CH 452

NORTH CAROLINA STATE UNIVERSITY.

Department of Chemistry

Measurements I Laboratory

Manual for Students

Contents

Basic Laboratory Experiments

Physical Experiments

Analytical Experiments

Computer tutorials

I Overview

This laboratory course in Physical and Analytical Chemistry has experimental and computational components. The computational components emphasize statistics, data analysis and methods that may be used for data processing. Many students in this course will have had little exposure to statistics. Students are encouraged to learn from the computational modules early in the course and consistently apply statistical methods throughout. A good foundation in statistics is extremely important for future work in industry or academia. Perhaps the most important aspect of the course is learning how to write and present scientific results. In this regard as well students are encouraged to start immediately to apply scientific writing skills. Self-criticism and criticism by others is important. It will help you to share your writing with a classmate and to correct or critique each other's work. There is a great deal of team work in this class. You may share many things, but please do your own writing. Please make sure to check text for plagiarism. If you cite other work, please use quotation marks and ensure that the references are properly cited. If you take information from the internet, try if at all possible to find the original source. Do not use websites as citations unless they are databases (e.g. the protein data base or the NIST data base for chemical properties).

II Safety

II.1 Equipment

a) Safety Glasses. Each student will be issued a pair of safety glasses for their personal used. *They must be worn at all times in the laboratory*
b) Fire extinguisher: Located next to the main door by the shelf of chemicals. It is useful for all fires except those involving alkali metals.
c) Fire blanket. Located next door to the door connecting to Dabney 612. It is used where clothing is on fire.
d) Eyewash. Located on lab bench opposite main door. It is used in case chemicals are splashed in the eye.
e) First Aid Kit. Located on the shelf next to the main door.
f) Safety Plan and Material Safety Data Sheets: Located in reagent shelf next to main door

II.2. Evacuation

When: Evacuate the laboratory on directions by either the TA or the instructor or when the alarm in the hall is sounding continuously.

Where: Proceed to the *right* and down the west stairwell to the ground floor. Gather under Williams Hall for a head count.

II.3 Spills

Strong acids or bases should be neutralized with bicarbonate solution (on side shelf next to first aid kit).
Other spills should be adsorbed in paper or cloth towels that are then placed in plastic bags labeled with the material spilled. Notify instructor.

II.4 Waste disposal

a) Always label chemicals used in laboratory experiments. When chemicals are dispensed into a beaker or vial, you should always label it in advance to avoid confusion.
b) Organic Wastes should be separated into those that contain halogenated *compounds* and those that do not (like I_2 in an non-halogenated organic solvent)
c) Sodium and potassium compounds and dilute acids and bases can be flushed down the drain.
d) Broken glassware and disposable pipettes should be placed in the cardboard glass waste box.

II.5 General Practices

a) Unplug equipment before making connections. Have the TA check your setup before beginning data collection.
b) Do *not* conduct unauthorized experiments. The student's initiative is welcome and appreciated, but only if the TA and/or instructor are consulted before modifying any procedures.
c) Use *bulbs* for all pipetting.
d) Do not attempt to move gas cylinders without consulting instructor. Gas cylinders are quite vulnerable and dangerous, *unless* the regulator is taken off and the protective cap is put back in place.
e) Volatile compounds must be transferred in the fume hood. Strong acids must be stored in the blue acid storage cabinet (under the bench by the window)
f) Close fume hood sash when not in use.
g) No eating or drinking in the lab.
h) Proper clothing is required: no open shoes or bare legs.
i) When opening sealed glass ampoules these need to be scored properly and wrapped in paper towels to prevent wounds from shattered glass.
j) Use proper (non-cloth) gloves when dispensing liquid nitrogen. Cloth gloves can freeze onto your skin and cause a burn if the liquid nitrogen spills on them.
k) Use tongs or non-cloth thermal gloves when handling dry ice.

III General requirements and good practice

III.1 Laboratory

The experimental facilities normally used in Chemistry 452 are on the **6th** floor in Dabney. The lab itself is Dabney 608. The computer tutorials in Dabney 121, 613 or other computer facilities in Dabney Hall.

III.2 Equipment and supplies

Most of the common chemicals required in the student's work are stocked in the laboratory. Should anything be missing, ask the TA. This also holds for ice, dry ice and liquid Nitrogen.

III.3 Data recording and documentation

Proper documentation is a vital part of experimentation, as the CH 452 lab will show.
A permanently bound **notebook** should be reserved for the use of the course, preferably an official version. A carbon copy will be made of each data page and left with the TA at the end of each lab period. The notebook need contain only original data with appropriate headings as to name of experiment, data (in tabular form) and partners. A proper record of the filenames and contents of electronic files should also be entered if appropriate. Calculations, preliminary graphs and any supplemental notes may be included. The calendar date must be recorded in the notebook at the top of every page and at the beginning of each day's entry. Please use an alphanumeric format for the month, 2-mar-01 or Mar-2 2001. Notation as 2/3/01 or 3/2/01 is confusing as it means different things in different parts of the world.
Use a ballpoint pen or other permanent marker that provides a *permanent* record as well as a good carbon copy. Errors are crossed out with a single stroke, never erased.

A (virus free) **USB drive** needs to be reserved for the use of the course as well. Some of the data are electronic and many of the calculations will require spreadsheet use.

The emphasis on documentation (period) is a *scientific* one, but there are *legal* reasons to insist on a format with permanently bound books using permanent markers, signatures, dates etc. They are needed to substantiate a *patent* claim and many industries therefore impose them as a standard. Unfortunately, that can make recording electronic data, graphs etc. a laborious task.

III.4 Calculations

Most of the intermediary calculations, e.g. calculating the molarity of a solution to be made up, require a **pocket calculator**. Most data work-up requires a spreadsheet and can be done in the computer facilities open to the students, e.g. on the 1st and 7th floor.

III.5 The need for spreadsheets and Excel analysis

A **laptop** is strongly encouraged for this lab. This is particularly useful for the second labs. The laptop needs to have Excel and the Data Analysis pack needs to be loaded. A strong emphasis of the lab is on the elaboration of data and their interpretation and report writing time can be much reduced if this is done on the spot rather than postponed till after lab, because the teacher's help is no longer available in that case.

The spreadsheet of choice for this lab is Microsoft Excel. There is a version present in the lab for in-lab analysis, needed for some of the experiments. A basic proficiency in Excel is assumed present for every student, but considerable additional use, particularly of the Statistics options of the spreadsheet, will be taught in the course of the lab.

III.6 Tabulation

An important part of the lab is to learn how to glean the essential information and present it as such. The use of tables is an important tool. They typically go in the Results section of the report. They need to be numbered, so that they can be easily referred to in the discussion.

III.7 Graphic representation and Statistics

A separate handout will be provided on the necessary statistics and error propagation topics that will be applied in this lab. The introduction phase will concentrate on the basic aspects of statistics. The rest will be taught as the need arises.

Graphs can all be made using Excel, but certain errors must be avoided.

a) Measured data points should be shown symbols only (no connecting lines)
b) Regression lines should be line only (no symbols) in overlay with the data.
c) Avoid color problems: some colors are lost when printing of photocopying in black and white.
d) The data should *fill* the graph. If necessary readjust the x and y scales. (The origin of the graph does not need to be at (0,0).
e) Properly label the axes with legends that indicate the units.
f) Choose the units sensibly. Avoid marker legends like 0.000001 0.000002 etc. preferably by switching to a decimal prefix like m,μ,n, and p in the units or

else by multiplying by a factor 10^a. This needs to be reflected in the axis legends

g) The graph title is *not* a repeat of the axis labels, but describes the process. E.g. if the axis as lnV and lnP the title could be “Adiabatic compression of CO_2.”

h) Be careful with font sizes. This particularly true for a presentation. When in doubt, try it out! Project it, go sit where the audience sits and see whether the result is legible or obnoxious.

i) If you struggle with a certain effect in the spreadsheet, *ask* the TA, the instructor or your fellow students.

IV Reports

The advancement of science depends strongly on proper *communication*. This is true on various levels; the most obvious one is that scientific work is not ‘finished’ (and will not be credited!) until it is published in written form. Formal and less formal forms of presentation, either orally or e.g. via a poster, are also important parts of a scientist’s work and indeed for many non-scientists as well.

All forms of reporting are preceded by an exercise in ***the organization of one’s thoughts***. After all, communication is the very transfer of one’s thoughts, convictions, and conclusions, what have you. Such a transfer is seldom a success, if those thoughts do not get organized beforehand. Therefore, the emphasis on good reporting in this (and other labs) goes much deeper than just the reports themselves: it trains people to think more clearly.

Although a written report and an oral one can very well cover the same subject matter, the two formats clearly have a different emphasis:

IV.1 Final presentations

An **oral** presentation revolves around ***clarity***. *Timing* is of the essence. The audience should be able to immediately understand what the speaker says. The same utterance made a few sentences too early or too late will impair the understanding. Time is also short, so that completeness is impossible and proper priorities must be made. The audience, however, can correct the priorities by asking questions.

Acquiring oral presentation skills is best done by practice. In the project phase, lecture time will be used by the students to give a brief *oral progress report* on their work in preparation for the final presentation. Students are required to be both speakers (on their A-day) and audience (on their B-day).

IV.2 Written reports

A written report revolves around ***retrievability.*** *Location* is of the essence. The information that the reader desires must be in the right spot (proper headers) and presented in a suitable format (tables, graphs). Some prioritizing is unavoidable for reasons of size, but completeness is a goal, because the reader has no recourse if something vital is left out.

Skillful use of the English (and Scientific) language is a great help in writing a good report and although personal traits and talents (and/or national origin) enter into this part of the equation, there are no skills without practice. Scientific writing will seldom qualify for a literary prize, if only because science is a global game and most global villagers have a different mother tongue. Nevertheless, there are important parameters like precision, clarity of expression, ease of reading, conciseness and correctness of rhetoric, grammar and spelling to consider. Impeccable logic and a willingness to step back and view the results through the eyes of the reader are necessary as well. These aspects can only be acquired by doing and comparing and being corrected.

A few general remarks:

a) The report must be an *original* piece of writing. All calculations including error estimation may be done as a team, but the report writing is *individual*.
b) Assume that the reader has the level of competence of an average CH 452 student, no more, no less.
c) Data or other material from an outside source must be referenced in full. (This does not apply to well-known constants like the Boltzmann constant, etc)
d) Overly lengthy reports are a sign of poor thought organization and will be graded as such.
e) The report should be well balanced, i.e. not overly focused on a minute detail, while skipping other important items. (Stated otherwise: authors must develop their sense of priority).
f) A scientific measurement must be reported in its complete form. In general it has four components: a sign, a magnitude, an uncertainty and a unit. More details will follow in the statistics handout.
g) Reports are typically written in an impersonal style. Despite its stylistic drawbacks, the use of passive voice is encouraged. (I.e. "The data are shown.." rather than "I show the data.."). As scientific practice stands today, the use of "we" and "our" in active voice is acceptable and on the increase, but that of "I" and "my" is not. If necessary use "the author" instead of "I" or pretend you are an anointed king and write a majestic "we".

The format of a report is very important for the sake of retrievability and reference. See the provided Formal Lab report rubric for the different sections of a lab report and what should go in each section.

V. Introduction to Statistics

Introduction.

A famous philosopher of science (K. Popper) states that what people do in science is to make *observations*. Most often they do so in a well-structured experiment. Subsequently they use their imagination to create a *model*. Then they compare the two and see if the model fits the observations. On this basis they make a conclusion regarding the validity of their model and/or their observations, which they communicate in report or publication. If the fit is bad, they may have to either amend or replace their model. Popper emphasizes that it is actually the latter *falsification* of the model that causes the steady progress in our knowledge. He also claims that the model always remains only a model, and never becomes the 'Truth'. Popper would say that it is not possible to prove that a model is correct, only that it is incorrect. Other philosophers, however, that it may also be difficult to prove that a model is incorrect. Perhaps a better approach is to assign a probability to the model based on our understanding of the state of knowledge and current data. This probability is a "prior" assessment of the model. After conducting experiments one can reevaluate the model and determine whether the measurements increased the probability of the model being correct. This type of approach is based on Bayesian statistics. In either case, modern science is based on comparisons of models to data or measured quantities. In this course we will focus mainly on understanding the quality of data and how to analyze measurements so that they may be compared with models.

Mathematical models have proven to be the most powerful type of models. There are other useful models, e.g. verbal ones, like the taxonomy of species in Biology or schematic ones, like the periodic table in Chemistry. While these non-mathematical models should not be underestimated, Physicists and Physical Chemists much prefer to work with mathematical models involving information of a quantitative numerical nature only. In fact, if possible, they try to translate other models into mathematical ones and find a way to make their observations numerical. The mathematical reduction of numerical data is therefore an essential skill in Physical Chemistry. Analytical Chemists are very helpful in this process because the concentrate on making sure the measured information is available and of high quality. In addition their work is often used for other purposes than the advancement of science, e.g. in forensic or medical science their measurement can well decide a person's fate.

Unfortunately, there are some serious pitfalls in both measurement and subsequent data reduction, because all numerical data are subject to *uncertainty* and can be fraught with other forms of error as well. We need to learn to recognize these error types and how to deal with them. In fact in many cases the latter often demands that we *anticipate* possible errors in the experimental phase, if we wish to be successful in our efforts.

1. Data reduction and error types

Experimental data can be subject to a number of different types of errors.

First of all there are **gross** errors, e.g. you type in the wrong number or you take the wrong sample flask etc. It can also be beyond your control, e.g. a surge in the voltage of the electricity you are powering your instrument with. Gross errors are typically *singular* events that can seriously upset the structure of the data.

Secondly, there are **systematic** or determinate errors, e.g. as a result of equipment being misaligned or wrongly calibrated. Such effects affect all data taken during the experiment in a systematic way, (although not necessarily all to the same extent.)

Thirdly, there are, what statisticians call **random** errors. Scientists often prefer to talk about the **uncertainty of the measurement,** because they cannot help making this 'error'.
A balance may indicate that your sample weighs 123.4 mg in the first measurement, but a second (**replicate)** measurement may well show 123.6 mg. The differences are totally

unpredictable and independent, like the throw of a die or the movements of molecules in an ideal gas.

A fourth kind of error results from **incompletely randomized** effects, chaos, etc. Drifting oven temperatures and electronic radio frequency (r_f) noise are a good example. This type of error is neither completely systematic nor completely random. As in the structure of a liquid there usually is short-range order either in time or in space, but no long-range predictability of errors due to incomplete randomization. These types of interference are notoriously hard to deal with. Unfortunately, the phenomenon occurs frequently when mixing is not done properly, when equipment drifts, when 'room temperature' varies with the weather or the fickleness of the air conditioning system etc.

2. Strategies to deal with errors

Each error type requires its own remedial strategy, as summarized below

Type of error	resulting in	strategies	statistical procedures	quality aspect
I: Gross	outliers	prevention documentation rejection	robust stats	'clean' data
II: systematic	bias	calibration	N/A	accurate data
III: Random	uncertainty 'white noise'	replication accumulation	averaging regression	precise and reproducible data
IV: incompletely Random	drift, chaos 'r_f noise', inhomogeneity	prevention, short duration mixing	(hardly any..)	'stable' data

Scientific data usually contain a combination of error types. Unfortunately, the presence of one type of error can interfere with the remedy of the other. For example, data need to be stable (IV) and clean (I) before using averaging or regression techniques to deal with uncertainty (III). Likewise data need to be stable (IV), before robust techniques can be used to reject outliers (I). Such robust statistical techniques do exist, but it is always preferable to have valid reasons other than the data themselves to justify rejecting a data point. This is why good documentation of what happened during an experiment is a powerful weapon against outliers or gross error.

Graphic representation and statistics are the only defense against random errors (type III). Graphics are powerful but often somewhat subjective. Statistics were specifically developed to deal with uncertainty in order to arrive at scientific conclusions objectively. This is why a scientist requires a reasonable understanding of basic statistics. However, even the best statistics in the world do not compensate for a calibration error (type II).

There are few cures for type IV errors, apart from prevention. Make sure your solutions are homogeneous by mixing properly before the start of an experiment. Improper mixing is a notorious cause of type IV error. (Remedy: try again…)

3. Physical measurements

There are many ways to measure something, varying from detecting light, sound, the force of gravity, a magnetic force etc. A device with which you measure a quantity is called an ***instrument*** (not a: machine) and not to get fooled by its values or destroy instruments you need to understand a whole bunch of things.

3.1 Measurement variables

A balance e.g. measures weight, which is the *force* (in N) that the earth attracts a mass. We can derive what the mass (in kg) is from that, but that is not the primary variable. This is often the case. We often measure something *indirectly*. Obviously, the questions of *dimension* and *units* are related to this issue. And yes, measurements need to be reported with their appropriate units.

A measurement detects *signal* of some sort.

One measurement is called a *datum*. It is a Latin word meaning *that which is given,* i.e. *the given* or *gift* or *yield*. It is a neuter word, derived from the verb *dare* meaning *to give.* The plural is *data*. In English the singular datum is in the process of becoming obsolete and often replaced by *data point*.

3.2 Measurement principle

A balance e.g. can involve a spring, a lever or an electromagnet. In this first case you measure the spring's deformation, the second the position of the lever and in the third you measure the electrical current needed to undo the deflection of a lever. Each principle leads to its own accuracy, precision but also cost.

3.3 Measurement requirements

How big must your sample be? How big can it be? How concentrated? How heavy? Can you tolerate an impurity to be there? Do you need a special sample container? Will your sample ruin the instrument? All these are important considerations before you do a measurement. The word *before* is the operative one. It is your job to find these things out before you do harm to your sample or someone else's instrument or both.

3.3.1 Measurement conditions

If an instrument works well at 25 ^{o}C in a dry environment, does it also work in a rainforest? Does it work at 30 K or at 1000 ^{o}C? Will it work an ocean going vessel? Does a balance work in space? Does your sample require refrigeration?

3.4 Measurement costs

Measurements are *not* free. People's time and effort is one cost, but reagents and instrument time are also important costs.

3.4.1 Instrumentation costs

Instruments are hard to make and therefore *costly*. Students need to be taught that their job is to do their best on using instruments wisely, *not* to be afraid to use them. In their careers it is inevitable that they will damage something at some point. That risk cannot be avoided only reduced by awareness and a willingness to prepare and be vigilant.

3.4.2 Consumable costs

Lamps wear out and must be replaced, weighing paper gets disposed and must be bought, solutions must be made up and this costs money for the chemical and time to make them. Afterwards chemical must be disposed of in an environmentally acceptable manner.

3.4.3 Time costs

The time it takes to perform a measurement is an important consideration, as we shall see there are statistical reasons for that as well as practical ones. If a measurement is fast and automatable it is often possible to take measurements while changing conditions, e.g. the temperature, the wavelength, the concentration, the reaction time etc. This leads to the question of dimensionality

3.5 Dimensionality

If I weigh myself in the morning, I have a measurement of a variable: my weight W. I could call this a single *datum*. If I do it five times I have a replicate *data set* of five points. This measurement is a *point* measurement, i.e. it is of dimension *zero* (0D). However, if I repeat the procedure every morning for a year, I get my weight as a *function* of time W(t). The measurement is now *one-dimensional* (1D) as W is a function of one *independent* variable t. W by contrast is called the *dependent* variable as it depends on time (and what I eat).

Often we measure 1D data like a spectrum consisting of the absorbance as a function of frequency (e.g. UV/VIS) or the scattered intensity as a function of scattering angle (e.g. XRD) or the heat capacity as a function of temperature (DSC).
Modern equipment easily produces multidimensional data. E.g. if we capture a series of X-ray diffraction images as a function of time with a CCD camera, each image has two independent variables, the pixel position x and y. In addition we have time. So the measured intensity is a 3D data set I as a function of x, y and t.

3.6 Accuracy

As we said measurements are often *indirect*. What I measure with my analytical balance is some electrical current (in mA) needed to keep my lever in balance. How does that relate to the mass (in mg) I put on my scale? We first need to establish that link by *calibration*. We first measure a couple of weights of *known* mass. From this we learn what value of the current corresponds to what mass. (Notice that this implies a unit conversion from milliamps to milligrams…)

Of course we cannot help making some error in the calibration, because it is itself a measurement. This means that our measurement will always have a limit on its ***accuracy***. (Need to know if that suits your purposes: is it accurate enough?).

3.7 Linearity

Ideally the relationship between the measured signal (say **A** in milliamps) and the meaning we give it through calibration (the quantity, say **q** in milligrams) should be ***linear***.

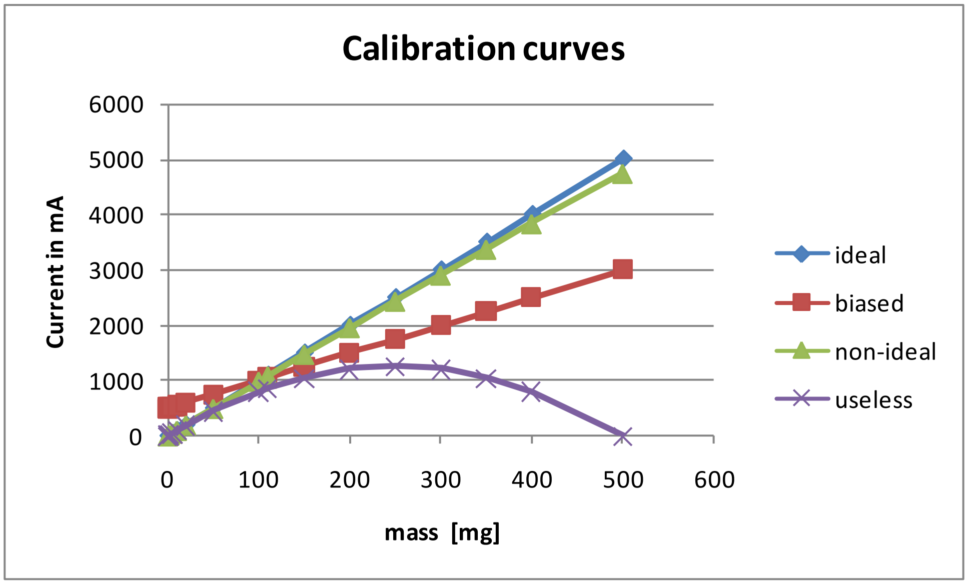
Signal = s. quantity+ b

$\mathbf{A}= s.\mathbf{q}+ b$

A small amount of non-linearity can at times be compensated for as long as the relationship remains at least ***monotonic***. If the latter is not guaranteed the measurement is pretty useless

3.8 Sensitivity

The slope s in the above equation or in general the slope of the calibration curve s=d**A**/d**q** is known as the ***sensitivity*** of the measurement. In general a large sensitivity is desirable because it means that we can measure small quantities with a reasonable precision. Note that if the calibration line is actually a curve that the sensitivity *varies* over the range of the line. If the curve is not monotonic the sensitivity becomes zero and then changes sign. This is why calibrations **must** be monotonic.



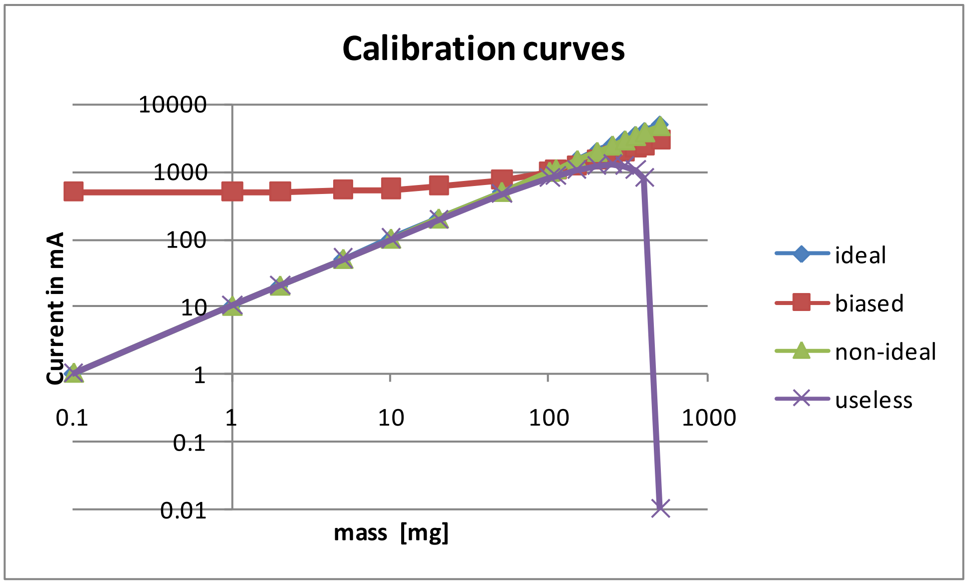
Notice that the ‘biased’ curve has smaller sensitivity (it is flatter) and also that it has a bias **b** because it does not pass through the origin. I.e. even if you put no weight on the scale you do need to pass a current.

3.9 Dynamic range

Often the linearity is only guaranteed over a certain limited range of values in **q**. That means we can only use the instrument in that range. E.g. the ‘non-ideal’ line could only be used up to about 300 mg before we would need to correct for non-linearity. The ‘useless’ scale could be used only below about m=50mg and even there we may have to correct.

Every instrument has a limited dynamic range between the smallest and the largest quantity we can measure with it. Don’t expect to go to a weighing station for trucks and weigh a pea on it or even yourself. And please do not put an elephant on an analytical balance. The weight will not be accurate, ever again…

A better way to look at this issue is to think on a double logarithmic scale. That too should be linear ideally. As you see the ‘ideal’ calibration above leads to a linear dynamic range from 0.1 up to about 500, i.e. about three and a half *decades*. The red *biased* curve looks terrible but that is because in the linear plot it does not go through the origin. It too is linear in the range 0.1-500 but there is a constant shift **b** (the bias) that needs to be dealt with before taking logarithms. The *non-ideal* curve starts to deviate a little as of 400, but it still has a dynamic range of at least three decades. The ‘useless’ curve is perhaps not as useless as it seems. It can still be used in the range 0.1 to about 50, i.e. it has one decade less.



Each instrument has its own *limited* dynamic range. Elephants are beyond the range of an analytical balance. Peas are below the one of a weighing station.

Many instruments have more than one scale that you can switch between with a button. Each of these scales is essentially another instrument with its own range and its own calibration and thus its own calibration and its own calibration errors. (This means it is not advisable to switch scales during an experimental run!)

3.10 Stability

Instruments tend to change over time. Calibrations have to be repeated off and on and changes may occur. The electrical grid is a source of much instability. Some equipment needs power stabilizers and/or good climate control in a room (humidity and temperature). All electronic equipment drifts to some extent. This is particularly important if the experiment takes a long time or is repeated many times for replication purposes.

3.11 Precision

This is *not* the same as accuracy as we have seen. It ties in with the dynamic range question because often the bottom of the dynamic range is an approximation of the random error in the measurement. E.g. if a scale of a balance goes down to 0.1 mg the random error in the measurement will be about 0.1-0.2 mg. This means that if you weigh something and it is 100mg the *relative error* will only be 0.1-0.2%, but if your weight is only 2mg the relative error will be as large as 5-10%. For many purposes the latter is not good enough. You need a more sensitive instrument with a more suitable dynamic range!

3.12 Destructiveness

If I determine the amount of iron in an organic sample by burning it and determining the iron oxide that is formed, I have measurement but my sample is destroyed in the process. If I can do the same by exposing it to X-rays and measure the fluorescence of Fe I also get a value but my sample is intact. (Or is there radiation damage?) Every technique has its own degree of destructiveness and its own requirements regarding how big the sample needs to be.

3.13 Data recording, processing, storage

The frontline of science is increasingly shifting to issues related to what happens to the data after they have been collected, because we live in the computer age. Many data are now stored electronically, but there are issues with durability of such records. Another problem is that instrument manufacturers will not always tell you what has already been done to the data the instrument provides. Some will not even allow you to access the raw data and you are forced to rely on the black box as given. A good scientist tries to prevent that at all cost: you remain responsible for the outcome!

3.14 How do I find out what I need to know about an instrument

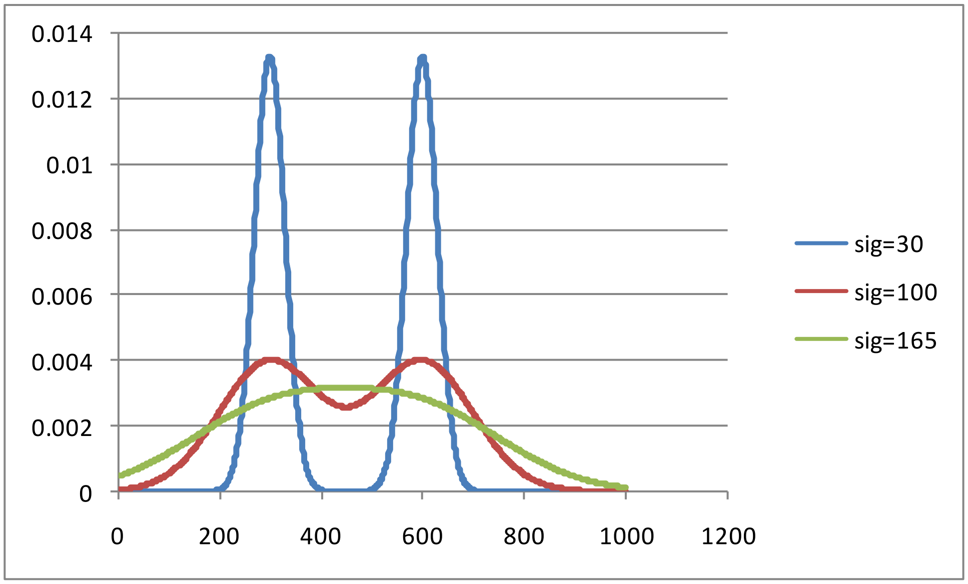
This is no trivial task. There a number of sources:

- books about the method
- people whose brain you can pick
- instrument manuals (hard to read and often missing)
- help files (often somewhat helpful)
- online sources
- publications that use the method
- your own common sense

Often people's brains are the easiest source, if treated gently and respectfully and not too often.

3.15 Resolution and broadening

This question is particularly important for measurements of 1D (spectra) and up. If a spectrum contains more than one signal they often appear as two peaks. Those two signals can only be distinguished properly if they do not overlap too much and that depends on how broad the peaks are.



The figure shows two Gaussian peaks, one at x=300, the other at x=600 (arbitrary units). The width is 30, 100 and 165 respectively. In the latter case it is not even clear that there are two signals because the resolution of the method is not enough to distinguish between them. A similar problem exists in 2D data (images).

3.16 What will the data tell me about my research topic, what do I learn from it? Who interprets what?

That is a 64 million dollar question…. Analytical Chemists usually do not worry about that part as much as they should. (*You wanted a number: here's your number! What is your account number again?*) Physical Chemists (and other researchers) often do not worry about that as much as they should until they have made the Analytical guy waste his time on a hard measurement. (*Why do I have to pay so much for a number that does not tell me anything?*)

This leaves managers with no scientific background to sort out the mess. (*Why are you guys wasting the company's money on useless measurements?*)

A bit of communication and preparative research will go a long way to avoid such situations. Another important point is that good science is typically done *by comparison*. It is often advisable to measure a blank, a known sample, something already studied etc. so that you can be sure what you are really after is not just a fluke.

4. Distributions of replicates

The basic idea underlying the statistical strategy against uncertainty is:

Just *repeat* your measurement

This ***replication*** strategy is as simple as it is counterintuitive. People (especially managers and undergrads) tend to consider doing exactly the same thing twice as a total waste of time or **du*plication***. The main message of statistics is that this is a fallacy, when dealing with uncertainty. Replication in fact yields useful information of **two** kinds: a better measurement value plus a value for the magnitude of the uncertainty. Provided the measurement procedure is repeated as exactly as possible, the latter is known as the **pure** error. It represents the **reproducibility** of your experiment. Interestingly, replication works no matter how complicated (or even arcane) the measurement procedure is. It can therefore be used to settle disputes about the appropriateness of a procedure.

Even though this is not always true, replicate values are often **assumed** to have a *normal* distribution, although many other random distributions do exist. The normal distribution is also known as the Gaussian or bell curve. The reason for its bold assumption is a theorem in statistics that states the following. Suppose your data are *not* normally distributed. Now take a lot of measurements and divide them in groups of say 5 or 6 (or any number n). Take the average over each group and collect the averages. The averages will typically have a distribution that looks more like a normal one than the original data. In other words for the number n going to infinity the distribution of the averages approaches normality. This is (part of) the **central limit** theorem.

In experiments we often average automatically over the whole sample volume, all the photons in the beam etc. A lot of data will display Gaussian behavior quite naturally. When in doubt we therefore assume normality as a default. Actually verifying what the distribution of our data is takes a lot of replicates and is often not feasible time wise.

The central limit theorem works for all but a few strange random distributions that do not *have* a mean. More importantly, the theorem does assume *complete* randomness; this means that data need to be *stable* for the theorem to work. This is a good reason why type IV (*in*complete randomness) errors are highly undesirable and best prevented.

(Stir well!). Below, we will assume that our data are stable at all times.

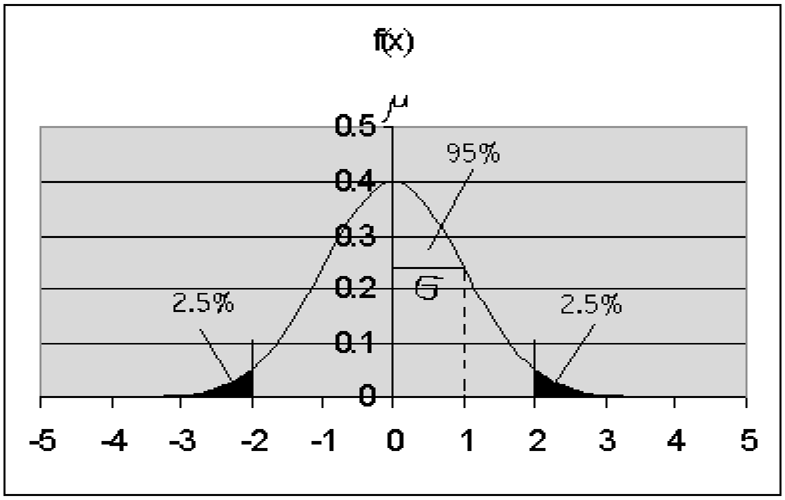


fig II

5. Prediction limits.

Looking at pure replicates is also known in statistics as **point estimation**. Say, we have a lot of time to spare and take many replicates. They will ideally follow the bell curve:

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

The measured quantity X is known as the ***variable***
The parameter μ is known as the ***mean*** and σ as the ***standard deviation***. . They are ***parameters*** The normal distribution is often written as N(μ,σ^2). Figure II shows the function f(x) for μ=0 and σ=1 (the *standard* normal curve), i.e. N(0,1).

The function f(x) is however *not* the probability itself but the **probability density.** The **probability** that a data point falls between ‘a’ and ‘b’ is found by integration over this function:

$$\Pr(a < X < b) = \int_a^b f(x)dx$$

Some facts are best memorized about the standard normal curve:

1) The function is symmetrical around x = μ, in other words there is no skewness.
2) The parameters μ and σ determine the position and the width of the curve respectively.

3) Each point ξ is known as a **percentile** of the distribution. E.g. $\xi = \mu$ is known as the 50th percentile (or ***median***) and the inflection points $\xi= \mu-\sigma$ and $\xi=\mu+\sigma$ are known as the 16th and the 84th percentiles, because:

$$\Pr(X < \xi) = \int_{-\infty}^{\mu-\sigma} f(x)dx \approx 0.16 \quad \Pr(X < \xi) = \int_{-\infty}^{\mu} f(x)dx = 0.5$$

$$\Pr(X < \xi) = \int_{-\infty}^{\mu+\sigma} f(x)dx \approx 0.84$$

4) The inflection points $x=\mu+\sigma$ and $x=\mu-\sigma$ ('half width') represent 68% limits.
5) Ca. 95% of the probability falls between the points: $\mu+\mathbf{2}\sigma$ and $\mu-\mathbf{2}\sigma$. (More exactly 1.96 rather than 2). These points are known as the 95% **prediction limits.**
6) In general a range $\mu+t\sigma$ to $\mu-t\sigma$ where t is some constant will have a certain probability associated with it

Note1: $x = \mu$ represents both the mean and the median for this *specific* distribution. (This is not so for skewed distributions, e.g.).
Note2: The percentiles represent a **one-sided** way of looking at probabilities. The prediction limits represent a **two-sided** case.

6. Estimates

When taking a sample of data both mean and standard deviation are (and *remain*) unknown, but we can obtain **estimates** for μ and σ from the sample. This can be done in a variety of ways. Let us start with the **safest** way: robust estimation, it involves *medians*: A median is found by first sorting your data by size and taking the middle one (or average the two middle ones if n=even). In Excel you can use the function = MEDIAN('range') where 'range' is e.g. D1:D15 containing your 15 replicates.

We estimate μ by **sample median (med):**

$$\hat{\mu}_{robust} = med(x_i)$$

Note that estimates are usually denoted with a ^ caret sign.

To calculate the **sample median absolute deviation (mad)** we subtract the estimate for μ from all our data points and use the =ABS(..) function to take the absolute value of each of the numbers that results. This produces the absolute *deviations* $|\delta|=|x_i\text{-med}|$ from the median. Lastly take another median, this time over the $|\delta|$ values. To obtain a good estimate for σ, we take 1.483 times the mad:

$$\hat{\sigma}_{robust} = 1.483\ mad(|\delta_i|)$$

The factor 1.483 can be derived from the shape of the bell curve.

The med and mad are called **robust** estimates, because they are largely insensitive even to *multiple* gross errors in your data. This makes them superior tools compared to e.g. the (older) Q-test in dealing with outliers. We will show how to do robust outlier rejection below. We can say that the removing the outliers "cleans" the data. It gets rid of some spurious points that are presumably just plain wrong for some reason that has to do with instrumental failure, environmental factors or operator error.

Once the data are clean, it is safe to use the Least Squares Estimates. They take optimal advantage of the information contained in your data, but only work properly, *if* the data are clean. The LS Estimates are rather well known: for μ is the ***sample* average,** defined as**:**

$$\hat{\mu}_{lse} = \bar{x} = \frac{\sum x_i}{n}$$

The LS Estimate for σ is the ***sample* standard deviation** defined as:

$$\hat{\sigma}_{lse} = s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum(\delta_i)^2}{df}}$$

In Excel we can use the functions =AVERAGE([range]) and =STDEV([range]) to find the LSE's.

Note that the **deviations** or **residuals** δ_i are once again obtained by subtracting the estimate for the mean μ (this time the LSE) from all your data. Note also that the **degrees of freedom (df)** are n-1 in this case. This lowering of the degrees of freedom results from the fact that we already used up one piece of information to estimate μ from the data.

7. The t-values and outlier rejection

As we saw earlier, we can predict that a certain percentage of our replicates will fall in the zone between the prediction limits at $x=\mu+t\sigma$ and $x=\mu-t\sigma$. We would know that probability, ***if*** we knew μ and σ, but unfortunately that is not so. All we really can do is work with estimates for μ and σ. Because the limited quality of the estimates introduces additional uncertainty, the relationship between the value of t and the probability will change: the ranges must become wider to achieve the same level of certainty. If we work with the usual LS estimates the boundaries become

$$\bar{x} - ts < x < \bar{x} + ts$$

where the value of the coefficient t now depends on the **degrees of freedom** as follows:

Df	1	2	3	4	5	9	19	10000
t(95%)	12.71	4.3	3.18	2.78	2.57	2.26	2.09	1.96
t(99%)	63.66	9.92	5.84	4.6	4.03	3.25	2.86	2.58

As you see, if you are interested in 95% prediction limits the t-value is only slightly above 2, unless you have very few replicates. The t(99%) values would give us 99% prediction limits.
In Excel the function =TINV(0.01,3) will give us the t(99%) value for df=3.

Example:

We measure the equilibrium pressure of a compound at 25^{0}C 5 times (df=4). The sample average is 34.56 Pa. The sample standard deviation s=0.12. Then the 95% prediction limits will be

$$34.56-(2.78)(0.12) < P < 34.56+(2.78)(0.12)$$
$$34.23 < P < 34.89.$$

We can say with 95% certainty that a new P value will fall in this range.

Of course, if our data are not clean we should not use LS estimates, but we can use the robust ones instead. In fact we can use robust prediction limits to **reject outliers,** because outliers do not belong to the normal distribution and will generally fall outside its prediction limits. To make extra sure the suspect points really do not belong to the normal distribution we typically take the t(99%) values (or even higher) and reject points, if they fall outside the range:

$$med - t(99\%)1.483mad < x < med + t(99\%)1.483mad$$

A convenient way to do this is to calculate ***studentized*** deviations t_i and deciding if they are larger than a certain critical value t(99%):

$$t_i = \frac{\delta}{\hat{\sigma}} = \frac{x_i - \hat{\mu}}{\hat{\sigma}} = \frac{x_i - \hat{\mu}_{robust}}{\hat{\sigma}_{robust}}$$

We should stress that rejected points should ideally also be explainable in terms of extraordinary events taking place during the experiment.

8. Confidence limits, improving precision and the accumulation game.

Prediction limits tell us where new replicate points are likely to fall. Apart from the outlier issue that is not really what we are after. What we really need is a way to say how good our point estimate for μ is. A good estimate for that is the **standard error s_e**

$$s_e = \frac{s}{\sqrt{n}}$$

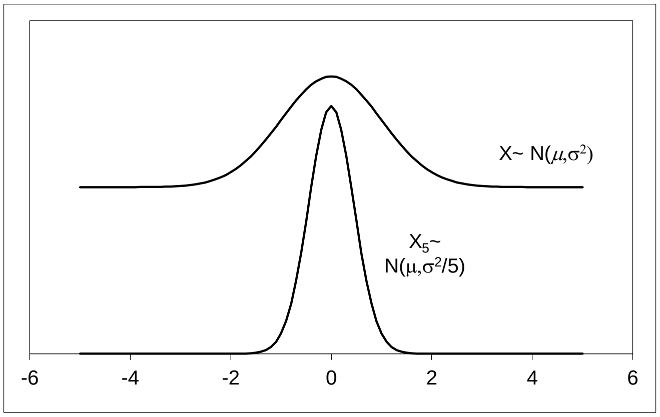


fig III

If the data points X have a normal distribution N(μ,σ^2) and we collect a great number of averages, e.g. over 5 points each, their distribution will have the same mean μ, but the variance is only σ^2/5 (see figure III). In general an average over n points is N(μ,σ^2/n), so that s_e is a good estimate for the uncertainty of the average.

We can define where we would expect the mean to fall by constructing ***confidence*** **limits:**

$$\bar{x} - ts_e < x < \bar{x} + ts_e$$

Example:
As above, we measure the pressure 5 times (n=5). The sample average is 34.56 Pa. The sample standard deviation s=0.12. The standard error s_e =0.12/√5 = 0.05 Then the 95% ***confidence*** limits will be

$$34.56-(2.78)(0.05) < \mu(P) < 34.56+ (2.78)(0.05)$$
$$34.39 < \mu(P) < 34.73.$$

We can say with 95% certainty that the true mean μ will fall in this range. If somebody else repeats the entire experiment, that's where they'll find the mean equilibrium pressure.

Obviously, the confidence limits (for the mean) are always narrower than the prediction limits (for a single replicate). Let's look at what happens if we take a lot of data, i.e. $n \rightarrow \infty$:

$$\lim_{n\to\infty} \bar{x} = \mu \qquad \lim_{n\to\infty} s = \sigma \qquad \lim_{n\to\infty} s_e = 0$$

If the number of data n is increased the prediction limits will change around a bit, but gradually stabilize to $\mu \pm 2\sigma$. The confidence limits, however, will shrink to zero, because of the extra factor $\sqrt{n}$. This is very important for science, because it provides a strategy:

To reduce uncertainty, simply increase n.

In other words, when fighting random errors: *accumulate* more replicates and average. This strategy is expensive, however: If we want to increase the precision by a factor of 1000, we have to take 1,000,000 times more data. If we let a robot do that for us, e.g. a computer that accumulates spectra, the data will improve with the square root of accumulation time

Warning 1: The accumulation and averaging strategy does not work, if the central limit does not work (unstable data (IV)) or if spikes (outliers (I)) pollute the data.

Warning 2: Prediction and confidence limits are often confused, as are s and s_e.

9. Rounding / the rule of 2 to 15

When you report data it is customary in the natural sciences to *round them off* in order to indicate how many **significant digits** a number possesses. This custom actually predates the emergence of statistics. Unfortunately, rounding can introduce errors of its own. Robust estimators for example are very sensitive to rounding errors. Rounding must therefore be done at the *very end* of your calculations. *To round off properly, use the standard error s_e as follows*.

If the first three non-zero digits of the s_e are ..155.. or larger, round s_e off to one digit
If the first three non-zero digits of the s_e are ..154.. or smaller, round s_e off to two digits
(Apart from the position of the decimal point, the result of this operation is always between 2 and 15, hence the *rule of 2 and 15*.)
Lastly, round off the mean value such that it ends in the same digit as the s_e.

An example: suppose you find an average value of 238.38746 with a standard error of 1.37673
The result is reported as: 238.4 (1.4) and has 4 significant digits.

Suppose you find an average value of 238.38746 with a standard error of 0.037673
The result is reported as: 238.39 (0.04) and has 5 significant digits.

It is customary to use scientific notation in multiples of 1000 (10^3, 10^6 etc), if possible incorporated into the units e.g. mM, μM, kJ, μm, nm etc.

E.g. 2,340,498,372 s_e : 449,302 becomes 2340.5 (0.4) 10^6

The factor √n in the formula for s_e implies that inappropriately rounding off one digit to the left obliterates 99% of your information. One digit to the right and you claim to have done 100 times more work than you actually did.
Warning1: A '±' notation like 238.4±1.4 usually refers to 95% confidence limits, not standard errors. This is ***not*** the same. In fact, it makes a difference of a factor of at least 2, because of the t-distribution values involved.

Warning2: Admittedly, the fact that this lab opts for the above procedure represents a somewhat arbitrary choice. Statisticians might even object that rounding is superfluous (and potentially harmful) once the standard error is known. For scientists, however, the important thing is that they do need to develop a discriminating eye for which digits are significant and which are not. Rounding is an excellent way to do just that.

10. Propagation of errors.

If we calculate a quantity Q from a number of measured values (A, B, C,..) each with their respective uncertainties (σ(A), σ(B), σ(C),..) we can calculate how the uncertainties *propagate* into the value of Q as follows
If

$$Q = f(x, y, z)$$

then

$$\sigma^2(Q) \cong \left(\frac{\partial f}{\partial x}\right)^2 \sigma^2(x) + \left(\frac{\partial f}{\partial y}\right)^2 \sigma^2(y) + \left(\frac{\partial f}{\partial z}\right)^2 \sigma^2(z)$$

Please note that we take derivatives versus '**x, y, z etc**' here. Each of the sigmas is defined as

$$\sigma(x) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N-1}}$$

The derivatives will be evaluated at the average

$$\left(\frac{\partial f}{\partial x}\right)_{x=\bar{x}}$$

This formula is an *approximation* that only holds true if the error sources σ(z),σ(y),σ(z), .. are independent (i.e. uncorrelated). If x and y represent e.g. the intercept and slope from the same regression this is generally not true. In addition we tacitly assume that we can replace the σ's by their estimates (i.e. s_e's)

If the function f(x,y,z,..) is of a relatively simple form, it can easily be shown (Try it!) that the formula reduces as follows.
Case 1: constant factors

$$Q = cx$$

$$\sigma(Q) = c\sigma(x)$$

The error in the function Q is a factor of c larger than the error in x.

Case 2: Addition and subtraction

$$Q = x + y \quad or\; x - y$$

$$\sigma^2(Q) = \sigma^2(x) + \sigma^2(y)$$

The quadratic nature of this addition is more ‘forgiving’ than a simple addition of σ’s would be. For example if $\sigma(x)=\sigma(y)= 1$ we get

$\sigma^2(Q)= 1^2 + 1^2 = 2$, so that $\sigma(Q)=\sqrt{2} = 1.4145$ (*not*: 2.)

It also means that there is usually only one dominating error source:

$\sigma^2(Q)= (100)^2 + (20)^2 + (10)^2 + (2)^2 + (1)^2$
$\sigma^2(Q)= 10505$
$\sigma(Q)\approx 102$

The first error term is really the only one that matters.
Note also that it makes no difference whether x and y are added or subtracted. This is no longer true, however, if the errors are correlated. Then we get:

$\sigma^2(Q)= \sigma^2(x)+\sigma^2(y) \pm 2\mathrm{covar}(x,y)$

The covariance term does depend on addition or subtraction. (In regressions, the covariance can be calculated from the off-diagonal elements of the same **(X’X)$^{-1}$** matrix that the standard errors are obtained from (see below), but unfortunately regression software seldom will actually display them.

Case 3: Multiplication and division:

$$Q = xy \quad or \qquad Q = \frac{x}{y}$$

Here we add the *relative* errors quadratically:

$$\frac{\sigma^2(Q)}{Q^2} = \frac{\sigma^2(x)}{x^2} + \frac{\sigma^2(y)}{y^2}$$

We can think of these in terms of the relative error, which is the important quantity in most reporting.

$$\frac{\sigma(Q)}{Q} = \sqrt{\frac{\sigma^2(x)}{x^2} + \frac{\sigma^2(y)}{y^2}}$$

The relative error in Q is the geometric mean of the relative error of x and y, which are its two variables.

Case 4. For other functions we may have to do a bit of algebra to find the appropriate formula. For example if Q = f(x) = x^n.

$$\frac{\partial f}{\partial x} = nx^{n-1}$$

$$\sigma^2(Q) = (nx^{n-1})^2\sigma^2(x)$$

$$\sigma(Q) = \frac{\sigma(x)nx^n}{x}$$

So the relative error is:

$$\frac{\sigma(Q)}{Q} = n\frac{\sigma(x)}{x}$$

In the above formulae when x and y appear one would use the best fit parameter in a non-linear fit of the average for each of these values in the derivative.

Propagation versus replication

With all the assumptions, it can be difficult to determine σ(Q) by propagation of errors, e.g. if we cannot estimate all of the error sources or it is not clear how they are correlated. Sometimes there may also be unknown (hidden) error contributions. In such cases it is better to go back to the basic strategy: *pure replication*. Simply repeat the entire procedure that leads to the determination of Q and average. The standard error will reveal the sum total of all error sources in its full ugliness. Sometimes both approaches can be used in comparison in order to identify hidden error sources.

11. Least Squares estimation.

We said that the average is the "Least Squares" estimate for a point, without explaining what that means. The term '*Least Squares*' refers to a very general estimation principle, useful for much more than point estimation. This is the general idea:

Let us start with a large set of (normal) replicate values x_i and make some wild guess for the mean μ of the distribution, say: *a*. We can now compute the deviations δ_i = x_i –*a*. Some of these may be positive, some negative and that is not useful for our purposes. Instead we take the *squares* and add them up: This produces the *sum of squares*

$$SS = \sum_i \delta_i^2 = \sum_i (x_i - a)^2$$

Obviously SS is a function of our guess *a*. If we are far off the mark, SS will be large, regardless of the direction of our blunder. A good guess on the other hand will result in a small SS value. Thus, the ('loss') function SS provides us with a way to optimize our guess. We look for that *a* value that produces the smallest 'loss'. This a_{min} value is called the **least squares estimate**. (LSE). We can easily find the LSE value for *a* by putting the derivative d(SS)/d*a* =0

We find:

$$\frac{dSS}{da} = \frac{d\sum_i (x_i - a)^2}{da} = -2\sum_i (x_i - a) = 0$$

$$-\sum_i x_i + \sum_i a = 0$$

$$\sum_i x_i = na$$

$$a_{LSE} = \bar{x} = \frac{\sum_i x_i}{n}$$

In other words, the sample average indeed minimizes the sum of squares. The median by contrast does not have this nice property.

Notice also that the sample *standard deviation* s and the sample *variance* s^2 are computed from the minimal value of the sum squares and are LSE's therefore:

$$s^2 = \frac{SS_{\min}}{df} = \mathrm{var}$$

$$s = \sqrt{\frac{SS_{\min}}{df}} = stdev$$

where df refers to the number of degrees of freedom.

12. Estimating lines instead of points

So far we did point estimation, on pure replicates i.e. the model we used to try and fit our data was that of a simple constant:

Model : measurement = constant + random error

$$x_i = a + \varepsilon_i$$

We typically assume that ε has a normal distribution $N(0,\sigma_\varepsilon)$. We now extend the least squares principle to more complicated models, e.g. a straight line fit through our data:

$$y_i = a + bx_i + \varepsilon_i$$

Note that our data are no longer (just) pure replicates, because we vary the value of x, but we can play the same game as before. We take some wild guess at slope *b* and intercept *a*, calculate deviations and SS. To minimize SS we must now take **two** derivatives (dSS/d*a* and dSS/d*b*) and put them zero simultaneously. Try it on a rainy Sunday if you like, but the math gets pretty messy. Luckily, matrix notation is a great help when dealing with this kind of problem. We can write the above model as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdot\cdot \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot\cdot & \cdot\cdot \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot\cdot \\ \varepsilon_n \end{pmatrix}$$

Or:

$$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{x} \cdot \boldsymbol{b} + \boldsymbol{\varepsilon}$$

The **X** matrix records for what values of x we choose to take a measurement, e.g. the temperature values we *set* our thermostat to. We generally assume that there is no error in these *set points* or *independent* variables. **Y** contains the *dependent* variable, the *measured* values of e.g. the absorption at the chosen temperatures.
The matrix $\boldsymbol{\varepsilon}$ contains the random errors that we assume to be normal ($N(0,\sigma_\varepsilon^2)$)
The matrix $\boldsymbol{\beta}$ contains the parameters we wish to estimate, the slope *b* and intercept *a* of our line.

Finding the LSE for $\boldsymbol{\beta}$ can be done quite elegantly in matrix notation. It takes a page of ugly math, but it can be shown that putting the derivatives of SS equal zero results in the same formula as above *minus* the $\boldsymbol{\varepsilon}$ term.:

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta}$$

Notice that the only unknowns left are in β. The **X** and **Y** matrices are known because they are either *set* or *measured.* Solving for β now requires some simple matrix algebra:

$$\boldsymbol{X^T Y} = \boldsymbol{X^T X} \cdot \boldsymbol{\beta}$$

$$(\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y} = \boldsymbol{\beta_{LSE}}$$

Interestingly, the latter (**regression**) formula minimizes the sum of squares for a great many different models: point, line, circle, parabola of polynomial. It is one of the most powerful equations in statistics. Let's first look at a simple straight line.

To construct the X matrix we take the derivative with respect to x of both of the variables in the equation for a line.

$$\boldsymbol{y} = \frac{\boldsymbol{\partial}}{\boldsymbol{\partial a}} \boldsymbol{a} + \frac{\boldsymbol{\partial}}{\boldsymbol{\partial b}} \boldsymbol{x} \cdot \boldsymbol{b}$$

The LINEST function

The easiest way of doing regressions in EXCEL: using the LINEST function. For a simple straight line you make a column of independent variables (X-range), e.g. the concentrations you made up and a column of your measured dependent variables (Y-range). Then you select a range of 5x2 cells and type: =LINEST(*Y-range,X-range*,1,1).

LINEST is an *array* function. Such functions need to be activated using Ctrl+Shift+Enter.

Excel gives you the following numbers

slope	intercept
$\mathbf{s_e}$ of slope(n)	$\mathbf{s_e}$ of intercept
R^2	***RMSE***
F*	df
SS(reg)*	SS(resid)*

1. Your slope and intercept appear on the first row.
2. Their s_e (not: std. dev!) values on the second row are the ones to be used to round off the parameter values on the first row.
3. The value of the Root Mean Square Error represents the estimate for σ_ε , the quality of a single data point.(Remember: The matrix **ε is supposed** contains the random errors that we assume to be normal ($N(0,\sigma_\varepsilon^2)$).

4. R^2 is a number between 0 and 1 that measures how closely your model correlates with your data.
5. The value of df (the degrees of freedom) will equal n-2, because you have determined two pieces of information out of n data points
6. *The statistics F,SS(reg) and SS(resid) will not be used in this lab

If you apply regression to a model with more than one independent variable, you need to select more columns and you get something like

slope(n)	slope(n-1)	….	slope(2)	slope(1)	intercept
s_e of slope(n)	**s_e** of slope(n-1)	….	**s_e** of slope(2)	**s_e** of slope(1)	**s_e** of intercept
R^2	***RMSE***				
F	df				
SS(reg)	SS(resid)				

Notice that the parameters run from right to left!

The limitations / assumptions of the regression formula are:

1) The model must always be written as **Y**=f(**X**,**β**) + **ε**, i.e. the dependent variable on the left, the independents and parameters on the right
2) The independent variables in **X** (*set points)* are assumed error free. In practice, any error in them will end up in the dependent variable **Y**.
3) The inverse **$(X'X)^{-1}$** must exist, i.e. the determinant det**(X'X)** cannot be zero.
4) The data must be clean. (Remember: LSE is *not* robust)
5) The degrees of freedom must be one or more (#data > #parameters+1).
6) The errors **ε** are normally distributed as N(0, σ_ε^2)
7) The size of the random error **ε** **(**σ_ε) is the *same* for all data. (The errors are *homoscedastic*)
8) The model function f(**X,β)** must be *linear* in the parameters a,b,c (in **β**)

The dangers of regression

It is important that all these requirements are fulfilled, because if they are not the numbers you get from, say the LINEST table cannot be trusted. This is why we will learn a number of checks that you ***must*** perform to be sure that you have applied regression *safely* later. Without these checks regression is *dangerous*, hence:

> *There are lies, damned lies and statistics!!*

Let's first look at an important application of statistics and *assume* all requirements are met.

13. Calibration

Calibration is the principle method in eliminating ***systematic errors***, or at least *trying to do so*. The principle is quite simple. If you are not sure whether your measurement produces an ***accurate*** measurement, you can check that by making a measurement on a sample for which the outcome is ***well known***. Such a sample is called a ***calibration standard.***

Balances e.g. are usually calibrated by putting certified weights of known value, 1 g, 10 g 100 g, 1 kg on them and comparing the response R of the instrument to the value V it should be.

Regression plays an important role in calibration procedures. Ideally, the instrumental response R should be a ***linear*** function of the variable V that is to be measured:

$$\boldsymbol{R = b + s \cdot V + \varepsilon}$$

Non-linear responses can at times not be avoided, but there are a number of rather unpleasant side effects associated with non-linearity.

The two parameters s and b of a calibration line have names. The slope s is known as the ***sensitivity*** of the measurement and the intercept term is known as the ***bias***, although this term is also used in a more general sense for any systematic (as opposed to random) effect that afflicts your outcome. To obtain high quality data the sensitivity should be large compared to both the random error and any residual bias remaining after calibration.

To obtain a calibrated value for an unknown sample, we follow the following procedure:

1. We measure a set of R_{cal} values for a number of standards with known values V_{cal}
2. We construct a regression line, i.e. determine the best s and b values.
3. We measure a $R_{unknown}$ for the unknown sample
4. We calculate V_{unk} = (R_{unk}- b)/s

Of course the calibrated value V_{unk} is subject to error. In fact its value is subject to two kinds of error:

1. The random error due to the measurement: ε_{unk}/s
2. Whatever residual systematic calibration error is left despite our calibration

We should note immediately that the first is a random error and can therefore be reduced by replication of the measurement of the unknown.

The second however will remain the same as long as we do not do a better calibration job, which means that it presents a systematic bias in the value of V_{unk}. This means we better do a good job on calibration; otherwise all subsequent measurements will be junk.
The calibration error can result from error either in the sensitivity s or the bias b. The latter does not depend on the value of V, the former does. In fact it equals zero in the center of gravity of the calibration data and diverges as a cone from there.

The calibration error can statistically be represented by drawing the 95% **confidence** limits around the calibration line. These limits form the two branches of a hyperbolic function. The total error (calibration + random measurement of the unknown) are given by the **prediction** limits. They also form a -somewhat wider- set of hyperbolic branches. The two sets of hyperbolas are given by:

$$Y_{confidence} = sX + b \pm t\ RMSE\ \sqrt{\frac{1}{N} + \frac{N(X - \bar{X})^2}{N\sum X^2 - (\sum X)^2}}$$

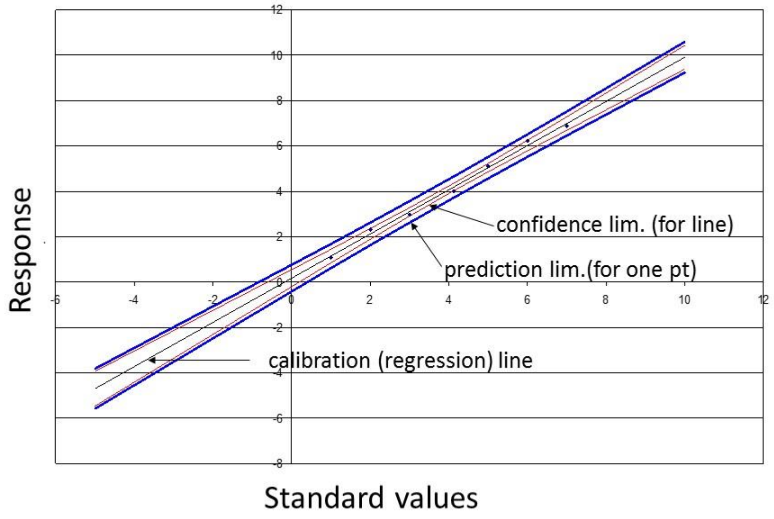
$$Y_{prediction} = sX + b \pm t\ RMSE\ \sqrt{\frac{1}{n} + \frac{1}{N} + \frac{N(X - \bar{X})^2}{N\sum X^2 - (\sum X)^2}}$$

$$D = N\sum X^2 - (\sum X)^2$$

The quantity D is actually the *determinant* of the **(X^TX)** matrix.
The value of N represents the number of calibration standards used. The value of t(p,df) represents the appropriate t-value at the given number of degrees of freedom (N-2) and the confidence level desired (usually 95% of p=0.05). The standard values are denoted by X. The center of the calibration set is given by the average of all X values. This represents the narrowest point where the error in the slope does not contribute.
If we take n replicate measurements of the unknown, the (outer) hyperbola becomes gradually narrower, eventually converging to the (inner) confidence limit as the 1/n term goes to zero. The inner limits represent the error due to calibration and can only be improved by doing a better calibration job. It is useful to note a few things

- If we continue to use the same calibration the calibration error is a *systematic* one (bias). The calibration error only becomes a random one if we apply many different calibrations.
- Subject to one particular calibration the slope error induces a positive bias on one side of the center, a negative one on the other side.
- The calibration line is best used around its center, because far outside this range the (systematic!) slope error starts to dominate. Extrapolations are not attractive.

Limits of calibration



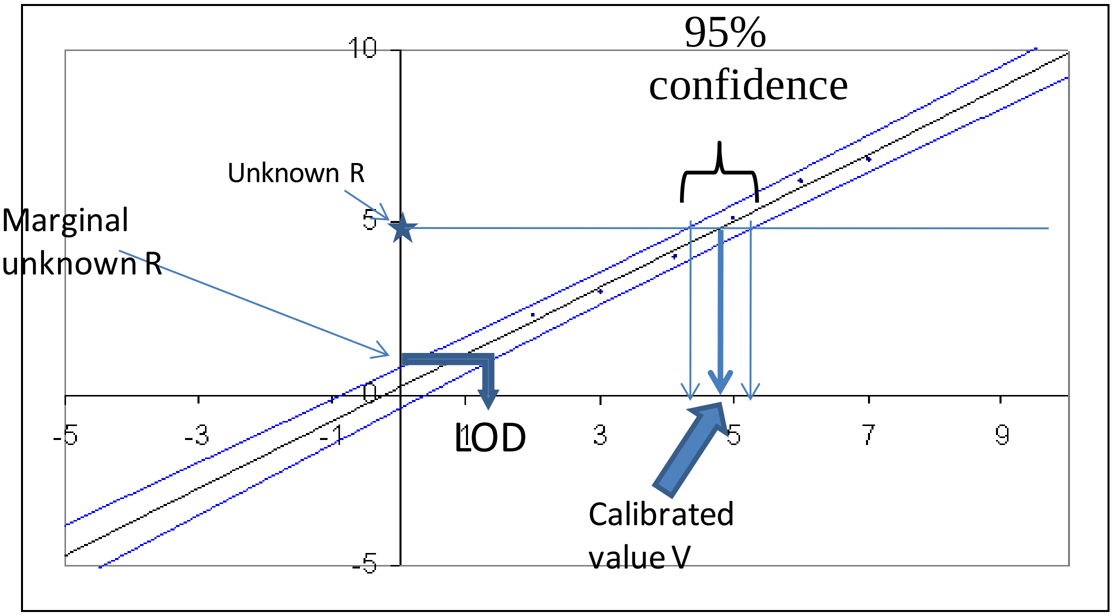
14. Reading back: the inversion problem

As we saw above we obtained a calibrated value for the unknown by taking the *inverse* function of the calibration line using the best estimates for s and b:

$$R_{cal} = s \cdot V_{cal} + b + \varepsilon$$

$$V_{unk} = \frac{R_{meas} - b}{s} + \varepsilon_{random} + \varepsilon_{calibration}$$

Graphically we can represent that as 'reading back' a value on the Y-axis (the measured R values) towards the X-axis (representing the calibrated V-values). Let us assume that the random error in each individual measurement is the same for all measurements (calibration and unknown alike). We can predict with say 95% confidence that a subsequent experiment of an unknown substance must fall within the outer hyperbolas. Since we know the response R (on the Y-axis) we can use the corresponding V values on the X-axis as confidence limits for our unknown V value.



The outer prediction limits ('trumpets') around the calibration line fix the LOD and the confidence limits of the calibrated value V

(If the vertical R value is *marginal,* the calibrated value obtained in the read back process is at the **limit of detection** because its confidence limits now include the value V=0.)

To obtain the complete error in the calibrated values we should therefore take the inverse of the hyperbolic prediction limits. Unfortunately, the analytical inverse of that function is too unwieldy to be of practical use. Also when we 'read back' the values on the upper and lower 95% prediction limits we get an *asymmetric* set of limits around our calibrated value. This even means that strictly speaking a calibrated V_{unk} value does *not* have a normal error distribution even if the measured R_{meas} does. This effect is generally ignored in science. Provided the error in the calibration slope is small and the calibration is used not too far from the center of the calibration data this is not too serious a problem. That does impose two requirements on calibration lines:

- The calibrated range should be larger than the range in which the calibration is used
- The calibration should be tight and linear (a high R-value is required with many nines like 0.9999998. R= 0.998 could well represent a *bad* calibration)
- We need to assume that the random error in the calibration experiment and the measurement of the unknown is identical
- Of course: all the assumptions and requirement of regression statistics must be met. **(examine your residuals!!)**

Under these conditions we can approximate the inverse of the hyperbolic functions with a ***symmetrical*** function:

$$\widehat{s_e} \approx \frac{RMSE}{|slope|} * \sqrt{1 + \frac{n.V^2}{D} + \frac{\sum(x_i{}^2)}{D} - \frac{2V\sum x_i}{D}}$$

- Once again D represents the determinant of the **(X^TX)** matrix as above. For a simple straight line that is equal $n.\Sigma x_i^2-(\Sigma x_i)^2$.
- The number of calibration samples is given as n,
- The x_i values are the independently chosen values, e.g. of the concentrations of the *standards*.
- V is the calibrated value of the unknown sample obtained by the read back process as shown above.

We are assuming there is only one such measurement available. If the unknown is measured k times and the average value used read back, the 1 in the formula is replaced by 1/k:

$$\widehat{s_e} \approx \frac{RMSE}{|slope|} * \sqrt{\frac{1}{k} + \frac{n.V^2}{D} + \frac{\sum(x_i{}^2)}{D} - \frac{2V\sum x_i}{D}}$$

Depending on its chosen level, the confidence limits should be calculated by multiplying with the appropriate t-value. Typically you take the t(df, 0.95) value for that. The degrees of freedom would be n-2 for a simple straight line.

15. Verdicts in the court of science: type I and type II error.

Scientific data are sometimes used to decide matters of life and death. Forensic science is a good example of that. The reason that confidence limits are so important is closely related to what happens in a court of law, where every verdict can be in error in two rather opposite ways:

	declared guilty	declared not-guilty
murderer	OK	type II error
no murderer	type I error	OK

Obviously it is OK to declare a murderer 'guilty' and someone who has done no wrong 'not -guilty', so the two possibilities on the diagonal of our table are not a problem. If the quality of the evidence is not so good however, the off-diagonal situations become more likely and no they are *not* the identical.

- When we commit a type II error we let a murderer loose on society to inflict more harm.
- When we commit a type I error we send an innocent man to death row ***and*** let the true murderer run free.

Obviously, neither is desirable but because our evidence is never 100% clear, we are confronted with an unpleasant choice. The pickier we are about the evidence the less likely we send an innocent man to his death, but the more likely we release a murderer. We must decide what is worse. Usually people decide to err on the side of caution: the judicial system is based on the premise that we would release a murderer in case of reasonable doubt or lack of evidence.

In science we basically make the same kind of choice. When we speak of a 95% confidence limit, we say that we are 95% confident that we are not committing a **type I** error (α) and we say nothing about the type II error rate (β). If we want to be pickier we can resort to the 99% confidence limits to further reduce our chances of committing a type I error. However, this *increases* the likelihood of committing a type II error, *unless* we improve the quality of our data. The latter is the only way to reduce the probability of error regardless of type. After the fact of measurement, all we can do is trade off and by being pickier we *increase* type II error

Warning
There is a lot of confusion on this point. People even advocate taking a 6-sigma (very picky) threshold to reduce both error types. This is really a fallacy of logic.

16. Limits of detection

This discussion brings us to a vital question in all Analytical chemistry, the limit of detection. It pertains to the question whether we can or cannot demonstrate the presence of a certain effect.

X=0 relative to. 95% limits	inside	Outside
means:	no effect demonstrated	effect demonstrated

No effect demonstrated can mean two things:

1. There is no effect
2. There is an effect, but the measurement is not good enough to demonstrate it.

The probability that #2 occurs is β. It depends on the quality of the data

Effect demonstrated means just that.

The probability that we falsely make this claim is $\alpha \leq 0.05$, depending on how far outside the confidence limits zero is.

A special point on the calibration graph is therefore the intersection of the upper hyperbola (of the prediction limits) with the Y-axis. It is known as the ***limit of detection*** (LOD). When we do a measurement of an unknown that produces a response right at this value and we use the calibration graph to read back we get a lower confidence limit for our V_{unk} value that is precisely zero. This means we must let the accused off the hook because our numbers tell us that it is quite possible that we did not measure anything. Then again: it could also be so that the V value is actually larger than zero and the man is a crook. The only way to reduce this type II error is to get better data. Maybe a replicate measurement under the same calibration will do the trick, but we might actually have to improve on our calibration line, particularly if we are far away from the center of the calibration design and the standard error of the slope is not so good.

17. Propagation of errors versus confidence limits

The propagation of errors as we have introduced it above cuts quite a few corners. One is that it completely ignores the effects of the degrees of freedom on the resulting confidence limits. For the proper reporting of analytical data this may not be sufficient, because legislation related to quality control (ISO, GLP etc.) often requires that they be reported to terms of confidence limits, not in terms of an estimated standard error. If the number of points n in the sample is large this roughly means that we are pickier by a **factor of two** (1.96 actually) but as sample sizes are often not large enough the difference gets more extreme.

Confidence limits are usually written using the symbol ±. If we have a simple *large* set of replicate values and we are using 95% confidence limits the following two notations are equivalent, because for n is large the t-value for 95% confidence is about 2:

120(3) corresponds 120 ± **6** at 95% confidence.

In the more common case that we have a sample of -say- four replicate values and that we are taking a simple average, the degrees of freedom df=3 and the t value is a little over 3. In that case:

120(3) corresponds 120 ± **9** at 95% confidence.

Of course we do assume that our replicates are normally distributed, so that we can use a t-distribution to link the variance (as estimated by standard error) to the confidence limits.

Unfortunately, error propagation that takes into account degrees of freedom quickly becomes pretty complicated. Often we need to determine the difference Δ between two rather small sets of replicates. If the data quality is equal (same variance) for the two sets, we can *pool the residuals (or: variances)*, i.e. subtract the respective average from each point

$$\sigma_{pooled} = \sqrt{\frac{\sum_{set\ 1}(x_i - \overline{x_1}) + \sum_{set\ 2}(x_i - \overline{x_2})}{n_1 + n_2 - 2}} = \sqrt{\frac{\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)}{df}}$$

We can compute the resulting confidence level from:

$$t_{calculated} = \frac{\overline{x_1} - \overline{x_2}}{\sigma_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

If the variances are *not* equal, e.g. because the data were taken on two different instruments by different people, things get already quite a bit more complicated:

$$t_{calculated} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The effective number of degrees of freedom is:

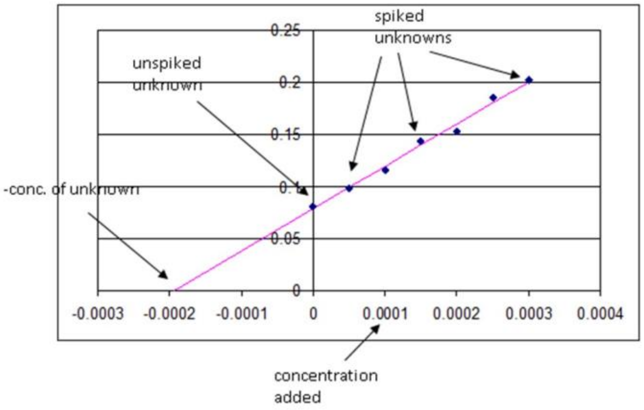
$$df = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\left(\frac{\sigma_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{\sigma_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

It will be clear that is errors are to be propagated from more complicated functionality than a simple subtraction it becomes impossible to keep track of the degrees of freedom and all we can do is calculate an *approximate* standard error, use the 2/15 method and forget about trying to convert that into ‘exact’ 95% confidence limits…

18. Standard addition

There is an interesting combination of measurement and calibration into one and the same procedure that involves regression. It is called ***standard addition***. Often unknown samples contain many compounds besides the one that is being analyzed (the analyte A) and this can affect the sensitivity of its measurement. These are called *matrix effects.* A good example is a fluorescence measurement, where part of the emitted light is partly absorbed by some other molecule present in the matrix.

One way to circumvent this problem is to add known quantities of A to the unknown sample (this is called 'spiking') and measure a number of spiked samples. We then plot the measured response against the *added* concentration of A. A regression line is constructed which will intersect the X-axis at a negative value -x. This point represents a hypothetical sample with [A]=0, so that x represents the actual concentration in the original mixture. The slope equals the sensitivity of the measurement *under the conditions of the sample itself.* We do need to assume that the matrix effects affect only the sensitivity, i.e. the slope and not the intercept. If the matrix *contributes* to the response the method does not work.



As the figure shows standard addition is an *extrapolation* method. It easily loses precision if the unknown concentration is large compared to the additions. The confidence limits of the measurement can be found from the intersection of the confidence (not: prediction!) limits of the regression line with the X-axis.

A standard addition line is essentially a calibration line in the matrix. We must arbitrarily set the value x of the unknown equal to zero. As shown in the figure as long as the spiked values have a greater x value (and therefore a greater y value) they will make a calibration line. When you extrapolate that line back to the negative x axis you have determined the difference between the unknown concentration and zero. Therefore, the concentration is equal –x (which is positive since x is negative). Note that the slope of the line is equal to:

$$slope = \frac{intercept}{-x} = \frac{rise}{run}$$

Therefore,

$$x = \frac{intercept}{-slope}$$

To find the approximate confidence levels we can use a similar formula as above, but the first term under the root symbol is missing, because we are dealing with the 'inner trumpets':

$$\widehat{s_e} \approx \frac{RMSE}{|slope|} * \sqrt{\frac{n.X^2}{D} + \frac{\sum(x_i{}^2)}{D} - \frac{2X\sum x_i}{D}}$$

What LINEST does not give you

Linest is wonderful to quickly to a regression but it does not give you any graphical output and that makes its application a little *dangerous,* because we have no check on the validity of our regression assumptions.

19. An example of multilinear regression in Excel.

Any model that is ***linear*** in the ***parameters*** (a, b, c, d,..) is a linear model.
That implies that ***quadratic*** or ***polynomial*** models like

$Y= a + b.x +b.x^2 + \varepsilon$
$Y= a + b.x +b.x^2 + c.x^3 + d.x^4 + \varepsilon$

are in fact *linear* models.

We can even go to more than one independent variable (***multivariate*** models):

$Y= a + b.x_1 + c.x_2 + \varepsilon$

This is still a linear model, because we do not have anything like a^2 or a/b in the model.

Suppose we wish to fit our data to the following model:

$Y= a + b.x + c.x^2 + d.\ln(x) + \varepsilon$

This model is *linear*, because it is linear in the parameters a,b,c,d. The fact that it contains x^2 and ln(x) is irrelevant, because the x values are known.

The data look like this:

X	=X^2	=ln(X)	Y values measured
0.1	0.01	-2.30259	1.04
0.4	0.16	-0.91629	1.44
0.5	0.25	-0.69315	1.46
0.8	0.64	-0.22314	1.59
1	1	0	1.53
1.2	1.44	0.182322	1.51
1.8	3.24	0.587787	1.45
2.1	4.41	0.741937	1.36
3.2	10.24	1.163151	1.16
3.5	12.25	1.252763	1.15

4	16	1.386294	1.04
4.1	16.81	1.410987	1.04
4.8	23.04	1.568616	0.9
5	25	1.609438	0.9
6	36	1.791759	0.81
7.3	53.29	1.987874	0.72
8.9	79.21	2.186051	0.81
10.2	104.04	2.322388	0.96

The 2nd and 3rd columns were calculated from the first. The 4th column represents the measurements (**Y**). Note that the first 3 columns represent the matrix **X,** except for a column of ones (for the intercept) that Excel adds automatically.

Go to Tools/data analysis and look for Regression on the pop-up menu. (If data analysis is not on the Tools menu, invoke add ins first and request the analysis pack)
In the Regression pop-up menu I specified D2:D19 as my dependent Y variables and A2:C19 (i.e. the first 3 columns) as independent ones. I checked the confidence level option and filled in 99% (I like to be picky). I told Excel where to put the output and opted for Residuals and Residual Plots. They are quite useful as shown below. This produces a lot of output. We will only explain the most important items. Look at the 1st table:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.997393
R Square	0.994794
Adjusted R Square	0.993678
Standard Error (*RMSE)*	0.022767
Observations	18

The Standard "*Error*" (of the Estimate) is a somewhat confusing name Excel uses for:

$$s_{\varepsilon} = \sqrt{\frac{SS_{\min}}{df}} = \sqrt{\frac{SSR}{df}} = RMSE$$

It is better known as the RMSE (root mean square error). It represents the spread of the data around the line (σ_{ε}, the 'noise level'), comparable to what the standard *deviation* does for point estimation. In regression you can skip tedious replication and still get a RMSE, but ***only if*** the model is correct. You don't always know that. This is a fundamental difference with point estimation. There you do pure replication and so you *know* that the (point) model is correct. It should be noted that df=18-4=14 in our example, because we estimate 4 parameters a,b,c and d. Please not that for regression to work df has to remain ≥ 1.

The R (correlation) statistics give you an overall idea of how well the model fits the data. R^2 is between zero (no correlation) and one (perfect correlation). R can in principle be between –1 and +1, because in general correlation can be negative. In regressions it is positive, because it describes how well the Y_i values correlate with the predicted ones $\hat{Y}_i$ ($=a+b.x_i+c.x_i^2+d.\ln(x_i)$). The disadvantages of R are that

- it tends to give numbers quite close to one even for mediocre fits. It is like a purity value: a metal that is 99.9% pure is not very pure: 99.9999% is much better.
- It is only useful if compared to another R value
- R also has a severely non-normal distribution. Therefore Fisher proposed a better measure:

$$z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

Fisher's z runs from 0 to infinity, as R runs from zero to one. It also has a normal distribution, in good approximation. For quality control comparisons it is a better tool than R. Warning: in Excel you can put a trend line in a graph which shows R, but not the much more important indicators like standard errors of the parameters.

We'll skip the second table that deals with ANOVA (Analysis of variance). It tells us how the sum of squares is explained by the various terms in the model.

The 3rd table is the most interesting for our purposes:

		Standard					*Lower*	*Upper*
	Coefficients	*Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*99.0%*	*99.0%*
Intercept	2.012576	0.020535	98.00724	2.9E-21	1.968533	2.056619	1.951446	2.073705
X Variable 1	-0.50492	0.01396	-36.1696	3.14E-15	-0.53487	-0.47498	-0.54648	-0.46337
X Variable 2	0.030357	0.000998	30.40967	3.46E-14	0.028216	0.032498	0.027385	0.033329
X Variable 3	0.402964	0.014554	27.68834	1.26E-13	0.37175	0.434179	0.359641	0.446288

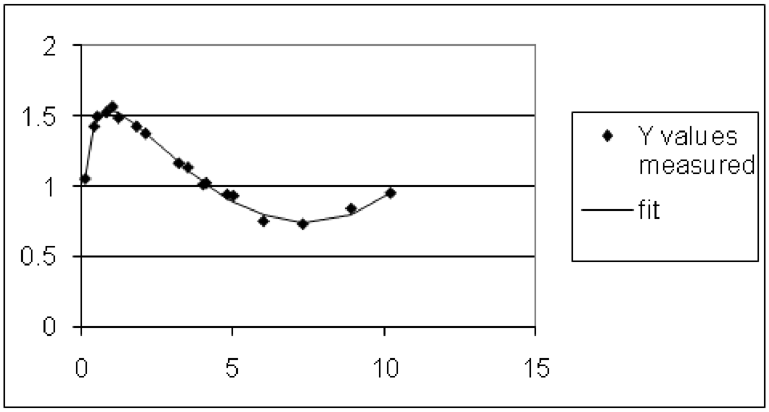
20. Analysis of fits and residuals

In order to establish that the model is not grossly wrong and gives us lies, damned lies and statistics we need to analyze the outcome of our calculation. There are two graphical ways to that:

a) Look at the fit.
b) Look at the residuals.

The 'coefficients' in the table above are the estimated values for the parameters a, b, c, d. They tell me that the best fit to my data is given by: $\hat{Y}=2.01 –0.505.x +0.304.x^2 +0.403.\ln(x)$

I could use this formula to calculate the best fit and produce a ***line-fit plot*** like this

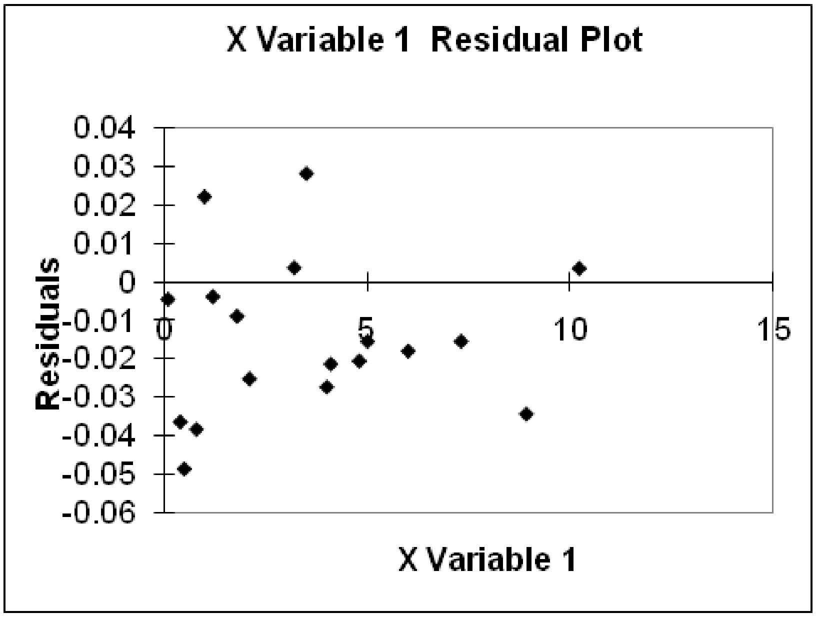


The rest of the 3rd table tells me how good (or bad) my estimates are. Column 3 gives the standard errors of the four parameters. They are calculated from the diagonal elements of the **(X'X)$^{-1}$** matrix and the RMSE. *They are the equivalents of the standard error s_e in point estimation, and are to be used for the application of the rule of 2 and 15.* They too dwindle when you add data to your set, although in a more complicated way that depends on where you pick your points (i.e. on **X**).

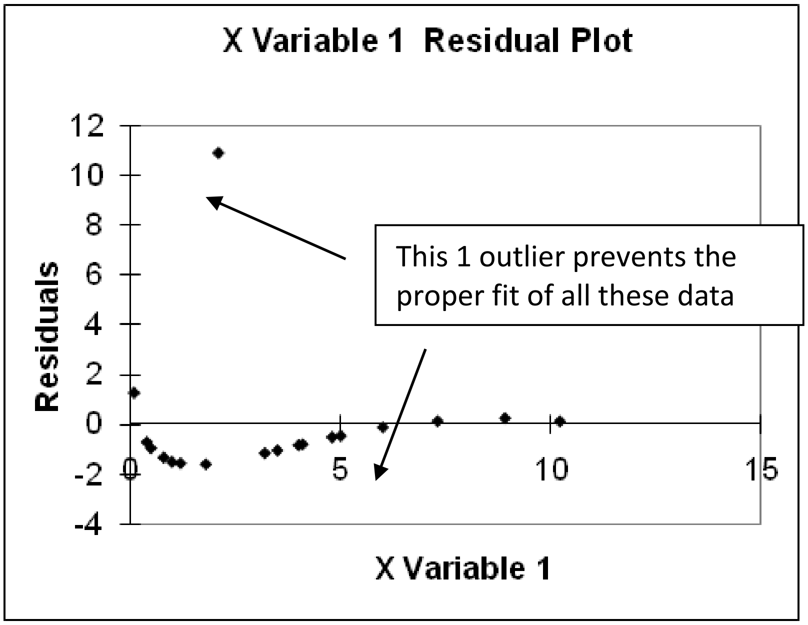
However, before you report anything, you should also look at the 95% confidence limits. Make sure that the value of zero is **not** included in the confidence limits of any of your parameters. If it is, it means that your data do not contain sufficient information to estimate that particular parameter and so your model is **wrong** (and so are all the numbers you obtain from it). You should try again by removing the 'dead' parameter from your model. (Alternatively, check that the P-values are smaller than 0.05 for all parameters).

The parameter values and their standard errors are only safe to report if you are sure you used the right model. There are number of ways to check that, the simplest one is to calculate the residuals δ_i by subtracting the fit from the data and have good look at them.

Ideally, the δ's should just represent the pure error $\boldsymbol{\varepsilon}$, i.e. they should look like random noise only.
The plot below is typically what you want: no pattern at all



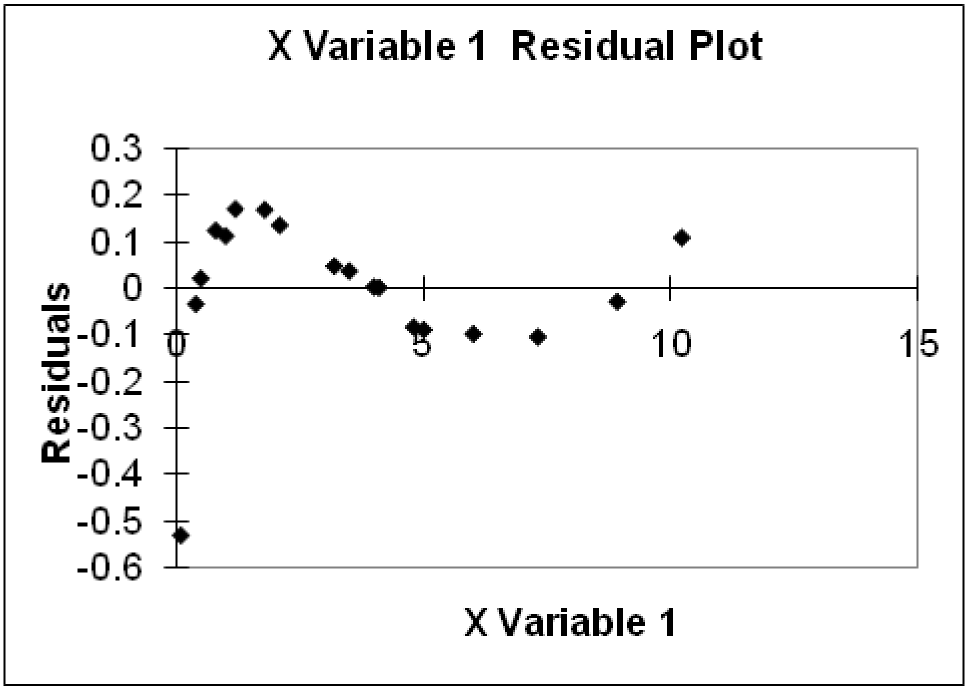
But sometimes you might have an outlier:



This 1 outlier prevents the proper fit of all these data

You *must* remove the outlier and redo the fit! Regressions tend to give ***complete nonsense***, if you leave in an outlier. (See topic 13 below).

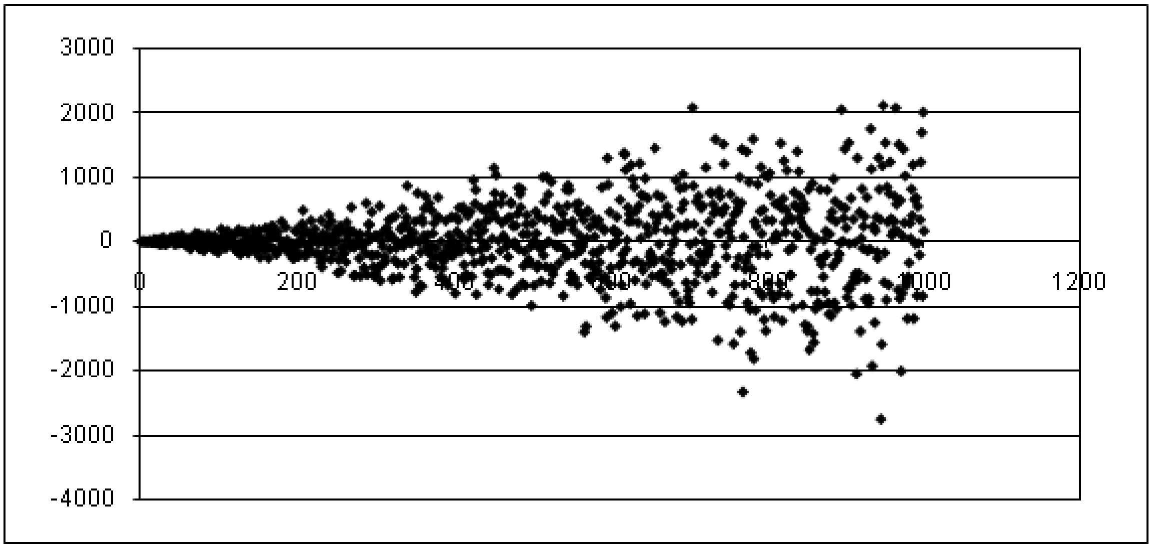
A wrong model might produce residuals like this:



There is a clear pattern here. You probably omitted a term in your model.
Note the different scales in the 3 above graphs: I used *simulated* data with the same random error **ε** for all of them. (I shifted the decimal point in one point to create the outlier and left out the logarithmic term to create a wrong model.)

As the Standard 'Error' of the Estimate value in the 1st table is computed from the residuals, its value only reflects the true random 'pure' error ε_{pure}, **if** there are no other error sources (left) in your data (see 1st plot). This means that if your model has *lack of fit* (as in plot 2 and 3, note the scale), the δ's will be much bigger than 'pure' random error ε_{pure}.(see plot 1). The values of s_e for the parameters a,b,c,d will be inflated accordingly. In principle the inflated size of the Standard 'Error' is a sure sign there is something wrong and comparing it to the pure random error ε_{pure} should tell you so. Unfortunately you don't know its value unless you determine ε_{pure} independently: Go back to the basic strategy and take *pure* replicates (more than one y value *without* varying the value of x) and take a local (average and) standard deviation.

Lastly, heteroscedastic residuals could look like this:



The data quality gets worse for higher x.

Reporting regression results

After you have convinced yourself that there is nothing untoward with your regression results you can summarize the first two columns of table 3 in a table in the results section of your report. The above data would give:

Parameter	Value
a	2.01(2)
b	-0.505(14)
c	3036(10) 10^{-5}
d	0.4039(15)

21. Robust regression, outlier rejection

Unfortunately regressions can be affected by one or more outliers quite seriously. Even the residuals do not always identify the culprit(s). Although the residuals do not look random in the presence of an outlier, the outlier is ***not*** necessarily the biggest residual.

Linear Regressions are Least Squares (LS) estimates. Even more than the sample average and the sample standard deviation, they are very sensitive to gross errors. The regression line may pass through the outliers rather than the 'good' points and the RMSE is usually grossly inflated, so it cannot be used to studentize residuals. Rousseeuw has proposed a solution to this problem. In Least Squares we minimize the Sum of Squares $\Sigma\delta^2$. That is the same thing as minimizing the *average* square deviation $\Sigma\delta^2/n$

Instead we could minimize the *median* square deviation med(δ^2). Unfortunately there is no simple analytical way to do that, but we can use an algorithm:

- Take two points from your data and define a straight line through them
- Take the residuals δ from this line and calculate med(δ^2)
- Do this for all possible (or a sufficient number of) lines
- Take the line with the lowest med(δ^2): the Least Median Square (LMS) line
- Use the med(δ^2) of the LMS line to estimate σ_ε (RMedSE instead of RMSE)
- Studentize the residuals of the LMS line with the RMedSE
- Remove all points for which the studentized residual δ/RMedSE is too large
- Do a Least Square regression on what is left

This whole procedure is known as Reweighted Least Square regression (RLS).

The robust version of RMSE is obtained from: RMedSe = $1.483.(1+ 5/n).\sqrt{}$med(δ^2)
Uncertainties in the parameters are not calculated for the LMS line, but that is not a problem, because in the end we revert back to a regular regression on the clean data, which does give us that information (the RLS output).

The Excel macro

A spreadsheet containing an Excel macro that performs the LMS/RLS procedure can be found on the CD. It suffices that this file is open together with the spreadsheet you are working on. Go to View/ Toolbars and make sure that the Visual Basic toolbar is visible. Select a suitable range with data, click on the blue arrow on the Visual Basic toolbar, select the RLSmacro and run it. Suitable ranges are either a single column with replicates (1D), two neighboring columns with X and Y values (2D), or three columns with X_1,X_2 and Y values (3D)(In this case X_2 can also be X^2).
The macro will ask you for a confidence level, 99% or higher is recommended. It will ask you whether to include or exclude an intercept in the 2D and 3D models and it will suggest a number of lines to try. For very large data sets you can lower this number if the macro takes too long.
The macro will generate a new sheet every time you run it. Both the LMS output (in blue) and the RLS output (in yellow) can be found in it. The latter is in LINEST format in the 2D and 3D case and can be used as the final result. The rejected points are shown in red. Two plots are generated: a line fit plot with both the LMS and the RLS fit and a plot of the LMS-residuals (studentized using RMedSE) and the RLS residuals (studentized using the RRMSE the Reweighted root mean square error as it comes out of the RLS LINEST results)

Warning: If your data do not fit for other reasons than outliers, e.g. there is a term missing in the model and you run it through the macro, it may well decide to throw away a lot of perfectly fine data points as 'outliers' to get rid of the lack of fit.

22. Nonlinear regression and refinement

The regression formula **β = (X'X)$^{-1}$X'Y** can also be applied to find LS estimates for the parameters in **β** if the model is not linear in these parameters. An example:

$$Y = \frac{abX_1}{bX_2 + 1} + \varepsilon$$

This model has two independent variables (X_1 and X_2), but more importantly it has two parameters (a and b). It is not possible to write the model as **Y=X.β + ε** in this case but we can fill a **J** matrix with the *derivatives* of the model versus the parameters, taken in each data point:

$$J = \begin{pmatrix} \left[\frac{\partial Y}{\partial a}\right]_1 & \left[\frac{\partial Y}{\partial b}\right]_1 \\ \left[\frac{\partial Y}{\partial a}\right]_2 & \left[\frac{\partial Y}{\partial b}\right]_2 \\ \vdots & \vdots \\ \left[\frac{\partial Y}{\partial a}\right]_n & \left[\frac{\partial Y}{\partial b}\right]_n \end{pmatrix}$$

The derivatives can e.g. be calculated from : $\left[\frac{\partial Y}{\partial a}\right]_1 = \frac{ab}{1+bX_2(1)}$ The quantity $X_2(1)$ is simply the first value we have chosen for variable X_2. For 'a' and 'b' however, we need to first make an *initial guess.* Then, if we calculate **(J'J)$^{-1}$J'Y** we get a better set of parameters with a lower sum of squares after regression SSR. If we repeat the process (*iterate)* the outcome usually converges, i.e. applying the formula will not produce much change in the parameters any more.
There is a problem though: the SSR is like a landscape with many valleys and ridges. If we start with wrong initial values, we may end up in a valley that is not the true minimum of the landscape. It is also possible that the matrix inversion may not work if the derivatives in **J** are ill defined.

In Excel itself non-linear regression is not possible, but the CD used to install the software contains a number of add-ins. One of these is the Solver. It can be used to do non-linear regressions. Unfortunately it does not produce standard errors of the optimized parameters, but the CD-ROM in the back of the Excel for Chemists book contains a macro (under Chapter 12) that remedies that problem. The macro estimates the derivatives in **J** by changing each parameter p_i a little bit to $p_i + \delta$ and calculating $[Y(p_1,p_2,\ldots\mathbf{p_i+\delta},\ldots,p_n)- Y(p_1,p_2,\ldots\mathbf{p_i},\ldots,p_n)]/\delta$ in each data point. As long as $\delta<<p$ this finite difference is a good estimate for the derivatives needed.

23. Some added remarks on Excel

We should be grateful that Excel makes a lot of useful graphical and statistical procedures readily accessible to us in the lab. It certainly is a useful educational tool for data exploration and reduction. It should be noted, however, that statisticians have found some serious flaws in the program. The matrix inversion routine used to calculate $\mathbf{(X'X)^{-1}}$ tends to produce nonsense if the determinant of this matrix is close to zero and does so *without warning*. This may happen if two variables are strongly correlated or more than one of them is 'dead'. Another problem is the use of data with very large number of significant digits; this can lead to round off errors in standard deviations etc. For serious publication work it is recommended to use better software.

The Regression tool under data analysis actually makes use of the same =LINEST array function, so that the results are identical. The tool has the advantage that it easily produces residual plots. The LINEST function on the other hand is 'alive'.

24. References

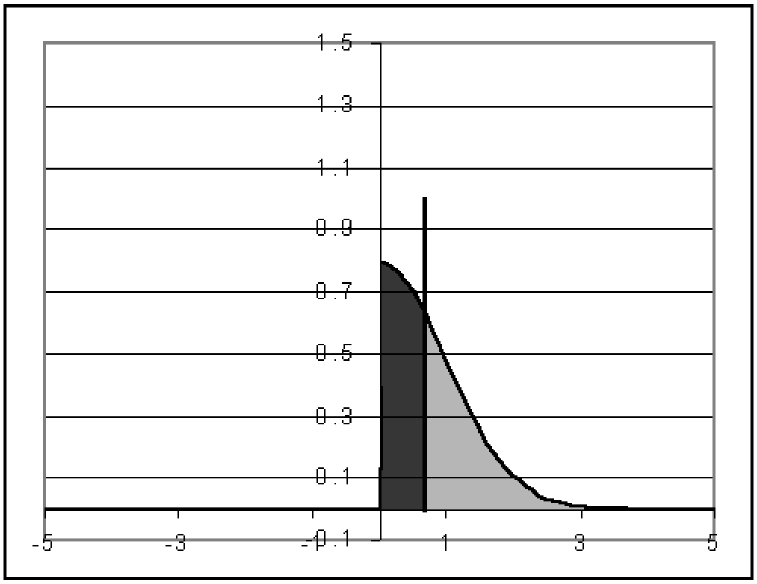
Although its description of regression is somewhat sparse, a useful book on the use of the statistical functions of Excel, is:

Beverly J Dretzke; Kenneth A. Heilman; *Statistics with Microsoft ® Excel*, Prentice Hall, Upper Saddle River, NJ 07458, 1998.

A more general description of Excel that includes a chapter on regression as well as on the use of array functions, macro's etc., is:

E. Joseph Billo; *Excel for Chemists,* Wiley-VCH, New York, 2001

Appendix I



The deviations δ from the median should ideally have a normal distribution around μ=0. The absolute deviations |δ| will therefore have a half-normal distribution h(x) where h(x)=0 for x<0 and h(x)=2.f(x) for x>0.

The median (i.e. 50th percentile) of this half normal distribution has the same x value as the 75th percentile of the normal distribution, i.e. that x=ξ value for which

$$\Pr(X < \xi) = \int_{-\infty}^{\xi} f(x)dx = 0.75$$

This point ξ lies at μ+σ/1.483, somewhat below the point ξ= μ+σ (the 84th percentile).

Appendix II

Other distributions

The normal distribution is but one of many different kinds of random distributions, some of which do occur in nature. A good example is the *exponential* distribution. In contrast to the normal distribution it does not have a square in the exponent. The Boltzmann distribution is an example of such a distribution. Another example is the *uniform* distribution. Its graphical representation resembles a rectangle rather than a bell curve. Rounding errors have such a distribution and most computer-generated random numbers do. The =RAND( ) function in Excel e.g. is uniform. To get a normally distributed number use =NORMSINV(RAND( )).) Lastly, events that have a rare occurrence, e.g. radioactive pulses in a Geiger counter, usually have a *Poisson* distribution. (This is just a small sample).

.

Tutorial I: Rounding and significance.

Rule of 2-and-15.

A. The rule of 2 and 15:

1. Compute the average and its standard error $s_e = s/\sqrt{n}$
2. Look at the first three non-zero digits of the standard error: 00xyz...
3. If xyz ≥ 155 round off to obtain **one** digit, else to obtain **two** digits
4. The resulting digits should always be between 2 and 15
5. Make sure that the average is rounded off, such that it ends in the same digit
6. Put the one or two digits of the standard error in (brackets) behind the average.

Problems: Round off the following numbers in accordance with the rule of 2/15
Note: the standard error s_e is computed from the standard deviation s as: $s_e = s/\sqrt{n}$

1. You compute an average of X= 15.837465 with a standard error of s_e= 0.14398.
2. You compute an average of X= 15.837465 with a standard error of s_e= 1.59398.
3. You compute an average of X= 158.37465 with a standard error of s_e= 15.9398.
4. You compute an average of X= 158.374 with a standard deviation of s= 1.29398; you have n=100 data points.
5. You compute an average of X= 158.374 with a standard deviation of s= 1.59398; you have n=10000 data points.
6. Is the end result better for 4) or for 5)?

B. Significant digits.

A number that is properly rounded as 213.6(3) is said to have four significant digits. In fact the major purpose of rounding is to determine which digits are significant or not. Suppose the original number was **213.5***92347.* The last digit (7) is entirely a matter of chance: a new measurement could produce any of 10 possible values. The first digit however (2) will not change in the next experiment. Therefore the digits **213.6.** are said to be significant, and the digits ...*92347* insignificant. In science we want to base our conclusion on significant data. That is why this difference is very important.

7. In each of the problems 1-6 indicate how many significant figures there are.

C. Relative errors

A relative error is the error expressed as a fraction (or percentage) of the averaged value.

For example, a number like 213.6(3) has a relative error of 0.3/213.6 = 0.0014 or 0.14%.

8. In each of the problems 1-6 calculate the relative error
9. What relation is there between significance and relative error?
10. Consider the following 2 data sets. Determine the average, the standard deviation and the standard error. Report the average in 2/15 format. Report the number of significant digits and the relative error.

Data A	4.38	4.81	5.28	5.27	5.75	5.51	6.5	6.33
Data B	15.9	16.89	16.73	14	15.54	14.34	13.92	14.02

D. Differences and propagation.

To make a cogent scientific argument we often have to demonstrate that two values we have measured are really different, i.e. they should differ *significantly.*

That means that we need to know what the error is in Δ= A-B, given s_e(A) and s_e(B). We can estimate that using ***variances.*** (i.e. the square of s_e):

$$[s_e(\Delta)]^2 = [s_e(A)]^2 + [s_e(B)]^2$$

This is called ***error propagation.*** We are cutting a few corners:

- we assume that the error $\sigma = s_e$ (We neglect degree of freedom issues)
- we assume that the data for A and B are independent (no correlation)

11. Using the values of problem 10 calculate your best estimate for Δ by subtracting the averages for A and B.
12. Estimate $s_e(\Delta)$ using the above equation
13. Estimate $s_e(\Delta)$ by simply adding the standard errors.
14. Report the value for Δ, (properly rounded) and calculate the relative errors with both estimates for $s_e(\Delta)$.
15. What is more forgiving, the linear addition or the quadratic one?

For a quantity Σ = A +B the propagation goes along the same lines.

16. Suppose D = 0.121(12) and E= - 0.129(14) . Is the sum Σ = D + E significant?

E. Products and ratios

For products (P = A*B) and ratios (R=A/B) we have to work with the squares of the ***relative error***.

$$[s_e(P)/P]^2 = [s_e(A)/A]^2 + [s_e(B)/B]^2$$

17. Report the properly rounded values of P and R based on the data of problem 10

F. Other functions

For any other function F = f(A,B,....) we must add ***weighted variances.*** The weight factors can be approximated by taking the derivatives $[\partial f/\partial A]$, $[\partial f/\partial B]$ etc.

$$[s_e(F)]^2 = |\partial F/\partial A|^2 [s_e(A)]^2 + |\partial F/\partial B|^2 [s_e(B)]^2 + ...$$

18. Report the value of 1/A and 1/B (see 10).
19. Assuming that the values in problem 10 are angles in degrees, report the value of Z= sin(A)/sin(B)

The equations for Δ,Σ,P and Q can all be derived from the general equation for F.

20. Derive the equation for the product P

Laboratory Experiments

In this phase students will do an experiment in the 608 lab one afternoon and a computer tutorial on the other. The experiments will be done in pairs or groups of three and the next time the student comes to 608 he/she will put a report on last period's experiment in the report tray.

Which experiments the student will do will be determined in lab by means of a rotation scheme.

Practice. Balances and calibration of volumetric tools

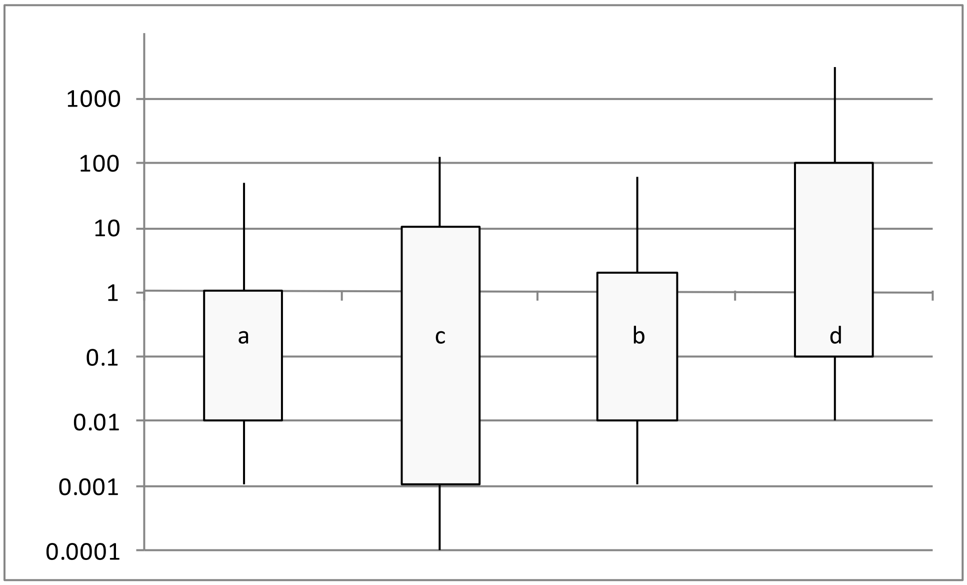
Balances are a very basic and very valuable tool in any chemistry lab and any chemist must understand their use, their proper treatment, their strengths and their vulnerabilities.

Dynamic ranges

The precision of a modern balance can be quite astounding, but their *dynamic range* (the ratio between the smallest and largest weight they can handle) is limited. This is why they come in different ranges. Some measure in tons, others in kg, yet others in grams, in mg or even in micrograms. In the lab we generally use two kinds of instruments:

- gram range (ca. 1000g down to ca .01 gram)
- milligram range (ca. 50g down to ca 0.1 mg)

The graph shows the dynamic ranges of the balances in Dabney 608 on a logarithmic scale in grams. It is up to you to *choose* the appropriate device for your measurement. Ideally your weight should be in the middle of the dynamic range



If a mass of 1 gram is weighed on a balance that goes down to .01 grams the *relative error* = .01/1 = 1%
If the same mass is weighed on a balance going down to .1 mg the relative error is .0001/1 = 0.01%
At the top of the dynamic range the precision gets even better, but then the accuracy is often no longer guaranteed.

Vulnerabilities

1. Air flow can affect measurements in the mg range
2. Buoyancy can affect measurements in the sub-mg range
3. Vibrations can affect measurements in the mg range
4. If the balance is not properly leveled this can affect the weight.
5. Suddenly dropping a heavy object of severe overloads can destroy the balance
6. uneven spread of the weight may lead to erroneous values
7. Leaving spilled chemicals on the balance corrodes it (at all weight ranges). ***CLEAN THEM UP!!!!***
8. To avoid corrosion, always use paper / aluminum / ceramic /glass containers or weighing boats.
9. Hygroscopic materials gain weight if left open. Either weigh quickly or weigh them in a flask you can close off

Introduction

Balances need calibration with standard weights, but once calibrated, they are used to calibrate other equipment. The purpose of this lab is (1) to calibrate two volumetric tools (a pipette, and an automatic pipette) and (2) to determine the density of a liquid using a burette. Ethanol or other liquids will be provided by your TA.

You will then use the Reweighted Least Squares technique (1D) to reject any outliers and calculate the average and RMSE of the volume delivered by the automatic pipette (micropipette) and the burette, respectively.

You will also use the Reweighted Least Squares technique (2D) to reject any outliers and calculate the density of the liquid from the burette measurements.

The dimension of measurement of the pipettes and graduated cylinder is *volume* rather than mass. This is why we need to know the *density* to be able to convert between the two:

mass [g] = volume [ml] * density [g/ml]

The density of a liquid depends a little on temperature. Therefore, before the measurement use a thermometer to *measure the temperature* of the liquid you are using and look up the density at that temperature. Compare the density determined from the burette measurements to the literature density of the liquid under the same temperature condition.

Experiment.

Caution: If you spill anything: clean up immediately, just use Kim wipes. Please do not ruin balances by leaving them dirty or wet! Don't push hard on a balance.

The basic procedure is essentially the same for the three volumetric instruments. Make sure you first properly rinse the volumetric tool with DI water then condition it with the liquid. There is DI a tap in the corner of Dabney 608 by the white board.

Write down all information about your volumetric tool: what volume can it measure? Do you know the tolerance (uncertainty)?

Select a suitable balance. Which one is suitable depends on the quantities you are about to work with, so you need to compare the weight you expect to get with the dynamic range of the balance. Take as sensitive a balance as you can afford under the circumstances. Use an appropriately sized beaker as the receptacle for weighing.

The measurement should be done using ten aliquots should be dispensed from your device consecutively.

There are few different strategies to weigh a sample:

1. You can either let the liquid accumulate in your weighing pan and tare each time you wish to dispense more
2. You can discard the liquid each time and then tare.
3. You can let the weight accumulate and not tare at all (keeping careful records of the values at each accumulation)

The last measurement may result in having to deal with greater quantities. (Is your balance capable of dealing with that?)
Make sure you write down the weight of the dispensed aliquot (or the total weight in strategy 3) in your lab notebook.

For the a**utomatic *pipette*** (micropipette)
Set the variable volume pipette to each volumetric value and write down which value you are measuring. For the calibration step, use strategy 2. above to record at least 5 different readings of the mass of the liquid delivered by your selected pipet at each volume. For the measurement of the density measure each of ten volumes ranging from 0.1 to 2.0 mL. Measure each volume and mass in triplicate.

For the ***burette***:
The burette will only be used for the density measurement using strategy 1. above. Fill your burette and record the initial volume (if you do not start at exactly zero, you will have to subtract this initial reading from all the subsequent readings). Using weighing strategy 3. above, tare a clean dry receiver flask and deliver approximately 1 ml aliquots (record the exact volume- bottom of the meniscus) and record the mass on the analytical balance after each delivery. Do not tare between measurements

Data work-up

Put your data in a spreadsheet. Make sure to convert the liquid masses to volumes with the right density for the pipettes. If you did accumulation compute the difference between a weight and the previous to determine the weight of each aliquot.

Determine the Relative and Systematic Error of the Micropipette

Use the Reweighted Least Squares technique to reject any outliers and calculate the average volume of the pipette and its uncertainty. This is like a calibration step, except that you may use a single volume, e.g. 1 mL for the large micropipette or 200 uL for the smaller variant and weigh a given volume of DI water at least three times. Based on the 95% confidence limit, what is the relative error? Relative error refers to the ratio of the 95% confidence limit to the nominal value used and may be converted to percent.

Also determine the ***systematic*** error of the pipette (the nominal value written on it minus your final average). Is this bias significant? Is there a tolerance given on the pipette? Is the bias within that tolerance?

Pipettes are often used in the first step of a dilution. Suppose you would use your pipette to dilute an aliquot of a 0.10000 molar solution to a volume of 1 liter in a volumetric flask. Calculate the final concentration of that solution *and* its precision (2/15 please) assuming that the molarity 0.10000 and final volume of 1000 ml are exact (error free). What is the relative error in the concentration?

Determine the Density of a Liquid Using the Micropipette and Burette

Graph ten measured weights against ten different volume values ranging from 0.1 to 2.0 mL that you read off from 1.) the micropipette and 2.) the burette. Do a regression (RLS if necessary). The slope value should equal the density of the liquid at your temperature. Use the trumpets Excel spreadsheet to calculate the slope, RMSE and 95% confidence limit for a line of regression. A. Is the difference in the measured and tabulated values of the density within precision? B. What is the RMSE of the slope of the line of regression? C. What is the t-value? D. Using the graphical method draw a horizontal line at a mass of 1 gram and determine the 95% confidence limit.

General questions

The accumulation method and the taring method each have advantages and disadvantages. The taring method limits the total weight, and that can be an advantage. It also has a drawback. Which one?

Laboratory Report

This laboratory is an exercise in measurement and calibration. Find articles in the Journal of Chemical Education or journals in Statistics, Physics or other quantitative sciences that describe methods for calibration and error determination. Write the paper as a general introduction to these methods using the categories in the rubric, Abstract, Introduction, Experimental, Results and Discussion. Discuss relative errors in measurement of density based on your data and any articles. Also, consider the errors in measurements of pH, heats of combustion, concentration (based on absorbance for example) or any other commonly measured quantity. Is there a consensus about how large a relative error is commonly observed or is acceptable for a scientific measurement?

Lab 1. Adiabatic and reversible compression of a gas

This laboratory requires a laboratory report.

Introduction

An ***adiabatic*** change is a change for which heat is prevented to exchange with the surroundings. This can be accomplished by insulating the system so that it cannot exchange heat with the surroundings. However, another way to accomplish an adiabatic change is to drive the process so quickly that there is no time for heat exchange. Sometimes, we therefore think of a sudden process as an adiabatic process. This is, in fact, the procedure we will use in this experiment.

A ***reversible*** change is a change that follows path consisting of states of rest. The idea that a process is a reversible is an idealization. In practice, a real process may be far from reversible. Clearly when a process is driven very quickly it is not likely to be reversible. Nonetheless, the model we will use assumes a reversible process. We will see how good that assumption is using the data in this experiment.

The initial and final states of an adiabatic and reversible volume change of an ideal gas can be determined by the First Law of Thermodynamics and this will be discussed extensively in CH433. The resulting equation for such a change relates volume and pressure:

$$P_i V_i^{\gamma} = P_f V_f^{\gamma} \tag{1}$$

Where the exponent:

$$\gamma = \frac{\overline{C_p}}{\overline{C_v}} \tag{2}$$

That means that γ is the ratio of the ***heat capacities*** at constant pressure and constant volume. You should review the derivation of this experiment starting from the first law of thermodynamics, ΔU = q + w with the condition q = 0. In addition, for an ideal gas there is an expression

$$\overline{C_p} = \overline{C_v} + R \tag{3}$$

From Eqns. (2) and (3), the molar heat capacity $\bar{C_p}$ can be determined from a measurement of γ.

Heat capacity is related to the number of ***accessible degrees of freedom*** system, such as rotations, vibrations and translations, because if the system has more degrees available we must spend more heat to make it go up one degree.

- Monatomic gases like Argon have only translational degrees of freedom. Statistical thermodynamics shows that $\bar{C}_v$ = 3/2R for such a gas
- Diatomic gases have 2 degenerate rotational degrees and one vibrational one. The latter is not active at room temperature, but the rotational ones add a term R to the $\bar{C}_v$.
- Triatomic molecules have an even greater heat capacity that is also somewhat temperature dependent

All gases in this experiment are sufficiently dilute that we can consider them ***ideal***

EXPERIMENT

The Pasco TD 8565 Adiabatic Gas Law Apparatus allows the rapid and simultaneous measurement of P, V and T for a gas sample confined in the cylinder of the apparatus. Two hundred data points can be stored on a computer for a compression lasting 200 milliseconds, fast enough to assume that the process is adiabatic, but slow enough that we can assume that the conditions are uniform throughout the gas sample, i.e. that the compression is also reversible.

It is advisable to first do a dummy run using whatever gas happens to be in the cylinder (air probably) to familiarize yourself with the PASCO TD 8565

Make sure the cylinder device is powered

- Start DataStudio
- Create experiment: the interface should be visible and active.
- On the picture of the interface on the “experiment setup” window there are three smaller yellow rings on the right
- Click on the left most of the 3 smaller rings
- Scroll pop up to bottom and opt for “voltage sensor”
- Click on the middle one
- Opt for “pressure sensor (absolute)”
- On the experiment setup window set the sample frequency to 500 Hz

NOTE: We will not use the temperature sensor

Calibration run

- On the left bottom window “Displays” click on graphs and make a graph for the voltage and one for the pressure.
- Open one of the gas taps and push the piston all the way down
- Click “start” on the computer panel

- Read off the position of the bottom of the piston on the centimeter scale on the cylinder. Bring the piston up stepwise 1 cm at the time. Write down how far the bottom of the piston is each time. The voltage curve should look like a stair case.
- Under file go to export data and export the voltage staircase to a suitable location and filename (create a directory for yourself).
- The exported files can be opened in excel.

Measurement run

- Click “sampling options” and then “delayed start” opt for data measurement = voltage and set the value to 4V. This means you will have to lift the boom of the instrument to the top to trigger a measurement
- Make sure your cylinder is filled with the right gas with both taps closed
- Have one partner at the cylinder, have the other click Start on the left top of the main panel
- Lift the boom to the top and quickly press it down. Let the other partner press stop
- Go to file (top left) and export the three data sets to disc. Make a separate directory for your data and decide upon sensible file names like argon-p1, e.g.
- The files can be opened in excel. They are tab delimited and you need to the combine the v,p and t files for the same experiment into one sheet.

Data collection

We will examine 3 gases: Ar, N_2 and CO_2. Use the valves to flush the cylinder five times with each gas. It is useful to have two people operate them (Open exit valve, move piston down, close exit valve, open entry valve to fill with gas, close entry valve, open exit valve, etc.) Close both valves at the end, but briefly open the exit valve to make sure the pressure in the cylinder with the piston up is equal to ambient. Make sure the piston is up, such that it is beyond the reset sensor point, otherwise the data collection is not triggered.

Once you have flushed the cylinder with the gas of interest it takes little time to make a measurement. Therefore, collect data for at least 10 consecutive runs. Make sure to compress the cylinder as quickly as possible. If you compress too slowly there will be heat transfer and the measurement will not be accurate.

Repeat the procedure for the other gases.

Calculations and Reporting

The data consist of spikes of pressure vs. volume (measured as voltage) that are digitized as the lever is pulled down to compress the gas. For regression purposes ***derive*** a suitable formula involving lnP versus lnV by taking a logarithm of Eq 1.(→ introduction). (What is the relationship between the slope of an lnP versus lnV plot to γ?)

The data files can be opened directly into Excel. The “text import wizard” menu will popup, select ‘Next’ and in the second popup tick the ‘comma option’, then select ‘Finish’.

The report need *not* contain all raw data in hard copy. Just list the file names (→ Data).

Calibration data

The output voltage from the sliding resistor on the side of the cylinder is a linear measure for the volume, but zero voltage does not correspond to zero volume. Take an average of the measured voltage for each plateau of the staircase and plot the height position against the voltage and determine the best regression line. Assume that the diameter of the cylinder is 4.45 cm to convert the height to volume.

Measurement runs

For each of the data sets do the following:

First you need to determine what the valid P and V data are. There are a lot of bad data at the beginning of the set and the point the piston hit bottom marks the end of the valid data. First inspect a plot of V against time. Use the intercept and the slope of the calibration line to convert the voltage to volume. (Include one such graph in Data section). If downward part of the spike does not look linear, then you have not pulled down the lever quickly enough. Note the *kink* in the graph; this roughly marks the end of the valid data. Put the cursor on the kink in the graph to find out up to which point the data is valid → Data).

Method for fitting the data: Fit the ln V vs ln P data and determine the slopes and standard deviations for each data set using linear least squares fitting. As above use the central region of the change in V and P for your linear fit. Once you have made the fits for calibration line and data sets use the t-test to determine what value to multiply calculate a 95% confidence limit for your data. Determine the heat capacity based with appropriate errors based on the slope and the known relationships between the parameters. Be sure to account for propagation of error. (see below for the math to complete this).

Use propagation of error to determine appropriate error estimates. Consider the fact that your calibration line for the volume contains a measurement error that is significantly larger than the pressure since you must convert from voltage (measured) to volume.

Complete the report as follows: Derive an expression that allows you to express the molar heat capacity at constant pressure $\bar{C}_p$ in γ and R (from Eqn 2 and 3). (→Introduction). Calculate $\bar{C}_p$ (using your formula) and its uncertainty (by propagation from γ, using your formula and its derivative). (Results in table please, together with literature values,→Results. Literature source →Reference. Sample calculation →Calculation).

Propagation of error: Derive expressions for the relative error in the pressure, P, the ratio of heat capacities, γ and the heat capacity at constant pressure, $\bar{C}_p$. As an illustration of how to carry out these steps the first step is done for you. To calculate the error in the pressure, you treat pressure as a function of volume.

$$P(V) = \frac{nRT}{V}$$

The error in the pressure is

$$\sigma(P) = \sqrt{\left(\frac{\partial P}{\partial V}\right)^2 \sigma(V)^2}$$

Once you have taken the derivative you could use a representative pressure and volume, e.g. standard temperature of 298 K at 1 atm of pressure to determine the relative error. Note that the relative error is

$$\frac{\sigma(P)}{P}$$

so it is unitless. We show that this relative error can be determined from the relative error in the volume

$$\frac{\sigma(V)}{V}$$

Which is obtained from the calibration line measured in the experiment. The propagation of error requires the derivative for an ideal gas.

$$\left(\frac{\partial P}{\partial V}\right) = -\frac{nRT}{V^2}$$

Substituted into the formula for propagation of error we have

$$\sigma(P) = \sqrt{\left(\frac{nRT}{V^2}\right)^2 \sigma(V)^2}$$

We can rearrange this expression

$$\sigma(P) = \sqrt{\left(\frac{nRT}{V}\right)^2 \frac{\sigma(V)^2}{V^2}}$$

To make it evident that it is a function of only pressure and volume

$$\sigma(P) = \sqrt{(P)^2 \frac{\sigma(V)^2}{V^2}}$$

which leads to

$$\frac{\sigma(P)}{P} = \frac{\sigma(V)}{V}$$

You should be able to carry out similar procedures for $\gamma(P, V)$ and , $\bar{C}_p(\gamma)$.

Compare the error from propagation to error estimates obtained from replication. You have ten replicates of each data set. Use these replicates to estimate the error in γ obtained from replication. Since ten measurements is not in the limit of large numbers you will need to apply to the t-test to estimate the error in terms of the 95% confidence limit. Note that if you determined the error in volume based on a 95% confidence limit then the propagated error for a single data set will also represent the 95% confidence limit.

DISCUSSION

Using the data points create a histogram of values of gamma obtained in a typical data set. Are your data normally distributed? Was this histogram 'expensive' to obtain? What values for $\bar{C}_p$ and γ would you theoretically expect for a monatomic ideal gas? Why is the $\bar{C}_p$ different for CO_2? Compare your measured $\bar{C}_p$ and γ values with theoretical and literature values. Discuss why the two temperature curves T_{real} and T_{meas} may not coincide. What would happen to the slope value for Ar if the flushing would not be entirely successful and your gas sample contains a small amount of either air or CO_2? Why is the cylinder made of a thick polymeric material rather than e.g. copper? Why is the intercept of the calibration curve more important than the slope?

Checklist for your report

1. Do all your measurements have a sign, a magnitude, a precision and a dimension (units!)?
2. Have you carried out a detailed error propagation?
3. Do all your tables and all your figures have captions? Axis labels? Units? Is their scaling and size appropriate? (Can the reader see what you want to show?)

REFERENCES

MacQuarrie & Simon, Physical Chemistry, University Science Books, Sausalito CA, 1997, pp754-756, 797-799
P.W. Atkins, Physical Chemistry, Freeman, New York, pp. 604-605
Pasco Manual.

* (=NORMDIST(x,μ,σ,false)) calculates the theoretical value of the probability density of normal curve N(μ,σ^2) at the point x. Because the bin-width is ½ we must scale this with N/2, to make sure the integral will sum up to N).

Lab 2. UV/VIS spectroscopy of d and f-electrons

This laboratory requires a laboratory report.

Introduction

The fourteen $4f^n$ orbitals are filled across the lanthanide series of elements (La-Lu), much like the ten $3d^n$ orbitals are gradually filled in the first transition metal series (Sc-Cu). Many of the lanthanides readily form divalent or trivalent ions in solution. Other oxidation numbers are possible too, but are less common. The lanthanides in particular have a strong preference for Ln^{3+}. Both series tend to produce colored salts because the partially filled shells facilitate electronic transitions in the visible. In a flame where we have isolated ions in a hot plasma these transitions lead to sharp absorption peaks that can be used to determine small amounts of the element by Atomic Absorption Spectroscopy. In solution or solid state this is different because the environment plays a role and *spectral broadening* is the result.

Broad absorption bands for d-d transitions

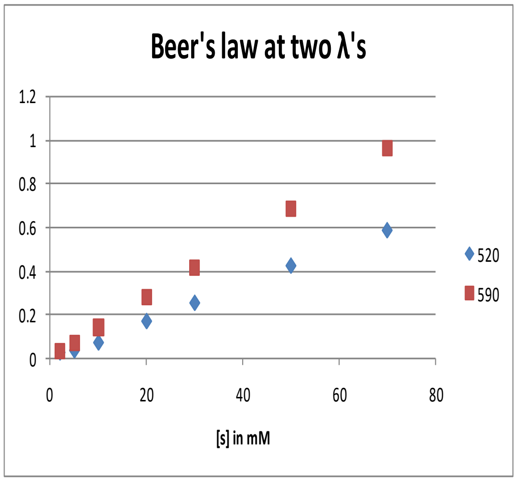
The transition metals have very broad d-d absorption bands that very little to do with the atomic spectrum and can only be understood in terms of *molecular* orbitals. The most important mechanism that causes such transitions to be broad is a strong coupling of electronic transition to the vibrationally excited states of the molecule (vibronic coupling).

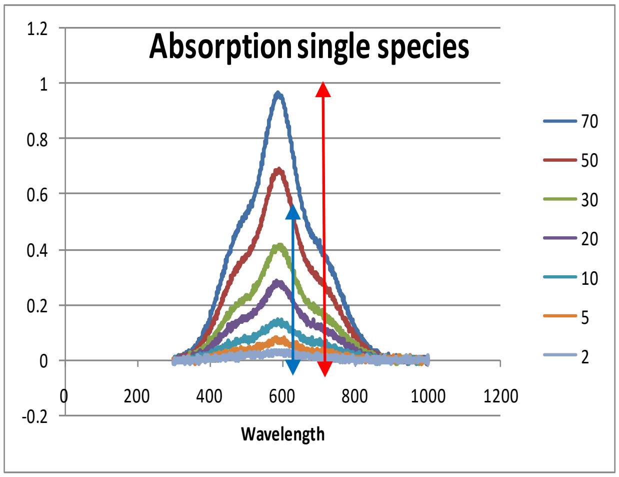
Sharp bands for f-f transitions

Lanthanide ions in solution have absorption spectra that strongly resemble the isolated atomic spectra, but the overlap with the environment is not negligible. Transitions between f-levels (*f-f* transitions) are relatively weak because the strongest mechanism (electron dipole coupling) is parity forbidden. Only magnetic dipole transitions are observed and they are generally much weaker. However through interaction with the environment (ligands) the parity rule can be broken a bit and this will enhance the absorption. So even f-elements are not entirely insensitive to the environment in solution

Lambert-Beer in mixtures

In a solution of a single component that absorbs over a range of wavelengths we can apply Beer's law at *any* wavelength that is absorbed by the solute s:
$A_\lambda = \varepsilon_\lambda.[s]$
It is usual to take the *maximum* of a peak in the spectrum as your working wavelength, because the *sensitivity* $S= dA/d[s] = \varepsilon_\lambda$ is maximal there. However, this is not necessary; we could take a different part of the spectrum. It just has a different value of ε_λ!





If the peak of the absorption exceeds A = 1.0 - 1.5, the measurement loses linearity. Remember that 90% of the light is absorbed if A = 1.0 so only 10% of incident light makes it to the detector. In such cases, it is actually better to use the side of the mountain, not the peak.

If you have two species, say s and t, in solution and the spectra overlap the measured absorbance at one wavelength is always a combination of two effects:

$$A_\lambda(\text{total}) = A_\lambda(s) + A_\lambda(t) = \varepsilon_\lambda(s).[s] + \varepsilon_\lambda(t).[t]$$

Unless the peaks happen to be exactly at the same wavelength the peak of one would be a hill side of the other and vice versa. It is often *not* possible to pick a wavelength where only one species absorbs (especially for d-ions!). Nevertheless we can still measure both concentrations [s] and [t] if we measure at (at least) *two wavelengths*, e.g. the peak of one and the peak of the other:

$$A_{\lambda 1=\text{S-peak}}(\text{total}) = A_{\lambda 1}(s) + A_{\lambda 1}(t) = \varepsilon_{\lambda 1}(s).[s] + \varepsilon_{\lambda 1}(t).[t]$$
$$A_{\lambda 2=\text{T-peak}}(\text{total}) = A_{\lambda 2}(s) + A_{\lambda 2}(t) = \varepsilon_{\lambda 2}(s).[s] + \varepsilon_{\lambda 2}(t).[t]$$

We can write these equations as a matrix product

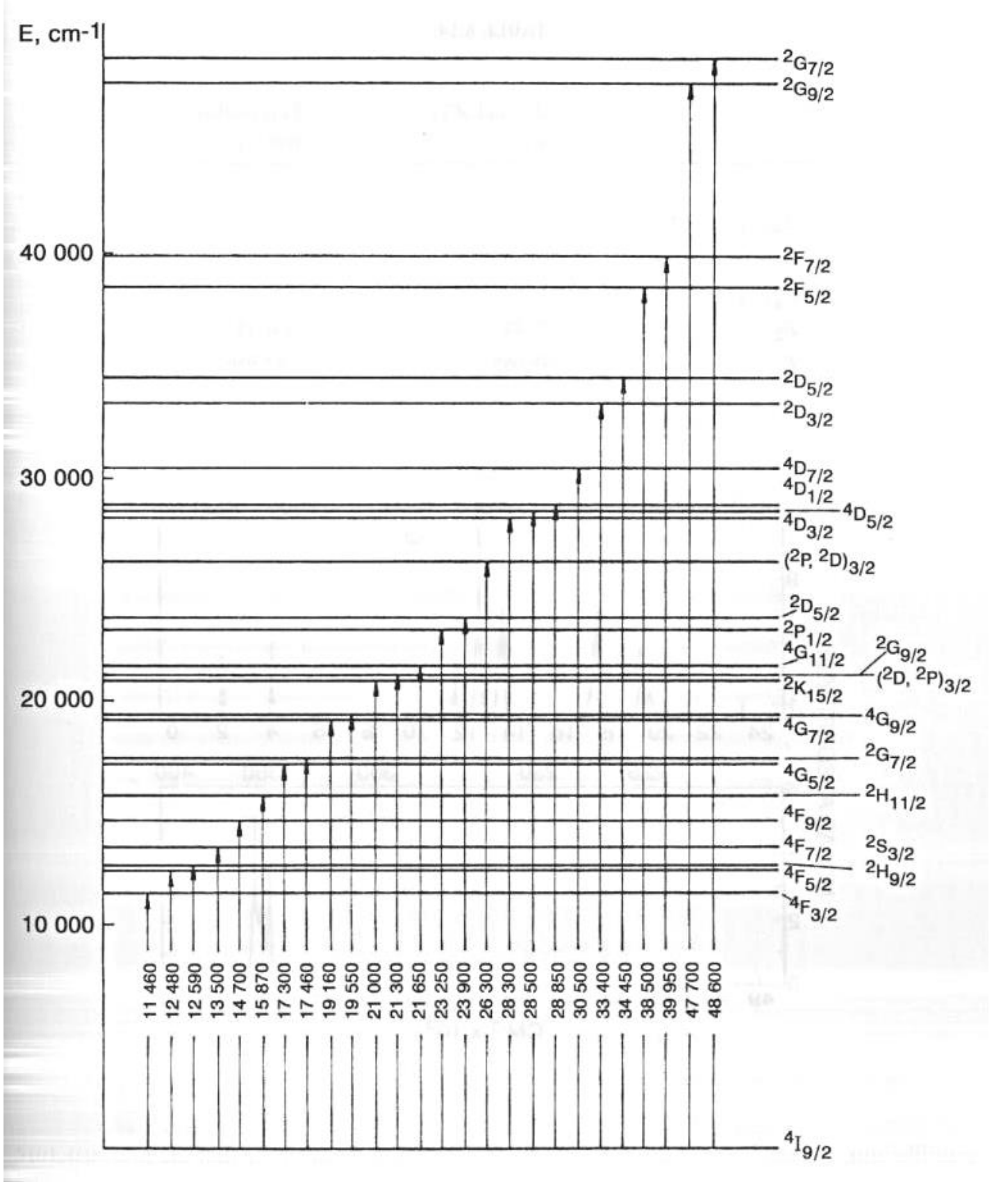
$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} \varepsilon_1(s) & \varepsilon_1(t) \\ \varepsilon_2(s) & \varepsilon_2(t) \end{pmatrix} \cdot \begin{pmatrix} [s] \\ [t] \end{pmatrix}$$

Provided we calibrate the four extinction coefficients $\varepsilon_\lambda(x)$ we can solve for the concentrations [s] and [t]:

$$\begin{pmatrix} [s] \\ [t] \end{pmatrix} = \begin{pmatrix} \varepsilon_1(s) & \varepsilon_1(t) \\ \varepsilon_2(s) & \varepsilon_2(t) \end{pmatrix}^{-1} \cdot \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

We could use more wavelengths (of which we have a spectrum full!) but then the matrix **ε** with extinction coefficients will not be a square, so we should use a generalized inverse like $(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon})^{-1}\,\boldsymbol{\varepsilon}^T\mathbf{A}$ rather than a simple $\boldsymbol{\varepsilon}^{-1}\mathbf{A}$**.** This would turn the calculation into a regression job.

These are the calculated energy levels for the f^3 ion Nd^{3+} and its f->f transitions. The energies are given in cm^{-1} (i.e. wavenumbers, note that wavelength[nm] = 10,000,000*wavenumber[cm^{-1}]). The transition to the state way in the UV cannot be observed because at those energies there are also other (much stronger f->d) transitions.



Lab instructions

The TA will provide you with stock solutions of a 0.2 M Nd^{3+} (aqueous nitrate) and a 0.2 Cu^{2+} (aqueous nitrate). These values assume you are using a 1.0 cm path length cuvette. A mixture of neodymium and copper nitrate of unknown concentrations will also be provided:

- From the two metal solutions prepare four dilutions: 1:1, 1:2, 1:5 and 1:10.

- Collect spectra from 400 to 800 nm. For the analysis we will use the absorbances at 521, 573, 640 and 680nm.
- Record the absorption spectra of the stock solutions and the four dilutions and write down the five absorbance values for each of the four wavelengths. Pay attention to the baseline. Plot the spectra in Excel or Igor and examine a portion of the spectrum that has little absorption (i.e. is relatively flat). If it differs from zero by more than 0.005 you should add (or subtract) a constant to the entire spectrum. Since absorbances may be small in this experiment, small offsets may affect the result. This procedure will eliminate such artifacts.
- Record the spectrum of the unknown and write down the absorbance at the same 4 values. Save this spectrum as well. You may want to compare the spectra later in your analysis.
- Export the data and save the files. Use suggestive shorthand nomenclature such as Nd02.dat, Nd01.dat, Nd005.dat for the 0.2 M, 0.1 M and 0.005 M solutions, respectively.

Data work up

Combine your absorbance values from multiple runs to determine the molar extinction coefficients of both ions at 521, 573, 640 and 680 nm by linear regression.

Prepare a table with the values of the extinction coefficients for the two species at the four wavelengths in column 1 and 2 and the absorbance of the unknown at those wavelengths in the third column. Then perform a regression of column 3 against 1 and 2. Report the concentrations of both species in proper 2/15 format.

Open the exported spectral files in Excel or Igor. Examine the spectra of the dilutions in the spectral region between 450 and 750 nm. Identify which transitions are responsible for the neodymium spectrum. These limits are chosen to include the 4 wavelengths and to frame the data nicely for making figures.

Determine the concentration of unknowns using the matrix method. For our purposes we will use a 2x2 matrix. Choose two of the wavelengths that are appropriate. Appropriate wavelengths should have relatively high absorbance for one species and low for the other, one for Nd^{3+} and one for Cu^{2+}. Use Excel to find the matrix inverse and solve the matrix equation. For a 2x2 you can always check your answer by hand, solving two equations with two unknowns. The brute force method does not work for a 3x3 or larger matrix, which is why the matrix method is so powerful. Here we are using the simplest case to learn the method.

Report the values of the extinction coefficients with 95% CI and the concentrations of the two species in the unknown mixture. Also report the errors in those concentration using appropriate propagation of error.

Lab 3. Analysis of the FTIR spectrum of HCl

This laboratory does not require a laboratory report. A worksheet must be completed for credit.

INTRODUCTION

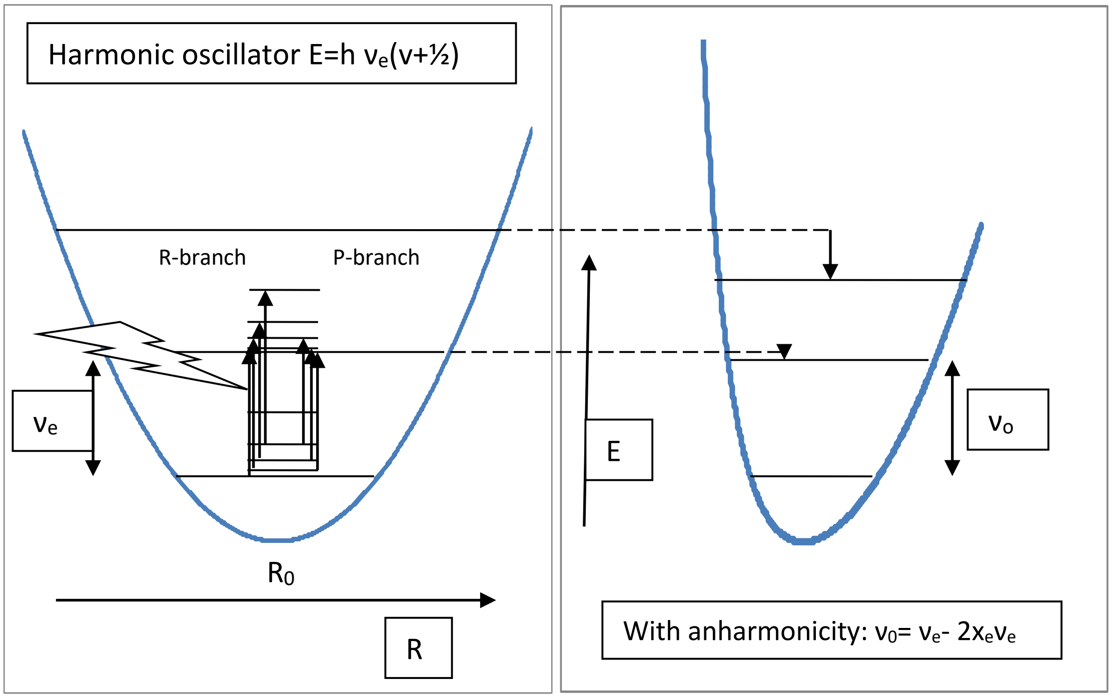

Harmonic oscillator E=h νe(v+½)
R-branch
P-branch
νe
R0
R
E
νo
With anharmonicity: ν0= νe- 2xeνe

Fig. E7.1 Energy diagram for the process of IR absorption by HCl

Using a Fourier Transform Infra-Red (FTIR) Spectrometer it is possible to resolve the rotational fine structure of the rotation-vibration transitions of a small linear molecule like HCl. Structural information about the molecule can be obtained from an analysis of this spectrum. Taking into account only harmonic and rigid terms McQuarrie and Simon derive two different expressions [13.12] for the R branch and [13.13] the P branch. However, they can be combined. Using the variable m, defined below McQ[13.12] and McQ[13.13] both yield:

$$\tilde{\nu}_{obs} = \tilde{\nu}_e + 2\tilde{B}m \tag{1}$$

m ≡ J_i + 1 for the R branch, where (ΔJ= +1), m ≡ - J_i for the P branch, where (ΔJ= -1).

If we include anharmonicity ($x_e\nu_e$) McQ[13.21] and rotation-vibration interaction (α) McQ[13.17] terms the same treatment yields:

$$\tilde{\nu}_{obs} = \tilde{\nu}_o + 2(B_e - \alpha)m - \alpha m^2 \tag{2}$$

where:
$\tilde{\nu}$ = the observed vibration frequency in *wavenumbers* cm^{-1} (as indicated by the tilde ~).
$\tilde{\nu}_o$= the frequency for a vibrational transition for $\Delta J = 0$, i.e. the (absent) Q branch.
B_e = the rotational constant for the equilibrium bond length R
α = vibration-rotation interaction.

The unit of cm^{-1} is the most common unit in spectroscopy. If you want to convert it to s^{-1} you need to use the speed of light, but the units must be cm/s ($c = 2.99 \times 10^{10}$ cm/s). To convert cm^{-1} to Joules use the conversion factor hc, where h is Planck's constant.

EXPERIMENT

In this experiment you will run the infrared spectrum of HCl in the vapor phase using an Excalibur Fourier Transform Infra-Red Spectrometer. The intensity of the absorption for each transition is a product of the population of the initial state and the absorption coefficient for the transition. You will use the integrated intensities to test the applicability of the Boltzmann distribution prediction of the populations of the initial states.

The HCl is introduced into a 10 cm quartz cell as two drops of hydrochloric acid, and the spectrum is taken of the vapor in equilibrium with this solution. Quartz transmits between 2500 and 3500 cm^{-1}. That allows the HCl fundamental to be observed, i.e. we will look at the frequency of light that causes the process:

$$\mathrm{HCl}(v = 0\ ;\ J_i\ ) \rightarrow \mathrm{HCl}(v = 1\ ; J_f) \text{ where } J_f - J_i = \Delta J = \pm 1$$

Since we will only look at ($v= 0\rightarrow1$), the fundamental frequency, we can only find the energy difference between the ground state and the first excited (vibr.) state, $\tilde{\nu}_o$, not $\tilde{\nu}_e$, the frequency corresponding to the curvature at the bottom of the parabolic potential curve (see McQ[13.22]). For the lines we study, the accompanying rotational change is either $\Delta J = +1$ (for R) or $\Delta J= -1$ (for P).

A. CALCULATION OF MOLECULAR PARAMETERS

First let us analyze the *frequencies (peak positions)* in the first column of the data.

The major part of this assignment is an analysis of the infrared vibration-rotation spectrum of HCl in terms of the theoretical model discussed above

First try a regression fit of the equation: $\tilde{\nu}= a_0 + a_1 m + \varepsilon$
Use a residual plot to show that it is necessary to add a term + a_2m^2 to the model. Compare this polynomial (actually quadratic) model with Eq. 2 to identity of the coefficients as:

$$a_0 = \tilde{\nu}_o \qquad a_1 = 2\,\tilde{B}_e - 2\alpha \qquad a_2 = -\alpha$$

This model allows for vibration-rotation interaction, but ignores the centrifugal distortion term $-D_eJ^2(J+1)^2$ (see McQ[13.23]) Show a plot of observed frequencies and the calculated ν *vs.* m and a residual plot.

If the residuals are not random, it may be necessary to add another term to the polynomial and investigate whether the centrifugal distortion $-D_eJ^2(J+1)^2$ could be responsible for such a term. Show a derivation along the lines of McQ[13.12/13] and use the definitions of m(→Introduction).

From these data you will calculate and report $\widetilde{\nu_0}$, $\tilde{B}_e$ and α (and perhaps D_e). From $\tilde{B}_e$ you will calculate the moment of inertia I and the bond length R. Include calculated uncertainties with all calculated quantities using the method of propagation of uncertainties (or errors). Ignore possible covariance between the coefficients a_i in the polynomial model.

Hint: To do the error propagation, first express B_e, I and R in terms of the parameters a_0, a_1, and a_2. Then calculate weights for the variances ($s_e^2(a_{0,1,2,3})$) by taking derivatives. Make sure you report and use the correct units.

B. FITTING THE BOLTZMANN DISTRIBUTION

Secondly, let us analyze the *intensities* of the peaks. Your data contains two values, both height and area, but one is clearly a better measure than the other, just look at a graph of area vs. frequency and height vs. frequency. One is smoother than the other, take the best one. From the above analysis the value for *m* and thus the initial value of J (J_i) is known. We will indicate the intensity value for a line with a certain initial J value as I_J below.

We will go beyond simply finding the properties of the molecules to examine the prediction of the Boltzmann distribution and the *relative* intensities of the rotational lines. The population, N_J, of the level with rotational quantum number J (with degeneracy g_J) is given by:

$$N_J = (2J+1)exp\left\{\frac{hc\tilde{B}J(J+1)}{kT}\right\}$$

Using this formula we see that N_0 = 1 (at J=0). Therefore, the relative population is:

$$\frac{N_J}{N_0} = (2J+1)exp\left\{\frac{hc\tilde{B}J(J+1)}{kT}\right\} \qquad (3)$$

The factor g_J is the degeneracy of the level J. Because rotation wave functions are identical to the rotational part of hydrogen wave functions g_J = 2J+1. (s, p, d, f, g,..) The intensity of absorption of a rotational line departing from rotational level J, I_J, is given by:

$$I_J = \epsilon_J N_J \tag{4}$$

For each line, the absorption coefficient ε_J is given by Herzberg as being *approximately* proportional to the line's frequency ν_J. Therefore the ratio of the intensity of a rotational line originating from rotational level J to that of a rotational level originating from the J = 0 level is:

$$\frac{I_J}{I_0} = \frac{\epsilon_J N_J}{\epsilon_0 N_0} \approx \frac{\nu_J N_J}{\nu_0 N_0} \tag{5}$$

And

$$\frac{I_J}{I_0} \approx \frac{\nu_J}{\nu_0}(2J+1)exp\left\{-\frac{hc\tilde{B}J(J+1)}{kT}\right\}$$

In this expression ν_0 is the frequency of the first transition (i.e. first rotational line in the R branch). The mean of the 0 here is that the transition originates from the zero state. It is a 0 -> 1 transition. In general, ν_J corresponds to a J -> J + 1 transition. From Equations 2 and 4 one can derive a formula to plot the relative intensities of different lines such that we should get a straight line. We can linearize the formula by taking the logarithm of both sides:

$$ln\left(\frac{I_J}{I_0}\frac{\nu_0}{\nu_J}\right) - ln(2J+1) = -\frac{hc\tilde{B}}{kT}J(J+1) \tag{6}$$

This means that if you plot:

$$ln(2J+1) - ln\left(\frac{I_J}{I_0}\frac{\nu_0}{\nu_J}\right) \quad vs. \qquad J(J+1) \tag{7}$$

You should get a straight line. From a regression you can extract the temperature. Be sure to include the error in the slope obtained from the fit and then to use propagation of error to obtain an error in your estimated temperature. The correct value should be less than 298 K since the building is usually not that warm during the winter months.

DISCUSSION

The data consist of two branches, the P and R branches. We will limit the analysis to the R branch.

1. Once you have obtained data converted to absorbance, you should make a table consisting of the rotational quantum number (column 1) and the wave number of the corresponding transition (column 2). Initially you can treat this as a linear regression and obtain a first estimate for the rotational spacing and the limiting wavenumber corresponding to $\Delta J = 0$.
2. Using the calculated rotational constant, calculate the H-Cl bond length.
3. Using the estimated vibrational frequency calculate the bond force constant for the H-Cl stretch.
4. Subsequently use a quadratic regression to obtain a more refined estimate of the rotational constant and the rotational-vibrational interaction.
5. What is the change in estimated bond length and force constant using these new estimates.
6. Estimate the temperature using the method described above by entering the appropriate values from Eqn. 7 into the excel spreadsheet and carrying out the linear regression.

Some Excel hints:

To do a quadratic regression create a column of the linear values of x (the A column), a column of the quadratic values of x^2 (the B column) and your y values (the C column). The select a range of 5 rows x 3 columns and type

=linest(C*first*:C*last,*A*first:*B*last*,TRUE,TRUE)

and use Ctrl+Shift+Enter to activate the formula. For the linear model you only need 2 columns, but for the quadratic model you will need three columns (e.g. for a model with ca term more use a column more.) The first row of the linest range contains the coefficients of the model *in reverse order*, i.e. for $Y=a + bX+ cX^2$ you get c,b,a. In two additional columns, use these parameters to construct a fit value for each data point and, by subtraction, a residual. Graph the data plus the fit versus m.

INSTRUCTIONS FOR THE WORKSHEET

Please turn in the Excel spreadsheet with spectrum and processed spectrum. Graph the absorption spectrum.

1. Provide the line spacing and center wavenumber, $\tilde{\nu}_o$, obtained from linear regression.
2. Provide the calculated rotational constant and show the calculation of the H-Cl bond length.
3. Using the center wavenumber show the work and provide the force constant for the H-Cl bond.
4. Repeat the above calculation using a quadratic regression.

5. Compare the values obtained from linear and quadratic regressions to each other. Compare calculated molecular parameters with those listed in Simon & McQuarrie p. 499 or another standard source.
6. Show how you estimated the temperature in the cuvette using the intensities (which you can assume are proportional to the peak value of each absorption line). Compare the calculated temperature with your best guess of the temperature in the sample chamber.
7. Why is there no absorption at $\tilde{\nu}_0$?
8. What is the difference between $\tilde{\nu}_0$ and $\tilde{\nu}_e$? [Cf. McQ & Simon Eq 13.21 (Remember v: 0→1) and see figure E7.1 above]

Some useful constants:
m_H= 1.007825 au, m_{35}= 34.968853 au m_{37}= 36.965903 au, 1 au =1.660540 10^{-27} kg.
h = 6.626076 10^{-34} Js, k_B= 1.38066 10^{-23} J/K. c = 2.99792458 10^{10} cm/s.
Careful: $\tilde{B}$ is in cm^{-1} (not m^{-1})

References:

1. Shoemaker, Garland, and Nibler, Experiment 38 and pages 758-763.
2. P. W. Atkins, *Physical Chemistry* (5th ed) 569-576.
3. McQuarrie & Simon, *Physical Chemistry,* Ch 13
4. Herzberg, G. *Infrared and Raman Spectra of Diatomic Molecules*
5. Pattacini, S.C., *J. Chem Educ*., **1996**, 73, 822.

WORKSHEET: Fourier Transform Infrared study of HCl

Please turn in the Excel spreadsheet with spectrum and processed spectrum. Graph the absorption spectrum. Please print this worksheet and upload the scanned or photographed completed work.

1. Provide the line spacing and center wavenumber, $\tilde{\nu}_o$, obtained from linear regression.

   Line spacing = __________________(cm^{-1}). $\tilde{\nu}_o$ = ____________________(cm^{-1}).
2. Provide the calculated rotational constant and show the calculation of the H-Cl bond length.

   $\tilde{B}_e$ = ___________________(cm^{-1}). $d(H-Cl)$ = ______________________(Å).
3. Using the center wavenumber show the work and provide the force constant for the H-Cl bond.

   k = ___________________(N/m).
4. Repeat the above calculation using a quadratic regression.

   Line spacing = ___________________(cm^{-1}). $\tilde{\nu}_o$ = ______________________(cm^{-1}).

   $\tilde{B}_e$ = ___________________(cm^{-1}). $d(H-Cl)$ = ______________________(Å).

   k = ___________________(N/m).

5. Compare the values obtained from linear and quadratic regressions to each other. Compare calculated molecular parameters with those listed in Simon & McQuarrie p. 499 or another standard source.

$d(H - Cl)$ = ______________________(Å) from the linear regression.

$d(H - Cl)$ = ______________________(Å) from the quadratic regression.

$d(H - Cl)$ = ______________________(Å) from literature.

What is the percent difference of each?

k = ___________________(N/m) from the linear regression.

k = ___________________(N/m) from the quadratic regression.

k = ___________________(N/m) from literature.

What is the percent difference of each?

6. Show how you estimated the temperature in the cuvette using the intensities (which you can assume are proportional to the peak value of each absorption line). Compare the calculated temperature with your best guess of the temperature in the sample chamber.

T = ________________________ K from data.

T = ________________________ K from your estimate of room temperature.

7. Why is there no absorption at $\tilde{\nu}_0$?

8. What is the difference between $\tilde{\nu}_0$ and $\tilde{\nu}_e$? [Cf. McQ & Simon Eq 13.21 (Remember v: 0$\rightarrow$1) and see figure E7.1 above]

Some useful constants:
m_H= 1.007825 au, m_{35}= 34.968853 au m_{37}= 36.965903 au, 1 au =1.660540 10^{-27} kg.
h = 6.626076 10^{-34} Js, k_B= 1.38066 10^{-23} J/K. c = 2.99792458 10^{10} cm/s.
Careful: $\tilde{B}$ is in cm^{-1} (not m^{-1})

Lab 4. Fluorescence of native tryptophan and accessibility to dissolved quenchers

Adapted from: Möller, M; Denicola, A. *Biochemistry and Molecular Biology Education* **2002**, *30*, 175–178.

Objective

The nature of chemical fluorescence provides many potential applications for the investigation of chemical and physical properties of analytes. From simple concentration measurements and reporter functionality to inter-/intra- molecular distances measurements and imaging, fluorescence has a wide range of applicability within analytical science. While some experiments require the introduction of fluorescent dyes, biological systems contain natural fluorophores, or are ripe for mutation to produce naturally occurring fluorescent molecules. As such, fluorescence has become an important tool for the investigation of biochemical and biological information. This experiment utilizes the native fluorescence of the amino acid tryptophan and the tendency of certain chemical additives to reduce fluorescence intensity (known as quenching) to gain information about protein structure.

Background

Spectroscopic methods can reveal a great deal about the nature of an analyte through the interactions of light and matter. The electronic transitions induced by light adsorption and emission reveal not only a great deal about the intrinsic electronic structure, but can also be used as tools to probe the relationship of the molecule with its surroundings. Fluorescence emission is of particular for studying environmental interactions due to the various mechanisms of molecular interaction involving the fluorescent excited state.

When a molecular absorbs incident radiation of the appropriate wavelength it promotes electrons from the ground state to one of many potential higher energy states. Vibrational and non-radiative relaxation can occur to return the molecule to less excited states, or a combination of relaxation and emission can occur. *Fluorescence* emission occurs through a radiative transition from singlet excited state to a lower ground state while *Phosphorescence* occurs as the result of an intersystem crossing from a singlet excited state to a triplet excited state, followed by a radiative transition to the ground state (Figure 1).

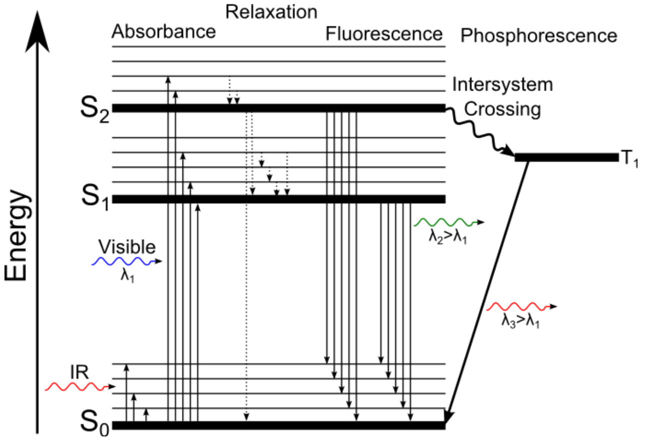


Figure 1. Jablonski diagram for molecular spectroscopic transitions

Because of the complexity of the electronic structure of molecules the both the absorbance and emission transitions form spectra over a range of wavelengths (Figure 2).

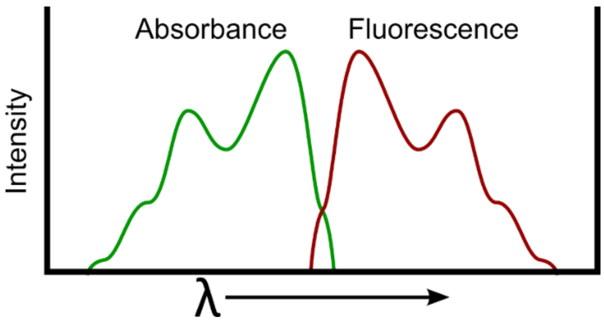


Figure 2. Overlaid absorbance and fluorescence emission spectra showing the characteristic Stokes shift and the “mirror-image” rule.

To obtain the *fluorescence emission spectrum* the sample is illuminated with light of a single wavelength and the emitted light is scanned through the entire spectral range. It is also possible to collect a *fluorescence excitation spectrum* by measuring the light emitted at a single wavelength while the incoming light wavelength is scanned through a spectrum. The excitation spectrum is very similar to the absorbance spectrum but can be collected using a fluorimeter setup rather than an absorbance setup.

When comparing the absorbance and fluorescence emission spectra for a single molecule the wavelength of maximum intensity, as well as the majority of the spectrum, is generally shifted toward longer wavelengths for fluorescence than absorbance; this phenomenon is known as the *Stokes shift*. The reason for this occurrence is apparent from a close analysis of the Jablonski diagram in Figure 1. Absorbance occurs from the ground electronic state (S_0) to an excited electronic state (S_n) and is accompanied by an increase in the vibrational energy level as well. However, fluorescence emission occurs only from the ground vibrational state and therefore some energy is lost between the absorption and emission, resulting in the observed shift in wavelength.

An additional feature of many fluorescent spectra is that they appear as mirror images of the absorbance spectrum. This is the result of the fact that the same quantum mechanical characteristics that make certain absorption transitions most probably also make the equivalent emission transition the most likely to occur. However, due to the fact that all emission takes place from the ground vibrational state the emission spectrum transition energies are Stokes-shifted, giving a mirrored emission spectrum.

The myriad transitions involved in absorbance and fluorescence occur very quickly (10^{-15}-10^{-12}s) but the lifetime of the excited state is many orders of magnitude longer (10^{-9}-10^{-7}s). While this timeframe is still very short from a human perspective, it is more than long enough for chemical reactions to occur or for the molecule to interact with nearby species in the environment. Therefore, it is possible for the fluorescent excited state to experience *quenching* wherein the state decays through a non-radiative pathway and fluorescence emission is lost. Quenching is a broad term that covers all interactions that reduce the fluorescence of a species; we will discuss only a simple quenching mechanism here.

Dynamic or *collisional quenching* is the non-radiative decay process that occurs as the result of energy transfer between the excited state fluorophore and a quenching agent in contact with it. The transitions of fluorophore X for absorbance (eq. 1), fluorescence (eq. 2), and dynamic quenching in the presence of quencher Q (eq. 3) are shown below.

$$X + h\nu \rightarrow X^* \qquad \text{(eq. 1)}$$

$$X^* \rightarrow \ X + h\nu \qquad \text{(eq. 2)}$$

$$X^* + Q \rightarrow XQ^* \rightarrow X + Q \qquad \text{(eq. 3)}$$

For this type of quenching the quenched fluorescence intensity (I) is related to the unquenched fluorescence (I_0) by the concentration of quencher ([Q]) and the lifetime of the fluorophore excited state (τ) in what is known as the *Stern-Volmer relationship* (eq. 4).

$$^{I_0}/_{I} = 1 + k\tau[Q] \qquad \text{(eq. 4)}$$

The value of k is a bimolecular quenching constant that can be experimentally determined only when the fluorescence lifetime is known. Because lifetime measurements requires an additional experiment, it is more common to use the Stern-Volmer constant (K_{sv}), which can be determined in a single experiment, and which is defined as the product of k and τ.

For most proteins of appreciable size there are multiple potential fluorescent residues. If these amino acids are not in equivalent positions the Stern-Volmer plot is non-linear and the relationship of total fluorescence to [Q] is defined based on the fluorescence of each species. An equation for a protein containing two non-identical tryptophans with unquenched fluorescence I_0' and I_0'' is shown (eq. 5) but can be extrapolated to three or more residues.

$$I = \frac{I_0'}{1+K_{SV}'[Q]} + \frac{I_0''}{1+K_{SV}''[Q]} \qquad \text{(eq. 5)}$$

The Stern-Volmer relationship has practical application in the study of natively fluorescing systems such as proteins. In proteins containing native aromatic residues, an intrinsic fluorescent signal can be obtained, and the addition of quencher can be used to modify the fluorescent intensity. However, not all fluorescent amino acids are equally exposed to potential quenchers due to the conformation of the protein and the relative size and identity of the quenchers. The Stern-Volmer constant can be used to estimate the degree of quencher accessibility to known fluorescent residues by measuring the fluorescent intensity under different conditions. For example, the denaturation of a protein unfolds its secondary structure and exposes residues that were buried in the protein core, inaccessible to solvated quenchers. A characteristic increase in the K_{sv} for a buried tryptophan would therefore be expected as a protein is measured in more denaturing solution conditions.

A wide variety of quenchers can be used, including molecular oxygen, halogen ions, acrylamide, and more. The choice of quencher can reveal additional data as well; large quenchers are unable to access residues in small pockets that are accessible to small quenchers, polar quenchers are unable to partition into hydrophobic protein cores but hydrophobic quenchers partition there preferentially, and charged quencher efficiency can be hindered by local ionic strength, all of which can yield useful information about protein structure with proper experimental design.

Pre-Lab

Answer the following questions in your lab notebook prior to the experiment.
Information gathered here will help you prepare for lab and for writing the lab report.

1. On a single set of axis sketch Stern-Volmer plots for a globular protein containing a single tryptophan residue in its core region under each of the following buffer conditions and justify the shapes you draw: 6mg/mL protein in pH 7.4 phosphate buffer; 6 mg/mL protein in 3M Urea; 3 mg/mL protein in 12M Urea.
   Note: Urea is a strong protein denaturant.
2. Lookup the expected emission wavelength maxima for free Tryptophan in solution vs tryptophan emission from OVA and BSA (using an excitation of 280nm)
3. Starting with the following provided solutions: 100mM pH=7 phosphate buffer, 3M NaCl, 1M KI, 50µM L-tryptophan, 5mg/mL Bovine Serum Albumin (BSA), 10mg/mL Ovalbumin (OVA), devise dilution schemes to prepare 4 sets of calibration standards directly in provided cuvettes (target final volume 3.5mL) as described in the procedure below.

Safety & Disposal

Potassium Iodide is a potential irritant and carcinogen, avoid contact and flush affected areas with running water. Samples containing protein should be disposed of in the labeled waste container.

Procedure

You will be provided with the following stock solutions: 100mM pH=7 phosphate buffer, 3M NaCl, 1M KI, 50µM L-tryptophan, 5mg/mL Bovine Serum Albumin (BSA), 10mg/mL Ovalbumin (OVA)

1. With no sample loaded into the fluorimeter, perform an excitation measurement using the default settings. The excitation spectrum of air should have its

maximum peak at 467nm. If the peak is not in its expected location, recalibrate the wavelengths.

2. Fill a fluorimeter cuvette with distilled water and measure the emission spectrum using the default settings (excitation at 350nm, slit widths 5nm). The sample should give a smooth curve with a maximum at 397nm. If the peak maximum is not in the expected location recalibrate the wavelength as described above.
3. Prepare the following solutions in 100mM phosphate buffer
    a. 2.5 µM Tryptophan
    b. 50 µg/mL BSA
    c. 100 µg/mL OVA
4. For each of the prepared samples collect an emission spectrum with an excitation wavelength of 280nm (slit width 1nm).
5. For each spectrum collected determine the wavelength of max emission
6. For each of the analytes prepare five additional samples with Iodine concentrations of 1-50mM and the same analyte concentration as in (3). For BSA prepare an additional set of five with the same Iodine concentration and a constant final 2M NaCl
7. Measure the fluorescence intensity for each of the iodine containing samples using the wavelengths determined in (5).
8. Graph the Stern-Volmer plots (I_o/I vs. [Iodine$^-$]) for each of the data sets and calculate K_{SV}

Post-Lab

Write a formal lab report and answer the following questions:

1. How does the value of K_{SV} for OVA compare to that of free L-Trp? Justify this relationship.
2. For a system containing multiple tryptophans it occasionally occurs that one or more is inaccessible to quencher. Using eq. 5, derive the following equation for a protein of interest that has two tryptophan residues with one that is completely inaccessible to quencher.

$$\frac{I_0}{I_0 - I} = \frac{I_0}{I'_0 K'_{SV}} \cdot \frac{1}{[Q]} + \frac{I_0}{I'_0}$$

Use your data to plot I_0/(I_0-I) vs 1/[I-] and use it to determine what percentage of fluorescence in BSA comes from quenchable tryptophan (the fraction of fluorescence is due to the quenchable tryptophan is the inverse of intercept) and to determine K'_{sv} of the quenchable tryptophan (intercept/slope). Do your findings make sense given what you know of the structure of BSA? Explain.

3. You should note that at low ionic strength the BSA Stern-Volmer plot does not show a perfectly linear relationship. Estimate the K_{sv} values for the curve assuming it is a combination of two linear functions.
4. Make sure to include table summarizing your solution preparations (volumes added of each ananlyte/solvent/reagent..etc.).
5. Report all calculated values to the correct number of sig figs (2&15 rule) and report errors..
6. include figures with the emission spectra and all generated straight lines (an trumpets)
7. Consult the following references and explain the observed behavior of BSA fluorescence including the effect of changing ionic strength and the source of biphasic quenching:

References

1. Skoog, D. A. *Principles of instrumental analysis*; 6th ed.; Thomson Brooks/Cole: Belmont, CA, **2007**.
2. Möller, M; Denicola, A. *Biochemistry and Molecular Biology Education* **2002**, *30*, 175–178.
3. Lehrer, S. S.; Leavis, P. C. In *Methods in Enzymology*; C.H.W. Hirs, S. N. T., Ed.; Academic Press, 1978; Vol. Volume 49, pp. 222–236.
4. Möller, M.; Denicola, A. *Biochemistry and Molecular Biology Education* **2002**, *30*, 175–178.

Lab5. Determination of the pKa of phenolic acids by reversed-phase HPLC

Reference: J. Chem. Educ. **2018**, 95, 310−314

Objective

In this experiment, you will consider a model that describes the effect of pH on retention in HPLC and use it to determine the pK_{a_1} value of a phenolic acid, ferulic acid (Figure 1). This requires the determination of chromatographic retention factors as a function of the mobile phase pH. You will use your determined retention factors and activity coefficients and mobile phase pH to perform nonlinear regression and calculate the pKa_1 of ferulic acid and its standard error.

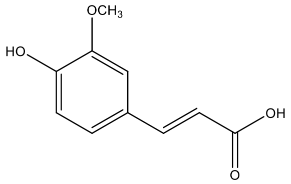


Figure 1. Chemical structure of ferulic acid

Background

The model used here[1] to determine the K_{a1} value by reversed-phase HPLC is based on the dependence of the chromatographic retention factor *k*, which is calculated by Eq. (1), on the pH of the mobile phase. The t_r and t_m in Eq. (1) represent the retention times of the retained and unretained components respectively.

$$k = \frac{(t_r - t_m)}{t_m} \qquad (1)$$

The retention factor of a ionizable compound at a given H^+ activity, a_{H+}, is a weighted average of the retention factors of the deprotonated and protonated forms, k_{A-} and k_{HA}, of the solute as described by Eq. (2) where x_i represents the mole fraction.

$$k = x_{HA}k_{HA} + x_{A-}k_{A-} \qquad (2)$$

Substitution of the expressions for the mole fractions x_{HA} and x_{A-} in (2) gives Eq. 3

$$k = \frac{[HA]k_{HA} + [A^-]k_{A^-}}{[HA]+[A^-]} \qquad (3)$$

Substitution of the acid dissociation constant, K_a, Eq. (4) into Eq. (3) and rearranging gives Eq. 5 which is used in this experiment to determine the K_a of ferulic acid where γ is the calculated activity coefficient.

$$K_a = \frac{(a_{A^-})(a_{H+})}{a_{HA}} = \frac{[A^-]\gamma a_{H^+}}{[HA]} \qquad (4)$$

$$k = \frac{k_{HA} + \frac{(k_{A-})(K_a)}{\gamma(a_{H^+})}}{1 + \frac{K_a}{\gamma(a_{H^+})}} \qquad (5)$$

The activity coefficient is calculated using the extended Debye-Hückel equation:

$$-log\gamma = \frac{A\sqrt{I}}{1 + a_0 B\sqrt{I}} \qquad (6)$$

The ionic strength, I, of the mobile phase will be estimated from the buffer preparation. Literature values of the Debye- Hückel constants A and a_0B in 30%CH_3CN will be used here.[2]

Pre-lab questions

Answer the following questions in your lab notebook prior to the experiment. Information gathered here will help you prepare for lab and for writing the lab report.

- Lookup the literature values of pK_a for ferulic acid (Fig. 1). Also, lookup the pK_{a1} value in 30%CH_3CN mobile phase system used in this experiment and in different percent organic compositions. How does the increase in percent organic affect the value of the pK_a? Can you suggest an explanation for this trend?
- What is the pK_a of formic acid? Describe how you will prepare 1L of 30mM formic acid buffers (e.g. show your calculations for the pH 4 buffer). The starting materials are solid sodium formate and 1M HCl solution. Specify how you will calibrate your pH meter.
- Show a sample calculation of the ionic strength, I, using the pH 4 buffer prepared above.
- The column that will be used in this experiment is a 150 x 4.6mm Agilent Zorbax S8-Aq C18 column with 3.5μm particle size. How will the retention time of ferulic acid change if (a) the pH of the mobile phase increases? (b) if 20% CH_3CN is used instead of 30% in the mobile phase? Why?

Procedure

Stock solutions of nominal 0.02M ferulic acid in acetonitrile and 0.03M KBr in deionized water will be provided.

Formic acid buffers (*ca*.30 mM) in the range pH 2.7-5.5 will be needed for the HPLC mobile phase. Use at least 5 buffer solutions and make sure you cover the full desired pH range.

For each HPLC injection, you will prepare a sample by measuring 100 μL of the ferulic acid with 300 μL KBr from the stock solutions into a 10mL volumetric flask and adjusting to the mark with the pH buffer solution for that injection.

Agilent 1200 HPLC system will be used with a binary pump and diode array UV/Vis detection and a (Agilent Zorbax S8-Aq) C18 column (4.6x150mm) packed with 3.5 μm particles.

Setup your HPLC method to pump 30%B and 70%A, Solvent B is acetonitrile and solvent A is the formate buffer. Select wavelength detection at 270 nm (ferulic acid) and 220nm (KBr), and a flow rate 1 mL/min.

For each buffer, record the pH of the solutions of that buffer mixed with water in the ratio 30%CH3CN:70% Buffer. Calibrate the pH electrode using the aqueous calibration standards available in the lab. Record the temperature at which the chromatographic runs are performed using a stick-on thermometer placed on the column.

The standard operating procedure for setting up and performing your HPLC runs will be available in the lab.

1) Follow instructions to purge and equilibrate the column with the first mobile phase (30% CH_3CN - 70% Buffer)
2) Half fill the autosampler vial with the first mixture solution prepared in the current mobile phase buffer as described above
3) Run an injection (30 μL) and save the chromatogram. (injections run from 2-5 minutes depending on the pH and the column). Repeat and save a replicate chromatogram. Verify that retention times are reproducible. this will be your indicator that the column has equilibrated at the pH of the mobile phase. If not, allow 5 more minutes of equilibration and repeat the injection.
4) Turn off the mobile phase flow and replace the line A bottle with the second buffer.
5) Turn the pump back on and equilibrate with (30% CH_3CN - 70% Buffer) for 15 minutes
6) Repeat steps 2-5 with the remaining buffers
7) Record all retention times in your notebook.
8) Exported data as .csv files to plot the chromatograms in Excel. Make sure to export the signals at 270nm and 220nm

Submit a full lab report with the usual sections, Abstract, Introduction, Experimental, Results, Discussion, and Conclusion. Include general considerations of the retention of ionizable compounds in reversed-phase HPLC, how the electrostatic and solvent effects impact the value of the acid dissociation constant and other fundamental aspects. Include all your experimental data (table with pH, a_{H+} and retention factors) and chromatograms collected at different pH (1 graph with clear legend identifying each chromatogram).

Briefly describe buffer preparation (if buffers already prepared for you, you will be provided with the procedure) and show calculations of ionic strengths and activity coefficients (use Debye Huckel coefficients from reference 3.

Use Solver in Excel to perform nonlinear regressions (k *vs* a_{H+}) and find K_{a_1}, k_{HA} and k_{A^-} values. Use the literature K_{a_1} value and the extreme retention factor, k, values as a starting point to fit the nonlinear model. Include the fitting graph and residual plot in your lab report. Use the SolvStat add-in in Excel to determine the standard error in the three regression parameters, K_{a_1}, k_{HA} and k_{A^-}. Calculate pK_{a_1} and its error. List values in a table and corresponding errors (using the 2/15 rule)

Compare to literature value in the same 30%CH_3CN system and comment on sources of error. Compare pK_{a_1} values to values determined in aqueous solutions and consider how electrostatic and solvation effects impact the value of the acid dissociation constant.

Consider in your discussion the values of k_{HA} and k_{A-} values, which one is larger? Is this what you would expect? And how did the retention times change with the pH of the mobile phase? Is that what you would expect?

Use the provided instructions for using Solver and SolvStat in Excel.

References:

1. Horváth, C.; Melander, W.; Molnr, I. Liquid chromatography of ionogenic substances with nonpolar stationary phases. *Anal. Chem.* **1977**, 49(1), 142-154.
2. Barbosa, J.; Sanz-Nebot, V. Autoprotolysis constants and standardization of the glass electrode in acetonitrile-water mixtures. Effect of solvent composition. *Analytica Chimica Acta* **1991**, 244, 183-191.

Lab 6. Plasticizer Analysis using Temperature Programmed GC-MS

Objective

Gas chromatography (GC) is a commonly employed separation method for the analysis of volatile samples in a variety of industrial and analytical applications. The separation mechanism consists of the differential partitioning of volatile species into the stationary phase but unlike liquid chromatography separations the composition of the mobile and stationary phases has only a minor effect on selectivity. Instead, separation is primarily accomplished through manipulation of temperature in a process known as temperature programming. This laboratory focuses on the use of temperature in GC separations; both an isothermal separation of closely retained compounds and a temperature-programmed separation of a homologous series will be performed.

Background

GC differs somewhat from its sister technique, liquid chromatography (LC), in that the flowing mobile phase is an inert gas, commonly helium, and all interaction is therefore between the analyte molecules and the stationary phase rather than being a competition between interactions with both phases. Furthermore, the samples analyzable by GC must be either inherently volatile, able to be vaporized without decomposition, or derivatized to volatilizable form since all analysis is carried out in the gas phase. For species that are able to be analyzed by GC however, the method offers great advantages over LC due to the increased resolution provided by the technique; the ability to use more universal detectors with better quantification than the detectors available to LC; and its easy coupling to additional dimensions of separation, including Mass Spectrometers. The efficiency of the technique can be explained utilizing the Van Deemter equation. Because GC separations take place in a long, open-tubular column, there is no packing material to yield an A term in the equation, reducing the plate height significantly; the open tube generates no significant backpressure, allowing for extremely long columns (50+ m in GC compared to 150mm in HPLC); and the lack of stagnant mobile phase within particles gives a reduced (but not eliminated) C term.

In gas chromatography the equilibrium partitioning is distinct from liquid-liquid partitioning in LC because it includes the liquid-gas phase transition. This means that, in contrast to liquid chromatography, temperature exhibits a large effect on the degree of partitioning into the stationary phase and specific intermolecular interactions with the stationary phase are only a partial player in partitioning behavior. Control of the retention time and selectivity, is somewhat limited in GC compare to in LC due to the limited number of interactions that are possible in partitioning. Changing the stationary phase is the only method for adjusting selectivity (α) to obtain different separations since the mobile phase is non-interacting, but controlling partitioning on a given column is achievable by adjusting the column temperature; with higher temperatures compounds have reduced partitioning into the stationary phase and faster elution times. Two compounds with different boiling points are thus easily separated by heating the GC column to a temperature where one species is significantly more volatile than the other. If no single temperature is capable of providing good separation while also allowing for a relatively speedy separation, a gradient of temperatures can be applied such that early eluting compounds experience a relatively low temperature for their separation, and later eluting compounds are more quickly removed by an increased temperature.

Information pertaining to the specifics of Gas Chromatography instrumentation can be found in literature reviews (see references), or in the appropriate chapter in an instrumental analysis textbook.

While reference data for α values for species in comparison to reference analytes are available for some analytes in the literature, the overall coverage is not sufficient for this to be an effective method for analyte determination. Retention index values are readily available on many common columns for certain analytes, but do little for the analysis of unstudied or poorly studied compounds additional methods of sample characterization are necessary. Thankfully, GC lends itself readily to coupling with mass spectrometry (MS) to form the hyphenated technique GC-MS, which is capable of eliciting a large amount of information from samples. MS is a separation technique in its own right and a thorough discussion of its capabilities and instrumentation are too extensive to detail, some relevant information is available below.

The fundamental mechanism of a mass spec is the separation of ions based on their mass-to-charge ratio. The basic layout of a mass spectrometer consists of an *ionization source*, which produces ions from an initial sample, a *mass selector*, which is the mass separating component, and a *detector*, which generates a signal for separated ions. The various combinations are extensive and we will limit our discussion of mass spec for this laboratory to a single ionization source *electron impact ionization* which is equipped to the GC-MS, and its mass selector, the *quadrupole*.

Electron impact (EI) is what is known as a *hard ionization* source, which means that it will heavily fragment analyte molecules to produce ions. The other forms of ionization, *soft ionization* sources, generate mostly molecular ions and few, if any, fragments as the direct result of ionization. The EI source functions by colliding analyte molecules with a high energy electron (~70eV) to remove an intrinsic electron and produce a charged radical ion for analysis:

$$M + e^{-} \rightarrow M^{+\blacksquare} + 2e^{-}$$

This molecular ion is often detected by the mass spec, but it also has the potential to further fragment itself and produce fragments that can be detected. The mechanisms of fragmentation are extremely varied and will not be covered in this background section but an example is shown (Fig 2).

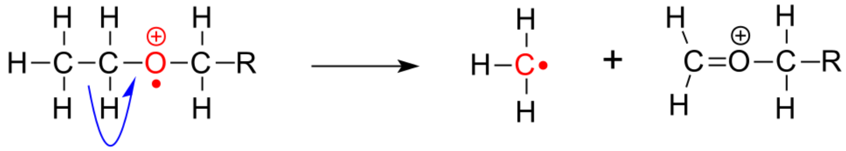


Figure 2. Mechanism of one potential fragmentation pattern for a simple ether. Two fragments are produced, but only the charged species can be detected.

Once ions are produced they are filtered using a quadrupole mass filter (Figure 3). The detected ions can be used to generate a total ion chromatogram (TIC) which shows the total number of ions detected at a given time, and is analogous to the chromatogram generated by a UV detector in HPLC, or a flame ionization detector in standalone GC, but the function of the quadrupole allows for much more information to be obtained than a simple chromatogram.

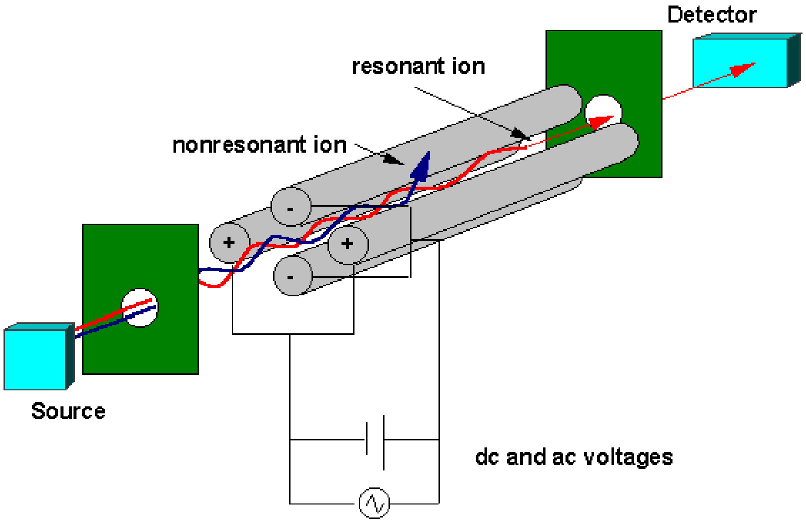


Figure 3: Quadrupole mass filter. Ions enter from the ion source and are channeled by charged robs. Ions within the filter range pass to the detector while those outside are excluded.

The basic operating principle of a quadrupole consists of the following: charged rods have a constant DC voltage applied, which is modulated with an AC current. The current modulation causes the poles to switch sign of the voltage, resulting in ions oscillating between rods as they pass down the column. When the voltages are tuned correctly, this oscillation allows only ions of a particular m/z through. Ions with masses that are too high or too low drift out of the quadrupole center and are wicked away from the detector. The quadrupole is capable of rapidly scanning through mass ranges and therefore isolated each m/z value in sequence. This allows for the rapid collection of an m/z spectrum for an incoming analyte.

The MS is able to identify the mass to charge ratio of the produced fragments with a high degree of precision and the fragmentation process produces distinct fingerprints for different compounds which enables the determination of analyte identity. The produced fragments and fragment patterns can be either analyzed by hand and compared to known fragmentation rules and patterns to determine the structure of the unknown analyte, or an algorithm can be used to search the mass spectrum against libraries of known structures. A samples mass spectrum of pentanol with a few peaks labeled is shown (Fig 4).

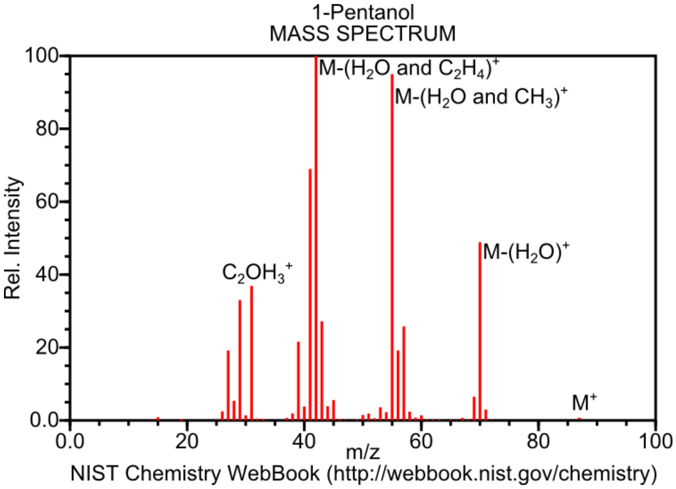


Figure 4. Mass Spectrum of 1-Pentanol after fragmentation with an electron impact source. The molecular ion (M^+) is weak while peaks corresponding to the M^+ with fragment losses are relatively abundant.

Pre-Lab

Answer the following questions in your lab notebook prior to the experiment.
Information gathered here will help you prepare for lab and for writing the lab report.

1. Lookup information on the HP-5MS (or DB-5MS) that will be used in this experiment; what is the stationary phase and what makes it a good choice for this experiment? What type of molecules will interact well with this type of stationary phase?
2. Look up the chemical structures, molar mass, density and boiling points of dibutyl phthalate (DP), and bis(2-ethyl hexyl) adipate (EHA)
3. Using the provided dilutions (25µL of neat liquid diluted to 25mL in cyclohexane), calculate the concentrations of the stock solutions (nominal 1000ppm). Then devise a dilution scheme to prepare a set of DP standards in the range 0.5-10ppm in cyclohexane using EHA as the internal standard (2ppm)
4. Lookup the retention indices of DP and EHA on the stationary phase used here (http://webbook.nist.gov/chemistry/) and show equation that will be used to calculate the retention index from data acquired in the lab.
5. Lookup the mass spec of DP and EHA from the same website http://webbook.nist.gov/chemistry/)

Safety & Disposal

Chemicals used in this laboratory should be disposed of in the organic waste container. Avoid contact with the instrument when working with the GC as the injector, detector, oven, and column are hot and will cause burns.

Procedure

In this lab, you will use gas chromatography coupled to mass spectrometry to identify and quantify select plasiticizers in a given sample.

Sample Preparation

- The following three stock standard solutions are provided: (a) C7-C30 saturated alkanes standard in hexanes (sigma Aldrich catalog # 49452-U, 1000μg/mL), (b) dibutyl phthalate (DP, 25μL diluted to 25mL in cyclohexane), (c) bis-(2-ethyl hexyl) adipate (EHA, 25μL diluted to 25mL in cyclohexane). Use the molar mass and density values of these compounds to determine the concentration of the provided stock solutions (these are nominally 1000ppm solutions but you need to determine the exact concentrations).
- Devise a dilution scheme to prepare 5 standards which contain nominally 0.5-10ppm of DBP and a fixed 2ppm concentration of EHA. The EHA will be used as the internal standard. (provided unknown will have 2ppm internal standard)
- A diluted solution of C7-C30 saturated alkanes (300μL diluted to 10mL in cyclohexane) ready for injection will be provided
- A solution which contains an extract of a shredded plastic in cyclohexane will be provided. The extract is prepared by taking a weighed sample (nominal 10g) of a shredded plastic bottle (tonic water) that has been soaking in 100mL dichloromethane for over 48 hours, filtering, rotovaping and re-dissolving in a small volume of cyclohexane (will refer to this solution as the 'unknown' and the exact mass of plastic bottle and volume of cyclohexane used will be provided). In the absence of a sample with a significant content of a plasticizer, a prepared unknown solution will be provided.

Instrument setup:

- The tuning and calibration of the mass analyzer is preformed once a week by the TA. This is achieved using decafluorotriphenylphosphine and p-bromofluorobenzene standards (the vial containing the standards is already inside the instrument, just need to run the autotune and save the file as atune.u).
- Set the GC-MS to utilize the following parameters: (**when injection the plasitizer sample, increase the final temperature hold to 6min.)**

1μL injection	Injector Temperature 250°C
Splitless injection with splitless flow 198 mL/min @0.75min	Detector Temperature 300°C
Column temperature: 50 °C for 1min Hold then ramp to 280 @30 °C/min. no temp hold ramp to 300 °C @ 15 °C/min. hold for 1 min	Column Flow Rate 1mL/min
Solvent delay 5min	MS Source 230 °C MS Quad 150 °C Aux-2 Temp 280 °C

- Open the mass spec parameters by selecting the quadrupole icon and selecting the edit scan option. Enter a 5 min solvent delay, that the scan range is 50-550 m/z, and the atune.u file is selected. It is very important NOT to bypass the solvent delay.

Analysis:
Identification of compounds using retention indices.

1. Make sure to select the scan mode in the mass spec parameters window
2. Perform one 1μL injection of the diluted alkanes mixture solution. Your TA will show you how to inject 1μL sample using a 10μL syringe.
3. Perform 1μL injection of the standard which contains the nominal 3-4ppm DBP (and 3-4ppm of the internal standard EHA)
4. Perform a 1μL injection of the prepared unknown solution. Compare area under the curves for the DP peaks in this step and in step 3 to make any final adjustments to the standard concentrations.
5. After analysis open the total ion chromatograms (TIC) in the mass spec data analysis and determine the elution time for each peak (you can start looking at completed data while another run is in progress). You will later use this data to calculate the retention indices of DBP and EHA– (you can compare your retention indices to values listed in the NIST webbook at http://webbook.nist.gov).
6. The mass spectrum can be seen by right click and dragging around a peak in the TIC. Once you see that your mass spec data is what you anticipated it to be, proceed to the quantitative analysis and you can export your TIC data during your wait times.
7. Export the TIC and the mass spectra to .csv file. The .csv files contain a list of identified mass peaks and their total abundances. Label the peaks in a meaningful fashion so that you can remember them later.
8. For each peak identified from the extract sample, right click and drag a box around the peak to obtain the averages mass spectrum of the peak. Compare

the mass spectra and the retention indices to the values determined from the standards.

9. Open each .csv file in Excel and convert the total abundances to relative abundances on a scale of 0-100 (the most abundant peak should have a value of 100 and others should have relatively less).
10. Go to http://www.massbank.jp/QuickSearch.html and perform a peak search on your generated peak list for each peak versus the EI database for MS data (select EI and only the MS checkboxes). Determine what you think each peak is based on the mass spectra. The NIST webbook link also provides mass spectra to compare to.

Quantitative analysis of plasticizers in in selective ion monitoring (SIM) mode

11. Make sure to select SIM mode in the mass spectrometer parameters window and select the m/z values to monitor (typically the following peaks should be present in your mass spectra and can be followed: 129, 147 and 241 (EHA), 149 and 223 (DBP)
12. Inject 1μL each of standard
13. Inject 1μL unknown
14. Record the areas under the peaks for DBP and EHA.
15. Perform least square analysis (areas vs concentrations) to find the best fit calibration curve and calculate the concentration of plasticizer(s) in the diluted unknown (use areas under the curve and method of internal calibration, dibutyl adipate is the internal standard). If a sample of plastic extract is analyzed, use dilution volumes and mass of shredded soda bottle to back calculate the w/w ppm of plasticizers in the soda bottle.
16. Calculate the detection limit.

Post Lab

Submit a full lab report with the usual sections, Abstract, Introduction, Experimental, Results, Discussion, and Conclusion. include all your TIC, SIM and mass spec results (representative total ion chromatograms, and mass spectra and table of areas under the curves. Also include chromatographic retention indices calculations. Identify DBA,DBP, EHA, EHP based on the mass spectra and based on the chromatographic retention and retention indices.
Use RLS macro to identify any outliers before performing least square analysis (use straight line with internal standard, i.e. y-axis is the ratio of the area of the analyte/area internal standard), use your calibration curve to calculate the amount of bis-(2-ethyl hexyl adipate) per g of plastic and report errors and give significant figures according to the 2/15 rules. Evaluate the results (retention index and concentration) and compare to expected value (always reference)
Determine the limit of detection.
Answer the following questions as part of your discussion:

Why is it important to use calibration with internal standard?
Why is the solvent delay used?
Why is it better to measure areas under the curve and perform quantitative analysis from SIM mode rather than scan mode (TIC)?

References:

1. Oteroa, P.; Sahaa, S. K.; Moanea; S.; Barronb, J.; Clancy, G.; Murraya, P. *Journal of Chromatography B*, **2015**, 997, 229–235
2. Guiochon, G.; Guillemin, C. L. *Review of Scientific Instruments* **1990**, *61*, 3317–3339.

Computer tutorials

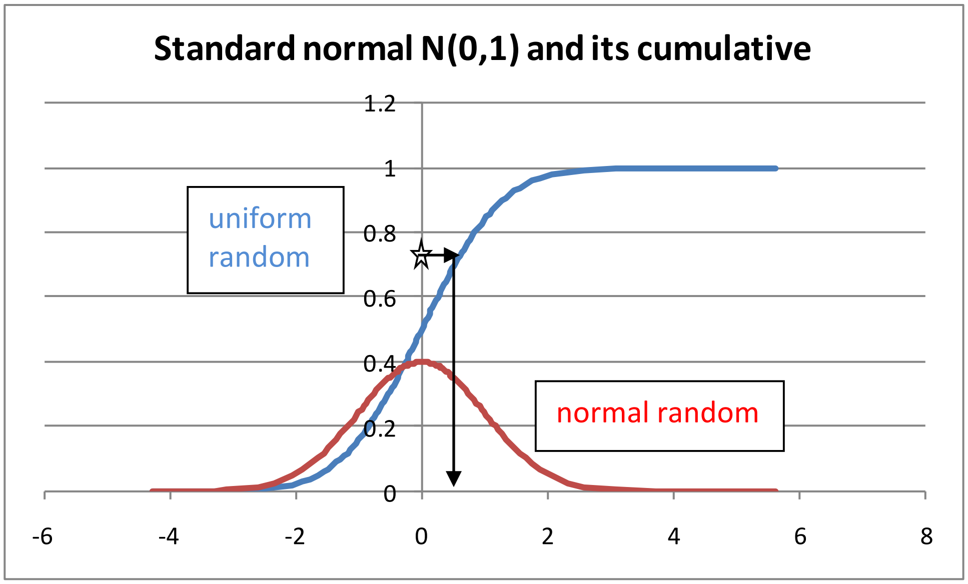
Tutorial 1. Random distributions

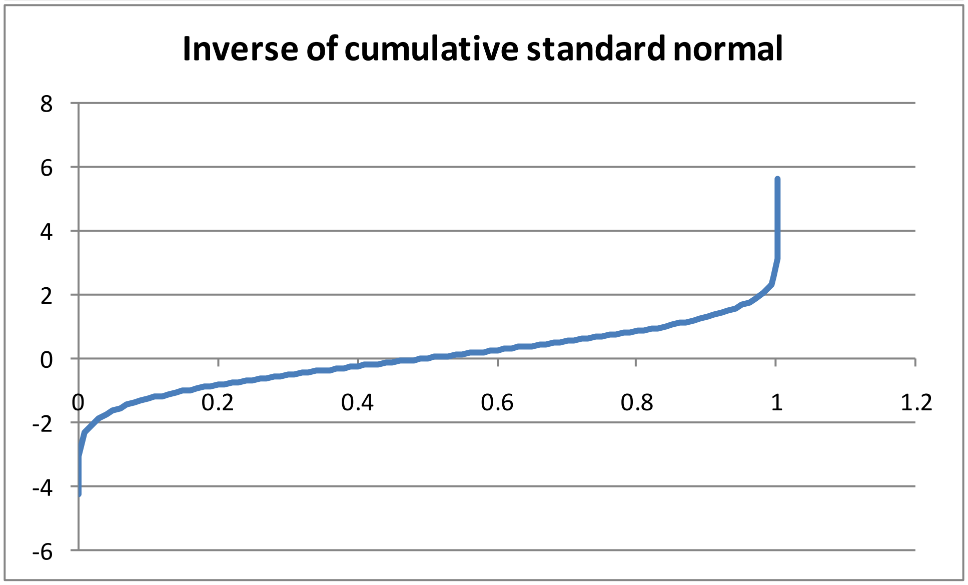
If you repeat a measurement you seldom get exactly the same number twice. In the mathematical sense what you measure is not a numerical variable, but a *stochastic one*. That means that there is a *random* component to the measurement.

Random variables are *distributed* over many possible outcomes and these distributions can have many different shapes.

Open a new spreadsheet.

1. Type in A1: =Rand()
2. Put the cursor on the right hand bottom corner of A1 until turns into a **+** Then drag this + to the right to copy the contents of A1 to B1.
3. While the range A1:B1 is selected put the cursor on the bottom right corner of B1 and now drag it down to fill all the way down to row 512. This will fill the range A1:B512 with random numbers
4. Now activate A512. Hold down the Shift key (that *selects* things). Now press RightArrow once, then End then UpArrow (note how you can quickly to the top or bottom of a range using End)
5. The range A1:B512 should be selected now and you can go to Insert to make a chart of the range. Make a scatter-plot with only markers (no lines). Stretch it a bit until it looks more or less like a square.
6. Press the F9 button a few times to recalculate the random numbers. As you see they fill a square between x=0-1 and y=0-1 homogeneously. This is a *uniform* distribution.
7. Now activate A1 and change the formula to =normsinv(rand()).
8. Activate A1 again and use the drag the **+** trick to copy the contents to B1.
9. Select A1:B1 and double click on the **+** symbol that appears on the bottom right of the range when you put the cursor on it. This should fill the new formula over the entire range A1:B512.
10. What happens to the chart? Press F9 a couple of times. The formula =normsinv(rand()) represents the inverse of the cumulative integral F over a normal distribution. As probability integrates to unity this function F increases from zero at x=negative infinity to one at x=positive infinity. If you pick a uniform random number (=RAND()) and read back on this function (i.e. use its inverse) you get a random number of a normal distribution.





11. Excel has more inverse cumulative functions that can be used to generate random numbers of various distributions from the uniform ones that RAND() produces.
12. Goto B1. Hold down Shift press End and ArrowDown. Now press Ctrl+x to cut. Press Ctrl+G and type in the address A513. This will take you to this cell (useful for big spreadsheets!). Now press Ctrl+V to paste. We now have 1000 replicate numbers with a normal distribution N(μ,σ^2)=N(0,1).
13. A good way to navigate back is: press Ctrl+Home
14. Type in D1 = average(A1:A1024). Type in E1 =stdev(A1:A1024). These values should be close to (but not exactly) 0 and 1, because they are (Least Squares) *estimates* for μ=0 and σ=1, the values for the standard normal distribution N(μ, σ ^{2}).=N(0,1).
15. Type in F1 =E1/sqrt(1024). This is the *standard error* that indicates how good the estimate for μ is. Press F9 a couple of times and examine the values in D1 (the

sample average) and G1 (the standard error of this average). The value in D1 should not differ from zero by much more than F1.

16. There are other ways of estimating μ and σ. Type in D2 =median(A1:A1024) to estimate μ.
17. Type the following *array function* in E2: =median(abs(a1:a1024-d2)). **Caution:** you need to use Ctrl+Shift+Enter to activate *array functions*.
18. *Excel hint:* The dollar signs in the formula mean that the row and column reference is made *absolute*, i.e. upon copying into another cell they remain the same. There is an easy way to modify references: type in D4: =a123 and while the cursor is still behind what you have typed in the cell press the F4 button a few times. Delete the contents of D4 when you are done.)
19. The two values in D2 and E2 are robust estimates for μ and σ: the sample median and the median absolute deviation (mad). Use F9 to study their behavior. You may see that the mad is too small. Edit its formula to
20. E2 =median(abs(a1:a1024)-d2)*1.483 and use Ctrl+Shift+Enter to activate the *array function* again.
21. A good habit is to *label* your numbers, otherwise you do not know what they mean anymore: Type in C1: Least Squares; in C2: Robust; in D3: Mean; in E3: Standard deviation; in F3: Standard error
22. Put the cursor on the border line between the header cells that say C and D. A symbol appears that looks like <-|-> Then double click. This should adjust the width of the column.
23. Now you will see that the Least Squares estimates for both μ and σ are close to the robust ones and close to zero and one.
24. Change the contents of A1 to 3000 and A2 to -100000.The values represent *outliers*.
25. What happens to the Least Squares estimates? And the robust ones?
26. Undo the changes in A1 and A2 (Use Ctrl+Z twice)
27. The random function is 'alive' in that it gets recalculated anytime you change the spreadsheet. This can slow down a lot of calculations, if you want to use the numbers for some purpose. Let's fix (freeze) one set of numbers. Select the whole A1:A1024 range (e.g. use the Shift, End and Arrow keys). Now press Ctrl+C to copy, right click and go to Paste Special, then opt for Values. Now the formulas are replaces by just numbers.
28. Let's create a suitable bin range to make a histogram. The bins are boxes (intervals) into which you sort your data. Then you count how many points are in each box and represent that in a bar graph. Choosing your bin ranges can be a bit tricky. Let's first determine the largest and smallest values in our set.
29. Type in D5: =max(A1:A1024) and in E: =min(A1:A1024). The values should be something like: 3.54837 and -3.123 or so. So we need boxes between say -3.5 and +3.5 and decide on a box size. If we make that 0.2 we'll get (3.5-(-3.5))/0.2 is 35 boxes. That should work fine for 1024 points
30. Type in D7: -3.5 and in D8: -3.3. Now select D7:D8 and put the cursor on the + corner and drag it down until you reach the value +3.5.

31. Go to Data on the ribbon (in Excel 2007) or the tool bar (in earlier versions). If the data analysis add-in is loaded there will be a data analysis option. Click it and opt for histogram. On the popup define the input data as A1:A1024 and the bin range as the one you just made. There is a button with a red dot in it that allows you to select the range by painting with the mouse, but you can also type in the range.
32. (If the analysis pack is not loaded follow the instructions in the Excel Appendix all the way at the end of this manuscript)
33. Opt for chart output at the bottom and go (OK button).
34. Do you get a Gaussian distribution?
35. Select the data in A1:A1024, copy them (Ctrl+c) and paste them in a new sheet (Sheet 2). Use paste special "Paste as Values" to make these numbers rather than the result of a function call.
36. While they are still selected go to the Data option of the ribbon or toolbar and use the sort option to sort your data by size (A->Z will do).
37. In B1 type =row() and double click the + of the right hand corner to copy the formula down to b1024.
38. As you see this function simply numbers the cell by their row (there is also a COLUMN() one). Change the function in B1 to =ROW()/1024 and double click the + to fill. We now have a fraction (essentially a *percentile apart from a factor of 100).*
39. Now select A1:B1024 and make chart with only straight line segments. The easiest way to do that is as follows. In Excel 2007 you go under Insert on the ribbon, click the *Scatter* icon in the chart group and then click the bottom icon that shows straight line segments rather than markers or both. There are also two icons that produce rounded lines (splines). *Never* use those for scientific data: the splines mess up your data in uncontrollable ways. In earlier versions of Excel there is a *chart wizard icon* that leads you to much the same choices.
40. By sorting your numbers, you get what is known as the *order statistics* of your 'measured' data. Plotting the order statistics give you an idea (estimate) of the *cumulative distribution function* of the data.
41. Let's check that! Type in C1: =normdist(A1,0,1,1). This gives you a Gaussian function with mean zero and standard deviation unity. Double click on the + at the right hand bottom of C1 to fill the formula down to the bottom. Now click on the graph. A blue box appears around the data in the B column. It should have handle at the bottom right corner and in newer versions also at the top right corner. Use that to drag and include the '*theoretical'* values in the C column into the graph. To compare these values to the random values in the A column copy columns A, B and C and use paste "special" (as values) to another location (e.g. E, F and G). Then sort the values in column G (those from the Gaussian function). The others should already be sorted, but you could also simply sort all three columns to make sure. Now plot all three columns, meaning you are plotting B vs. A and C vs. A to show that the normalized distribution of the random numbers is equivalent to a Gaussian.

42. The two curves should be very similar but not identical. In fact, it is possible to subtract the two and take the largest deviation. There are tables that will tell you if that deviation can still occur just by chance or that the curve really does not fit. Such a test is called the Kolmogorov-Smirnov test. Unfortunately, Excel does not have a probability table for this built in (as it does for the t-test). Real statistics software like SAS etc. does have that. Such a test allows you to see if your data are actually normally distributed or not.
43. Go to a new sheet (Sheet 3) now and fill two columns, say A1:B512 with normally distributed numbers using =NORMSINV(RAND()) and make a chart as before.
44. In D1 type =correl(a1:a512,b1:b512). This calculates the *correlation* between the values in the A and the B column. Pressing the F9 key will show you that the correlation is quite small. In fact it should be (close to) zero.
45. Now replace the formula in B1 by: =$C$1*A1+ NORMSINV(RAND()) and use the + trick to distribute the formula over the whole B1:B512 column. Delete whatever is in C1. As long as C1 is empty it should not make much of a difference, but start making the value of C1 larger: first 1 then 5,10 and 100. What happens to the graph, what happens to the correlation?
46. Now make the formula into =A1^2+ NORMSINV(RAND())^2. Why is the correlation close to zero now? Also try =A1* NORMSINV(RAND()). Again there should be hardly any correlation. Does that also mean that there is no relationship between the two rows?
47. What is the correlation between A and B for these data and why? (Make chart!)

1	0
-0.5	0.866
-0.5	-0.866

Tutorial 1 Quiz

Let's understand the relationship between the binomial distribution function and the Gaussian.

1. Using a new Excel document construct an array from 0 to 10 in A1:A11
Type 0 in A1 and then A2:=A1+1. Copy A2, select A2 to A11 and paste the formula in the cells to complete the array.

2. Calculate the binomial distribution in cells B1 to B11 for each point using the syntax FACT(10) as the factorial of 10.

The formula you need is

$$B_n = \frac{10!}{(10-n)!\,n!}$$

Construct this formula in B1, copy B1 and then select B1 to B11 and paste the formula into the entire array.

3. You can plot this by selecting A1:B11 and clicking the ScatterPlot under Chart

4. What does the function look like? When is the area underneath the curve you have created? You may use the SUM function to calculate the sum of the points B1 to B11. In B12 type B12:=SUM(B1:B11). What value did you find? Can you see any relationship between this value a power of 2^n?

5. Let's create a normalized function. In C1 type C1:=B1/$B$12
The $B$12 takes the fixed value in B12, which is the sum (i.e. in this case the area underneath the curve. This is the area since the spacing between each point on the x axis $\Delta x = 1$. Copy C1 and paste it into C1 to C11. What is the sum of C1 to C11?

6. Repeat these steps for an array of 60 points. Now let's compare the binomial function with a Gaussian. The Gaussian you need has the form.

$$G_n = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(F1-30)^2}{2\sigma^2}\right\}$$

Assuming that you have an array from 0 to 60 in F1 to F60 then the Gaussian function can be constructed using the above formula. Plot the normalized binomial distribution for 0 to 60 and the above Gaussian on the same plot. But, what value should we use for σ? Here σ is a constant that determines the width of the Gaussian. By trial and error match the Gaussian to the binomial distribution. Hint: Try σ = 1 and then σ = 10. Does it look reasonable? What value of σ looks the best. Final question: Is the above Gaussian normalized?

7. What is the normalization factor for the binomial distribution function?

Tutorial 2. Calibration lines

Calibration is a very important procedure because it is the standard way to remove *systematic errors* from measured data. It is also the only way to make sure that your scale of units corresponds to everybody else's. The basic idea is very simple:

1. Take a sample with ***known*** properties (a standard)
2. Measure it with your instrument
3. If your measurement gives a wrong value, correct it

Of course you are not interested in one value but in a *whole scale* within the dynamic range of your instrument and so in general we have to use a *series* of standards and correct the whole scale. This is typically done using *simple linear regression* although there are far more elaborate schemes possible. In this lab we will explore some of the properties of calibration curves

Instructions

- Open up the workbook trumpets.xls. Make sure macros are ***enabled***. Microsoft tends to disable everything.
- Sheet 1 contains two buttons and a two column range of data points that represent a series of measured standards.
- Select the range and click the ***trumpet*** button. The button activates a macro that calculates a simple linear regression using the Linest function. The output of this procedure is summarized in the block in the J and K column. At the top of this block e.g. you will find the values for the slope and the intercept of the calibration line.
- The formulas underneath the heading: ***fit, Conf95%+, Conf95%-, Pred95%+, Pred95%-*** should be selected. Put the cursor on the bottom right corner of that range until it changes into a **+** and then double click. This should fill down to the last calibration point.
- Now select the entire data block including the empty cell above the first data point and the headers on the first row. Make a chart: a scatter plot with *only markers*
- While the chart is active, click the ***decent trumpets*** button.
- There is a green straight line. This is the ***calibration line***. It is what you use to correct you measurement with.
- Typically what is on the horizontal (X) axis here are the ***calibration values*** (the 'right' ones). Vertically you have the measurement (Y). The dimensions are *not* necessarily the same.

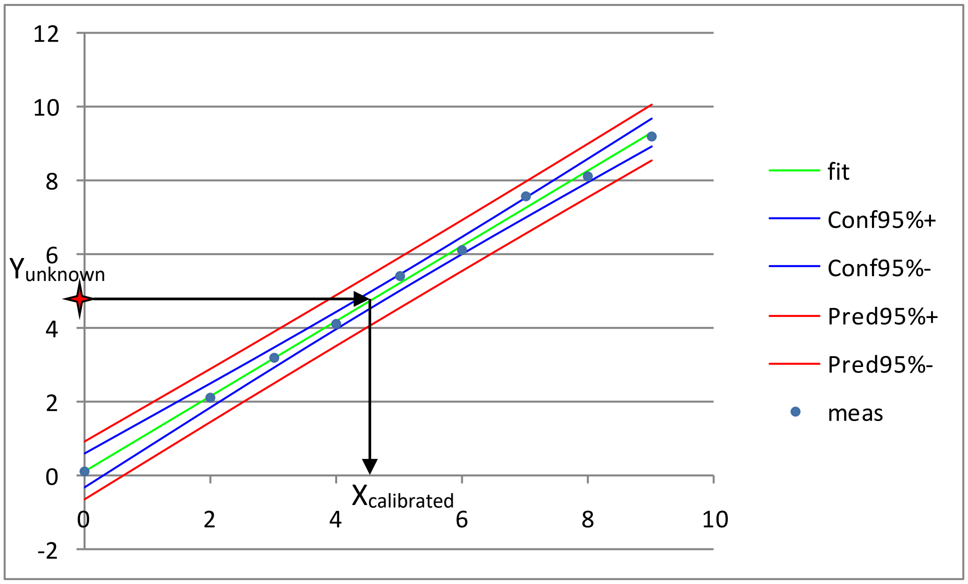
Questions:

1. Suppose we are calibrating a UV/VIS spectrophotometer and measure absorbance at a wavelength of 400 nm. We want to know if an accused

person has actually put a red poison in someone's drink. We have made up a number of solutions of that poison with known molarity and measured them. What are the units on the vertical and on the horizontal scale? What does Y represent?

2. The calibration line can be written as $Y_{cal} = b_{intercept} + m_{slope}.X_{standard}$. What are the units for the intercept and the slope in this case?
3. Inspect the formulas in E11, F11, G11 and H11. Activate each cell and then click behind the formula that appears in the formula bar above the sheet. Write out the formulas in mathematical format in terms of the quantities given in the statistics table. Compare to the statistics handout in the CH452 manual. What is the difference between the formula for the prediction and the confidence hyperbolas?

- Notice that if I measure an ***unknown*** sample, what I do not know is the poison concentration $X_{unknown.}$ All I can do is measure its Y absorbance value. To arrive at a concentration value I have to read back, i.e. ***we need to invert***: $(Y_{unknown} - b_{intercept})/ m_{slope} = X_{calibrated}$



- Suppose $Y_{unknown}$ is measured to be 4.342. Use the slope and intercept values in the Linest block to calculate $X_{calibrated}$. (The intercept is on the top right of the range; the slope is on the left).

Of course the above totally ignores the fact that both in the calibration measurement and in the measurement of the unknown there are *inevitable uncertainties* (read: random errors). This is why I have added the red and blue 'trumpets'. They may look like straight lines but they are really hyperbolas.

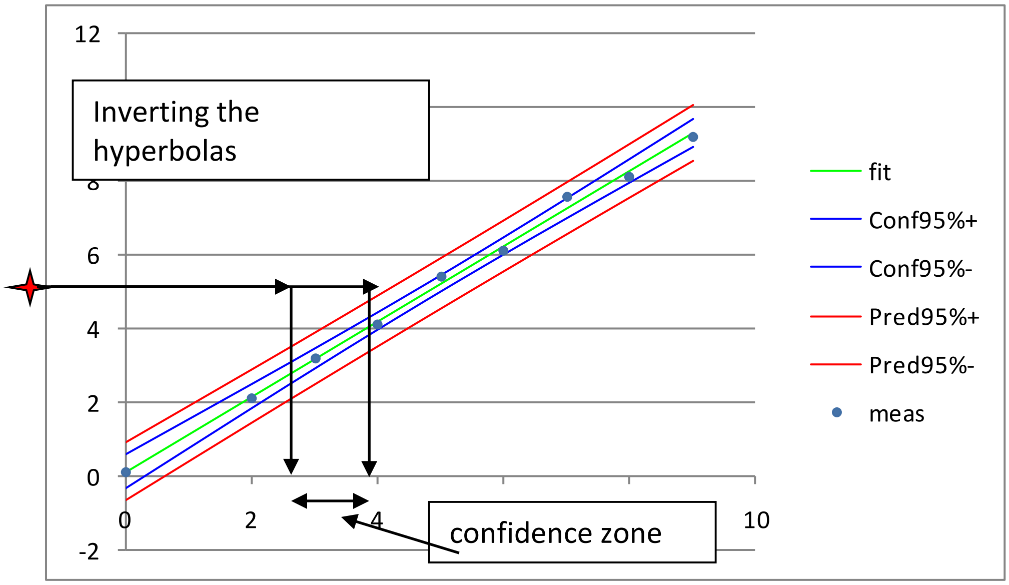
- The calibration set contains a *blank* measurement, i.e. one where X=0. Let's make a gross error there. Change its measured value to 10. As you see that really screws up things: the calibration line no longer passes through the data points but the hyperbolas become much clearer. Change the value at X=0 back (Ctrl+z).

The two trumpets

There are ***two sets*** of hyperbolas:

1. The area between the ***inner*** blue curves, known as the ***confidence limits (of the line)*** represents the zone within which you would expect ***any new calibration line*** to appear, if you measured the same standards again.
2. The area enclosed between the ***outer*** red curves, known as the ***prediction limits (around the line)*** represent the zone within which you can say ***any new data point*** will appear and be right about it 95% of the time.

So, whatever measurement we do we expect it to come within the outer trumpets, as long as the data quality and instrument settings etc. do not change. Therefore, we can use the outer curves to find the error in the calibrated X value of an unknown by a read back procedure much like what we did above:



We know that the point on the Y scale must have come from between the outer trumpet, so if we ***invert*** the hyperbolas we should get the lower and upper 95% confidence limits of the calibrated value we found above. ***Caution***: the nomenclature is very confusing: you use the *prediction limits for a point (around the line)* to find the *confidence limits (of the point)*.

You may wonder what the inner limits mean: they represent the **systematic component** or **the calibration error.** If we were to replicate our unknown measurement, we can improve our uncertainty by averaging and thus reduce the width of our confidence zone, but the calibration error would remain the same as long as we keep using the same calibration line. Thus the inner blue part does not average out no matter how many replicates we measure.

Question:

You may notice that the band between the trumpets is not equally narrow everywhere. Where would you get the best results? What happens when you move to higher or lower concentration values? What happens to the systematic component?

Unfortunately the exact inversion of a hyperbola leads to horrible algebra, but in a spreadsheet you can do it graphically or by preparing a look up table

- Type in O11: 0; type in O12: 0.001
- Select O11:O12 and put the cursor on the corner until **+** appears
- Drag the **+** down to O1011. This should fill O11:O1011 with numbers increasing in steps of 0.001
- Go back to where the calibration data are and select the formulas in the first row beneath the heading : ***fit, Conf95%+,Conf95%-,Pred95%+.Pred95%-*** (starting under ***fit***). Hit Ctrl+c to copy
- Go to Q11 and paste (Note: it is important to skip a column for consistency in the formula)
- Use the **+** double click trick to copy the formulas down to the bottom of the region.
- Select O11:U1011 and make a scatter plot with only markers
- Use the ''Decent trumpet" button to clean it up

We now have values for the calibration line and the trumpets that are not limited to where we took our calibration standards.

- Select O11:O1011 (Go to O11; hold down *Shift*; press *End*; press *Arrow-Down*)
  Or on newer version use *Shift; Ctrl; Pgdn* then copy (Ctrl+c)
- Goto V11 and paste

We are now ready to look up values in order to determine the width of the distribution in the x-direction. See the figure above called "Inverting the hyperbolas". What you are doing is graphically trying to find the place where a certain value of y cuts across the line and across the outer trumpets. For example, suppose we wish to know the inverted values for a measured value of 0.44. In other words, what is the predicted concentration (if we assume that 0.44 is absorbance)? And what is the 95% prediction error? For that we will need to look at where the number 0.44 appears in the table in three places. Problem: Determine the concentration (x-value) when y = 0.44 and determine the prediction error.

You can also use a lookup function to find the values.

- Type in W7 = 0.44

We can use a function called Vlookup. You give it the value you want it to look up, then a range (table) in the first column of which it looks up your number and then the column with the values you want returned:

- type in X7: =VLOOKUP(w7, $Q$11:$V$1011, 6)
- type in Y7: =VLOOKUP(w7, $T$8:$V$1011, 3)
- type in Z7: =VLOOKUP(w7, $U$8:$V$1011, 2)

You can see that the values are close to those you found by inspection. However, the lookup function cannot see things such as the fact that the value of x is halfway between to values of y-calc.

We now have the calibrated values in the X column and the lower and the higher 95% confidence limits in Y and Z. Unfortunately there is a bit of a problem. Use the AA and AB columns to calculate the distance δ_+ and δ_- from the calibrated value $X_{calibraed}$ to the upper and lower limits. (=X4-Y4 and =Z4-X4). As you see the error margins δ_+ and δ_- are *not* quite the same. This implies that the statistical distribution around $X_{calibrated}$ is no longer strictly Gaussian! This is an inconvenient truth that is conveniently *ignored in science*. Just remember: almost all data in science are obtained through calibration, so that this would mean that scientific data is generally *not* normally distributed. Fortunately the deviation from symmetrical is pretty small, particularly if the trumpets are narrow and typically people use a *symmetrical approximation formula* that can be computed from the statistics in the statistics block:

$$Approx.std.error = \frac{RMSE}{|slope|}\sqrt{1+\frac{Nx^2+\sum_i x_i^2 - 2x\sum_i x_i}{DD}}$$

Where

$$DD = N\sum_i x_i^2 - \left(\sum_i x_i\right)^2$$

$\delta_+ = \delta_-$= t-value*approximate standard error.

$$x = X_{calibrated} \ \ and\ N = df + 2$$

- Use the statistics in the regression block to compute these error margins

Notice that we use the Student t value in this approximation formula to multiply an approximate estimate for the standard error of $X_{calibrated}$ (the rest of the formula), happily assuming we can treat it as normally distributed. We then can report the result either as

: $X_{calibrated}$(approx. st. error) in 2/15 format

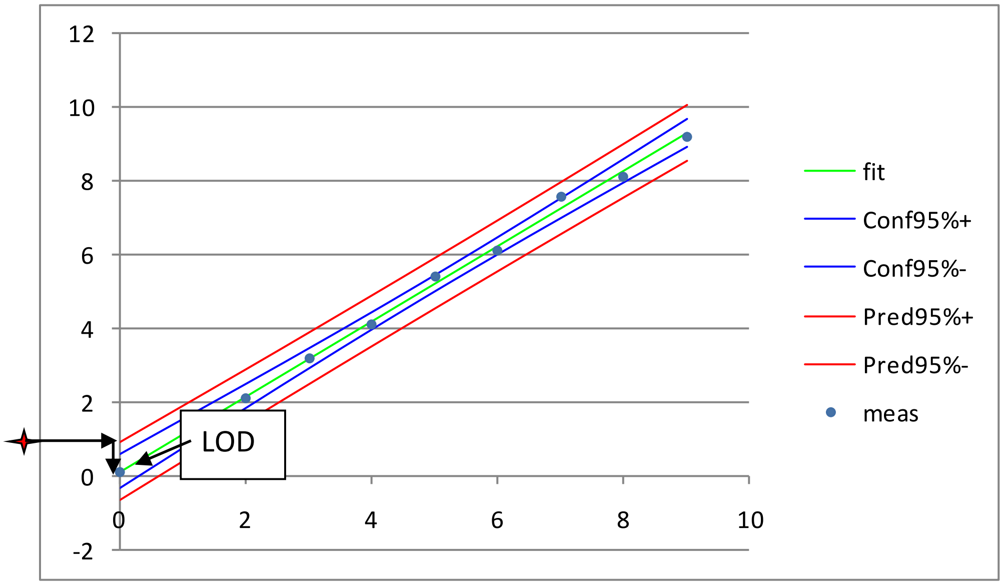
Or, and the ISO 9000 laws often *prescribe* that it must be given as:

: 95% confidence limits are: $X_{calibrated}$± δ.

Whether you get 95% limits or say 99% depends on what t-value you use. That is easy to change in the cell labeled as t-value: change =TINV(0.05,*xxx*) into =TINV(0.01,*xxx*) and watch what happens.

The limit of detection

Examine what the intersection point of the upper prediction hyperbola is with the Y-axis. This value has a special meaning. It is called the limit of detection (LOD). Note that you can estimate it using the 95% confidence limit as shown using the graph below.



When you measure such an absorbance value you cannot really say much about your sample. The confidence limits now contain X=0. That means they also contain X=0.1 of 0.001 or 0.0001 or 10^{-10}. That is to say that:

1. you do not even know if your poisonous compound is actually there
2. *if* it is there, you do not even know at what scale it is there
3. all you know it is not more than the Limit of Detection

This puts you in the position of a jury trying to decide whether an accused person should be condemned or not on insufficient evidence. They can make two different kinds of errors: they can send an innocent man to jail of they can let a crook go. In either case there is a crook on the loose!

The limit of detection obviously depends on the confidence level I take. If I opt for 99% the bands will be broader. So what do I take? There is a *trade off* here. If I am pickier and insist upon 99% or 99.9% confidence my chances of sending an innocent man to jail will diminish (LOD will be higher), but I'll let more crooks go. The only way to diminish both types of error is to get better data.

Question

Let's use the data in this worksheet to determine the limit of detection. Repeat the steps above used for the estimate of the confidence zone, but use a spacing of 0.001 (instead of 0.01). To determine the LOD you need to determine the value of the upper outer trumpet when x = 0.0. This can easily be seen from the table. Once you have found this value now you need to find the value of x when the lower outer trumpet has this same value. Using this approach estimate the LOD.

Standard addition

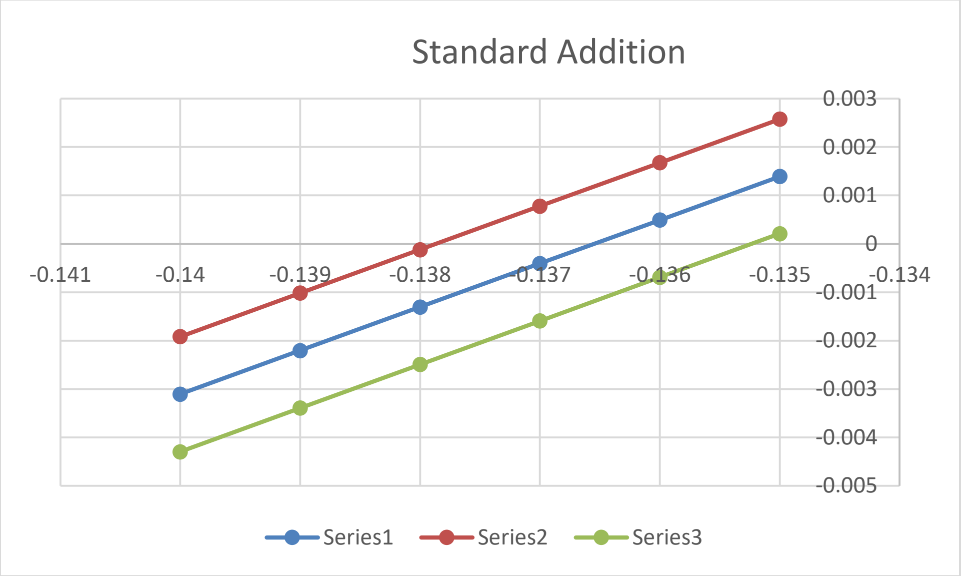
Go to Sheet 2 of the workbook. It contains a number of measurements of the atomic absorption of calcium in milk. A known amount of calcium solution was added to the sample, a method called ***standard addition.*** The concentration given is the *added* concentration. In addition each sample contains a contribution from the sample itself. Select the data range and use the two buttons to make a calibration graph with decent error trumpets.
To find the concentration of the unknown you need to back extrapolate the calibration line to where it intersects with the concentration (x-) axis, so:

$$intercept + slope\ x = 0$$

$$x = -\frac{intercept}{slope}$$

Calculate this value of x. Now to find the errors in x (i.e. the errors along x axis) we will need to extend the trumpets down to that value and even past it until both trumpets cross the x-axis. To do this generate a column with x values in small steps around it. You will not necessarily know how far to go so just try some values. Then copy the functions for the fit line and the two inner confidence limit trumpets and use them to calculate their values and make a graph. The idea is to find the points where the *inner* trumpets

cross the horizontal axis graphically (i.e. where they cross the x-axis) to find the 95% confidence limits of the concentration of our unknown sample.
The resulting graph will look something like this.



Notice the blue line (fit) and the trumpets (red-brown and green) all cross the x-axis. Now you can read the errors in the estimated value for standard addition by determining where the trumpets cross the x-axis.

$\delta_+ = \delta_-$= t-value*approximate standard error.

There is also a symmetrical approximation formula you can use.

$$Approx. std. error = \frac{RMSE}{|slope|}\sqrt{\frac{Nx^2 + \sum_i x_i^2 - 2x\sum_i x_i}{DD}}$$

Where

$$DD = N\sum_i x_i^2 - \left(\sum_i x_i\right)^2$$

Use it to check your results against the graphical method. All the statistics are already in your sheet.

Question: why would we use the inner trumpets in this case?

Question: Why do you need the value of the lower inner trumpet at x=0 to be positive?

Question: Why is it undesirable if the slope of the line is small?

Tutorial 2 Homework

Part 1. multiple choice questions.

1. **Ordinary calibration**. You may notice that the band between the trumpets is not equally narrow everywhere. You get the best result around the center of gravity of your calibration points.
Question: What happens when you move to higher or lower concentration values, i.e. away from the center?

A. The resulting uncertainty gets larger because the vertical distance between the two outer trumpets gets larger

B. The resulting uncertainty gets smaller because the horizontal distance between the two outer trumpets gets larger

C. The resulting uncertainty gets larger because the vertical distance between the two inner trumpets gets larger

D. The resulting uncertainty gets larger because the horizontal distance between the two outer trumpets gets larger

2. **Ordinary calibration.** You may notice that the band between the trumpets is not equally narrow everywhere. You get the best result around the center of gravity of your calibration points.
Question: What happens to the *systematic component*, i.e. the *calibration error* when you are far away from the center?

A. The widening of the outer trumpets indicates that the calibration error becomes less significant

B. The systematic component, indicated by the inner trumpets becomes the dominant contribution so that the calibration error dominates any random contribution

C. The outer trumpets become wider but this is the result of random errors in the measurement of the unknown

D. The systematic component of the error is constant for all values of the calibration line

3. **Standard addition. Question:** What should we use to find the confidence limits of the final measurement?

A. Because we measure the same unknown multiple times we cannot use either of the trumpets because we need to construct an outer trumpet for more than one replicate

B. We are measuring the same one unknown sample over and over, we should use the outer trumpets

C. We are using all the data at once to find the intercept of the calibration line with the Xaxis. The inner trumpets determines where we expect the line to be

D. We should use the standard errors of the slope and the intercept and do error propagation

4. One requirement for Least Squares regression to be successful is that the matrix ($\mathbf{X}^T\mathbf{X}$) has an inverse. This depends on:

A. The software you are using

B. The quality of the data set

C. Whether or not the data are homoschedastic

D. The design of the data set, i.e. your choice of independent variables

E. Whether or not the data set contains outliers

F. Whether the equipment is properly calibrated or not

5. In matrix notation a set of data points can be written as $\mathbf{Y} = \mathbf{X}.\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
However this equation reduces to $\mathbf{Y} = \mathbf{X}.\boldsymbol{\beta}$

A. when the parameters ε are chosen such that the sum of the squared residuals is minimal

B. when the parameters are chosen such that the sum of residuals is minimal

C. when the calibration is applied

D. when the parameters β are chosen such that the sum of the squared residuals is minimal

E. when the parameters are chosen such that the squared residuals are minimal

6. In matrix notation a set of data points can be written as

Y = X.β + ε

The symbol **ε**

A. Stands for the random error component; it is assumed to be normally distributed as N(0,σ2)

B. Stands for the estimated parameters

C. Stands for the bias due to the calibration error

D. Stands for the random error component and can have any symmetrical distribution

Answers

1.________ 2. ________ 3. _________ 4. _______ 5. _______ 6. ________

Part 2. Numerical Analysis

We will study the fundamentals of Ordinary Least Squares. Let's make some data.

1. First generate an array of 0.0 to 0.5 in A1:A6 using the method used in the last homework. Type A1:=0.0 then A2:=A1+0.1. The copy (crtl c) A2 and paste (crtl v) into the array A2 to A6.

2. Then let's make a line with noise in B1 to B6.

Type B1:= -5 + 12*A1 + NORMSINV(RAND())*0.4

Then copy B1 and paste it into B1 to B6.

$$\boldsymbol{y} = b\boldsymbol{x} + a$$

where b = 12 and a = -5. We have included a noise function with an amplitude of 0.4.

3. The method for obtaining the slope an intercept is as follows. For a line,

$$\boldsymbol{y} = \beta\boldsymbol{x} + \alpha$$

The vector y is the dependent variable while x is the independent variable. Here the **x** vector in A1:A6 and the **y** vector is B1:B6.

The averages are

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

and

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

You can obtain the averages using the function AVERAGE. In A7 type A7:=AVERAGE(A1:A6). You can copy A7 and paste it into B7 and you will also have the average for that array.

4. The slope is calculated using the formula

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}$$

In C1 type
C1:=(A1-$A$7)* (B1-$B$7)
and in D1 type
D1:=(A1-$A$7)* (A1-$A$7)
Then copy these two cells and paste them into C1:C6 and D1:D6, respectively.
Form the sum of each of these arrays.
C7:=SUM(C1:C6) and D7:=SUM(D1:D7)
Now, in E9 write "slope" in F9 type
E9:=D7/C7

5. The intercept is obtained from

$$\alpha = \bar{y} - \beta\bar{x}$$

In E10 write "intercept" and in F10 type
F10:=$B$7-F9*$A$7

Keep in mind that this method is only one route to solve this problem. We could think of this approach as “two equations and two unknowns”, where the unknowns are α and β. The matrix approach solves for α and β as a column vector with two elements. While it is not required you might try following the procedure in the lecture for using the matrix approach in Excel to obtain the values for α and β.

6. Once you have obtained α and β, you can calculate the correlation coefficient R^2.

$$R^2 = \frac{\sum(\hat{y_i} - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(\hat{y_i} - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Note that $\widehat{\boldsymbol{y}_i}$ are the calculated values from the values $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that you obtained from the ordinary least squares decomposition and the $\boldsymbol{y_i}$ are the original values obtained using a and b. We have also written this as $\boldsymbol{f_i}$, so $\boldsymbol{f_i} = \widehat{\boldsymbol{y}_i}$. Explicitly this means that we can calculate

$$f_i = \alpha + \beta x_i$$

And then use this calculated function to define three important quantities in the least squares approach. These are the sum of the squares of the residuals

$$SS_{res} = \sum (f_i - y_i)^2$$

The sum of squares of the regression line

$$SS_{reg} = \sum (f_i - \bar{y})^2$$

and the sum of square of the total

$$SS_{tot} = \sum (y_i - \bar{y})^2$$

Note that

$$SS_{tot} = SS_{res} + S_{reg}$$

Note the difference is that f_i is calculated and $y_i$represents the data. Using these definitions

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

You may calculate R^2 using either of these approaches using the data.

7. Plot the original line from A1:B6 using a scatter plot. The LINEST function (or click on trendline) to obtain a fit. How does the fit obtained using LINEST compare with your fit.

8. Calculate the 95% confidence limit for this line. Calulate the line again using a noise amplitude of 0.8 (instead of 0.4). How does the 95% confidence limit change?

Tutorial 3. Matrices and Complex Numbers

Actually both matrices and complex numbers become a lot more interesting (and fun) in a spreadsheet than they are on a blackboard or in a math test. And yes they are quite useful in quantitative science. In future labs we will see some more applications and you may need them in your project phase.

Rotation

- Start a new sheet and put the following data in A1:B5

1	0
2	0
3	0
4	0
5	0

- Type in C1: =COMPLEX(A1,B1) (This creates a complex number A1+B1.i)
- Type in D1: =IMPRODUCT(C1,$G$1) (*Hint:* Use F4 to put in the $ signs). This function multiplies two complex numbers.
- Type in E1= IMREAL(D1). (Guess what this does?)
- Type in F1= IMAGINARY(D1). (Guess what this does?)
- Select C1:F1 and double click the bottom right corner of F1 to fill down to F5
- Type in G1: =IMEXP(COMPLEX(0,H3)) (More about this in a minute..)
- Type in H3: =PI()*I3
- Type in I3: =0.25
- Select A1:B5 and make a scatter plot with only symbols
- Select E1:F5 and make a scatter plot with only symbols
- While the second chart is active, copy it using Ctrl+C
- Activate the first chart and *paste* using Ctrl+V
- Delete the second chart
- In the first chart change the axis range to -6 to +6 fixed values for both X and Y and stretch to make the graph look square.
- Press Alt+F11. This takes you to the VBA IDE
- On the left you should see all the VBAprojects currently open, including the workbook you are working in. Click that project to activate and go to the Insert menu. Insert a *module*
- In the middle pane a new window should open. Type the following program in it

```
Sub rotate()
a = Range("I3").Value
a = a + 0.1
Range("I3").Value = a
End Sub
```

- Either press Alt+F11 again or click on the small Excel icon top left to go back to your workbook
- Click your chart to activate and right click.
- On the popup assign the *rotate* module to the chart
- Click on an arbitrary cell, then on the chart and again and again.

Roots of one

The above demonstration makes use of a set of complex numbers with very special properties. You could call them the nth roots of unity, because they are the (complex) solutions of the equation:

$$x^n=1$$

There is always n such roots. These are known as the" roots of unity", which is the keyword you should use if you want to read more about these numbers in Wikipedia. Question: What are these roots for n=2? For n=4? To answer this question we should introduce the unit circle of radius 1 in the complex plane. The complex plane has a real axis (x-axis) and an imaginary axis (y-axis).

The *absolute value* or *magnitude* of the mth root: $|x_m| = \sqrt{}(x_m*x_m)$ of these numbers is always one and that means that in the complex plane they lie on the *unit circle* because the radius represents the magnitude |x|. In complex exponential notation the roots are easy to write:

: $x_m = \exp 2\pi i[m/n]$ where m = 1,2,3,4, …n

It is this function that we used to make things rotate in the demonstration above, because *multiplying* with such a number corresponds to a *rotation* in the complex plane. According to the *Euler rule* for complex exponentials,

$$e^{\pm ix} = \cos(x) \pm i\sin(x)$$

the real part of our root-of-one is cos2π[m/n] and the imaginary part sin2π[m/n].
To find the roots we will also need the De Moivre formula:

$$(\cos(x) \pm i\sin(x))^n = \cos(nx) \pm i\sin(nx)$$

Let's have look at these numbers using an Excel spreadsheet.

1. Go to a new sheet
2. Type in A1: 0; type in A2: 1;
3. Select the two cells, put the cursor on the bottom right corner of the selected range until it changes into a +. Then drag down to A15. (This should *fill* the range with the numbers m.)
4. In B1 type: =A1*2*PI()/15; (This gives 2π[m/n] with n=15)
5. In C1 type: =IMEXP(COMPLEX (0,B1)) (giving exp2πi[m/n]; COMPLEX(a,b) gives: a+b.i)
6. In D1 type: =IMREAL(c1) (The Re part cos2π[m/n])

7. In E1 type: =IMAGINARY(c1) (The Im part sin2π[m/n])
8. Now select the range B1:E1, put the cursor on the bottom right corner of E until a + appears and double click
9. The formulas should now have filled done to row 15. In the B column we have calculated 2π[m/n]. This value is also known as the ***phase angle***.
10. In the C column we have exp2πi[m/n] and in the D and E columns we have its real and imaginary parts cos2π[m/n] and sin2π[m/n]
11. Now select the range D1:E1 and make a scatter-plot with only markers, no lines. Provided you stretch the chart a little is should look like a perfect circle. (Make it so!)
12. In A17 type =CORREL(E1:E15,D1:D15) These are the roots of x^{15}=1. Look at the angles between them. How big are those and why are they so regularly spaced? What is the meaning of the contents of A17? Are the numbers random?
13. In C16 type =IMPRODUCT($C$3,C1). This calculates the product of the two complex roots in C1 and C3. Drag the contents down to C30 to fill. Select d15:e15 and double click + to fill down to E30. Make a second graph of D15:E30.
14. The graphs look identical but if you look at the numbers in rows 15 to 30 you'll see there is a difference. (What?). Activate the second graph and press Ctrl+c to copy, then click on the first graph and press Ctrl+V to paste.
15. Change the formula in C16 to =IMPRODUCT(COMPLEX(0,1),C1) and double click the right bottom corner + to fill.

    The function complex(0,1) represents the complex number 0+1i = i. Was this number one of the original roots? What happens if you multiply by it?

    In next week's lab we will see an important application of these numbers the Fourier Transform.

Let's do rotation in a different way by using a ***rotation matrix***.

- Go back to your first sheet
- Type in H5: =cos(H3)
- Type in I6: =cos(H3)
- Type in I5: = -sin(H3)
- Type in H6: =cos(H3)
- Select the range E1:F1, type the array function: =MMULT(A1:B1,$H$5:$I$6) (*hint:* you can type =MMULT(and then use the mouse to select the A1:B1 range, then type a comma, then use the mouse to select the H5:I6 range, then use F4 to put in the $$ signs). Activate this array function with Ctrl+Shift+Enter
- We have now multiplied our (x,y) coordinates with our rotation matrix. Use the bottom right + double click to copy the new function over all five rows
- Click on the graph to see what happens

As you see 2D rotations can be done both ways: complex or matrix. In 3D it gets more complicated. There are no 3D complex numbers, but there are 4D ones, known as quaternions and they do the trick in 3D. Most computer games use quaternions.

Euler matrices in 3D also exist but have a mathematical problem known as ‘gimbal lock’ which makes your program crash.

Using Matrix Commands in Excel

General Advice and Background

Matrices are used in many practical problems. For example, when you produce a diagram, most of the time, matrices appear in the diagram. More importantly, many calculations are most easily performed by using matrices. In this class, we will learn to enter and to perform various operations of matrices in Excel. We will use these skills to some practical problems in the next class. It is important that you know the basic properties of matrices in order to deal with them in Excel.

Entering and Naming Matrices

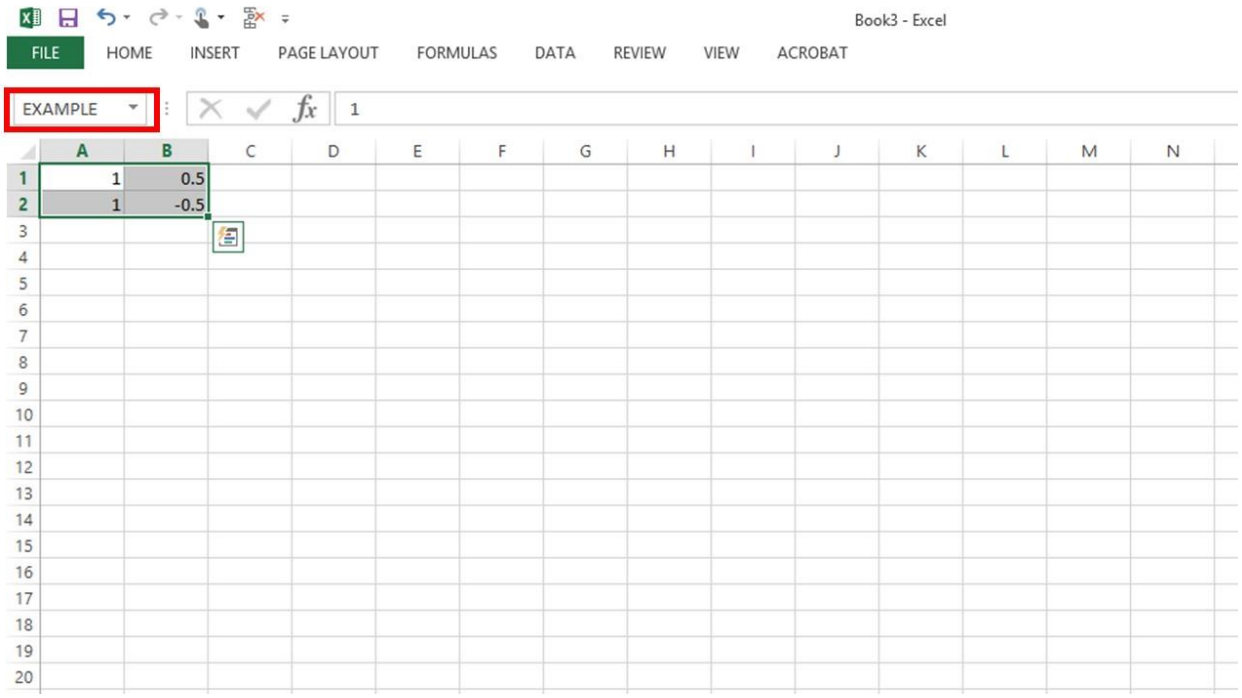
Before entering the elements of any matrix, you should always enter the **Matrix Name**, such as P, Q, etc. You may also want to enter the **Description** of the matrix, like Coefficient Matrix, Product of P and Q, etc.
Let

$$P = \begin{bmatrix} 1 & 0.5 \\ 1 & -0.5 \end{bmatrix}, \qquad Q = \begin{bmatrix} 2 & -1 \\ 1 & 4 \end{bmatrix}$$

To enter P, we can begin by entering the name in cell A1 and the elements in the block of cells A1:B2. To name the block of cells as matrix P, we do the following:

(a) Highlight A1:B2 with the mouse.

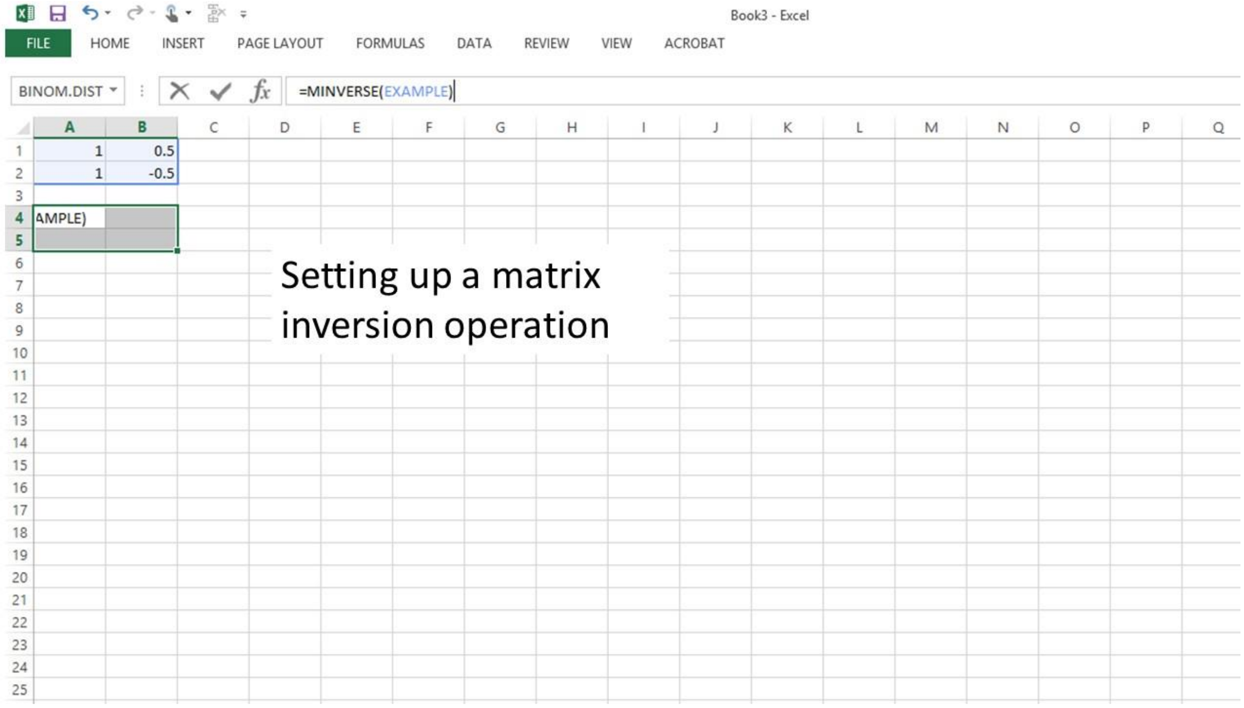
(b) Write the name of the matrix in the cell definition box. This is shown in the illustration below. The cell definition is in the red rectangle. In this case we can call the matrix EXAMPLE. We could have called it P or Q. However, some letters and even some combinations are reserved in Excel. Sometimes you will get errors if the name you choose conflicts. If you suspect such an error just choose another name.



EXAMPLE | 1

	A	B
1	1	0.5
2	1	-0.5

We can demonstrate the use of matrix commands in Excel using this matrix. In the figure below we have set up the matrix inversion. The command is MINVERSE() and the argument is the name of the matrix. Notice that Excel highlighted the matrix in blue because it recognized EXAMPLE as the name of the matrix (as we designated above).



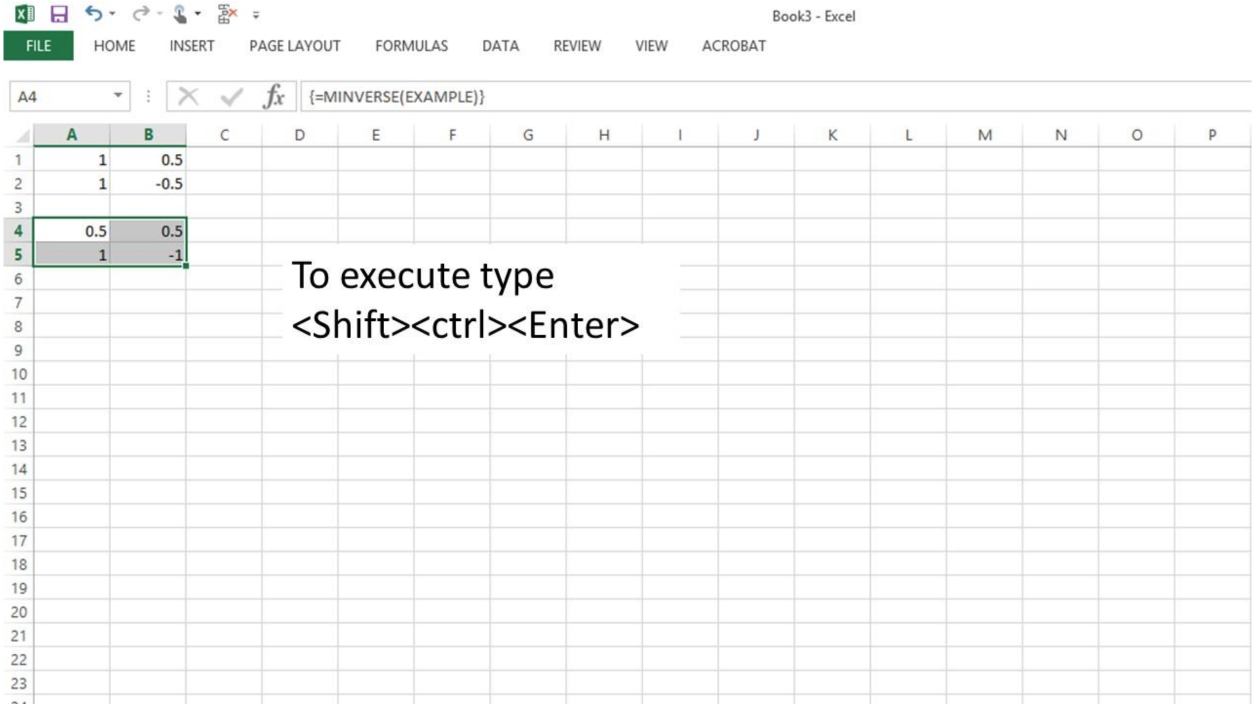
BINOM.DIST | =MINVERSE(EXAMPLE)

	A	B
1	1	0.5
2	1	-0.5
3		
4	AMPLE)	
5		

Setting up a matrix inversion operation

The execute the command and invert the matrix we need to type <Shift><ctrl><Enter>.

This is shown below. We see the result is a different matrix.



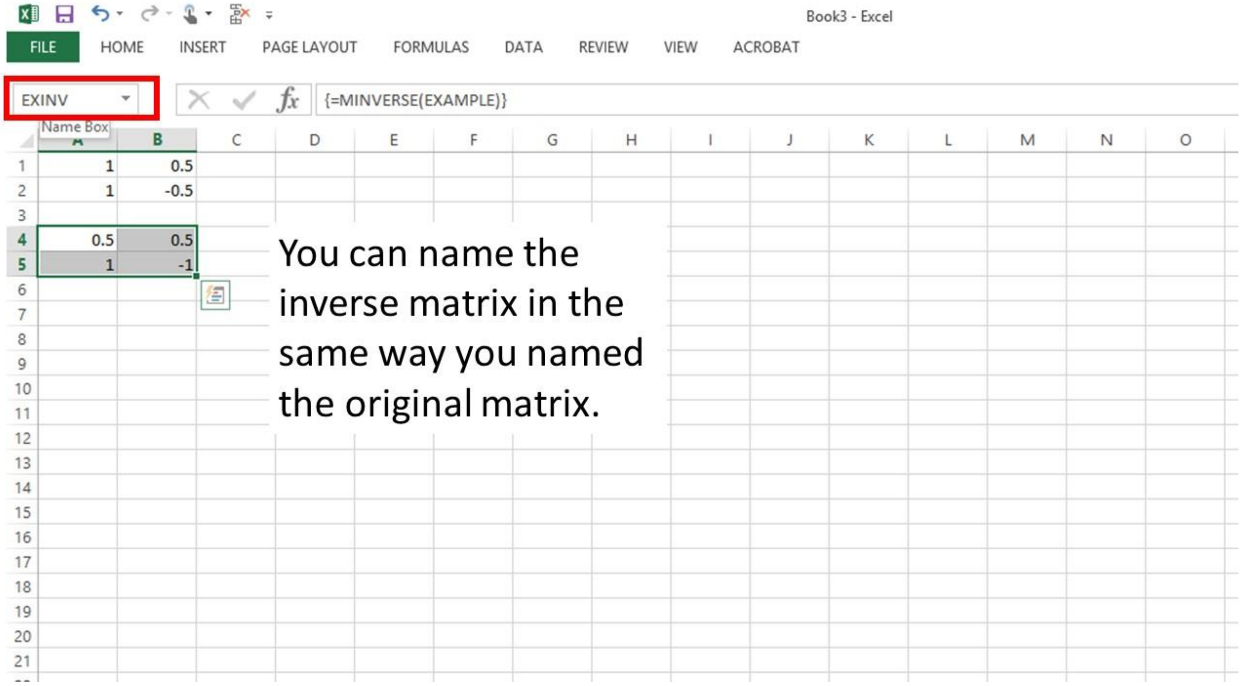
A4

{=MINVERSE(EXAMPLE)}

	A	B
1	1	0.5
2	1	-0.5
3		
4	0.5	0.5
5	1	-1

To execute type <Shift><ctrl><Enter>

First we will name it EXINV to represent that it is the inverse of the matrix EXAMPLE.



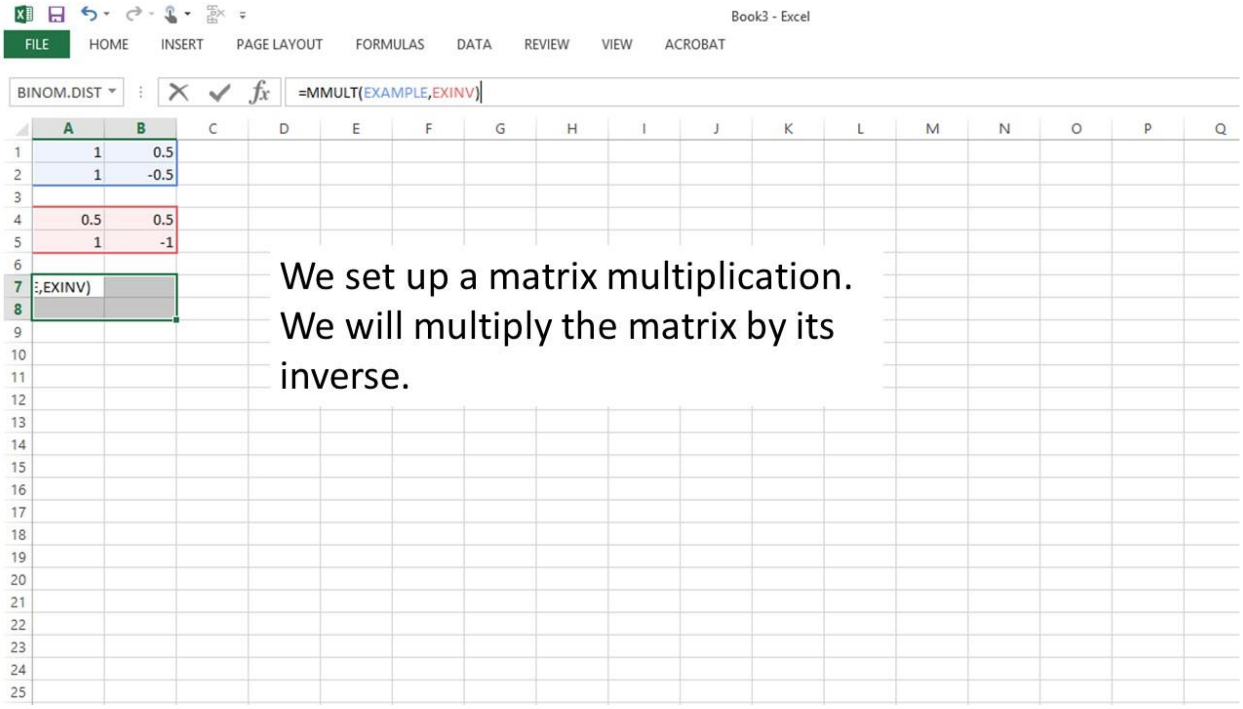
EXINV

{=MINVERSE(EXAMPLE)}

	A	B
1	1	0.5
2	1	-0.5
3		
4	0.5	0.5
5	1	-1

You can name the inverse matrix in the same way you named the original matrix.

Next we will test this matrix to see if it really is the inverse matrix. We will use the matrix multiplication function to multiply the matrix EXAMPLE by its inverse. Below we can see that we have set up the calculation and both matrices are highlighted, which means that Excel has recognized both of their names.

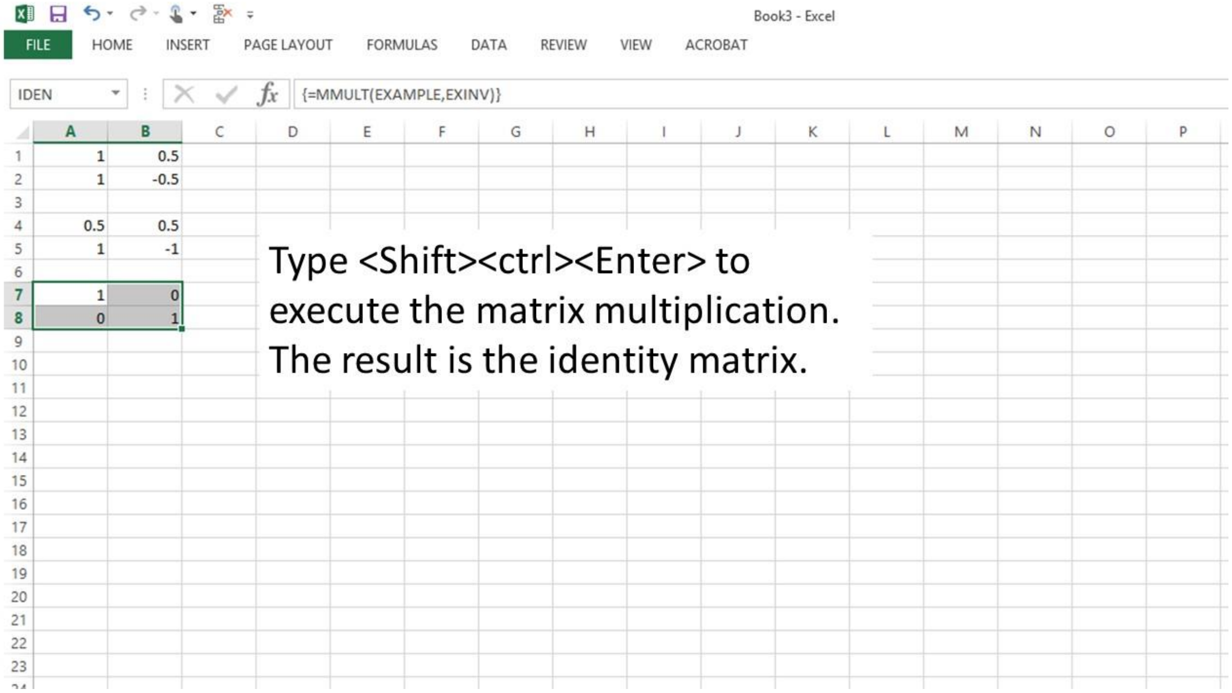


=MMULT(EXAMPLE,EXINV)

	A	B
1	1	0.5
2	1	-0.5
3		
4	0.5	0.5
5	1	-1
6		
7	E,EXINV)	
8		

We set up a matrix multiplication. We will multiply the matrix by its inverse.

Now we can execute the matrix multiplication by typing <Shift><ctrl><Enter>.



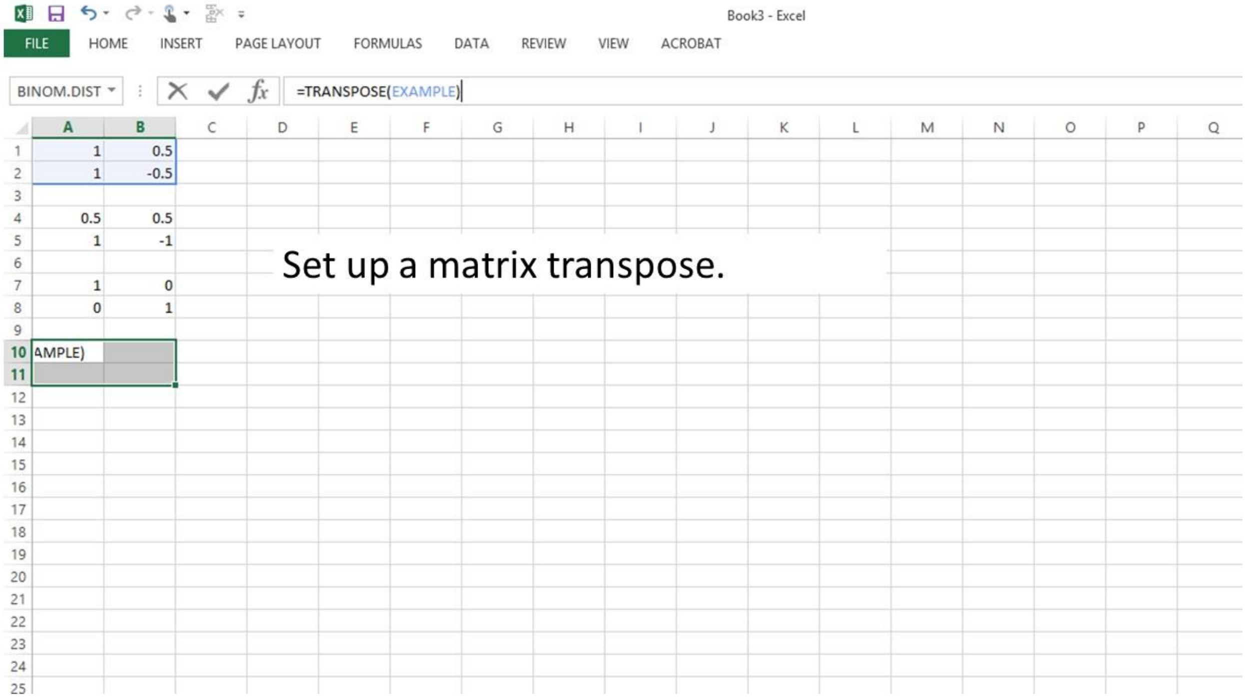
{=MMULT(EXAMPLE,EXINV)}

	A	B
1	1	0.5
2	1	-0.5
3		
4	0.5	0.5
5	1	-1
6		
7	1	0
8	0	1

Type <Shift><ctrl><Enter> to execute the matrix multiplication. The result is the identity matrix.

It is gratifying to see that the product matrix is the identity matrix. Thus, the matrix EXINV really is the inverse matrix. This is not so difficult to check by hand for a 2 x 2 matrix. But, try doing it for a 6 x 6! In Excel it is easy using the same commands we have used here.

We can also illustrate the matrix transpose function.

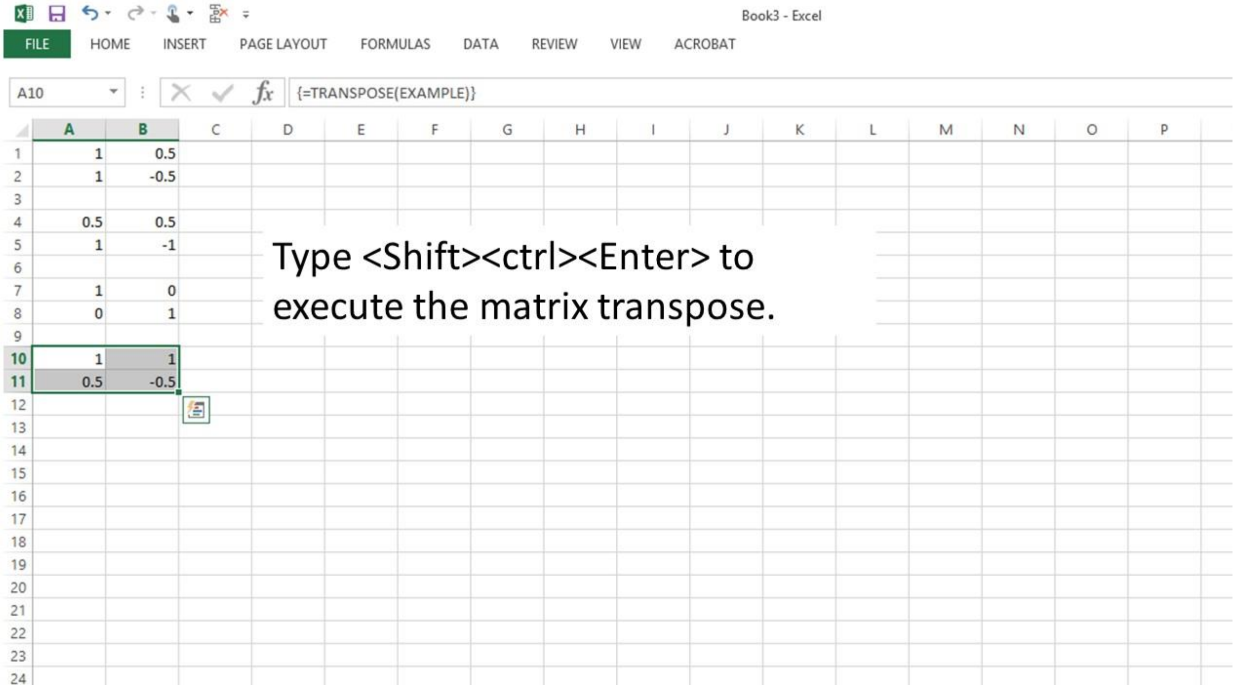


=TRANSPOSE(EXAMPLE)

	A	B
1	1	0.5
2	1	-0.5
3		
4	0.5	0.5
5	1	-1
6		
7	1	0
8	0	1
9		
10	AMPLE)	
11		

Set up a matrix transpose.

To implement we type <Shift><ctrl><Enter>.



{=TRANSPOSE(EXAMPLE)}

	A	B
1	1	0.5
2	1	-0.5
3		
4	0.5	0.5
5	1	-1
6		
7	1	0
8	0	1
9		
10	1	1
11	0.5	-0.5

Type <Shift><ctrl><Enter> to execute the matrix transpose.

In order to play around with the matrix operations, enter the two matrices above (P and Q). You can enter them anywhere you like. To keep things clear we should probably keep some spaces between them. So, for example we could enter the P matrix in A1:B2 (as above) and then the Q matrix in A4:B5. Please do this and then use the information below to implement various matrix operations.

Matrix Operations

We need to know the following rules on matrix operations:

P + Q *or* P - Q *can be performed if* P *and* Q *are of the same size.*

The multiplication PQ *can be performed if the number of columns in* P *is the same as the number of rows in* Q.

The ***transpose*** *of* P *can always be performed.*

The ***determinant*** *of* P *is defined only is* P *is a square matrix.*

The ***inverse*** *of* P *is defined only if* P *is a square matrix with nonzero determinant.*

As the matrices P and Q are of the same size, we know P+Q can be performed. To do this in Excel, we use the following steps.

(a) Enter P and Q (which have been done already).

In the cells above those which will contain the answer we enter a description and the name of this answer. For example, we can enter **sum of P and Q** as the description, and enter RS as the name.

(b) Use the mouse to highlight a 2 x 2 block of cells (as P +Q will be a 2 x 2 matrix) such as B9:C10 in which the answer is to be stored.

(c) Enter the Excel function

= P + Q

followed by pressing <Shift><ctrl><Enter>

(d) Name the new matrix RS as we did above for the example.

Note: *You may like to name* P + Q *as* R*, but you will find this is not allowed as R is reserved in Excel for other use.*

For the other operations, you follow the same steps with the appropriate changes. The Excel function for

A-B is	= A - B
AB is	= MMULT(A,B)

A^T (Transpose of A) is	= TRANSPOSE(A)
A^{-1} (Inverse of A) is	= MINVERSE(A)
det(A) (Determinant of A) is	= MDETERM(A)

Task: Find P – Q, PQ, P^T, Q^{-1}, det(P) and PT where

$$T = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

Tutorial 3 Homework

Part 1. Answer the following questions to verify your progress.

1. What does the function =IMREAL(X) do?

A. It calculates the product of the imaginary and the real parts of X

B. The syntax is really =IMREAL(a,b) and it produces a complex number with a as imaginary and b as real part

C. It renders the real part of the complex number X

D. It swaps the real and imaginary parts of X

2. What are the four roots of the equation $x^4=1$?
(Write out your answer in the space provided)

3. In step 12 you calculated the root of $x^{15}=1$. You could consider each root as a vector originating from the origin. The set of solutions then looks like the spokes of a bicycle wheel. What is the vector sum of all solutions? (Integer please)

Vector sum = _________________.

4. In last week's computer lab we have seen the hyperbolic 'trumpets' around a calibration line. Make a sketch showing how you would determine the errors in the value of an unknown using the line with trumpets.

Part 2. Excel Spreadsheet Assignment

An analyst determines the absorbance of a solution known to contain 4 organic compounds X=A, B, C and D at four different wavelengths λ= 435,472, 513 and 570 nm.

The extinction coefficients $\varepsilon_\lambda(X)$ for the 4 compounds at these 4 wavelengths are known (units lit/mol):

	A	B	C	D
435	325	3.5	0.1	100
472	50	200	590	0.1
513	1290	700	4.3	12
570	2	0.1	24	1350

The values she finds for the absorbance in a cuvette with L=1cm are:

435	0.070259
472	0.480401
513	0.791769
570	0.433884

What are the four concentrations?

(*Hint*: Absorbance A= ε_λLc is an additive quantity, so you can write out the problem as a set of linear equations. Then write this as a matrix formula and see if you can solve it by matrix algebra.

Hint: the equations have the form

$$A_1 = \varepsilon_{11}c_1 + \varepsilon_{12}c_2 + \varepsilon_{13}c_3 + \varepsilon_{14}c_4$$

$$A_2 = \varepsilon_{21}c_1 + \varepsilon_{22}c_2 + \varepsilon_{23}c_3 + \varepsilon_{24}c_4$$

$$A_3 = \varepsilon_{31}c_1 + \varepsilon_{32}c_2 + \varepsilon_{33}c_3 + \varepsilon_{34}c_4$$

$$A_4 = \varepsilon_{41}c_1 + \varepsilon_{42}c_2 + \varepsilon_{43}c_3 + \varepsilon_{44}c_4$$

We can write these compactly in matrix form as:

$$\boldsymbol{A} = \boldsymbol{\varepsilon c}$$

Where the knowns are the vector A of absorbances and the matrix of the extinction coefficients. We can solve for the concentrations using the matrix inverse:

$$\boldsymbol{\varepsilon^{-1}A} = \boldsymbol{\varepsilon^{-1}\varepsilon c}$$

Which tells us that

$$\boldsymbol{c} = \boldsymbol{\varepsilon^{-1}A}$$

Tutorial 4. Fast Fourier Transforms

Phase factors

There are functions that *produce* roots-of-one as a function of time (t) or place (x). A good example is a Bloch function ϕ(x) = exp(i kx) or the phase factor ϕ (t)= exp(2πi**v**t)= exp(i**ω**t). The first is a function of location (x), the latter of time (t). In both cases the function ***runs around on the unit circle*** we have seen before.

Notice for the latter that there are two conventions for the frequency. If we use **ω** the factor 2π is usually considered *included* in the frequency ω. (The same holds for the *wave vector* k in the Bloch function). You probably have run into these functions before because they are used a lot in science.

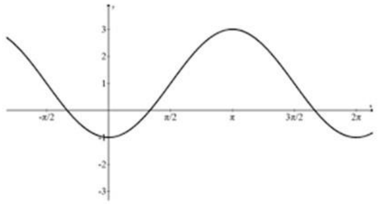
The phase factor is exactly what the name says: If I multiply by such a factor I *leave all magnitudes intact* but I impart a certain phase in the complex plane to my value. Thinking in the complex plane all I do is: rotate along the unit circle, not stretch or contract its radius. This property is the basis for the Fourier transform. If I have a measurement f(t) as a function of time I can analyze it *by frequency* by multiplying with a phase factor ϕ (t; ω) = $e^{-i\omega t}$ and integrating it over all time:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt$$

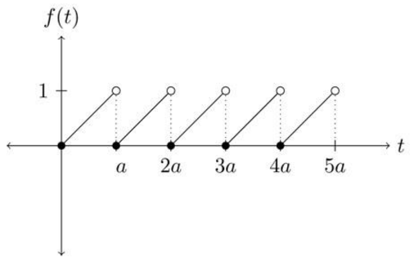
Essentially all you do here is label each measurement with a phase angle leaving all magnitudes intact and see what that gives over your whole data set for a given frequency. In quantum mechanical terms: I am computing an overlap integral to see if my function f(t) ***contains*** ϕ (t; ω). (Yes, my phase factors are an *orthogonal* set: no overlap between them ever). Another way of looking at it is to think of my data as a moving string frozen in time and now I decompose all motion in its normal modes (like you do with vibrating molecules). Each frequency represents a normal mode.

Wave forms

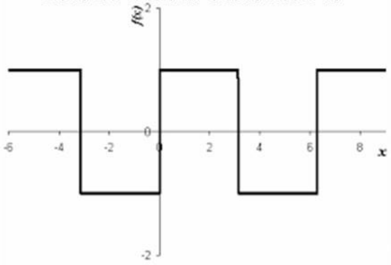
A Fourier series is a decomposition of a repeating wave in terms of sinusoidal functions. A Fourier transform (FT) is the integral representation of the type of decomposition. An FT can also be carried out on a non-repeating wave form e.g. Gaussian or Lorentzian functions. Four important types of wave forms are shown in the figure below for reference as we proceed with the lab.



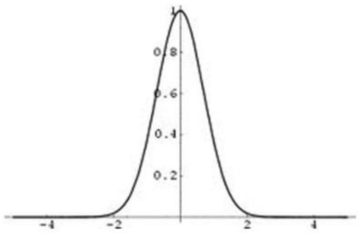
Sinusoidal function



Sawtooth function



Squarewave function



Gaussian function

Fourier transforms

This remarkable decomposition operation is known as ***the Fourier transform*** and it can be shown that (under certain conditions) the new function F(ω) (in the frequency domain) contains the same information as the original function f(t) (in the time domain). There is also an inverse operation that brings F(ω) back to f(t). It involves a very similar phase factor ϕ ‘(t; ω)= exp(-iωt). Thus Fourier transformation allows you to look at your data in a different way without altering the information content.

Discrete and Fast Fourier transforms

Mathematically the Fourier integral runs from -∞ to ∞, but data are typically more limited than that. If you take your data at regular time intervals (a constant *sampling frequency*) it is also possible to do a Discrete Fourier Transform, i.e. one where all integrals are replaced by sums:

$$F(\omega) = \sum_{t=-n}^{n} f_t e^{-i\omega t} dt$$

and there are algorithms to do that fast. The most famous is the Cooley-Tukay algorithm. That is also known as the Fast Fourier Transform (FFT). The crucial thing to understand about the FFT is that it uses the symmetry (think of the unit circle above) to divide the process into two processes, one with even and one with odd terms. Because of this division by two, the number of points must be a power of two. If you do not have a factor of 512, 1024, 2048 etc. you can use zero filling to get the number of points you need. You just add a bunch of points to your data that are all zero.

Computer lab

The data analysis pack of Excel contains a Fast Fourier Transform option based on a famous algorithm, that of Cooley and Tukey. In its original form these authors showed that the fastest way to do a transform on a discrete data set (a Discrete Fourier Transform DFT) is attained if the data set contains a number of points that is a binary power: n= 2^N. That is n should be 2, 4, 8, 16, 32, 64, 128, 256, 512, etc.
If the number of data points n contains factors of 3^x, 5y and higher primes it was later shown that transformation is possible but takes more computer time. If n contains large prime numbers the algorithm gets pretty slow, cumbersome and complicated.
Excel contains the original algorithm and thus ***requires*** n to be a binary power. We will use n=512.

- Open up spreadsheet ***FFTlab*** and make sure the macros are active. Excel tends to deactivate them for security reasons. To activate them go to File / Options / Add-ins. Then select Analysis Toolkit. Click on Go at the bottom of the menu. Actually you need to activate both Analysis Toolkit and Analysis Toolkit VBA, but they will both appear on a menu. Select both and click OK.
- Put a 0 in A1 and 1 in A2. Select A1:A2 and put the cursor on the bottom right corner of the selected range until a + appears.
- Now drag the + sign down until you reach A512. The cells should fill with integers and the last one should read 511. This is your time axis (say in seconds). This means that your sampling rate is 1 point per second.
- Type in cell B1: =A1*2*PI()/512 and Enter.
- Put the cursor on the bottom right corner of B1 until it turns into a +. Then double-click. This should copy and fill the formula down to B512
- As you see from the formula the B column now runs from 0 to 2π in 512 steps. This is handy because we are going to work with sines and cosines
- The cell C1 is where you can experiment with functions. For the moment let's just put in some normally distributed noise. Type in C1: = NORMSINV(RAND()). Use the same copy and fill trick to spread the formula over the C range you did before for the B range.
- Now select B1:C512 (Activate B1, hit End; hit Down-arrow; hold down shift; hit Right-arrow; hit End; hit Up-arrow)
- Make a scatter-plot of these data with a solid line only. Should look pretty messy. (All you have is noise.) This is your ***time domain plot***
- I made a button for you. What is does is copy the formulas in the C column and past them as values (so that they do not get recalculated all the time and slow the spreadsheet to a crawl). Then it runs the Fourier Transform option of the Analysis Pack (You can also use that directly)
- In I1 type =A1-1 and then double click to fill I1:I512
- Scroll down till you see E257. As you see it is also a *real* value (the imaginary component is zero). This 'half-way' frequency is known as the ***Nyqvist frequency*** and it represents the sum of all odd points *minus* the sum of all even ones (-+-+-

+-+-+ etc.) . That is: the phase angles *alternate* between 0 and π from point to point. This is actually the highest frequency your data provide any information on. If you want to measure something that happens faster you should have measured with a higher sampling frequency than 1 point per second. In that case your data set would be larger, (say 1024 points if you sample twice as fast). In that case the Nyqvist frequency would be down at row 513 not at 257.

- If what you study actually fluctuates at a frequency a bit faster than the Nyqvist frequency, say at ω_{Ny}+ δω (i.e. you did not sample fast enough) you will get something known as *aliasing*. A false signal will appear at a frequency ω_{Ny}- δω in your analysis. If this happens with a sound recording you get ugly distortions.
- Look at the contents of cells E256 and E258. As you can see they are each other's complex conjugates. The values below the 257th row represent *negative* frequencies and do not contain any new information. The symmetry around the Nyqvist frequency comes from the fact that the original data are real numbers. (The symmetry needs to be preserved when operating on the data in the frequency domain otherwise the inverse transform will not yield real numbers).
- Change the value in I258 into -255 and in I259 into -254. Then select the two cells and double-click the bottom right corner to fill. All the way at the bottom the last number should read -1.
- Type in J1: =IMABS(E1)^2. This calculates the value of $|Z|^2 = Z^*Z$ of the complex numbers that the FFT produced. This value is known as the *power spectrum* or the ***intensity*** of your signal.
- Use the bottom corner trick to fill the formula down to J512.
- Make a scatter plot of I1:J512 and make the Y-scale logarithmic (Click on one of the points, right click and go to the format menu). This is your ***frequency domain plot.*** Note that the plot looks symmetric about 0, which tells you that the information content in the last 256 points is the same as in the first 256.
- The graph should look pretty random, because the transform of random is random again: random (*white*) noise contains all frequencies equally. (Think of white light!)
- Let's do something less random: Type in C1: =COS(B1*4) and do the + trick to fill
- Push the FFT button. Graph A1:E512 (i.e. select columns A and E and then make a scatter plot selecting the line option).
- I recommend that you save this spreadsheet and a new spreadsheet where you type in the function COS(B1*4) again. The spreadsheet before you perform the FFT function can be called LAB4A. Then the spreadsheet after you press FFT you can save the sheet as LAB4A_FFT. The reason for this procedure is that the FFT function replaces the C column with values and the original function is lost. If you do this for each function you will have a record of your work that you can return to. To show that you have done the experiment you could either submit your figures pasted into a PPT document or simply submit the last step in this series of functions.

- Compare the two graphs in the time domain and in the frequency domain. As you see all the frequencies are zero now except harmonic number 4. (see cell J5, but also look at E5).
- Do the same for =SIN(B1*4)
- Did anything change? Not in J5, but what happened to E5?
- The problem is that by calculating Z*Z we have thrown away the *phases* of the complex numbers in the E column (the Fourier components or harmonics) and sines and cosines only differ by a phase shift of π/2= 1.5708. We will not need phase information here, but if you want to extract it you can use (=IMARGUMENT(E1)).
- We can add some random noise to the cosine function. To do this type in C1: = cos(4*b1)+ 0.3*normsinv(rand()) . Fill the C column and recalculate the FFT.
- As you see the data generated in the C column are a pretty noisy cosine wave, but the fourth harmonic still stands out nicely above an ocean of noise. Its intensity is still about 65000, and its phase is only a little different from zero. All the other harmonics have about the same intensity and their phases are randomly scattered between - π and + π. If we could throw away all harmonics but number four and transform back what would we get?
- Let's take another example. We can call this a linear combination of sinusoidal functions. Type in C1:=cos(b1)+0.8*sin(2*B1)+0.7*cos(4*b1+0.1)+0.2*sin(8*b1) and fill the C column. Then save as LAB4B. Then FFT and save as LAB4B_FFT.
- As you see the 'data' in the C column are now a pretty complicated function and you would never have guessed how many components there are just by looking at the graph. The FFT however flawlessly picks up how many components there are, how strong they are and what their phases are. This is the main use of Fourier transforms: ***analyze data by frequency***. This is particularly useful if you think of colors of light or pitches of sound.
- Other periodic functions include the sawtooth function.
  Fourier analysis works for any periodic function. In fact, Fourier series is just a way to decompose any repeating function into a linear combination of sines and cosines. One simple function that we can create is the sawtooth function. Type in C1: =A1/512 and fill the C column. The function looks like a straight line. However, you must remember that this is a repeating function. At end of the rise it returns to zero and rises again. Call this spreadsheet LAB4C. Study the function with FFT by clicking on the FFT button and call the spreadsheet LAB4C_FFT.
- Change both the x- and y-axes of the frequency domain plot to logarithmic. NOTE: That plot will only show the positive values since you cannot take the logarithm of a negative number. Using this representation the FFT has the appearance of a straight line. Using a Fourier series a saw-tooth function can be written as f(t)=2[sin(t)- sin(2t)/2+ sin(3t)/3-sin(4t)/4+…..] = 2Σ±(sin(ft)/f) This means that the intensities $I=f^2$ should drop off with the square of the frequency f. Thus, the ln(I) = -2 ln(f), which is evident in the log-log plot. The slope is negative because the intensities decrease as the frequency increases.

- Let’s look at a squarewave function: =IF(A1<256,1,-1) . This time the slope of the double logarithmic plot is -4 (the amplitudes now drop off as the square of the frequencies). However, because the block function is an odd function (antisymmetric around the midpoint) all even harmonics are empty. Thus, the values of the log-log plot have an oscillatory appearance. You can call this pair LAB4D and LAB4D_FFT.
- You can represent a Gaussian function in Excel using C1: =NORMDIST (A1,244,15,0). Fill the C column. This should give a Gaussian peak around x=244 with a width of 2x15. As you see the FFT is also a Gaussian. Notice that the intensity drops to zero pretty fast at higher frequencies. You can call these LAB4E and LAB4E_FFT. Now you have a record of each wave form and the FFT.
- Change the axes of the frequency domain plot to linear-linear if they are not already. Then take the FFT. The Fourier-transformed Gaussian is also a Gaussian, but now it is in frequency space. This is a unique property. The Gaussian is the only function whose FT has the same functional form.
- Do this again but replace the standard deviation =15 by 5. What happens in the frequency domain? This domain is often called *reciprocal space*. Why?
- What happens if we put the peak somewhere else: =NORMDIST(A1,380,5,0)?
- Let’s add some noise to the original Gaussian. Use C1: =NORMDIST(A1,244,15,0)+ 0.005*NORMSINV(RAND()) and then fill.
- Do you still see the Gaussian in reciprocal space? Where does the intensity in the low frequencies come from? The peak? The noise? Both?

  Some examples of FFTs:

  1. Scattering of light is essentially mother nature’s way of doing a Fourier transform, so all X-ray diffraction in based on it
  2. Any regularly sampled 1D data set can be analyzed for its noise spectrum and operation in the frequency domain allow noise suppression and deconvolution (removal of peak broadening)
  3. Interferometry is based on inducing path (and thus phase) differences. It is used in e.g. FTIR as an alternative for a grating monochromator
  4. Pulsed techniques like pulsed NMR or pulsed voltametry hit a sample with a block wave, i.e. a mixture of frequencies, the response of each of which is unraveled by FFT
  5. Mechanical spectroscopy hits samples with block function like deformations, again:FFT.
  6. Fourier transforms are a standard trick in solving diff-eqs.

Tutorial 4 Homework

Part 1. Questions related to the FFT lab

1. After adding the noise to the Gaussian does the intensity still drop to zero at higher frequencies?

yes ___________ no _____________

2. Do you still see the Gaussian in reciprocal space?

yes ___________ no _____________

3. Where does the intensity at low frequencies come from

A. mostly from the Gaussian peak but some of it also from the noise

B. the Gaussian peak only

C. mostly from the noise but some of it also from the Gaussian peak

D. the noise only

4 Prior to World War II the technology of making movies was not advanced enough to take more than 16 frames per second or so. In old silent movies this is often visible when you see a wheel or a propeller starting to turn. At first it seems to start turning the right way but at some point the propeller seems to stand still and then start to turn backwards.

Why is this related to the Nyqvist frequency?

Part 2. Numerical computations using Excel

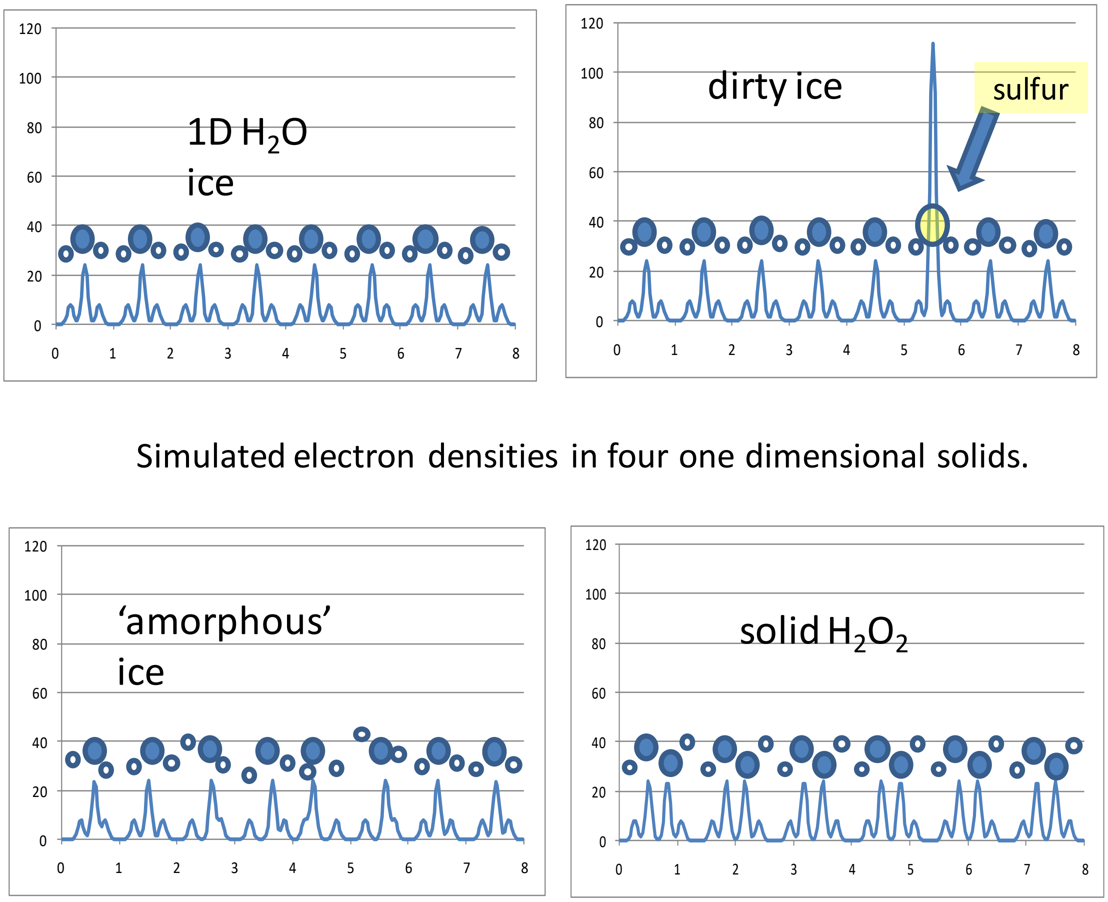
2.A. Modeling NMR spectra

Using the FFT worksheet activated with the same methods used for the laboratory, let's create some Free Induction Decays (FIDs) that represent the kind of signals seen in NMR spectroscopy. In NMR a 90o pulse rotates the magnetization due to the nuclei into the x-y plane. Then this magnetization rotates at a characteristic frequency and relaxes back to the vertical configuration. The FID is a cosine (or sine) function multiplied by an exponential function.

1. First type into C1 := COS(32*B1)*EXP(-B1). Fill the C column. Take the FFT
2. Determine the position and estimate the line width of the resulting spectral feature. To determine the line width find the full width at half maximum (FWHM). What is functional form of the line shape?
3. Next type into C1 := COS(32*B1)*EXP(-2*B1). Fill the C column. Take the FFT

4. Determine the position and estimate the line width of the resulting spectral feature. We can call the exponent the relaxation rate. For example, in the first part the relaxation rate, Γ, is $\Gamma = 1$ and in the second trial the relaxation rate is $\Gamma = 2$. What can you conclude about the relationship between the relaxation rate and the FWHM?
5. If there are two spins connected through a bonding pathway (H-C-C-H) then they can interact by a scalar coupling. In this case one spin's magnetic field will cause a splitting in the magnetic field of the connected spin. To model this effect type into C1 := 0.5*(COS(B1*30)+COS(B1*34))*EXP(-B1). Fill the C column. Take the FFT. Describe what you see. What is the scalar coupling J?
6. In the following aspect we will consider the appearance of the scalar coupling as the relaxation rate increases. To compare with the above scalar coupling type:
   A. C1 := 0.5*(COS(B1*30)+COS(B1*34))*EXP(-2*B1).
   B. C1 := 0.5*(COS(B1*30)+COS(B1*34))*EXP(-4*B1).
7. Describe the observed line shapes as the relaxation rate increases from $\Gamma = 1$ to $\Gamma = 4$. Can you state a general rule for the appearance of line shapes in terms of the relative magnitude of J and Γ?

2.B. Modeling crystal structures



Simulated electron densities in four one dimensional solids.

Scattering, including X-ray diffraction is nothing but a Fourier transform of the electron density function of your sample, e.g. an ice crystal. Consider how the electron density (the local concentration of electrons) fluctuates inside an ice crystal. We shall pretend it is just a one dimensional row of ordered molecules and project all the electron density onto the x-axis.

Open up spreadsheet: icefft.xls. You will find the functions with and without S and a "liquid" (amorphous) one. Also a 1D crystal of hydrogen peroxide is simulated.

Make a graph of the four functions. They should look like above. You can think of the x-axis as Angstroms, although I did not adhere to ionic radii. Notice that the third function is not quite as regular as the others.

- How many molecules are there in 8 Angstroms?
- What is the repeat unit in Angstrom?

Now transform each of the functions by FFT. You can transfer the data to the other sheet and use the button, but you can also go to the data analysis pack and invoke Fourier transform. There is a popup where you specify where your data are and where you want the result.

We will not consider the zeroth harmonic. In X-ray diffraction that corresponds to the incident beam passing straight through the sample and you cannot measure anything if there is no diffraction. Now make a plot of the intensity = $(IMABS)^2$. The graph you get looks very much like a powder pattern, which you will encounter later in the demonstration portion of the course. However, here you may consider these peaks reciprocal space to be due to the periodicity of the structure. For example, the peak at in reciprocal space represents the minimum periodic "wavelength" of 1 molecule per Angstrom. In that case how would you interpret "frequencies" of 1/2 or 1/4? Perhaps more challenging is to understand the meaning of the peak at 3/4. Plot your output in reciprocal space (i.e. using the $(IMABS)^2$ of the FFT and try to interpret the peaks that result.

The substitution of one molecule of H_2S in the lattice of H_2O is called dirty (or impure) ice. Interpret the $(IMABS)^2$ of the FFT of this structural file.

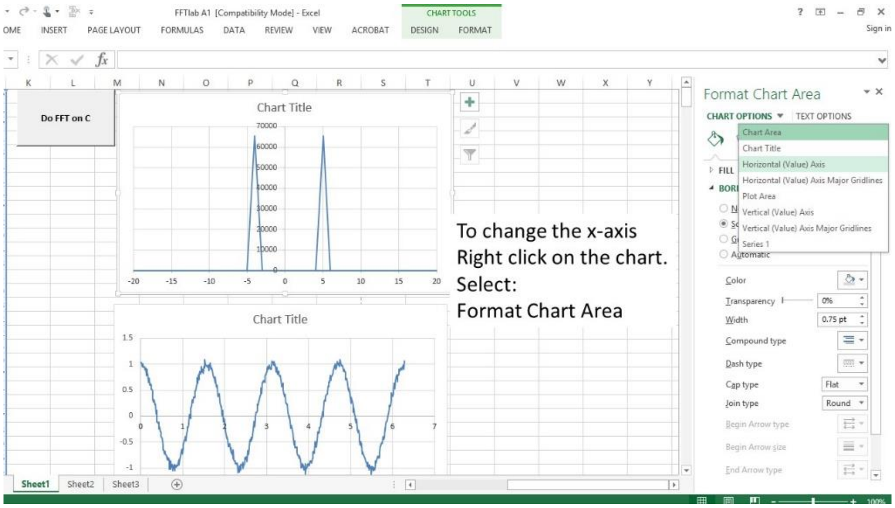
- Does the substitution of O by S affect the peak positions?
- Does it affect the peak intensities?
- What happens to the 'empty' harmonics between the peaks, when the strict translation order is broken in the dirty ice?

The amorphous ice structure has greater disorder. Take the FFT of this structural file and then plot the $(IMABS)^2$. Describe the change in structure. Can you account for this change?

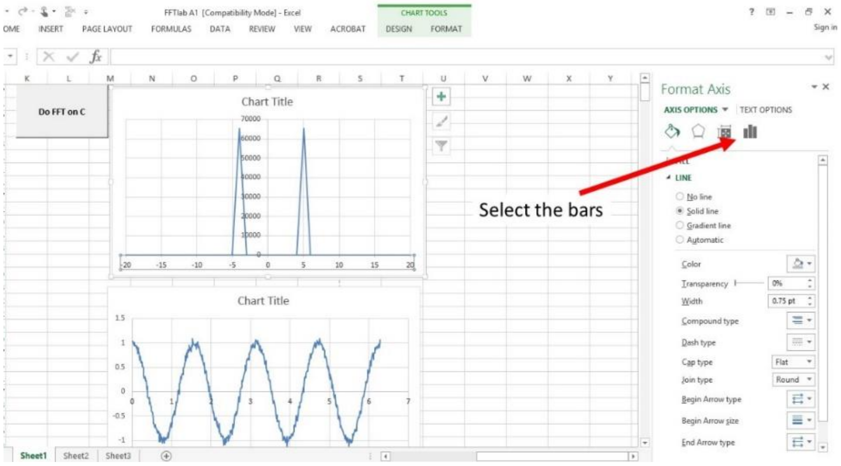
In what ways is the hydrogen peroxide reciprocal lattice (given by the FFT) different from the ice reciprocal lattice?

Appendix: Plot functions in Excel

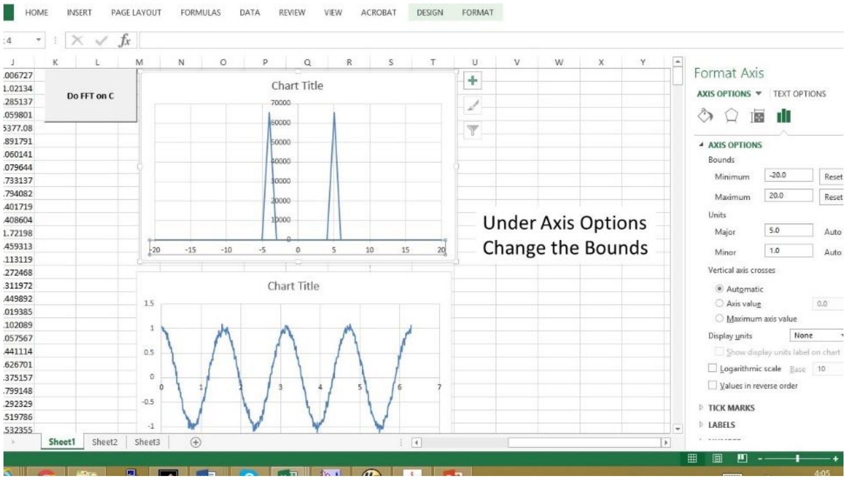
Changing plot limits (also called bounds) in Excel involves finding menu options buried under several layers of menus. The way to find these is illustrated in the figures shown in this Appendix.



To change the x-axis
Right click on the chart.
Select:
Format Chart Area



Select the bars



J
.006727
1.02134
.285137
.059801
5377.08
.891791
.060141
.079644
.733137
.794082
.401719
.408604
1.72198
.459313
.113119
.272468
.311972
.449892
.019385
.102089
.057567
.441114
.626701
.375157
.799148
.292329
.519786
.532355

Format Axis

AXIS OPTIONS ▾ TEXT OPTIONS

AXIS OPTIONS

Bounds

Minimum -20.0 Reset

Maximum 20.0 Reset

Units

Major 5.0 Auto

Minor 1.0 Auto

Vertical axis crosses

● Automatic

○ Axis value 0.0

○ Maximum axis value

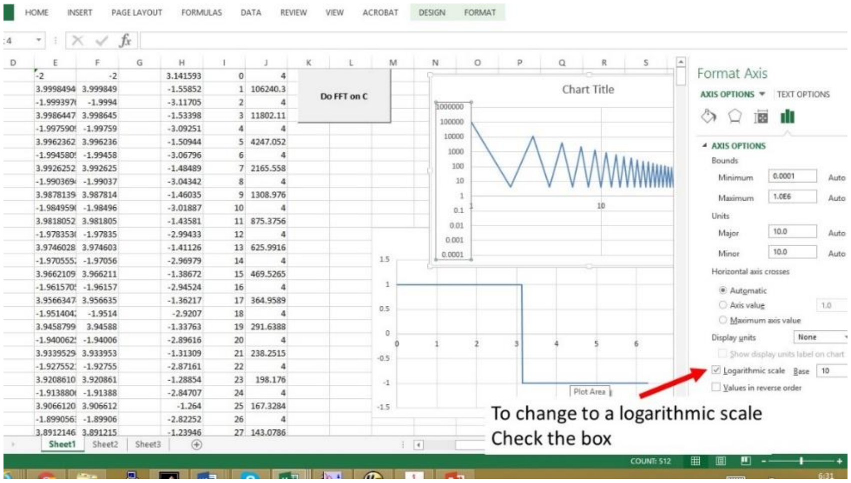
Display units None

□ Show display units label on chart

□ Logarithmic scale Base 10

□ Values in reverse order

Under Axis Options
Change the Bounds



D	E	F	G	H	I	J
	-2	-2		3.141593	0	4
	3.9998494	3.999849		-1.55852	1	106240.3
	-1.999397(	-1.9994		-3.11705	2	4
	3.9986447	3.998645		-1.53398	3	11802.11
	-1.997590!	-1.99759		-3.09251	4	4
	3.9962362	3.996236		-1.50944	5	4247.052
	-1.994580!	-1.99458		-3.06796	6	4
	3.9926252	3.992625		-1.48489	7	2165.558
	-1.990369	-1.99037		-3.04342	8	4
	3.9878139	3.987814		-1.46035	9	1308.976
	-1.984959(	-1.98496		-3.01887	10	4
	3.9818052	3.981805		-1.43581	11	875.3756
	-1.978353(	-1.97835		-2.99433	12	4
	3.9746028	3.974603		-1.41126	13	625.9916
	-1.970555:	-1.97056		-2.96979	14	4
	3.9662109	3.966211		-1.38672	15	469.5265
	-1.961570!	-1.96157		-2.94524	16	4
	3.9566347	3.956635		-1.36217	17	364.9589
	-1.951404:	-1.9514		-2.9207	18	4
	3.9458799	3.94588		-1.33763	19	291.6388
	-1.940062!	-1.94006		-2.89616	20	4
	3.9339529	3.933953		-1.31309	21	238.2515
	-1.927552:	-1.92755		-2.87161	22	4
	3.9208610	3.920861		-1.28854	23	198.176
	-1.913880(	-1.91388		-2.84707	24	4
	3.9066120	3.906612		-1.264	25	167.3284
	-1.899056:	-1.89906		-2.82252	26	4
	3.8912146	3.891215		-1.23946	27	143.0786

Format Axis

AXIS OPTIONS ▾ TEXT OPTIONS

AXIS OPTIONS

Bounds

Minimum 0.0001 Auto

Maximum 1.0E6 Auto

Units

Major 10.0 Auto

Minor 10.0 Auto

Horizontal axis crosses

● Automatic

○ Axis value 1.0

○ Maximum axis value

Display units None

□ Show display units label on chart

☑ Logarithmic scale Base 10

□ Values in reverse order

To change to a logarithmic scale
Check the box

Tutorial 5. Non-linear fitting

Non-linear fitting

As we have seen the matrix formula $\mathbf{(X^TX)^{-1}X^TY}$ allows us to calculate the least squares estimates in a variety of models, provided these models are *linear in the parameters* $\boldsymbol{\beta}$. In many cases we cannot linearize our fitting problem. Fortunately you can still minimize the residuals (actually their sum of squares SS) with a very similar formula $\mathbf{(J^TJ)^{-1}J^TY}$**.**

The difference between the two is that **X** only contains information on where we take our data (our *independent* variables). **J** however also depends on the parameters themselves. In fact **J** contains the derivatives of the fit function f(x; $\boldsymbol{\beta}$) versus each parameter in each chosen data point.
This means that we need to have an idea of what $\boldsymbol{\beta}$ is before we can compute **J**. It also means that $\mathbf{(J^TJ)^{-1}J^TY}$ will only give us a *better* estimate of $\boldsymbol{\beta}$, not the *best*. That's no problem: we can keep applying the process until no more improvement is observed. This iteration process is called ***refinement***.

1. make guess of parameters
2. calculate the **J** matrix based on that guess
3. calculate $\mathbf{(J^TJ)^{-1}J^TY}$ to get better parameters $\boldsymbol{\beta}$,
4. if this is an improvement go to step 1; if not stop process

What refinement does is look for the minimum in the SS function. However this function is now like a landscape with hills and valleys, not a single well. Therefore the procedure will only work ***if*** your initial guess for $\boldsymbol{\beta}$ is close enough to the final minimum. Otherwise the procedure gets lost in the hills.

The final $\mathbf{(J^TJ)^{-1}}$ matrix and Sum of Squares tells you the uncertainties in the final parameters, just like its cousin $\mathbf{(X^TX)^{-1}}$ did. Because you have to terminate the refinement process somewhere (if improvement is less that some criterion) these values are known as *asymptotic standard errors* and $\mathbf{(J^TJ)^{-1}}$ produces an asymptotic variance covariance matrix.

Excel contains an add-in that will do all this for you and we will fit some data with it. The add-in is called the ***solver***. The instructor/TA will help you to make sure it is loaded. Unfortunately the Excel solver does not produce values for the uncertainties, but the book Excel for Chemist by J. Billo has another module on its CD that remedies that.

The data

The data we will be fitting comes from the TGA, in fact it is a decomposition run on Calcium oxalate monohydrate ($CaC_2O_4.H_2O$). It represents a series of decomposition reactions the oxalate as the temperature is ramped to 1000°C.:

$CaOx.H_2O \rightarrow CaOx \rightarrow CaCO_3 \rightarrow CaO$

The first step produces water vapor, the second CO and the third one CO_2.

Open the Non-lin excel spreadsheet. Select the data block and make a graph (line only) of the data in the C column (weight) versus the A column (time). We will work in time, but as the oven was ramped at constant heating rate we could easily translate time into temperature (in the B column).

The Model Function

The biggest problem with fitting data is always to formulate a reasonable *model*. More often than not you do not know '***the'*** model and this is a good example: a good physical model is not known for this kind of data. The graph certainly makes it obvious a straight line will not do! Deviations from straight can often be modeled by adding higher order terms (a polynomial) but this is not recommended for this type of data.
Sigmoidal (S-shaped) step functions are notorious for requiring an infinite number of terms to fit well. This violates the ***parsimony*** principle: always try to retain as many degrees of freedom as you can, or stated differently: is you can fit something with three variables, do not fit it with 300. If you throw in enough parameters you can fit even the kitchen sink! This is why we fit this with a function that already has a sigmoidal shape, the ***logistic function lgt(x):***

$$lgt(x) = \frac{e^x}{1+e^x}$$

We can fit every decomposition event as:

$$W_o lgt(a - bt)$$

As you see there are three parameters (not 300!) per event: *W_o*, *a* and *b*. W_o stands for the amplitude of the change (the entire weight loss of the event). The time around which the event happens is t_{event}= -a/b and the value of b represents the slope in the weight curve at this time (how sudden the event takes place). As we are losing weight its value is always negative in our case. As we have three events, but do not lose all weight the total fit function becomes

$$W(t) = W_{o,1}lgt(a_1 - b_1t) + W_{o,2}lgt(a_2 - b_2t) + W_{o,3}lgt(a_3 - b_3t) + W_{baseline}$$

As you see we have a total of ten parameters: 3+3+3+1.
It is advisable to build up such a fit job systematically in your sheet and not write out a function like this in one cell as you are bound to make mistakes.

- First copy the entire data block (three columns) and open a new sheet and paste it in cell A8, B8 and C8. It consists of 1452 points so that the three columns should be filled down to line 1460.
- Now put the following initial guesses for the parameters in the range C1:G4 including the labels. The numerical parameters should be in D2:G4.

	first	second	Third	Final
W	2	3	5	7
A	15	40	35	
B	2	2	1	

1. Type in C6: t-event
2. Type in D6: -D3/D4
3. Copy D6 over D6:F6
4. Type in D8: =D$3-D$4*$A8 (this calculates *a-bt* for the first event)
5. Copy D8 over D8:F8
6. Type in G8: =EXP(D8)/(1+EXP(D8)) (The lgt function)
7. Copy G8 over G8:I8
8. Type in J8: =$D$2*G8+$E$2*H8+$F$2*I8+$G$2 (This computes the W(t) fit function)

9. For plotting purposes we will copy the relevant columns:
10. Type in K8: =A8; (time)
11. Type in L8: =C8 (the measured weight)
12. Type in M8: =J8 (the fit function)
13. Type in N8: =L8-M8 (the residual)
14. Now select D8:N8 and use the double click on the + symbol that appears on the bottom right corner of N8 to double click and fill all your calculations over the whole data set.
15. Select the K,L and M functions and make a chart with only lines of the measured and calculated data.

16. You can now change the values in the parameter block to make the function fit a bit better. It is useful to change them by hand to get a feel for what they do. If the new guess is really bad, just hit Ctrl+z to correct your mistake. Here is a trial example using the input data set from the website.
17. n e.g. M4 type =SUM(N8:N1459^2)/1450/STDEV^2. To activate this formula use <Shift> <ctrl> <Enter>. This calculates the a value known as chi-squared,

$$\chi^2 = \frac{\sum(y_i - y_{model})^2}{N\sigma^2}$$

Chi-squared is the sum of the squares (SS) divided by the number of data points and STDEV, where STDEV is the standard deviation in your data. You may estimate STDEV using a flat section of the data to generate a series of numbers. Then you can calculate the STDEV for this section. Remember that STDEV is equal to,

$$\sigma = \sqrt{\frac{\sum(y_i - \langle y \rangle)}{M}}$$

where M is the number of data points in the short section you are using to calculate the STDEV. M is not at all the same as the value of N you used above, which consists of all of the points in the data set.

18. Try a change in one of the parameters and look at the value of SS. It should get lower as the fit gets better. When you have obtained an initial guess that is reasonably close to the shape of the data then you can use the solver to get a better fit.
19. Now run the solver. It should be under Data.
20. On the pop up click the icon with the red dot of the *set target cell* option and select the cell that contains the SS (M4)
21. Then click the *Min* option of the *Set equal to* set. (We want to minimize SS!!)
22. Click the red dot icon of the *By changing cells* block an select the cell that contains the final weight. Then click solve.

You can choose which parameters you want to run first. Often it is wise not to take too many parameters at once, but once you are close enough to a good fit you can select then all at once. E.g. you can click the red dot icon and select D2:F4 then type a comma and then click G2 to get them all. The fit is not bad but not ideal either. Make a plot of the residuals to see how bad it is!

In programs that use non-linear least squares fitting the errors in the parameters can be output using the second derivative of the $\mathbf{J}^{\mathbf{T}}\mathbf{J}$ matrix to estimate the root-mean-square error in each parameter. It is important to have the correct weights for the estimate (i.e. the error should be calculated for displacements of the fitted curve relative to the σ you calculated above for a flat portion of the data). This type of analysis has numerous assumptions that may be incorrect. The error in the data may be larger than the σ you calculate. The curvature of the $\mathbf{J}^{\mathbf{T}}\mathbf{J}$ matrix may have significant distortions from a quadratic shape. And so on. For this reason we will use another method to estimate the quality of the fit using a comparison of the the parameters to a physical model. This is often a useful approach and has more significance for the scientist than a number that comes from a statistical analysis that has possible numerical inaccuracies.

Let us make a hypothesis: we have obtained four weights from the regression, let us assume they correspond to the compounds H_2O, CO, CO_2 and finally CaO.

1. Compute the molar masses for these four compounds and plot the weights against the molar masses. This should give a straight line. The weight of the CaO is only correct if the instrument was properly calibrated.
2. Do a linear regression of the weights against the M_w. How many moles of Ca did the sample contain?
3. Using the slope and intercept compute an estimated weight for each compound and determine the absolute value of the residual from the regression line. These residuals should correspond to the asymptotic standard errors in the weights, which we did not calculate (it can be obtained from the covariance matrix $\mathbf{J}^\mathbf{T}\mathbf{J}$ as discussed above). To get a better feeling for how important the residuals for each point (i.e. $\sqrt{(y_i - \langle y \rangle)^2}$) you could divide the residual by the standard deviation for the straight line fit.
4. What does the result say about the chemistry?

Assignment in peak fitting

Go to sheet 2 of the nonlinear spreadsheet.

I generated some data for you. They consist of two partially overlapping Gaussian peaks. I made a version with two different noise levels. As these data are *generated* ones you may expect the final residuals to be random noise only. (Check residuals. Real data may not always be so nice.)

The data represent a problem that is often encountered in science, that of ***peak resolution***. In many techniques we get a pattern consisting of a series of signals each in the form of a peak. This is true for spectra, for thermograms, for chromatograms and many other types of data alike.

The peaks can have a variety of shapes. Gaussians are the simplest one and we will only do those today. Unfortunately peaks may *overlap*. The more they do the harder it is to separate the two signals and derive independent information from them. *Limited* overlap can be overcome by peak fitting.

Caution:

Peak fitting works reasonably well ***if***:

- the overlap is not too large,

- there are not too many peaks involved,
- you know how many there should be,
- you know what shape they should have (Gaussian, Lorentzian etc.)
- the noise level is low
- the peaks are not too broad
- the peaks are (more or less) symmetrical

If any of these requirements are violated, peak fitting notoriously yields various different solutions that are fit equally well but are ***quite incorrect***. In other words: you can get out what you want by changing the model….

This is why in e.g. chromatography people try to ***avoid*** overlap, e.g. by choosing a different internal standard that does not overlap. Instrument builders also try to make their broadening factor as small as possible to avoid or diminish overlap trouble. However overlap cannot always be avoided and fitting may be all you can do. The ***parsimony*** principle applies here: always fit with as few parameters as possible while demonstrating the fit is as perfect as random noise allows. (Look at residuals!)

There are other statistical methods that ***do*** allow you to use overlapping data, but they typically require that you do not have one spectrum, but a series of them, e.g. taken as a function of time

The data

For each of the 'spectra' that I generated, use non-linear fitting to determine the *amplitude*, the *position* and the *width* of the two overlapping peaks, i.e. six parameters in total. In Excel a Gaussian function f(x) is easily generated by the function:

=*amplitude**NORMDIST(x,*position*,*width*,0)

Add two of these terms to generate a fit model. Note that in Excel the NORMDIST function is ***normalized***. This means that it integrates to unity. Thus the *amplitude* parameter is automatically also the integration value of each peak.

In order to use solver you will also need to generate a target. Our target is chi-squared (χ^2), which is defined as above.

I used integer values to generate the data for all six parameters. How large is the bias, i.e. the difference between the parameters you find and the nearest integer? How does the noise level of the data influence the bias? What would happen if the noise level would get larger? What if the peaks got broader? Closer together?

Tutorial 5 Problem set

Part 1. Test your knowledge of the concepts.
Answer the multiple choice questions.

1. Parsimony means:
A. We need as many parameters in a model as we have data points
B. We can only find the precision of our parameters in an asymptotic way in nonlinear regression
C. We should try to fit with as few parameters as we can
D. We have to have a good initial guess of our parameters otherwise the nonlinear fit does not work
E. We should use refinement to fit nonlinear models

2. If the model is linear in the parameters we do not need to iterate because
A. The dependent variables do not depend on the parameters
B. The independent variables do not depend on the parameters
C. The parameters do not depend on the dependent variables
D. The elements of the J (or X) matrix do not depend on the parameters
E. The random errors in the measurement do not depend on the parameters

3. The logistic function has an important property: both for very low and for very high values it approximates a constant value ('goes flat'). This is important because:
A. This is actually detrimental because it causes lack of fit
B. This allows us to measure the weight loss before and after an event even when it overlaps with another
C. This allows subtraction of the constant weight of the platinum basket
D. This causes the function to have only three parameters

Part 2. Numerical Calculations using an Excel spreadsheet

1. Open the spreadsheet exponential_functions_1.xlsx
2. Create a non-linear fitting function that consists of a single exponential function, which will have this form =EXP(-$X$1*A1). The cell X1 will have the rate constant parameter. Make an initial guess for X1 by plotting the model function and data on the same plot. Once you have found a reasonable value for X1 proceed to the next step.
3. Determine the noise in the data by calculating the standard deviation of the last 50 points. We will call this value σ_{noise}, i.e. this is an estimate of the noise in the data.

4. Define a chi-square function in a cell, e.g. Y1. Recall that

$$\chi^2 = \frac{\sum(y_{fit,i} - y_{data,i})^2}{N\sigma_{noise}^2}$$

5. Use solver by minimizing on the χ^2 function. What should the value of χ^2 be for a good fit to the data? What is the value of χ^2?
6. Plot the residuals. Do they have any structure? Explain your observations.
7. We will use a fit of trial and error to determine the error in the fit parameter (i.e. the rate constant). The error limits can be found by manually changing the value of the parameter until χ^2 increases by 1 unit. Try this both by increasing and by decreasing the parameter. Please keep track of the original value and the bracketing values. Why does χ^2 increase regardless of whether you increase or decrease the parameter value? Please explain.
8. Now we will compare the fit using a linear method. Create a new spreadsheet by copying the exponential_function_1 spreadsheet. You may call it exponential_linearized_1.xlsx.
9. Linearize the function in column B (the exponential) by taking the logarithm. =LOG(B1). Double click to fill the entire C column. Plot C vs. A using a line scatter plot. Note that there is a region that looks nicely linear and then the data start to look messy. Why is this? For the purposes of fitting we can select the linear part of the data. Delete all of the cells below cell 400. Now the plot should look fairly linear (although is still gets noisy near the end).
10. Use LINEST to fit these data to a linear model. Note that you want to set the stats flag to TRUE. The function call has the form = LINEST(B1:B399,A1:A399,TRUE,TRUE). Use <Shift><ctrl><Enter> to execute LINEST.
11. Compare both the fitted value and the magnitude of the standard error of the fit to the values obtained from the non-linear fit. Note that the LOG function in Excel is $\log_{10}$ (log to base 10) so you will need to convert the value for the slope using the appropriate factor in order to compare the linear and non-linear fits. The factor is the same as that needed to convert a parameter in the exponent of base 10 to base e.
12. Calculate R^2 for the linear fit.
13. Tabulate R^2, m and the SE for the linear fit as well as χ^2, the rate constant parameter and its error from the non-linear fit.

Literature

1. Salvador Naya; Ricardo Cao; Ignacio Lopez de Ullibarri; Ramon Artiaga; Fernando Arbadillo; Ana Garcia, *Journal of Chemometrix* **2006**, 20, 158-163

Projects

The following pages list examples of projects attempted in the past and some reference material.

Dehaloperoxidase enzyme kinetics

Introduction

Enzymes are widely used in the chemical, biochemical and biotechnological applications. The range of chemical synthesis that is possible using enzymes is quite broad and new methods are under development that will greatly expand the field of enzyme use. Enzymatic catalysis is also one of the best understood types of catalysis. Many catalysts are difficult to characterize because they function at low concentration in complex mixtures and the intermediates are nearly impossible to isolate. In many chemical applications the addition of catalysts is understood as kind of “pixie dust” that just works, but no one knows why. Biological catalysis or enzymatic catalysis is consists many very well characterized reactions, in which we know the structure of the enzyme, the details of the active site, and often we know how the substrate binds and precisely what aspects of protein structure are responsible for lowering the transition state energy. For these reasons enzymatic catalysis can be considered a model for how we would like to understand all of chemical catalysis.

The importance of enzymatic catalysis has been appreciated for more than 100 years. An early approach to characterization of the kinetics of enzymes is attributable to Menten and Michaelis. While there are many variations of Michaelis-Menten catalysis the simplest version treats and enzyme, E, and substrate, S, that combine to form a complex known as ES, the enzyme-substrate complex and then to form product, P, and reform the original enzyme. We can write this mechanism as follows:

$$E + S \underset{k_{off}}{\overset{k_{on}}{\leftrightarrow}} ES \overset{k_{cat}}{\rightarrow} P + E$$

Often the Michaelis-Menten equation is plotted as the initial rate, V_0, which is equal to d[P]/dt. In this form we have,

$$V_0 = \frac{V_{max}[S]}{K_M + [S]}$$

Where the Michaelis constant is:

$$K_M = \frac{k_{off} + k_{cat}}{k_{on}}$$

There are several special regimes that can be useful to understand the Michaelis-Menten equation:

Maximal rate: If there is excess substrate present the rate is limited by the rate at which the ES complex falls apart. The rate of formation of products is a maximum and $V_{max} = k_{cat}[E]_0$ is called the maximum velocity.

Half-maximal rate: If $V = V_{max}/2$, then $[S] = K_M$.

Second order regime: If [S] << K_M then the rate of formation of products is d[P]/dt = k_{cat}/K_M $[E]_0[S]$. The rate depends on [S] as well as $[E]_0$.

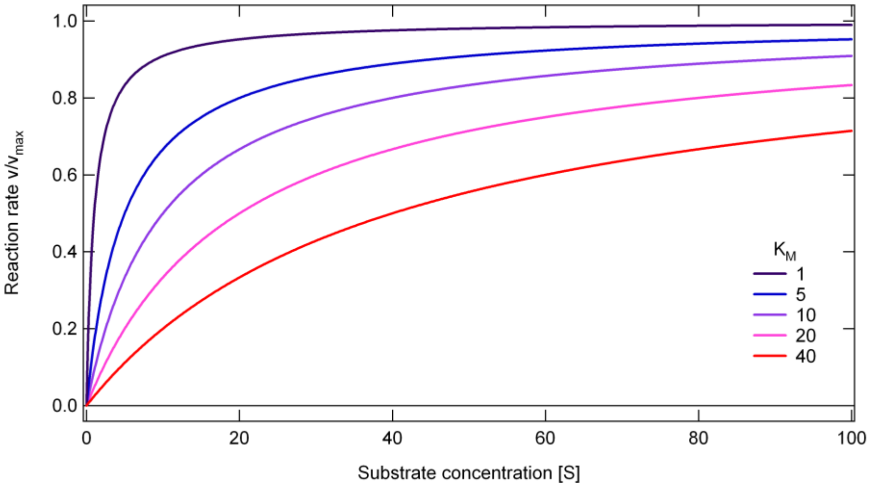


Figure 1. Michaelis-Menten curves showing the saturation of the kinetics at high [S]

Our experiment will determine the Michaelis-Menten parameters for the enzyme dehaloperoxidase (DHP). DHP is a highly versatile and interesting enzyme in its own right. DHP was first discovered as the hemoglobin of a marine organism known as *Amphitrite ornata*. Nearly 20 years later it rediscovered as a peroxidase capable of degrading 2,4,6-tribromophenol, which is a known naturally occurring pollutant in shallow coastal waters. In the subsequent 20 years several unique properties of DHP have been revealed by enzymatic studies. DHP is also a peroxygenase and an oxidase. It has four functions and appears to have activity to oxidize a range of substrates including bromophenols, brominated pyrroles and indoles. In this study we will investigate the kinetics of DHP or one of mutants in the pH range from 5.0 to 7.5. We will choose one set of conditions and one mutant in the interest of time. However, our goal will be to compare the values to published values and understand some new feature or prediction about the reactivity. While we can be quite certain that the experiment will work the interesting aspect is that the exact outcome may be a new result.

The mechanism of the peroxidase reaction involves activation of DHP by H_2O_2. We can think of the reaction with H_2O_2 as a preliminary step that creates active form of the enzyme that we call compound ES. It is compound ES that binds to the substrate oxidizes substrate in two steps. Since these aspects are discussed in detail in the references [1-5], we will not discuss the mechanism further in this protocol. Rather we will ask the student to read the publications and to make the discussion of mechanism part of the laboratory report.

Experimental

You can measure the time-resolved kinetics of enzymes using a photo-diode array spectrophotometer. This type of spectrophotometer reads all of the wavelengths from 200-1000 nm simultaneous in less than 1 second. The advantage of this technology is that the instrument can be set up to run in kinetics mode, in which successive spectra are obtained each 3 seconds. If the kinetic change during catalysis by an enzyme has an optical signal in the range of the instrument and a time course that is longer than a few seconds, but shorter than a few hours it is appropriate for measurement using a photodiode array. One of the nice features of enzyme kinetic experiments is that the overall rate can be tuned by changing the enzyme concentration. Since we measure rates relative to the maximal rate, V_{max}, and

$V_{max} = k_{cat}[E]_0$

we see that we can get V_{max} to have a range of values by changing the enzyme concentration. Of course, there are limitations since the enzyme concentration cannot be higher than the solubility of the enzyme and there are practical limitations to how dilute the enzyme can be to function in a reliable way.

In order to obtain data appropriate for a Michaelis-Menten analysis, you will need to make 6 dilutions of the substrate with constant concentrations of enzyme and co-substrate hydrogen peroxide. The hydrogen peroxidase concentration should be high enough that complete conversion to product can be achieved at the higher substrate concentration. The volume of solution will need to be sufficient that the optical path of the light in the photodiode array passes only through solution and there is no air space on the top that could give rise to spurious absorption. In a small volume cuvette (with a 0.4 cm pathlength) this volume is 1.2 mL. We can summarize the requirements for this experiment as follows:

$[E]_0$ is constant (usually $[E]_0 = 2.4 * 10^{-6}$ M)

$[H_2O_2]$ is constant ($[H_2O_2]$> [S] for all measurements)

[S] ranges from zero to a maximal value of approximately 1 mM.

Note finally that [S] will need to have a higher coverage at low concentrations since the Michaelis-Menten curve has a greater rate of change at lower concentrations. For an enzyme that has an unknown catalytic rate, we will need to make a run to estimate the kinetics. Then once we have an idea what the value of kcat and Km are we can determine the values. Typically we will want 3 values below Km and two values above Km. An approximate distribution of substrate concentrations is:

[S] = 0.1, 0.2, 0.5, 1.0 1.5 and 3.0 K_m

You will need to make stock solutions of the H_2O_2 and substrate. These solutions should be made fresh for each experiment since H_2O_2 tends to react at room temperature and phenols also degrade by slow light and oxygen-dependent reactions.

Setting up the data acquisition

The data are acquired on a PC using HP 845x UV-Visible System software. You set the HP 845x UV-Visible System software to Kinetic mode. You will want to monitor at 272 and 314 nm. The 272 nm wavelength is used to monitor the appearance of the quinone product. The 314 nm wavelength is to monitor the disappearance (consumption) of substrate 2,4,6-TBP. Note that the instrument will record all wavelengths from 200-1000 nm on each acquisition (i.e. every 3 seconds in our experiment). Thus, for a 3 minute data acquisition 60 spectra will be acquired. We will extract all of these spectra for analysis using Singular Value Decomposition. We will use the data at 272 nm for fitting the initial slope to obtain V_0 as needed for the Michaelis-Menten protocol discussed in the introduction.

Stock solutions

The concentration of DHP protein stock solution is determined by using UV-Vis in the Standard mode. The absorbance at 406 or 407 nm which is the λ_{max} of the soret band is recorded and used to calculate the concentration of DHP stock solution according to Beer's law: c=A/(ε_{406}*b). The path length b of the quartz cuvette is 0.4 cm, and ε_{406} for DHP is 116,400 cm^{-1} M^{-1}. The 2,4,6-TBP stock solution in 100 mM KPi buffer (~ 1 mM TBP) can be made by heating the solution to ~ 100 ^{o}C for about 5 min in a water bath. The hydrogen peroxide stock solution is prepared freshly before the kinetic experiments. For a typical DHP A peroxidase kinetic reaction with the final hydrogen peroxide concentration at 1200 μM. You can prepare the hydrogen peroxide stock solution by having 10 mL KPi buffer mixing with 7.4 μL of 30% concentrated hydrogen peroxide solution (from Sigma-Aldrich).

Mixing protocol

Calculate the volume of each component (protein, substrate, buffer) you need for each kinetic assay, add the buffer to the cuvette first, then substrate solution and finally mixed with protein solution in the cuvette. The volume for hydrogen peroxide solution is fixed at 200 μL and will be added in the end to initiate the reaction. Place the solutions in the cuvette, wait 3 min for it to reach the desired temperature. Make a new file for each kinetic assay and set the experimental parameters. When you are ready to start the reaction, press F7 to start the experiment while at the same time add the H_2O_2 solution, mixing the solution once or twice quickly with the syringe tip.

Analysis

Data transfer

The data can be extracted from the software by copying and pasting into an Excel spreadsheet. The spreadsheet can be transferred to your UNITY account using Secure Shell software. Secure Shell is a windows based program that is based on the LINUX sftp (secure file transfer protocol) command. You will find the Secure Shell 3.2.9 Icon on the desktop. You can set up the path for transfer using the software. The procedure is shown on the website using the screen shots of the SSH software.

Uploading the data into IgorPro

To upload data from an Excel spreadsheet into IgorPro the easiest method is the cut and paste method. You may open the Excel spreadsheet and select the rows and columns you would like to analyze. Then these values can be copied (<ctrl c>) and pasted (<ctrl v) into the table in IgorPro. When IgorPro is opened there is always a table presented as a default. When you paste the data into this table the data columns will have the labels wave0, wave1, wave2 etc. As long as you keep careful records there is no need to rename all of these columns. For example, if you have a typical spectroscopic data set with wavelengths from 400 – 600 nm, there will be 200 columns. It would be a bit painful to change the names of 200 data waves. If you do need to change a name of a wave you can do it on the command line of IgorPro.

- Rename wave0 time

- ➢ Rename wave1 lamda400

Note that IgorPro waves must have names that begin with a letter and not a number. Plot all the time courses of the absorbance at the given wavelength. Select all the corresponding absorbance waves as the y axis and the only wave “time” as the x axis at Windows -> New Graph and go ahead to plot them.

Fitting the kinetic data using the method of initial rates

Igor has a number of standard fitting functions. The fit to a straight line is a standard fitting function. As you have learned, fits to a straight line are known as linear least squares fitting and there is a unique solution for the slope and intercept. In this problem the intercept is not important for the kinetic analysis, but the slope tells you how ⍰A changes with time. Once you know this you will need to convert ⍰A to ⍰[P], the change in concentration. For this step you will use Beer’s law.

To fit to a straight line you will need to plot the kinetic data and then use the cursors to select the first few points (6 to 10 points). The data are only linear over a very small range of time. Make sure that you have selected a short enough range that it is linear. However, you will need a minimum number of points to make the fitting meaningful. Experience suggests that 6 points is the minimum.

The procedure for plotting, selecting data with the cursors and defining the fitting function are given on the website. Use the website to guide you in this step. Record the values of your fit for each of your kinetic runs at each of your concentrations. Make a table with the following entries

	⍰A_1/⍰t	⍰A_2/⍰t	⍰A_3/⍰t	⍰⍰A/⍰t>	⍰⍰⍰A/⍰t)
$[S]_1$	0.00478	0.00512	0.00499		
$[S]_2$	0.00839	0.00812	0.00806		
…					

Constructing and fitting the Michaelis-Menten plot

Once you have obtained the average values of the initial rate, V_0 = ⍰⍰A/⍰t>, for each substrate concentration you will need to construct a plot of the initial rates vs. substrate concentration, V_0 vs. [S]. This is the Michaelis-Menten plot. The data in this plot will be fit using non-linear least squares fitting. In IgorPro this can be done by adding a macro to the software. IgorPro has a number of standard fitting functions for non-linear least squares fitting, but the Michaelis-Menten model is not one of them. You may use the macro below, which is available for download on the website. You will need to add these lines of text to the Procedure Window of IgorPro. When you close the Procedure Window it will automatically compile the macro and make it available in the Analysis menu. You may select this fitting function in that menu.

Non-linear least squares fitting differs from least squares in that an initial guess for the parameters is required. In the case of M-M fitting you will need to input the V_{max} and K_m values. How can you “guess” these values? It seems a bit tautological since the whole point of fitting is to obtain the values. In the case of M-M you can estimate V_{max} since that is the limiting value of the initial rate at large [S] concentration. You can inspect the plot and either simply use the largest value or perhaps a somewhat larger value based on your intuition of how much the curve is increasing over the observable range. Km can be estimating finding the value of [S] the corresponds to $V_{max}/2$, since K_m = [S] gives $V_{max}/2$ in the M-M formula. You will still need to do the fitting in order to obtain accurate values of these parameters. The fitting menu allows you to input an estimated value of the standard deviation (also called the weight). When this is entered the fitting function will return a value of chi-squared, ⍰^{2}. In non-linear least squares fitting this is the figure of merit that is most frequently used to indicate the goodness of it. If the errors are properly estimated a good fit should have ⍰^{2} = 1.

References

1. **Ma, H.; Thompson, M.K.; Gaff, J.; Franzen, S.** "Kinetic analysis of a naturally occurring bioremediation enzyme: Dehaloperoxidase-hemoglobin from *Amphitrite ornata*" J. Phys. Chem. B, **2010**, 114, 11283-11289
2. **Zhao, J.; Franzen, S.** "Kinetic study of the inhibition mechanism of dehaloperoxidase-hemoglobin A by 4-bromophenol" *J. Phys. Chem. B* **2013**, 117, 8301-8309
3. **Zhao, J.; Zhao, J.; Franzen, S.** "The Regulatory Implications of Hydroquinone for the Bifunctional Enzyme Dehaloperoxidase- Hemoglobin from *Amphitrite ornata*" *J. Phys. Chem. B*. **2014**, 117, 14615-14624
4. **Thompson, M.K.; Davis, M.F.; de Serrano, V.; Nicoletti, F.P.; Howes, B.D.; Smulevich, G.; Franzen, S.** "Two-site competitive inhibition in dehaloperoxidase-hemoglobin" Biophys. J. **2010**, 99, 1586-1599
5. **Thompson, M. K.; Ghiladi, R.; Franzen, S.;** "The Dehaloperoxidase Paradox" Biochim. Biophys. Acta – Proteins and Proteomics, **2012**, 1842, 578-588

The clock reaction

In this project the kinetics of the following redox reaction are studied in aqueous solution:

$$S_2O_8^{2-} \quad + \quad 2\, I^- \;\rightarrow\; I_2 \quad + 2\, SO_4^{2-}$$

persulfate + iodide → iodine + sulfate

A small amount of starch in the solution will produce and adduct with iodine, which results in an intense blue color. To follow the kinetics of the reaction a small amount of a sacrificial reducing agent, thiosulfate, will be added. This reacts very quickly with any liberated iodine to form back iodide:

$$2I^- \quad + \quad S_4O_6^{2-} \leftarrow I_2 \quad + \quad 2S_2O_3^{2-}$$

iodide + tetrathionite ← iodine + thiosulfate

As long as the thiosulfate has not been depleted the blue color will therefore not appear. In this time the concentrations of both iodide and persulfate are essentially constant. Iodide is being formed back, whereas persulfate is in excess. That means that the first reaction is running at constant rate and this rate can be measured from the amount of thiosulfate originally present and the time it takes for it to be consumed, i.e. the time it takes for the blue color to appear. This time can be measured with a stopwatch.

The rate of the reaction can be studied as a function of a variety of variables, such as the initial concentrations of iodide, persulfate, thiosulfate, temperature, ionic strength and the presence and concentration of a catalyst like Cu^{2+} or Fe^{2+}.
The hydrolysis of ethyl acetate

The hydrolysis reaction of ethyl acetate can be studied by monitoring the conductivity of an aqueous solution of the ester with various amounts of sodium hydroxide.

$$\mathrm{EtAc} + \mathrm{OH}^- \ (+\mathrm{Na}^+) \rightarrow \mathrm{EtOH} + \mathrm{Ac}^- \ (+\mathrm{Na}^+)$$

Because the mobility (Λ) of the acetate ion is lower than that of the hydroxyl ion, the conductivity of the solution will decrease as the reaction progresses. The sodium concentrations does not change but gives a constant contribution to the conductivity Depending on the extent of reaction ξ, the conductivity will follow the expression:

$$\sigma = \{\Lambda_{Na} + \Lambda_{ac}(\xi) + \Lambda_{OH}(1-\xi)\}[Na^+] = \{\Lambda_{Na} + \Lambda_{OH} + (\Lambda_{ac} - \Lambda_{OH})(\xi)\}[Na^+]$$

$$\text{Thus } \sigma_{\xi=0} = \{\Lambda_{Na} + \Lambda_{OH})\ [Na^+]$$
$$\text{and } \sigma_{\xi=1} = \{\Lambda_{Na} + \Lambda_{Ac})\ [Na^+]$$

When we measure the conductivity as a function of time, we can deduce the extent of reaction from:

$$\xi(t) = \{\sigma(t) - \sigma_{\xi=0}\}/\{\sigma_{\xi=1} - \sigma_{\xi=0}\}$$

Unfortunately the extremes $\xi=0$ (no reaction) and $\xi=1$ (reaction complete) cannot be achieved and so the two values $\sigma_{\xi=0}$ and $\sigma_{\xi=1}$ have to be estimated using a parameterized regression.

The reaction is reportedly first order in both $[OH^-]$ and [EtAc], so that integrated rate laws can be used to fit the data.

The rate of the reaction can be studied for various concentrations $[OH^-]$ and [EtAc] as well as the temperature. From a latter study an energy of activation can be derived.

The reaction of iodide and iodate

At low pH iodide and iodate undergo a redox reaction to form iodine in aqueous solution

$$6H^+ + IO_3^- + 5I^- \rightarrow 3I_2 + 3H_2O$$

The reaction rate can be studied photometrically by monitoring the absorption A in the visible using a UV/VIS spectrometer as a function of time

The method of initial rates is used, by extrapolating the slope of the A(t) data back to the origin. The rate of the reaction can be studied as a function of the initial concentrations of iodide, iodate, the pH and temperature.

The binary phase diagram of organic solvents.

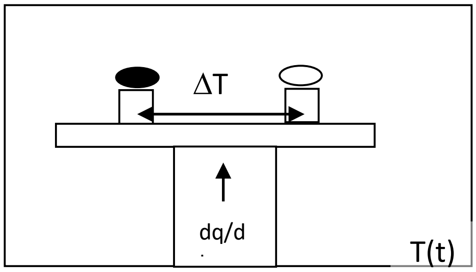
Many organic solvents –such as cyclohexane, mesitylene or octane- are perfectly miscible at room temperature. MDSC will be used to study the low temperature behavior of a binary mixture of two such solvents. In many cases, when the temperature is dropped to –90^{o}C, the solvents will be present as separate solids with very limited –if any- mutual solid solubility. The calorimeter will be used to follow the melting behavior of such mixtures and a phase diagram constructed. The van ‘t Hoff relationship that describes melting point depression will be used to fit the liquidus lines of the diagram as a check on the assumption that the liquid miscibility is ideal and the solid solubility negligible. If applicable, solid-solid transitions will also be studied.

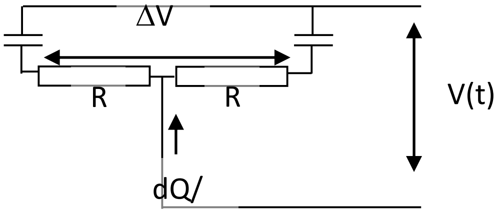
Differential scanning calorimetry of polyols

Introduction

One of the most advanced applications of thermodynamics is Modulated Differential Scanning Calorimetry (MDSC).[1]

Calorimetry: any measurement of heat (flow). (Since 18th century)
Differential: the *difference* between the sample and a reference is measured. (Early 20^{th})
Scanning: the calorimeter is carefully engineered such that a temperature *program* can be imposed. Traditionally this program is a linear ramp: $T(t) = T(0) + \beta.t$ with constant heating rate β. (late 60's)
Modulated: the temperature program T(t) consists of a linear ramp *plus* a sinusoidal *oscillation* of a certain frequency, (1992)
i.e. $T(t) = T(0)+ \beta t + \alpha \sin(\omega t)$





The sample and the (empty) reference pan both sit on a slab of a material of known, low heat resistance R_q (i.e. high heat conductivity). As the oven program T(t) heats both sample plans up, heat conduction makes heat flow from the slab into both pans. If the two heat flows are not equal, the difference will cause a small temperature difference ΔT across the slab, according to the (Newton's) caloric version of Ohm's law:

$$\Delta T= K\, dq/dt \qquad \text{(cf.: } \Delta V = R.I = R.\, dQ/dt; \text{ where Q is charge).}$$

The value of K of the slab is measured by running a calibration sample upon installation of the instrument. By measuring ΔT across the slab[2], we can then measure the *difference* in heat flow into the sample compared to the reference. Because the difference is taken, the heat flow due to the two aluminum containers (more or less) cancels. As long as the sample does not undergo changes (melting, decomposing,

[1] Unfortunately all manufacturers of this type of device try to lay commercial claim to their own acronym. The technique is also known as AC DSC, ADSC, TMDSC, MTDSC etc.

[2] Actually the formula for ΔT is a bit more complicated as it also involves heat capacities for the slab and the pans, but that too can largely be calibrated out.

reaction etc) the heat flow should be proportional to the heat capacity C_p of the sample at T(t):

$$\Delta T = K\, dq/dt = K\, C_p.\, dT/dt \qquad (= K.C_p.\beta)$$

To measure C_p therefore we need to do two experiments (sample and calibration sample) at the same constant heating rate (β). Unfortunately, there are few more elements in the circuitry, e.g. the heat conductivity of the sample, the contact resistance between the sample and the slab and the fact that the temperature difference ΔT is measured at a finite distance from the actual samples. In general the relationship will become:

$$\Delta T = K.\, dq/dt = K\, C_p\, .\, dT/dt + f(T,t)$$

Once the system reaches a stationary state, the function f(t,T) tends to become a constant, so we could overcome the problem by doing experiments at different values of dT/dt and compare. This is very tedious.

In the modulated experiment one run suffices however. The instantaneous heating rate dT/dt oscillates:

$$\begin{aligned} T(t) &= T(0) + \beta t + \alpha \sin\omega t \\ dT/dt &= \quad\;\; + \beta \;\; + \alpha\omega\cos\omega t \end{aligned}$$

You could say you are experimenting at a range of heating ranges [$\beta \pm \alpha\omega$] at once. The heat flow will in general also fluctuate with the same frequency. By comparing the amplitude of the two oscillations, C_p's can be measured directly. One initial calibration run (sapphire) at installation suffices.

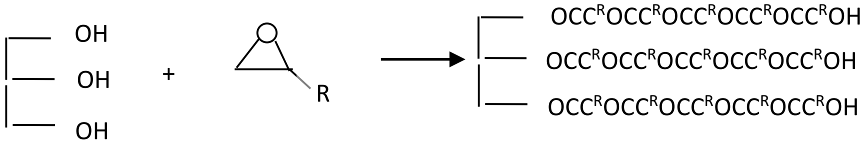
Because the heat flow dq/dt oscillates, we can distinguish an overall trend (total heat flow) and the amplitude of the oscillation (the reversing heat flow). This means that we can decompose the dq/dt response into two components a *reversing* or capacitive and a *non-reversing* or kinetic one. The latter is calculated from the difference between the total and the reversing heat flow.

The term revers*ing* is related to the thermodynamic term revers*ible* but they are not identical: a process is reversing if it is reversible on the time scale of the oscillations or faster. A non-reversing process is either reversible but too slow or simply irreversible (at all time scales).

Often the graph of C_p (T) is quite boring: a gently sloping, almost flat base line. Things get more interesting as the sample undergoes change, e.g. in a phase transition.

Phase transitions in polymers

We will compare two polymers with similar chemical structures, but rather different thermal behavior. They are both made from a multifunctional alcohol (e.g. glycerol) and an organic oxide:


OH
OH
OH
+
R
OCC^ROCC^ROCC^ROCC^ROCC^ROH
OCC^ROCC^ROCC^ROCC^ROCC^ROH
OCC^ROCC^ROCC^ROCC^ROCC^ROH

Glycerol + propylene oxide -> polyol (polyglycol) PPG
ethylene oxide PEG

We will compare R=H (polyethylene glycol) PEG and R=CH_3 (polypropylene glycol) PPG

Losing degrees of freedom

The latter is a liquid that solidifies into a glass at low temperatures. Although the molecules do not order during the glass transition, certain degrees of freedom do become inactive. The loss of degrees of freedom (DF) implies a change in C_p. The result is a shift in base line in the thermogram, usually broadened out to a sigmoidal shape. Glass transitions are typically reversing in character.

Losing degree of order

On the other hand, when a PEG melt is cooled down from high temperatures, it *crystallizes*. This means that there is not just a loss of DF, but also an *ordering* of molecules into a regular lattice. On top of the ΔC_p effect there is also a *sudden* change in entropy ΔS corresponding to the loss of order. Because we are at equilibrium at the melting point T_m we can say

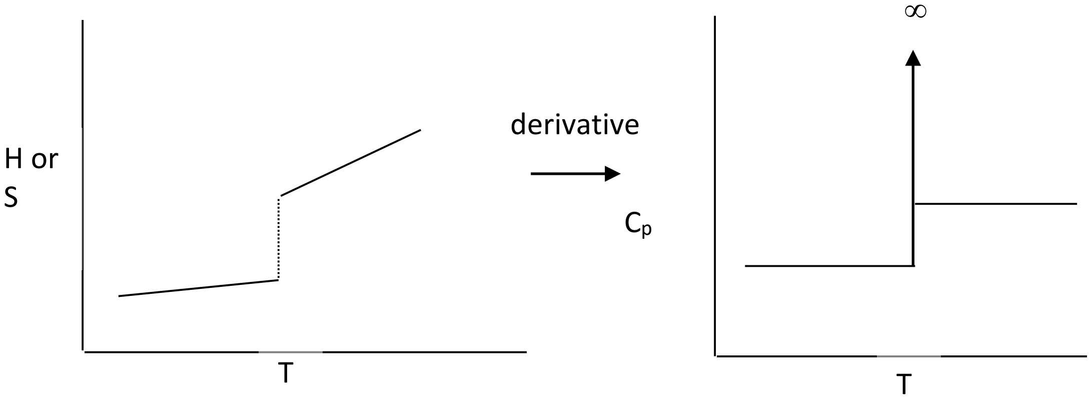
$$\Delta G = 0$$
$$\Delta G = \Delta H - T_m \Delta S = 0$$
$$\Delta H = T_m \Delta S$$

Both T_m and ΔS are finite, thus there must be a sudden jump in enthalpy, known as the enthalpy of fusion ΔH_f. A finite amount of heat q_{latent}=ΔH_f is either liberated (solidifying) or taken up (melting) instantaneously at *one* temperature. The heat capacity (that we are essentially measuring because it is linked to the heat flow) is related to the enthalpy by a derivative:

$$\left(\frac{dH}{dt}\right)_P = C_p \sim \frac{dq}{dt}\frac{1}{\beta}$$

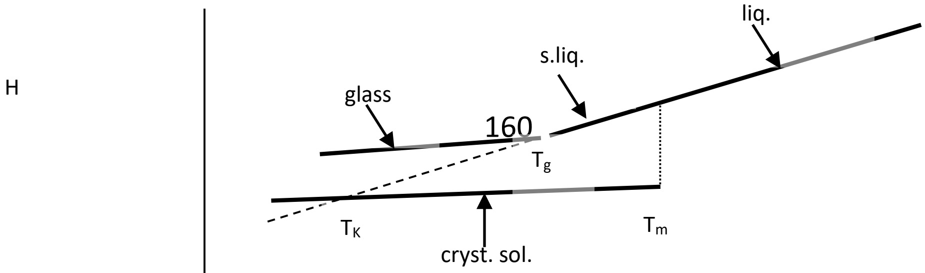
This means that the measured heat flow dq/dt should go to infinity (ΔT=zero!) but for instrumental reasons the 'spike' in the heat flow actually becomes broadened into a ***peak***. The peak area corresponds to q_{latent}.

The broadening process causes the size and the width of the peak to change with β. The onset temperature is however not very sensitive to these effects.



Melting events are typically endotherms, but they can be so instantaneous that they are over before one dT/dt oscillation has been completed. In that case MDSC cannot dissolve the heat flow into reversing and non-reversing properly. The total heat flow can still be used, however. In polymers, melting often occurs over a broader range and MDSC can then sometimes be helpful in unraveling what is happening, but we will not go into details.

When a liquid is cooled to below its melting point it sometimes fails to crystallize, in that case we get a supercooled liquid. As the slope of the H function for the liquid is generally steeper than that of the crystalline solid this leads to a paradox, named after Kauzmann. The extrapolated line for the supercooled liquid should intersect the line for the solid at what is known as the Kauzmann temperature T_K. However this leads to a physically impossible situation where a disordered liquid would be denser and the interaction inside it stronger than in the packed crystal. This cannot be. Therefore the H-curve *must* at some point flatten out before the Kauzmann temperature T_K is reached. The flattening however *implies* a loss of C_p, i.e. a loss of degrees of freedom. The temperature at or around which this happens is known as the glass transition point T_g.



For a crystalline solid like PEG we can calculate T_K once we have determined T_m, ΔH and ΔC_p. Assuming that heat capacities are constant over the temperature range we can write

$$H_{liquid} = H_{liquid}^{o} + C_{p,liquid}.T$$

$$H_{solid} = H_{solid}^{o} + C_{p,solid}.T$$

Subtracting these two expressions, we find an expression that links the three values we measure at the melting point:

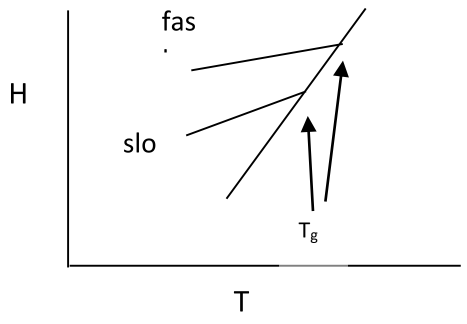
$$\Delta H =\Delta\ H^{o} + \Delta C_{p,}.T$$

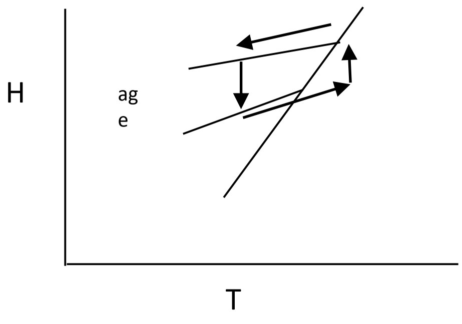
From this we can calculate T_K.

Glass transitions

Glass transitions are in a sense simpler than crystalline ones: there is df-loss but no ordering (entropy-loss). They also have a complication; they are time (or frequency) dependent. This puts them at the edge of the (static) classical thermodynamics you learn in CH433.

If the melt is cooled down *slowly*, the glass transition occurs at a *lower* temperature and the resulting glass will be denser than if you quench the melt rapidly. Schematically we can represent the enthalpy of the liquid and glassy states as follows:





If we cool down fast, but let the sample 'age' just below T_g, the system slowly relaxes to the lower enthalpy. If we then heat up fast, it 'overshoots' the 'slow' T_g and the missing enthalpy is 'recovered' at the upper 'fast' T_g point. In classical DSC this will look like an endothermic peak on top of the ΔC_p sigmoid. It has often misled people into believing they were dealing with a crystalline solid that melts. MDSC, however, nicely resolves the two effects: the recovery peak is non-reversing whereas the glass transition is *reversing*, so we will see the sigmoid and the recovery peak on two different ‘channels’.

Experiment

The instructor will show you how to open the instrument. Make sure the sample chamber is at ca. room temperature before opening the cover via the touch screen. Note the position of the reference pan and the sample pan.

First make sure the N_2 flow is on 20 psi. Make sure the RCS unit is running. (See the green switch on the refrigeration unit.) First prepare the PEG sample and make sure to get it running. In the mean time your partner can clean the syringe. The PPG sample can be prepared once the PEG is running

PEG

Carefully weigh ca 4-6 mg of PEG sample into a hermetic pan, put on a lid and crimp them together. The instructor will show you how by crimping an empty reference pan. Record the weights.

Make sure not to touch the pans with fingers, use the precision tweezers. (Don’t lose or damage the tweezers, put them back in the box afterwards).

Run conditions:

1. Equilibrate at 20^0C.
2. Modulate temperature $\pm$0.5K and a period of 60 seconds (i.e. ω=$2\pi/60$ s^{-1}).
3. Isothermal for 3 min
4. Ramp at 5^o/min to 100^0C

Data work up.

Once the isothermal statement is executed you can observe the data in the analysis software. Open your file and do a right-hand click on the graph. Select *signals* and change the selection boxes to modulated temperature (the signal you impose) and modulated heat flow (the raw signal that you measure). Go to graph and on the bottom of the dropdown opt for refreshing data. You will see the sinusoidal pattern of the signals. The software will separate that into two signals: the overall Heat Capacity (the trend line through the oscillations) and the reversing heat capacity (rev C_p). The latter gives a better measurement of the C_p anywhere but at the phase change.

Once the melt event has been observed, change the signal to heat capacity and rev C_p. Click the *reversing* C_p curve. Use the Δ icon above the data to measure the change in C_p before and after the phase transition. Let the instructor explain what this value means and how to put it in the diagrams you are supposed to include in your report. Then integrate the peak of the other curve (total heat capacity) with the first whitish integration icon.

PPG

For PPG you can use the Hamilton syringe to bring a drop of ca 10 μl in the pan. Make sure the syringe is clean and cry (use water and acetone, then blow dry with Ar). Run against empty pans. *Do weigh the sample!*

Run conditions:

For PPG: Use MDSC program.

1. Equilibrate at –92°C.
2. Modulate with amplitude ±0.5K and a period $2\pi/\omega$ of ***X*** seconds
3. isothermal for 3 minutes
4. ramp at 3K/min to –45°C.
5. Equilibrate at –92°C.
6. Modulate with amplitude ±0.5K and a period $2\pi/\omega$ of ***Y*** seconds
7. isothermal for 3 minutes
8. ramp at 3K/min to –45°C.

Use ***X=60, Y=40 or X=70, Y=50 seconds***

Methods of Analysis

For PEG, include a graph of the total heat flow. Report ΔH_{fusion} and the (onset) temperature of fusion by making a schematic drawing of the enthalpy H as a function of temperature for the PEG fusion event. Compare your diagram with the above schematic diagram for the glass transition and prepare a diagram for C_p versus T in both cases. Calculate the ΔS_{fusion} for PEG.
Estimate the ΔC_p between the liquid and the crystalline solid for PEG from the reversing C_p and use ΔH and ΔC_p to calculate the Kauzmann temperature for this compound.

For PPG : Include a graph for the total, reversible and non-reversible C_p for one of the runs, also include an overlay graph of the reversible C_p's for the two frequencies. Report the change in ΔC_p and T_g for the two values of ω (Table). Discuss the trend in T_g you see in terms of the time dependence of the glass transition. Include a graph of the non-reversible heat-flow and determine ΔH of relaxation.

Sketch a diagram for H and C_p as a function of temperature and

Why is the thermal behavior of these two related materials so different? What kinds of degrees of freedom become inactive at T_g? Why does that correspond to a ΔC_p? Compare the ΔC_p for PEG and PPG. And compare ΔH_{fusion} to $\Delta H_{recovery}$. Which comparison is closer and why is that?

References

Folmer JCW, Franzen S
Study of polymer glasses by modulated differential scanning calorimetry in the undergraduate physical chemistry laboratory
JOURNAL OF CHEMICAL EDUCATION 80 (7): 813-818 JUL 2003