Statistical and Numerical Techniques for Astronomers

1. Decision

ASTR509 Jasper Wall Fall term 2013

Decision time

Science is decision.

A list of what it is **NOT** may be infinite (and arguable):

- Building instruments
- Observing
- Reducing data
- Making graphs
- Writing code
- Writing papers
- Reading the literature
- Learning the trade: physics/astronomy/maths

These may be tools of science: but as for science itself, only decision counts.

We decide by comparing

Example: Is the faint smudge on an image a star or a galaxy?
- Measure FWHM of the point-spread function.
- Measure full-width-half-maximum, the FWHM.
- The data set, the image of the object, is now represented by a *statistic*

► Decision!

Statistics are there for decision against a background.

Every measurement, parameter or value we derive requires an error estimate, a measure of range (expressed in terms of probability) that encompasses our belief of the true value of the parameter.

No measured quantity or property is of the slightest use in decision unless it has a `range quantity' attached.

What is or are statistics? Why?

A **statistic** is a quantity that summarizes data; it is the ultimate data-reduction.

It is a **property of the data** and nothing else. It may be a number, a mean for example, but it doesn't have to be.

It is a basis for using the data or experimental result to make a decision.

We need to know how to treat data with a view to decision, to obtain the right statistics to use in drawing **statistical inference**.

It is the latter which is the branch of science; at times the term is loosely used to describe both the **descriptive values** and the science.

How not to decide - I.

Do not use the data on which a hypothesis was proposed to verify the hypothesis.

Example 1: the golf ball lands on a blade of grass.

There are 10^7 of these on the golf course.
Therefore this event is impossible? 1 chance in 10^7 ?

Example 2: *The Black Cloud* (Hoyle 1958)

It is headed for the Earth!
Therefore it is intelligent!   (?)

How not to decide - II.

The essence of classical statistical analysis is

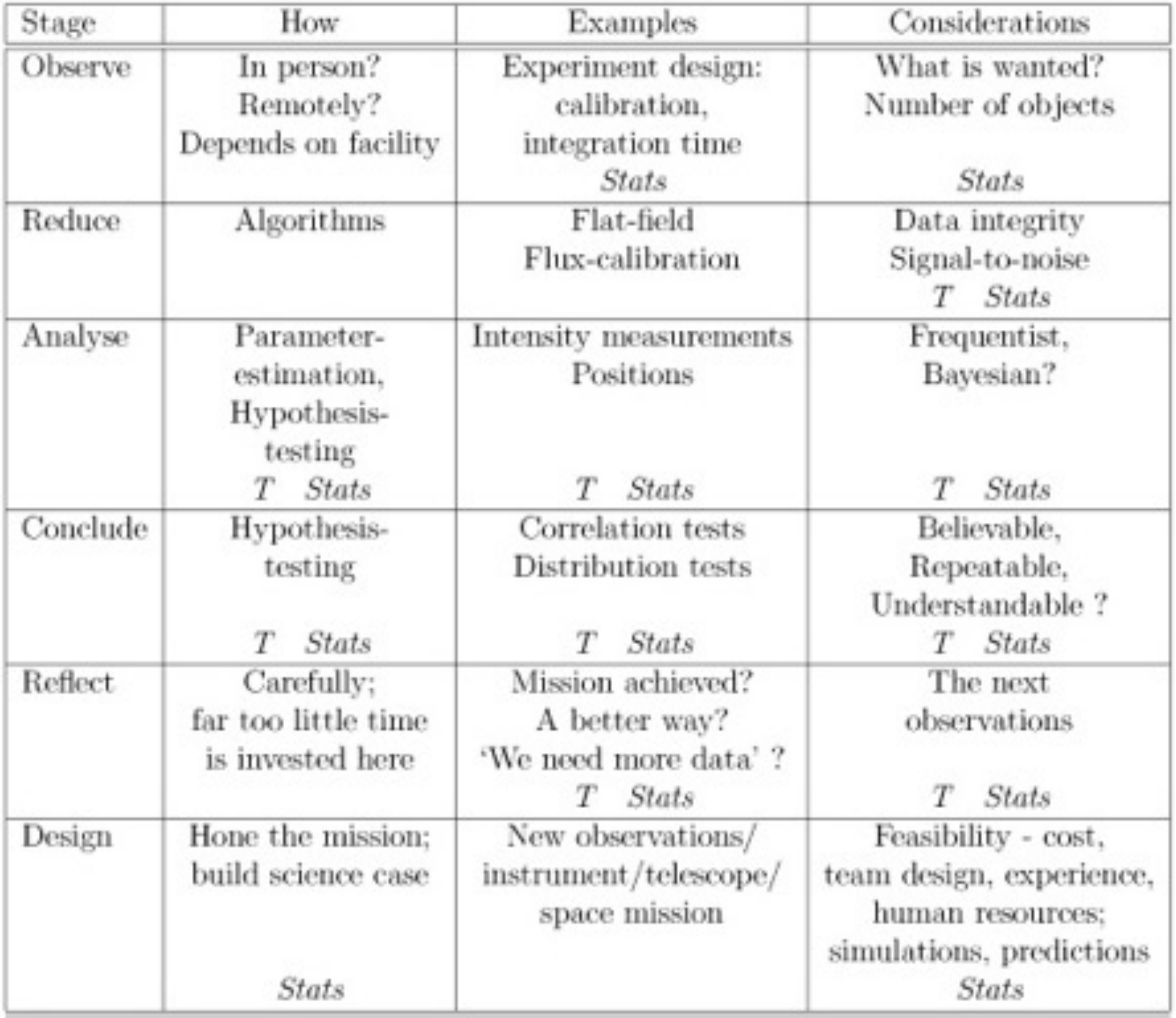
(i) the formulation of hypothesis,
(ii) the gathering of hypothesis-test data via experiment, construction of a test-statistic.
(iii) comparison with the sampling distribution.

But we can't `rerun our experiments'.
Thus we don't know the underlying distributions of the variables:

- Small samples
- Poor experimental control

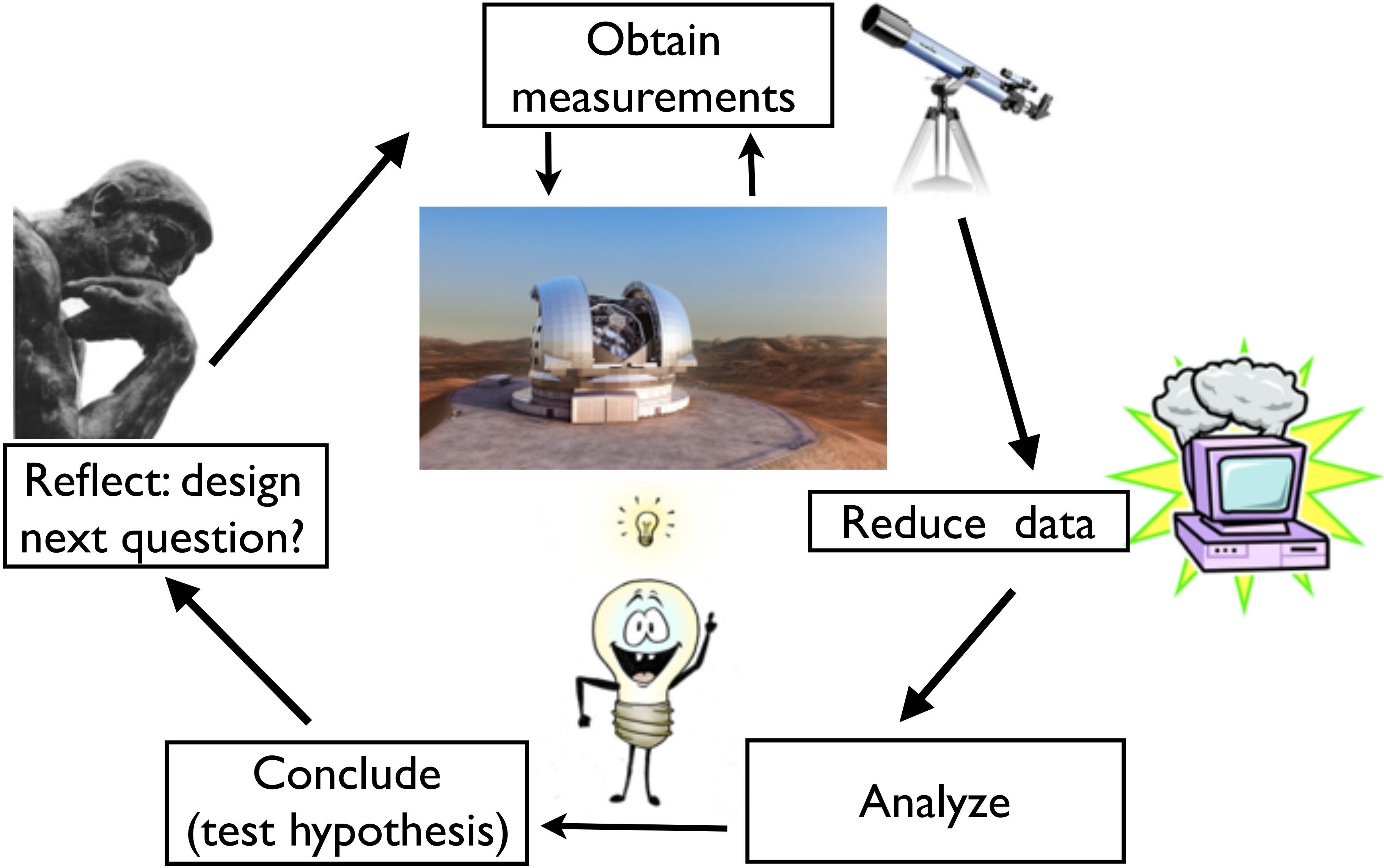
▶ **We have to be smarter than this**

The six stages of a project or experiment



Stage	How	Examples	Considerations
Observe	In person? Remotely? Depends on facility	Experiment design: calibration, integration time *Stats*	What is wanted? Number of objects *Stats*
Reduce	Algorithms	Flat-field Flux-calibration	Data integrity Signal-to-noise *T* *Stats*
Analyse	Parameter-estimation, Hypothesis-testing *T* *Stats*	Intensity measurements Positions *T* *Stats*	Frequentist, Bayesian? *T* *Stats*
Conclude	Hypothesis-testing *T* *Stats*	Correlation tests Distribution tests *T* *Stats*	Believable, Repeatable, Understandable ? *T* *Stats*
Reflect	Carefully; far too little time is invested here	Mission achieved? A better way? 'We need more data' ? *T* *Stats*	The next observations *T* *Stats*
Design	Hone the mission; build science case *Stats*	New observations/ instrument/telescope/ space mission	Feasibility - cost, team design, experience, human resources; simulations, predictions *Stats*

The process of science



Obtain measurements

Reduce data

Analyze

Conclude (test hypothesis)

Reflect: design next question?

We cannot avoid statistics...

...and there are several reasons for this unfortunate situation:

1. Error (range) assignment - ours, and theirs - what do they mean?

2. How can data be used best? Or at all?

3. Correlation, testing the hypothesis, model fitting; how do we proceed?

4. Incomplete samples, samples from an experiment which cannot be rerun, upper limits; how can we use these to best advantage?

5. Others describe their data and conclusions in statistical terms. We need some self-defense.

6. Above all, we must decide. The decision process cannot be done without some methodology, no matter how good the experiment.

Common uses of statistics

- Measuring a quantity (“parameter estimation”) : given some data, what is our best estimate of a particular parameter? What is the uncertainty in our estimate?

- Searching for correlations : are two variables we have measured correlated with each other, implying a possible physical connection? Graphics important!

- Testing a model (“hypothesis testing”) : given some data and one or more models, are our data consistent with the models? Which model best describes the data?

But wait ... 'bolt-on' statistics?

- are stats just 'bolt-on', a technological feature? `Just add water'?

- necessary, kind of unpleasant, but solved technology? **Or do we need to think?**

- Risk vs Uncertainty: widely discussed in 2008 > after the near financial meltdown

- applying statistics implies gross assumptions : recognized by Knight (1921):

    - successful firms dealt in uncertainty

    - run-of-the-mill ones (or failures) dealt in risk (known)

- Risk : known probabilities e.g the roulette wheel at a casino

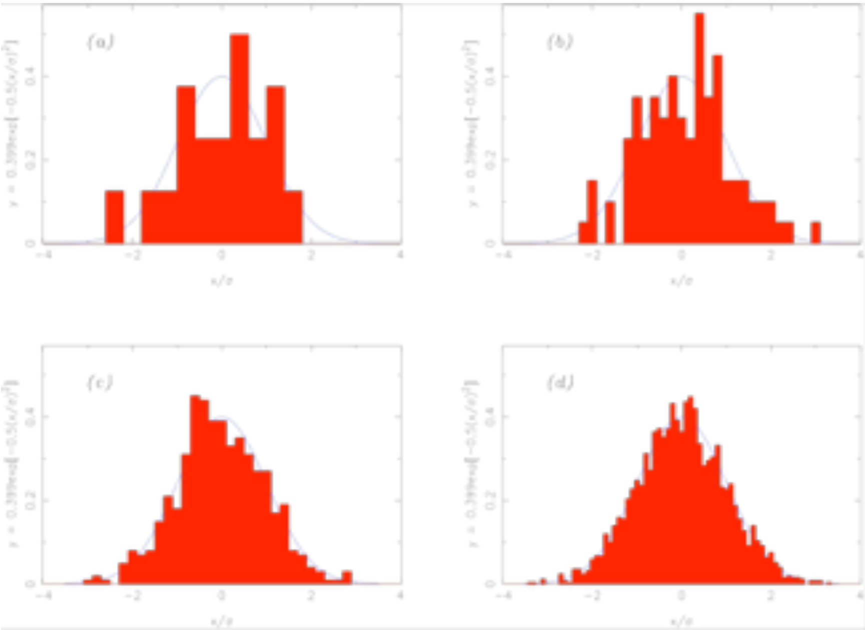
- Uncertainty: the tiger example – the greatest known loss a casino ever made.

Probability: Distributions: Statistics: Inference 1

‘Probability’ is crucial in the decision process

We have a built-in sense of probability

- from distributions or frequencies, which we ‘know’
- from experience
- from data



Consider the eye-brain system observing an approaching person .....

.....It carries out a complete scientific experiment and makes a decision

Probability: Distributions: Statistics: Inference 2

Statistics are combinations of Data – and Nothing else

Example: **average**

- we expect it bears some relation to the true mean
- we calculate the *sampling distribution* ≡ the probability of various values it may assume if we (hypothetically) repeat the experiment many times.
- we then know the probability that some range around our single measurement will contain the true mean.

This is precisely the utility of statistics: they are laboriously-discovered combinations of observations which converge, for large sample sizes, to some underlying parameter we want to know.

Probability: Distributions: Statistics: Inference 3

- the Bayesian way

A radically different way of making inferences focuses on the probabilities immediately, and to hell with statistics

Invert the reasoning just described: The **data** are unique and known!

Example: in the previous example it is the **mean that is unknown**, that should have probability attached to it. We instead calculate **the probability of various values of the mean**, given the data we have.

The approach comes far closer to answering the questions that we actually ask. Of course it allows us to make decisions.

So we should **always** use it, but for the buts:

1. but the brain works the other way
2. but other people work the other way, and we've gotta check them out
3. but the data may not be given to us in a form we can Bayesiate it
4. but there may not be a model

Non-parametric statistical methods

There are four reasons why we need these:

Θ We are measuring in experiments being run out there in the Universe, not by us. Do we know the **underlying distributions**? We can only make safe statistical inferences with non-parametric statistics, methods that do not require knowledge of the underlying distributions.

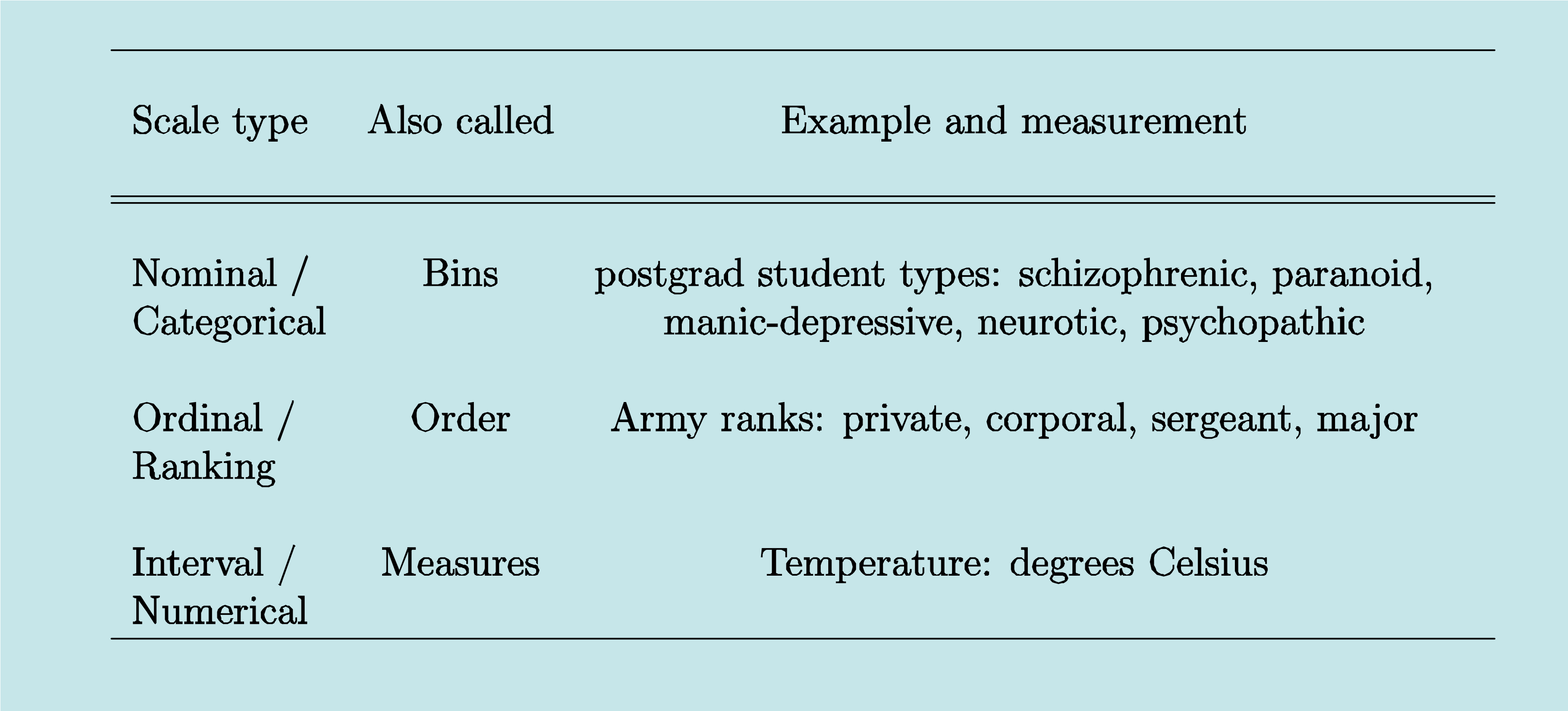
Θ We may have to deal with **small(ish) samples**, like N=3. Non-parametric techniques have the power to do this.

Θ There are different observation scales. Each such scale has formal definition, formal properties and admissible operations. Use of scales other than numerical (`interval') requires in most (but not all) cases that we use non-parametric methods. **We may wish to make statistical inference without recourse to numerical scales.**

Θ Others use such methods to draw inference. We need to understand what they are doing.

Non-parametric methods **enormously increase the possibilities in decision-making** and form an essential part of our process.

Measurement Scales



Scale type	Also called	Example and measurement
Nominal / Categorical	Bins	postgrad student types: schizophrenic, paranoid, manic-depressive, neurotic, psychopathic
Ordinal / Ranking	Order	Army ranks: private, corporal, sergeant, major
Interval / Numerical	Measures	Temperature: degrees Celsius

The point of statistics

- It allows us to formulate the logic of what we are doing and why. It allows us to make precise statements.

- It allows us to quantify the uncertainty in any measurement, which should always be stated.

- It allows us to avoid pitfalls such as confirmation bias (distortion of conclusions by preconceived beliefs)

- DECISION

Computing - some pointers

- **believe nothing**.
- **believe nothing**, especially if program compiles successfully.
- **believe nothing**, except that it's your fault, not the prog/hardware/bugs etc.

- always **declare all variables**, i.e. ‘implicit none’, at outset of program or subroutines. This keeps 'variable discipline', and minimizes mistyped variables.

- write at most **5 lines of code** at a time, eg a single simple read-in loops.
- check that every bit does the right thing, using dummy data if necessary.

-write logical constructions with **complete syntax** before filling in what you want it to do, e.g. do-loops in Fortran.

- **comment** everything.

- have **program skeletons** lying around.

- write in **modular bits**, preferably subroutines. This builds you a **library** as you go, e.g. conversion of ra, dec in (hms, dms) to (deg, deg), or sorting of an array on a given column/row. You’ll need to do it again!

Computing - more do’s and don’ts

- use **traps** in read statements to catch running off the end of file or errors in data.

- don't **write long statements**! Break long expressions up into bits. Pretend you're writing for a three-yr-old - in simple steps that you can understand when you come to look at it later.

- don't use **computed goto.**

- if you want speed, avoid **if (branching) statements**. With a bit of thought, you can do it.

- apparently impossible results/failures - totally inexplicable: commonest cause is hidden memory problems. **You've overwritten an array or a variable**.

- **don't make it look beautiful**; use lots of (commented-out) write statements, or ! statements. You think you will never forget how it works/what it does?

- don't get **programmitis. (Do you have a comprehensively-checked answer? Then MOVE ON!)**

- make sure that basic linux commands like **cp** and **rm** have a built-in check in your bash or tcsh shell so you don't accidentally overwrite things.

- put your compile/link into a **command file** so you don't have to try to remember every time how to do this incantation.

Assignments and assessment

- usually 3 or 4 problems, issued every Tues for next Tues.
- you need 80%; marking is ~linear; QED you need to do all.
- email me (1) a file containing your coding, (2) a file with results e.g. the resultant tables or diagrams, and (3) a file with explanation and summary of your results. This latter is like a README file (but call it something sensible!), and it may not always be necessary. It should not exceed a page. NO MARKS FOR STYLE, spelling, glorious colours...but if I can't understand your summary in one read, then.....
- USE SOME IDENTIFIERS IN FILE NAMES, your initials say, and the assignment number.
- if you can't meet the deadline for any particular week, e.g. illness or conference or observing: CONTACT/CONSULT ME before the deadline.
- I am looking to arrange a 'surgery' time for chats as req'd.