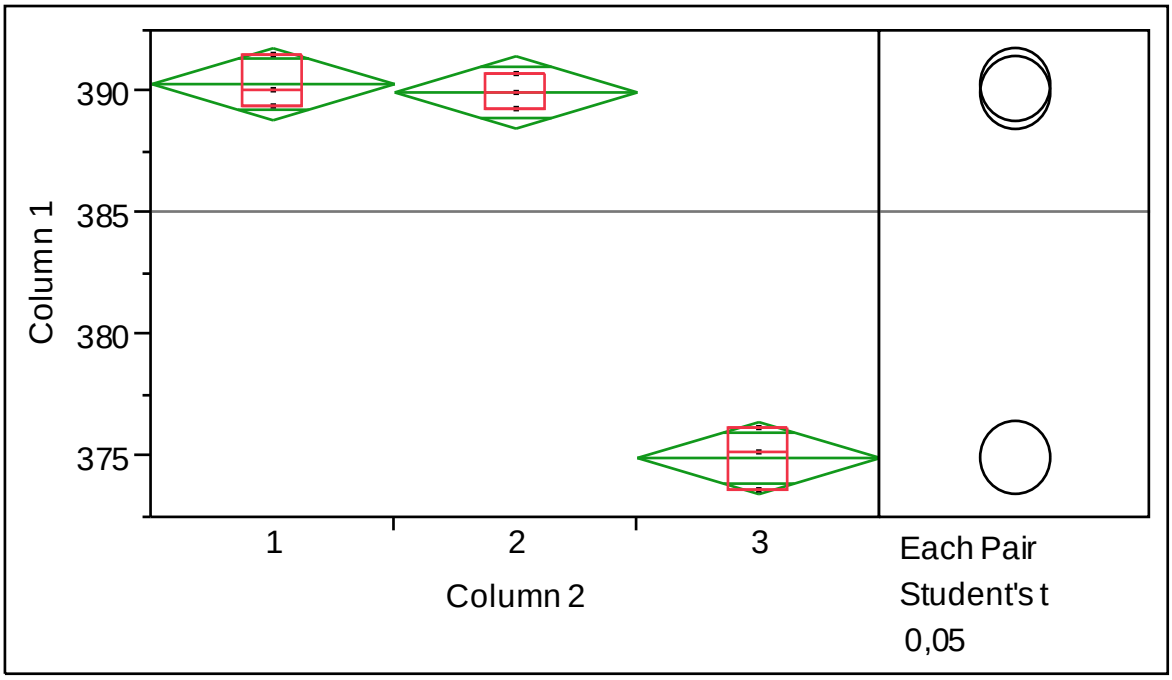
We now check the second day with a one-way ANOVA like above as well. The data is put into a table like the first time:

Second day		
Column 1	**Column 2**	**Column 3**
389,43	390,76	373,60
391,49	389,26	376,14
390,11	389,99	375,17

As there are no apparent outliers, a data table from SAS JMP is then set up as before:

Analysis of Variance table					
Source of variation	**Degrees of freedom**	**Sum of squares**	**Mean square**	**F-ratio**	**Probability > F**
Treatments	**2**	**462.18**	**231.09**	**209.29**	<0.0001
Error	**6**	**6.625**	**1.104**		
Total	**8**	**468.80**			

A visual representation of the data:



This time, the value of F has a much lower probability. This is caused by the much lower variance in the data sets of this test, and the fact that one of the means is so far away from the others numerically.

Two of the means are seen to be very close to each other, meaning that there either could have been some mistake done in relation to the third mean.

8.4. Spiking

The concentrations of the spiked samples are calculated based on the concentration of the mixtures. As described earlier, 15 spiked samples are made; 5 that are not spiked for comparison, and 5x2 that are spiked at different levels. The spiked samples are created by mixing 3 mL of the stock solution of 1 g/L aspartame to 30 mL of diluted (by a factor 5 as usual) Coca-Cola Zero for the first set of 5 samples, and 6 mL of the stock solution of aspartame to 30 mL of diluted Coca Cola Zero for the next set of 5.

It is assumed, that if *not* spiked, the concentration of spiked colas would be equal to the total mean of the 18 samples done during the repeatability and reproducibility tests: $\overline{x} = 76.44 \frac{mg}{L}$. We can calculate the expected concentrations of the spiked samples by simple math:

Spike 1:

Total amount of mg mixed:

$$1000 \frac{mg}{L} * 0.003\ L + 76.44 \frac{mg}{L} * 0.030\ L = 3\ mg + 2.29\ mg = 5.29\ mg$$

Total concentration:

$$c = \frac{n}{V} = \frac{5.29\ mg}{0.003\ L + 0.030\ L} = 160.40 \frac{mg}{L}$$

Spike 2:

Total amount of mg mixed:

$$1000 \frac{mg}{L} * 0.006\ L + 76.44 \frac{mg}{L} * 0.030\ L = 6\ mg + 2.29\ mg = 8.29\ mg$$

Total concentration:

$$c = \frac{n}{V} = \frac{8.29\ mg}{0.006\ L + 0.030\ L} = 230.37 \frac{mg}{L}$$

The results obtained from the spike tests can be seen in the table below:

	Concentration [mg/L]		
	Spike 0	Spike 1	Spike 2
	77,88	166,40	238,00
	78,29	165,82	238,99
	78,11	165,92	238,98
	77,81	165,74	239,59
	78,31	165,73	239,66
Mean:	**78,08**	**165,92**	**239,04**
Variance:	**0,053**	**0,078**	**0,441**

The concentrations here are calculated from the peak areas in the following tables. From each set of data, we can compare the amounts measured to the expected amounts, and find the percentage of (genfundet) Coca-Cola Zero in each sample:

Spike 0	Area [mAU*s]	Concentration [mg/L]	Expected concentration [mg/L]	Difference	Recovery [%]
1	2381,64	77,88	76,44	1,44	101,9
2	2393,18	78,29	76,44	1,85	102,4
3	2388,03	78,11	76,44	1,67	102,2
4	2379,4	77,81	76,44	1,37	101,8
5	2393,96	78,31	76,44	1,87	102,5
Mean:	**2387,24**	**78,08**	**76,44**	**1,64**	**102,1**
Standard deviation:		0,23042			0,30144

Spike 1	Area [mAU*s]	Concentration [mg/L]	Expected concentration [mg/L]	Difference	Recovery [%]
1	4914,58	166,40	160,4	6,00	103,7
2	4897,95	165,82	160,4	5,42	103,4
3	4900,69	165,92	160,4	5,52	103,4
4	4895,6	165,74	160,4	5,34	103,3
5	4895,2	165,73	160,4	5,33	103,3
Mean:	**4900,80**	**165,92**	**160,4**	**5,52**	**103,4**
Standard deviation:		0,27981			0,17444

Spike 2	Area [mAU*s]	Concentration [mg/L]	Expected concentration [mg/L]	Difference	Recovery [%]
1	6963,42	238,00	230,37	7,63	103,3
2	6991,56	238,99	230,37	8,62	103,7
3	6991,48	238,98	230,37	8,61	103,7
4	7008,79	239,59	230,37	9,22	104,0
5	7010,77	239,66	230,37	9,29	104,0
Mean:	**6993,20**	**239,04**	**230,37**	**8,67**	**103,8**
Standard deviation:		0,66405			0,28825

As mentioned earlier, we expect the unspiked Coca Cola Zero to have a mean concentration of ca. μ=76.44 mg/L. As we can see from the table of spike 0, we find a mean of 102.1 % of what we expected to find. We can then expect to find more in the spike 1 and spike 2 samples as well, so that the ideal recovery from these also would be at a few percentages higher than 100.

We can now find the mean of the recovery from both spike 1 and spike 2 combined, and the standard deviation of these samples. We can also calculate the CV% to give a better idea of the deviation:

	Recovery [%]
	103,74
	103,38
	103,44
	103,33
	103,32
	103,31
	103,74
	103,74
	104,00
	104,03
Mean	**103,60**
Standard deviation	**0,2819**
CV%	**0,2721**

So we end up with a mean recovery of 103.60 %, which means that this is what can be expected for future attempts of spiking. As mentioned, there was also minor increase in the amount measured in relation to the amount expected in the spike 0 tests. It might be possible that this increase also affects spike 1 and spike 2, even though we cannot confirm this.

We have a CV% of 0.27 %, and this corresponds to the variation that can be expected if more tests were made. It is also important to remember, that the cola we use for spiking is diluted by a factor of 5 before being spiked.

8.5. Cola samples

So far, we have only looked at the tests of a single type of cola from the same bottle. We also tested three other brands of colas, with the purpose of finding out whether the aspartame contents in these matched each other or not. Furthermore, we tested the same four brands of cola once again, this time from new bottles, to see whether the contents also differ from bottle to bottle, instead of just from brand to brand.

The resulting concentrations found are listed in the table below. We start by analyzing Pepsi Max:

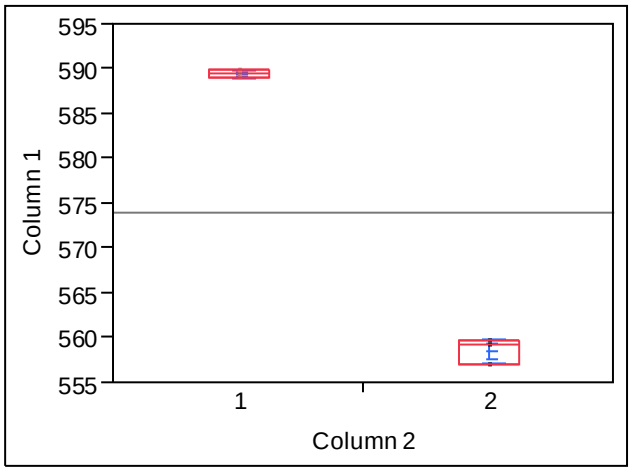
Old Pepsi Max			New Pepsi Max		
Area [mAU*s]	**Concentration [mg/L]**	**Real concentration [mg/L]**	**Area [mAU*s]**	**Concentration [mg/L]**	**Real concentration [mg/L]**
3523,89	117,80	589,01	3340,15	111,38	556,91
3529,47	118,00	589,99	3353,32	111,84	559,21
3526,22	117,88	589,42	3355,32	111,91	559,56

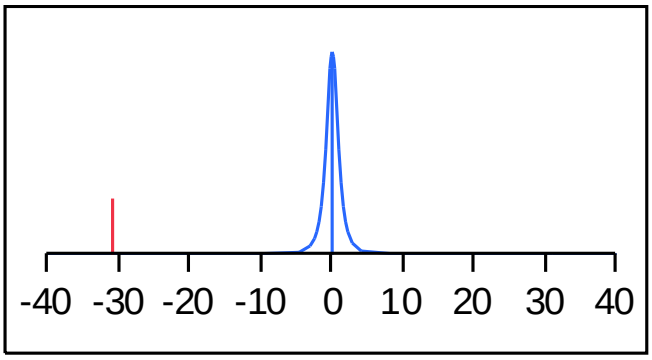
Immediately, it is possible to see a difference between the measured concentrations in the new and old Pepsi Max's. We want to test the means by using a t-test, so we start by setting up a null hypothesis and an alternative hypothesis:

$$H_0: \mu_1 - \mu_2 = \delta_0 = 0$$

$$H_1: \mu_1 - \mu_2 \neq \delta_0$$

We then input the data in SAS JMP, and run a t-test:





From the graphics, it is possible to see that the two means are different, based on their low variance. We now set up a table of data from the t-test:

	Old Pepsi Max	New Pepsi Max
Mean	589.47	558.56
Standard deviation	0.4922	1.440

	t-test data
t-ratio	-35.193
Prop > |t|	0.0002

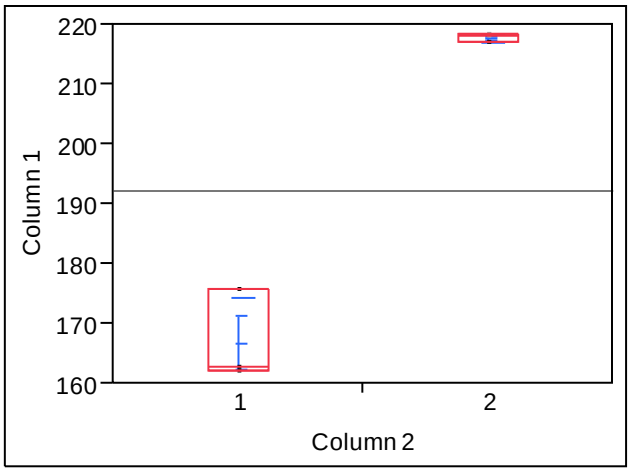
The parameter “Prop > |t|” gives the probability for the t-ratio to be of the same or a more extreme value, provided we assume the null hypothesis is correct. If this value is lower than the value of our chosen level of significance, then we can reject the null hypothesis. With 99 % confidence, this null hypothesis can easily be rejected.

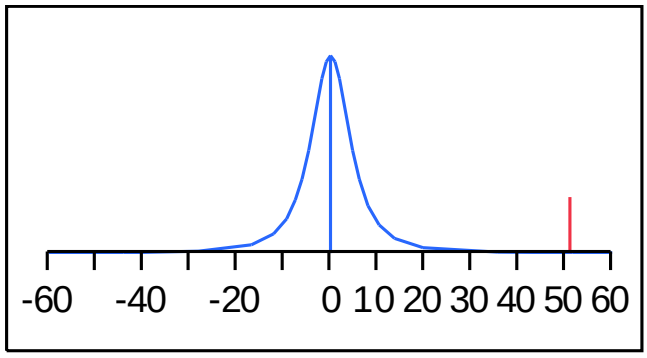
So we conclude that the two means are different, meaning that there is a significant difference in the concentrations of the different colas.

We then look at Coca Cola Light:

Old Coca Cola Light			New Coca Cola Light		
Area [mAU*s]	**Concentration [mg/L]**	**Real concentration [mg/L]**	**Area [mAU*s]**	**Concentration [mg/L]**	**Real concentration [mg/L]**
1080,25	32,40	162,02	1394,69	43,39	216,97
1158,21	35,13	175,64	1400,25	43,59	217,94
1083,25	32,51	162,55	1401,8	43,64	218,21

With the same hypothesis, we also test whether the means here are similar or not:





	Old Coca Cola Light	New Coca Cola Light
Mean	166.74	217.71
Standard deviation	7.715	0.6521

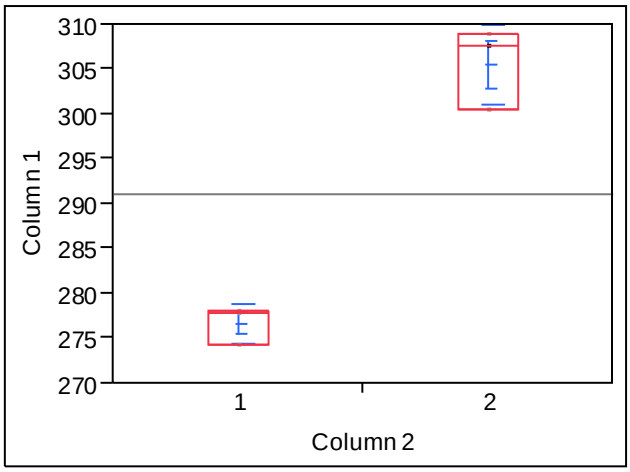
	t-test data
t-ratio	11.402
Prop > |t|	0.0072

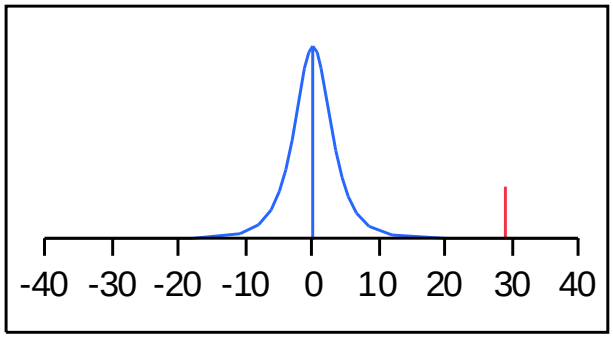
The probability for the t-ratio is higher this time, but can still be rejected with 98 % confidence.

We also test Harboe Minus:

Old Harboe Minus			New Harboe Minus		
Area [mAU*s]	**Concentration [mg/L]**	**Real concentration [mg/L]**	**Area [mAU*s]**	**Concentration [mg/L]**	**Real concentration [mg/L]**
1744,58	55,62	278,10	1872,15	60,08	300,39
1742,75	55,56	277,78	1920,31	61,76	308,81
1722,39	54,85	274,23	1912,56	61,49	307,46

Results of the t-test:





	Old Harboe Minus	New Harboe Minus
Mean	276.70	305.55
Standard deviation	2.148	2.611

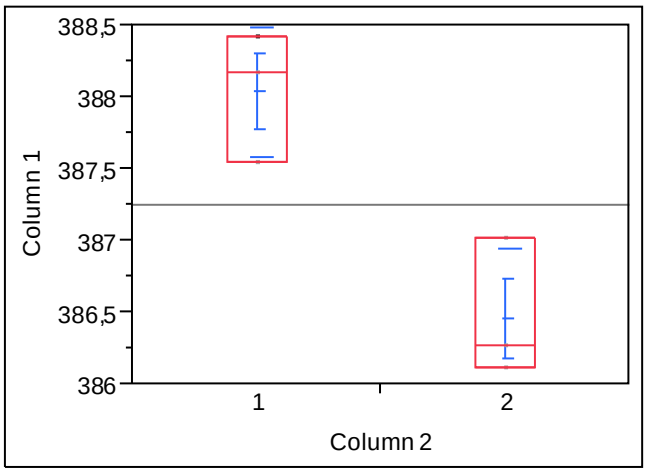
	t-test data
t-ratio	9.981
Prop > |t|	0.0026

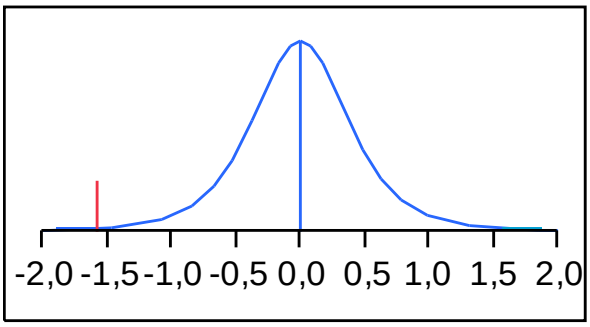
We see higher standard deviations for this brand of cola. However, the null hypothesis can still be rejected with 99 % confidence.

Lastly, we can look at Coca Cola Zero. As we did a lot of tests on the first Coca Cola Zero we had, we randomly choose a set of three data points from the same sample preparation to compare with the three data points from the new Coca Cola Zero. We then proceed with t-testing in the same way as the above three brands:

Old Coca Cola Zero			New Coca Cola Zero		
Area [mAU*s]	**Concentration [mg/L]**	**Real concentration [mg/L]**	**Area [mAU*s]**	**Concentration [mg/L]**	**Real concentration [mg/L]**
2374,42	77,63	388,16	2367,84	77,40	387,01
2375,9	77,68	388,42	2362,70	77,22	386,11
2370,88	77,51	387,54	2363,57	77,25	386,26

t-test:





	Old Coca Cola Zero	New Coca Cola Zero
Mean	388.04	386.46
Standard deviation	0.4521	0.4822

	t-test data
t-ratio	-4.140
Prop > |t|	0.0145

Here, we can reject the null hypothesis with 95 % confidence. As we showed in the repeatability and reproducibility, it is possible to say that two samples from the same cola will give the same results. However, we can conclude for all of the four tested cola brands that there is a difference in the concentrations of aspartame between production dates.

9. Discussion

The aim of the project was to find a method with which we could find the concentration of aspartame in soft drinks. We wanted to be able to separate aspartame from the other components in the cola via HPLC analysis, which we would accomplish by changing different analysis parameters. We based our analysis on a method from a scientific article.

At first, we ran several samples of different cola brands to get an idea about the levels of aspartame present in the colas. Afterwards, we ran a standard solution to check what retention we could expect aspartame to have under the chosen conditions.

It is a good idea to try to determine the concentration range we can expect to see our cola samples in, by running a few samples once the details of the method have been confirmed by good chromatograms that meet the requirements. From these samples, we can determine the range in which we need the concentrations of our standard solutions to be.

The retention time of aspartame in all the samples highly depends on the buffer; this can be the pH of the buffer, or the contents of the buffer other than the bases/acids. We have been very careful in the laboratory with preparing the buffer, to avoid having these changes in the retention time. The change in retention time could also be caused by other unknown factors.

We used a shorter column than the one given in the article, since we had to share the HPLC with another group. This meant that we had to find a column that both groups could use. This did not seem to affect our results in a negative way, but seemed to give us an acceptable separation of the different components of the soft drinks and a shorter retention time of aspartame. We use another wavelength than given in the article, with which we achieve a higher absorbance, meaning that the results are more reliable.

We achieved separation of the wanted products, but the resolution of aspartame in relation to the nearest peak was very high. Therefore, it would be an option to optimize the method even more. This could be done by changing the gradient flow, which could result in reduced analysis times. It would also be possible to run standards and test for the other component that appears strongly in every chromatogram; this component is believed to be another sweetener, and a simultaneous determination of two sweeteners would prove more useful.

Because of the limited time in the laboratory, we choose to keep the method as it is, and start the measurements. Since every time a change to the method is made, we would have to start all over with all analyses, and we do not want that.

We found that the data is normally distributed, and we produced a standard curve with a range of 0-500 mg/L of aspartame, with good linearity. We find a coefficient of determination of 0.999, which is very satisfying. We then choose to use this standard curve to calibrate the concentration of all our future samples. To get an acceptable calibration curve, it is important to be very careful when preparing the samples in the laboratory. We measure each standard 3 times to be statistically sure that the data is useable. Since the standard curve does not go through (0,0) on the graph (even within the confidence intervals) there must be some error involved with the measurements. This could be caused by human error, a continuous error in the detector of the HPLC apparatus, or perhaps if the curve is not entirely linear. This last problem would cause the trendline to deviate from (0,0) if some of the other points are off course.

From the ANOVA calculations, we can see that human error has a big influence on the sample preparations. This can also be caused by sample preparations done by different people. This might also have had an influence on the standard curve.

We determined via a t-test that the method has both repeatability and reproducibility, but only for the same cola. As we tested over several sample preparations, and over two days, the method is valid for the same cola at any times.

To calculate the recovery, we use the calibration curve to find the concentration of the spiked samples. We compared this to the expected concentrations and find a mean recovery of 103.60 %. This could be caused by several factors. A problem could be the variation in the calibration curve, along with error in the machine or human error, as described before.

We tested cola samples from different bottles, and determined via t-tests that the difference from bottle to bottle of the same brand is significant. This would mean that the producers do not measure the exact amount put into every batch. In reality, this has no significance as long as the concentrations of aspartame stay under the legal limits. We found that there is a huge difference in the amount of aspartame between the different brands.

In general, our repetitions have a very low standard deviation. This is the standard deviation of the machine. The standard deviation between the sample preparations is much higher, and this is caused by the human error. This proves that the machine is the most reliable part.

10. Conclusion

Choice of method

When we first started the project, we were looking at a method that included a long sample preparation. Unfortunately, long sample preparation are not only time demanding, but also provide a higher probability of human error in the laboratory. However, we later found a newer method that required less sample preparation, and we decided to use this method as a starting point for our own experimental work. We improved on the method based on what resources were available in the laboratory, and ended up with a solid method that we would then validate using statistics.

The HPLC (high performance liquid chromatography) machine we use has the following specs:

Merck HPLC system with quaternary pump L-7100, auto sampler L-7200, column thermostat L-7360, Diode Array Detector (DAD) L-7455, interface L-7000, solvent degasser L-7612 and HMS manager software LiChroCART 250-4 LiChrospher 100 RP-18 (5 µm) Merck column

The method gave retention times for aspartame in the area of just over **12 minutes**. In total, the cola samples were run for 30 minutes, with a post run of 5 minutes included to clear the column of any residuals. We observed perfect, Gaussian peaks for both aspartame and other UV-absorbing content of the cola samples.

Analytical parameters

We obtained the first chromatograms from our cola samples, and used these to calculate some of the most important parameters related to the chromatograms; retention times, selectivity, number of plates and resolution. We conclude that the yielded chromatograms have excellent plate numbers and resolution – for the selected chromatogram, the plate number (column efficiency) is **28535**, giving an HETP (height equivalence of a theoretical plate) of **5.2566** µm. Furthermore, we calculated the resolution of the aspartame peak in same chromatogram in relation to the nearest large peak, and obtain a value of **27.5379**. Based on the similarity of all the chromatograms, we conclude that these results can be transferred over to all of the aspartame peaks of the chromatograms we obtain during the entire project.

Statistical parameters

The standard curve was created from the data of the standard solutions, and resulted in a linear curve with a coefficient of determination of **0.999**, and a linear range of about 0-500 mg/L of aspartame. From this curve, we obtained the equation used to calculate the concentration of all of our cola samples. With the parameters α and β (the intersection and the slope respectively), we got $\alpha = 153.00 \pm 123.38$ and $\beta = 28.61 \pm 0.48$. With the confidence intervals with α=0.05 also calculated, we can see that there is very little variation in the value of the slope of the curve. However, the confidence interval for the intersection is large, but does not contain the value 0 as it should have, which could be caused by several factors.

We conclude that the method used has repeatability and reproducibility; by t-testing, we found that the concentration measured in the same cola was the same from one day to another, and between different sample preparations. We found the variance of the different aspects of the tests; the variance of the machine, the variance of the sample preparations and the variance between days.

The precision of the method was tested by ANOVA, from which we can conclude that the precision of different sample preparations highly depends on the person performing the analysis.

In the spiked samples, we find a percentage of recovery of **103.60 %**. This is above the expected value of 100 %, and could be caused by error in the sample preparation. In the unspiked samples, we find 102.1 %, which backs up the above hypothesis that it could be caused by human error.

For the 4 different cola brands tested, we conclude by t-test that the contents differ from bottle to bottle. The contents of the different colas can be seen below:

	Old Cola concentration [mg/L] (bottle 1)	New Cola concentration [mg/L] (bottle 2)
Coca Cola Zero	382.21	386.46
Coca Cola Light	166.74	217.70
Pepsi Max	589.47	558.56
Harboe Minus	276.70	305.55

The concentrations shown are for non-diluted cola samples.

So all in all, we acquired the results we were after, and were able to develop a method that works for finding aspartame in soft drinks.

11. Sources

[1] http://www.springerlink.com.proxy1-bib.sdu.dk:2048/content/tg5m343k8758p780/fulltext.pdf

[2] http://www.pharmgkb.org/views/index.jsp?objId=PA448493#tabview=tab2&subtab=31

[3] http://www.afssa.fr/ftp/basedoc/aspartamgb.pdf

[4, p. 2579] http://ac.els-cdn.com.proxy1-bib.sdu.dk:2048/S0008621508004229/1-s2.0-S0008621508004229-main.pdf?_tid=c298448c4727ac6fd154e85599384ecb&acdnat=1336570730_9fd6d510c7d1406dbe110079ced1d621

[5] http://www.efsa.europa.eu/en/topics/topic/aspartame.htm

[6] Halldorsson TI, Strøm M, Petersen SB, Olsen SF, 2010. Intake of artificially sweetened soft drinks and risk of preterm delivery: a prospective cohort study in 59,334 Danish pregnant women. American Journal of Clinical Nutrition 92, 626-633.

[7] Soffritti M, Belpoggi F, Manservigi M, Tibaldi E, Lauriola M, Falcioni L, Luciano Bua L, 2010. Aspartame Administered in Feed, Beginning Prenatally Through Life Span, Induces Cancers of the Liver and Lung in Male Swiss Mice. American Journal of Industrial Medicine 53, 1197-1206.

[8] http://www.efsa.europa.eu/en/efsajournal/doc/1996.pdf

[9] http://www.fsai.ie/uploadedFiles/Dir94.35.pdf

[10]http://www.foedevarestyrelsen.dk/SiteCollectionDocuments/25_PDF_word_filer%20til%20download/06kontor/endelig%20rapport%20slankekost%202005-20-64-00453.pdf

[11] http://www.fvm.dk/english.aspx?id=14541

[12] Textbook: Modern HPLC for practicing scientist. Michael W. Dong. Ed.Wiley-Interscience

[13] Textbook: Johnson, R., Freund, J., Miller, I., Probability and statistics for engineers (2011), eighth edition, Pearson.

[14] http://musikkrom.wikispaces.com/normbasert+vurdering

[15] SAS JMP Statistical Software

[16] http://en.wikipedia.org/wiki/Student%27s_t-distribution

[17] http://en.wikipedia.org/wiki/F-distribution

[18] http://www.analyse-it.com/products/standard/correlation.aspx

[19] http://udel.edu/~mcdonald/statregression.html

[20] http://mathworld.wolfram.com/CorrelationCoefficient.html

[21] http://www.creative-wisdom.com/computer/sas/regression_assumption.html

[22] http://maths.fallibroome.cheshire.sch.uk/coursework.htm

[23] Textbook: Blom-Hanssen, J., Statistik for praktikere (2002), first edition, first issue, Ingeniøren|bøger, ingeniøren A/S

[24] http://www.standardbase.com/tech/HPLC%20validation%20PE.pdf

[25] http://irmm.jrc.ec.europa.eu/activities/food_additives/Documents/eur22727en.pdf

[26] http://fabad.org/fabad.org/pdf/volum32/issue1/15-22.pdf

[27] http://www.cvg.ca/images/HPLC_Method_Development.pdf

[28]http://www.thairohs.org/index.php?option=com_docman&task=doc_view&gid=66&PHPSESSID=a32b3787a8773f7baefd136f409460bb

[29] Wróbel, K., Wróbel, K., Determination of aspartame and phenylalanine in diet soft drinks by high-performance liquid chromatography with direct spectroflourimetric detection (1996), Journal of Chromatography A 773.

[30] Nseir, W., Nassar, F., Assy, N., Soft drinks consumption and nonalcoholic fatty liver disease (2010), World Journal of Gastroenterology, volume 16, issue 21.

[31] Croitoru, M. D., Fülöp, I., Kincses Ajtay, M., Balogh, C., Dogaru, M. T., Direct HPLC-UV determination of cyclamate, saccharine and aspartame from soft drinks (2011), Acta Alimentaria, volume 40.