Notes for Research Methods

Michael Marder

January 15, 2007

Contents

1. Curiosity and Research

Science originates in curiosity about the world. However, while curiosity in its raw state is necessary for science, it is not enough. Science relies upon systematic patterns of thought, organized investigations, mathematical and computer models, specialized instruments, and many other things. The subject of this course will be the research methods that turn curiosity into science.

Although this course is intended for prospective teachers, it has little to do with courses by similar names often taught in education colleges. The goal is to obtain a feel for research the way it is carried out by active scientists, mathematicians, and computer scientists.

The course will be built around four elements: lecture sessions, homeworks, readings, and laboratory inquiries. Some of the lecture sessions will contain direct instruction on statistics, or other topics listed on the syllabus. Others will be closer to discussion or activity sessions. You will be evaluated on attendance at these sessions, but not otherwise. The readings will not be especially long, and will mainly be presented as resources, or to stimulate discussion. We will be giving out homework to emphasize certain basic skills we hope you will acquire.

The inquiries will take most of your time. In addition to two hours a week working with equipment or in discussion, you will need to spend additional time doing research or writing up results. You will find, either to your delight or to your dismay, that the assignments are less structured than in other classes. There are two reasons. First, real research lives within a context of existing knowledge, and evolves subject to many constraints, but there is no manual for it, and there is no place to look up hints for the answers. We want you to have some sense of what it is like to face a blank sheet of paper, given the task of formulating simultaneously both questions and answers. Second, as teachers you will need to develop class activities. Secondary science teachers must spend 40% of class time in laboratory activities. The task of developing your own way to present material to your own students has many elements in common with basic research.

Here are the main goals for the course:

1. Design small-scale research investigations, interacting with instructors, but without the explicit guidance typically provided by laboratory manuals.

2. Make use of mathematics to summarize and model experimental findings.

3. Employ statistics in a research setting.

4. Evaluate scientific conclusions critically.

5. Become familiar with the various forms of research, ranging from deductive to inductive, from controlled environments to observations, from pure to applied, and from theoretical to experimental and computational.

6. Become familiar with the social context of research, including the research literature, research communities, the funding process, the publication process, and the way scientific communities react to new ideas.

Here are some of the main questions that the course should help you address:

- How can I design my own inquiry and laboratory activities?

- How can I estimate what the results of the inquiries will be using simple mathematics?

- How can I explain data gathered by me or others using simple mathematics?

- How can I find the importance of chance events beyond my control in my inquiry?

1.1 Scientific Methods

Most beginning science courses describe the scientific method. They mean something fairly specific, which is often outlined as

Test a Hypothesis

1. State a *hypothesis*; that is, a falsifiable statement about the world.
2. Design an experimental procedure to test the hypothesis, and construct any necessary apparatus or human organization.
3. Perform the experiments.
4. Analyze the data from the experiment to determine how likely it is the hypothesis can be confirmed or disproved.
5. Refine or correct the hypothesis and continue if necessary.

This model of scientific investigation focuses upon starting with an hypothesis in order to prevent aimless wandering that occupies time and equipment without proving anything. It is particularly suited to medical research, where the hypothesis often concerns some new course of treatment that may be better or worse than conventional ones.

However, there is a problem with this description of science. Most of the time it simply does not describe what practicing scientists do. Since they are busy doing what they do, not describing it, the gap between conventional description and reality causes them no difficulty. Philosophers of science are perfectly aware of the discrepancy, and write about it at great length. One point of view holds that there are no rules to science at all: "Anything goes!"[1]. This point of view is certainly controversial, but there is plenty of evidence for it, and it cannot be dismissed out of hand.

I think that there are rules to science. They are difficult to articulate and include more variety than simply setting out to test an hypothesis. But if scientists themselves cannot be bothered to articulate what they do, and if philosophers of science cannot agree, why should prospective teachers care? The reason has to do with kids and curiosity. If a kid starts investigating a question in a fashion like a research scientist and a teacher notices that his or her methods do not fit the hypothesis–testing mold, the teacher may order the kid to stop or change direction. It would be a pity if that were to happen simply because of a hasty application of the dominant model of science.

Therefore I will provide below a list of some additional common modes of research. They are not certainly not an exhaustive list, but constitute the broadest list I have seen presented.

Measure a Relationship

1. Observe phenomenon
2. Identify *control variables* and *response functions*.
3. Design an experimental procedure to vary the response functions through the control variables keeping other factors constant.
4. Perform the experiments.
5. Analyze the relation between control variables and response, and characterize mathematically.

Much experimental physics proceeds along these lines. For example, if water is placed between two counter-rotating cylinders, there is a range of rotation speeds where the fluid develops a series of rolls. A research project might try to determine the boundary of the region of rotation speeds where these rolls occur, and the amplitude and wavelength of the rolls as a function of rotation speed, keeping the shapes of the cylinders fixed.

Measure a Value

[1] Paul Feyerabend, *Against Method*

1. Identify a well-defined quantity.
2. Design a procedure to measure it with increased precision.
3. Perform the experiments.
4. Analyze the accuracy of the results.

This procedure seems conceptually simpler than many of the others, but it underlies much of the most expensive large group projects science. For example, one could describe the Human Genome Project in this way, or much of experimental particle physics. The Human Genome Project declared success when it determined a string of letters that make up the sequence of the average human gene, while experimental particle physics often aims simply to measure the mass of a particular particle, such as the neutrino. Government agencies love to fund projects of this type for the simple reason that the success of the project is almost guaranteed. The researchers will come up with a collection of numbers, and those numbers are a deliverable that the government agency can display to show that the money was well spent.

While this type of investigation seems trivial if one sets out to measure the weight of a rock, it is much less trivial if one sets the goal of quantifying something more complex. For example, in testing to see whether a teaching method works, it is difficult to identify just what one is trying to measure as an “outcome,” and very difficult to design a procedure to measure it even halfway well. Nonetheless scores on standardized tests are often given as proof of success or failure of educational methods and student learning, without examining whether the tests really measure the learning that is intended.

Pure Mathematics

1. Learn the vocabulary and concepts of an existing area of mathematics.
2. Establish a relationship between two apparently different statements that can be expressed with this vocabulary.
3. Develop new vocabulary and proceed.

This model of research tries to capture what happens with pure mathematics. Sometimes the “relationship” takes the form of a conjecture that is easy to arrive at from many examples, and the hard part is proving it. At other times, the relationship could not possibly be imagined beforehand, and only emerges from the path that constitutes its proof. (In the case of Ramanujan, relationships that could not possibly be imagined without the proof are imagined anyway.) The origin of pure mathematics is particularly baffling since it seems to come out of nothing, and yet creates the most secure knowledge we ever have in any branch of the sciences.

Algorithm Development

1. Learn the vocabulary and concepts of an existing area of computer science.
2. Develop a new conceptual method for solving a problem in this area.

Computer science has its roots in formal logic and discrete mathematics, but has now become a separate discipline with its own goals and standards. Algorithm development lies on the theoretical side, and can include very abstract principles, such a proofs that a program is correct or can terminate, or ideas that underlie new classes of computer languages.

Programming

1. Learn an existing computer language.
2. Develop an application program in this language to solve a problem.

Programming bears roughly the relation to computer science that engineering bears to mathematics. It is the side of the discipline concerned with solving problems that matter to people. Programming is a slow and error–prone process, since a single mis-typed character in two million lines of code can have catastrophic effects. Much of computer science is concerned with finding ways to approach programming tasks that will maximize the generality of the solution, and decrease the likelihood of error.

When most people think of using computers, they think of using programs other people have written — computer applications. I would not classify this activity as scientific research, although it may be part of a different mode of scientific research, such as analyzing data.

Construct a Model —Applied Mathematics and Theoretical Sciences

1. Identify a regularity or relation discovered through experimental investigation.
2. Build mental pictures to explain regularity, and develop hypothesis about origin of phenomenon.
3. Identify basic mathematical relations from which regularity might result.
4. Using analytical or numerical techniques, determine whether experimental regularities result from the starting mathematical equations.
5. If incorrect, find new mathematical starting point.
6. If correct, predict new regularities to be found in future experiments.

This mode of research describes much of applied mathematics, theoretical physics, theoretical chemistry, theoretical geology, theoretical astronomy, or theoretical biology. For example, the experimental observation might be intense bursts of γ-rays. A hypothesis might be that they emerge from gravitational collapse of certain stars. A lengthy process of modeling the collapse of stars, trying to calculate the radiation that emerges from them, would be needed to check the hypothesis.

In variants of this mode of research, the modeling takes place without any experimental input, and emerges with experimental predictions. In other variants, this type of research can lead to new results in pure mathematics.

Improve a Product or Process —Industrial and Applied Research

1. Identify desired product or process.
2. Design procedure with potential to create desired outcome.
3. Build apparatus.
4. Determine whether proposed method produces desired result.
5. If not, modify until some approximation of desired outcome is achieved, until one gives up, or loses his job.
6. If so, optimize procedure with respect to speed, cost, environmental effects, and other market factors.
7. Bring product to market and continue.

Many large companies have one or more divisions devoted to "Research and Development." The research carried out in the corporate setting is usually more closely tied to an immediate profit-making goal than research in an academic setting. Here is a discussion of corporate research written by Ralph Bown, Director of Research at Bell Telephone Laboratories in 1950. Bell Labs was for around 40 years the greatest industrial laboratory in the world. The quotation is the preface to William Shockley's book *Electrons and Holes in Semiconductors*, which provided the scientific base for the creation of electronics:

> If there be any lingering doubts as to the wisdom of doing deeply fundamental research in an industrial laboratory, this book should dissipate them. Dr. Shockley's purpose has been to set down an account of the current understanding of semiconductors... But he has done more than this. He has furnished us with a documented object lesson. For in its scope and detail this work is obviously a product of the power an resourcefulness of the collaborative industrial group of talented physicists, chemists, metallurgists and engineers

> with whom he is associated. And it is an almost trite example of how research directed at basic understanding of materials and their behavior, “pure” research if you will, sooner or later brings to the view of inventive minds engaged therein opportunities for producing valuable practical devices....
>
> In the course of three years of intensive effort [an] amplifier has been realized by the invention of the device named the transistor.
>
> It would be unfair to imply that any and every fundamental research program may be expected to yield commercially valuable results in so short a time as has this work in the telephone laboratories. To achieve such results, careful choice of a ripe and promising field is prudent and a clear recognition of objectives certainly helps; but there should be no illusions about the necessity of a large measure of good luck.

Observational and Exploratory Research

1. Create instrument or method for making observation that has not been made before, or choose some location that has never been investigated.
2. Carry out observations, recording as much detail as possible, searching for anything unexpected objects or relationships.
3. Deliver results to other modes of research as appropriate.

This mode of research covers an enormous range of possibilities. It describes the expeditions that revealed the different continents to our European predecessors and mapped the globe. It describes the first investigations whenever a new scientific tool is developed. It describes the increasingly accurate maps of the night sky created by new generations of telescopes, or new particles discovered in particle accelerators. It describes some geological field work. One point of the conventional idea of the scientific method is to prevent this mode of research from proliferating once the techniques employed and objects found are no longer new.

Many research projects contain a purely exploratory phase, which produces something of interest that then becomes the subject of another mode of research. For example, a large portion of the current research in the Center for Nonlinear Dynamics at UT Austin stems from the results of putting a plate full of sand on a loud speaker and vibrating it. There was no hypothesis or specific goal originally; the graduate student was just curious to see what would happen.

Proof of Principle

1. Pick any of the preceding modes of research.
2. Design a study.
3. Carry out a small portion of the study so as to establish the technical challenges to be faced in the full study.
4. Determine whether or not the full study is feasible.

While this type of activity is not normally considered research in its own right, most scientists spend a lot of their time engaged in it. For one thing, one needs to carry out these sorts of truncated studies in applying for funds. Grant proposals are always full of references to “preliminary results.” Sometimes the investigator has actually carried out a full study already, and is just going after funding for what he or she has already done. More often, the investigator has carried out proofs of principle, and is hoping for money to complete the job. In addition, as part of every ongoing research project, the investigator will use try numerous methods on a small scale that prove to be ineffective.

Other modes of research try to establish facts about the world. Proofs of principle establish facts about the competence of a researcher, or the viability of a method.

Library Research

1. Pose question or hypothesis.

2. Search for answer in existing information sources.
3. Evaluate quality of results, decide on reliability, proceed to other forms of research or stop as a appropriate.

Because of the vast quantity of research that has already been performed, it is irresponsible to move very far through a project without attempting to determine whether the answer is already known. Searches are becoming easier and easier through the internet, although such searches do not cover everything. It is still difficult to access Soviet contributions from the 1950's and 1960's, although they may be very significant, particularly for mathematics and physics. The drawbacks of relying too heavily on literature searches are that they can lead to a sense of despair as one contemplates the mass of prior work one must understand before beginning something new, the process of studying old results can stifle new ideas, and correct and incorrect work can be difficult to distinguish.

2. Dealing with Error

2.1 Reasons one must deal with error

One of the most challenging features of making measurements in the real world is that rarely if ever does one obtain the same number twice, although one is measuring what seems to be the same thing. There are two different sorts of reasons this can happen.

2.1.1 Measurement error

The first reason is that every device ever built by humans to measure quantities in the world has some limits to its accuracy. If one tries to measure the height of a person with a ruler, it will always be difficult to tell exactly where the top of her head really lies, and if one makes a pencil mark on the ruler, it is impossible to read the height off from this mark with infinite precision. Two different people looking at a pencil mark on a ruler will in general read off different numbers. The general name for errors of this sort is *measurement error*.

2.1.2 Distributions

The second reason is that quantities we choose to measure usually cannot be described by a single number. Even so simple a quantity as the height of a single person is of this type. The height of a person can differ by more than a centimeter from morning to night. It certainly varies by more than that during a person's lifetime. The height of all freshman college students in the US is another example. It is easy to ask a question such as "Are freshmen taller in the US or in France?" To answer the question requires representing many numbers — the heights of many people — with one number — usually an average. When the quantity one wishes to measure in fact corresponds to many numbers that vary in space and time, one says it is *distributed*.

2.1.3 Precision and accuracy

One way of reducing the variation from one measurement to another is to come up with some rules that restrict the possible answers. Returning the example of measuring a person's height, one could require that it always be measured to the nearest foot. This way, one could measure someone's height repeatedly, morning, noon, and night, with many different observers and always find that Sheldon Lariviere is exactly six feet tall. This would be a very *precise* result, because the number is specified to infinite precision. However, it would not be very accurate, because his height is really much closer to 6 feet, 2 inches, give or take half an inch. This result is more *accurate* but less precise.

In scientific writing, the precision with which one measures a quantity is usually indicated by the number of digits one records. So if someone measures the period of a pendulum, and writes "I found that the period of the pendulum is 2.82934724 seconds" it is fair for the reader to ask about the marvelous instrument that was able to measure time to within a billionth of a second. In short, it is usually wrong to record all the digits your calculator or computer gives you in a writing up a scientific inquiry.

2.1.4 Improving experiments

One can always do a better or worse job of measuring numbers. So, if one wants to find the height of a four–year-old boy, it does make life much easier if one can convince him to stop jumping up and down first. There is no experiment that cannot be improved; noise and uncertainty can always be reduced by improved experimental design. Still, there will always come a point at which the improvements stop, and one pushes ahead and starts gathering numbers. It is not much help at this point to observe that with more work they might stop varying so much. For distributed quantities, the variation is truly unavoidable. So the task in this chapter is to learn how to deal with the fact that measurements of real quantities vary from time to time and place to place. Learning to do so is one of the central skills that defines a scientific approach to the world.

2.2 Coin Flips

2.2.1 Goals

We will begin by discussing a measurement that consists in nothing but flipping a coin. This example is quite rich. It contains within it many of the ideas that run throughout all analysis of error. If you understand this one example well, you should be in a good position to understand:

- Why one makes repeated measurements in determining experimental values,
- How to describe ones confidence after determining those values,
- How to design an experimental procedure to obtain a desired confidence,
- How many people from a large group to query in order to have a reliable sense of their preferences
- How to use a wide variety of special mathematical tools that have been developed to deal with tricky and complex cases.

What is there to measure about a coin? Flip it and it may come up heads, and it may come up tails. Is there anything else to say? Yes, there is. We are going to discuss measuring the *fairness* of the coin. By that, we mean the *probability* that it comes up heads versus the probability it comes up tails. If pressed to define precisely what that means, we might say that we imagine flipping the coin 100 trillion times, and the fairness is the fraction of times it comes up heads out of 100 trillion. We will view the fairness as being a property of the coin just like its size, weight, and color, and will set out to measure it. Coins do not have to have the same fairness. Magicians' shops sell coins that appear to be normal currency, but come up heads 60% of the time. One of the questions one might pose is this: "How can I tell what the fairness of my coin is, and if someone gives me two different coins, how can I tell whether they are equally fair or not?"

2.2.2 Setting up the problem

It will probably help in following the discussion to make things as real as possible.

Activity 2.1: Each student in class flips a fair coin 24 times and records the number of times each coin comes up heads.

The results in a class of 20 students might be

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Heads	13	12	4	13	8	13	15	14	15	12	15	16	15	17	14	10	11	9	14	9

2.2.3 Histogram

A compact way to summarize the results is through a *histogram* or *bar graph*, which describes the *number* of times each particular outcome for heads occurred. The histogram from the students above

Number Heads	Number Occurrences
4	1
5	0
6	0
7	0
8	1
9	2
10	1
11	1
12	2
13	3
14	3
15	4
16	1
17	1

In graphical form, the histogram appears like this:

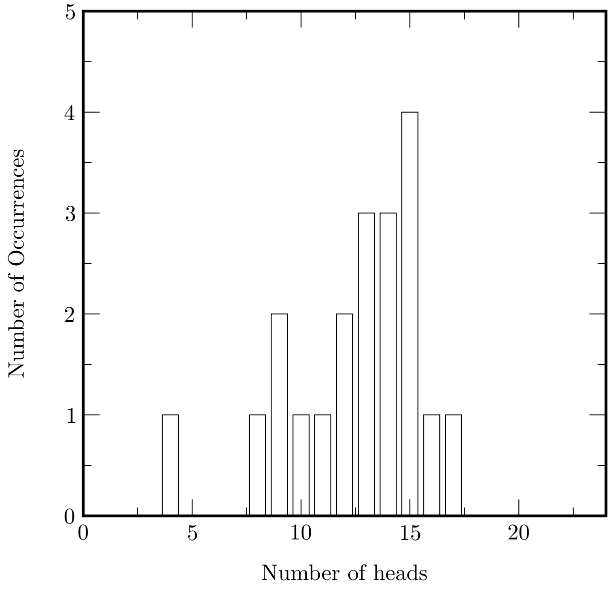


Figure 2.1. Histogram showing numbers of occurences of different numbers of heads.

2.2.4 Probability distributions

Flipping a coin is produces a distribution. There is no one right answer to the question, “Does a flipped coin come up heads or tails.” There is no one right answer to the question, “How many heads and tails will I get if I flip a coin 24 times?” However, there is a single right answer to the question, “What histogram for 24 coin flips should we get if an infinite number of people perform the experiment with fair coins?” This answer is the *probability distribution*. For a single fair coin, the probability distribution is given by saying that the probability of getting heads is $1/2$, and the probability of getting tails is $1/2$. For 24 flips of a fair coin, the probability of getting m heads, where m lies between 0 and 24 is given by the binomial distribution, and is

$$\text{Probability of } m \text{ heads} = \frac{1}{2}^{24} \frac{24!}{m!(24-m)!}. \tag{2.1}$$

The exclamation mark means to compute the factorial: $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$. This result is derived in courses on probability, and will be discussed in a bit more detail later in the notes. The main lesson to take away for the moment is that probability distributions exist, and describe quantities that can have different values.

An important property of a probability distribution is that it should describe *all possible outcomes* of an experiment, and tell how likely each of them is to happen. Whenever one makes a measurement there is some outcome, so *probability distributions must sum to 1.* For example, flipping a coin one gets heads, tails, coin balances on edge, or coin is lost and cannot be measured. The last two outcomes are so unlikely that it's fair to ignore them, and just focus on the first two, with probabilities $1/2$ each. The probability of measuring heads plus the probability of measuring tails sums to 1. One can check that Eq. (2.1) sums to one as well. This means that

$$\sum_{m=0}^{24} \frac{1}{2}^{24} \frac{24!}{m!(24-m)!} = 1 \tag{2.2}$$

2.2.5 Mean

Histograms and probability distributions provide a complete representation of the results of flipping coins 24 times, but they are still too complicated to think about very easily. Most people would like to reduce a question such as "Is this coin fair?" to one orn two numbers. Further progress toward this goal comes by reducing all the data still further into two numbers, the *mean* and *standard deviation.*

The *sample mean* $\bar{x}$ is obtained by adding up the numbers of heads obtained by all students and dividing through by the number of students:

$$\bar{x} = \frac{\begin{pmatrix} 13+12+4+13+8+13+15+14+15+12+15+16 \\ +15+17+14+10+11+9+14+9 \end{pmatrix}}{20} = 249/20 = 12.45 \tag{2.3}$$

There is a second way to get the same result, One can

- group together all the times the different integers appear, and then
- multiply each integer number of heads by the number of times it shows up.

This is the same as

- running through the histogram,
- multiplying the integer on the horizontal axis by the height of the vertical bar, and
- dividing at the end by the total number of students:

$$\bar{x} = \frac{\begin{pmatrix} 4\times1+5\times0+6\times0+7\times0+8\times1+9\times2+10\times1 \\ +11\times1+12\times2+13\times3+14\times3+15\times4+16\times1+17\times1 \end{pmatrix}}{20} = 249/20 = 12.45 \tag{2.4}$$

2.2.6 Standard Deviation

The *sample mean* or *sample average* does as good a job in condensing all the information from all the coin flips as a single number can do. There is information in the histogram, however, which this number does not convey. In particular, the mean gives no sense of how wide the histogram might be. There is a conventional way to describe the width of a histogram. It provides a *standard* way to capture in an additional number the typical amount by which measured data *deviate* from the mean, and is called the *sample standard deviation.*

Like all quantities that are now conventional, the sample standard deviation was originally invented by someone, and it may be helpful to try to retrace the steps that led to this invention. Here is a first prescription for finding how much a collection of values differs from their mean:

- Take each of the values
- Subtract the sample mean from each one in turn
- Add up all the results
- Divide by the total number

We try it:

$$\frac{\begin{pmatrix}(13-12.45)+(12-12.45)+(4-12.45)\\ +(13-12.45)+(8-12.45)+(13-12.45)+\\ (15-12.45)+(14-12.45)+(15-12.45)+\\ (12-12.45)+(15-12.45)+(16-12.45)+(15-12.45)\\ +(17-12.45)+(14-12.45)+(10-12.45)+\\ (11-12.45)+(9-12.45)+(14-12.45)+(9-12.45)\end{pmatrix}}{20}=0/20=0 \qquad (2.5)$$

The result comes out to be zero. It is no accident. If one has a collection of numbers, they lie above and below the sample mean equally, so if one adds them up after subtracting the sample mean from each one, the result vanishes. This first idea for how to characterize the way a group of numbers jiggles about a mean value is *wrong*.

So we return to the task of characterizing how a collection of numerical values differs from their mean. The problem with the procedure outlined above was that values above and below the mean canceled out. We must do something so that the number characterizing distance to the mean is positive, whether our number lies above or below it. All right, we can take each number, subtract the mean from it, and *square* the result. All entries in that sum will be positive. They cannot cancel each other out. Then we average all the squares together to find their typical value. Finally take the square root. These ideas have led to the following procedure:

- Denote the number of data points, or *sample size* by N.
- Subtract the mean from each data point in turn
- Square the result so that it is positive; note that if the quantities one is measuring have units, those units now are squared.
- Add up all the results
- Divide by sample size minus one, $N-1$.
- Take the square root, so that the units of the result are the same as the units with which one began.
- The result is called the *sample standard deviation* s.

Dividing through by sample size minus one, $N-1$, rather than by sample size N is motivated by theorems showing that this procedure provides the least inaccurate estimates one can have. See Section 2.3.8. The difference between N and $N-1$ is only significant when N is small, say less than 10, and one is trying to extract as much information as possible from a very small number of measurements.

We try it:

$$s=\sqrt{\frac{\begin{pmatrix}(13-12.45)^2+(12-12.45)^2+(4-12.45)^2\\ +(13-12.45)^2+(8-12.45)^2+(13-12.45)^2+\\ (15-12.45)^2+(14-12.45)^2+(15-12.45)^2+\\ (12-12.45)^2+(15-12.45)^2+(16-12.45)^2+(15-12.45)^2\\ +(17-12.45)^2+(14-12.45)^2+(10-12.45)^2+\\ (11-12.45)^2+(9-12.45)^2+(14-12.45)^2+(9-12.45)^2\end{pmatrix}}{(20-1)}}=\sqrt{190.95/19}=3.17 \qquad (2.6)$$

Why include the final step of taking the square root? The motivation for including this step can be understood by consider a sequence of one million numbers 0, 6, 0, 6, 0, 6, 0, 6..... We can work this example out in our heads. There are equal numbers of 0's and 6's. So the mean is 3.

For reference, the definitions of sample mean and sample standard deviation are written out in conventional mathematical form in Section 2.3.3.

Exercise 2.1: A class of 15 people flips coins 24 times and obtains the values 8, 10, 9, 18, 15, 12, 11, 12, 14, 10, 13, 10, 16, 11, 11. Compute the sample mean $\bar{x}$, the sample standard deviation s, and draw a histogram

Exercise 2.2: Find for yourself a set of 10 numbers, no two of them the same, that within 10% have sample mean of 6 and a sample standard deviation of 2.

Exercise 2.3: Invent an alternate procedure that also could be used to find characteristic widths of histograms, but does not involve the operations of squaring or taking a square root.

2.2.7 Samples versus Ideals; Statisticians versus Scientists

We have been referring to sample mean $\bar{x}$ and sample standard deviation s. What is the word "sample" doing there; why not simply talk about mean and standard deviation?

In fact, almost all scientists do simply talk about mean and standard deviation when referring to the calculations we have described above, but statisticians object because they want to reserve these words for something else. In the view of statisticians, behind every finite collection of data points there is hiding a probability distribution that is permanent and precise. If only one could take an infinite number of measurements, ones measured histogram would converge to this true probability distribution, the sample mean $\bar{x}$ would converge to the true mean μ, and the sample standard deviation would converge to the true standard deviation σ. The reason for the factor of $N-1$ rather than N in the definition of the sample standard deviation is that statisticians have proven that s defined in this way comes as close as possible on average to σ.

For example, flipping a coin 1000 times, it is very unlikely that heads will come up precisely 500 times. Actually carrying out the experiment, one will get a value such as $\bar{x} = 0.476$. This value needs to be contrasted with the ideal $\mu = 1/2$ that describes fair coins.

2.2.8 Uncertainty in the Mean — Standard Error

In the logic of statistics, there are perfect and precise values hiding behind fluctuating random measurements, and if one could only make an infinite number of measurements, one would find them. Return to the example of the coin; the coin really has a value of fairness, which is the fraction of times it would come up heads if one flipped it an infinite number of times. We found above that when 20 people flip a coin 24 times, the sample mean was 12.45. Does this result show that the coin is not fair, and comes up heads more than tails? Or does it show that the coin is fair and would come up heads 50% of the time if flipped enough?

To answer this question, one has to know how sample means $\bar{x}$ approach true means μ as one takes more and more data.

We will just report here the result of a moderately elaborate calculation. The answer is that the uncertainty in how well one knows the fairness of the coin becomes less and less as more and more people flip. To compute the uncertainty in the sample mean $\bar{x}$, take the sample standard deviation s and divide by the square root of the number of students whose measurements contributed. The result is the *standard error*,

$$s_{\bar{x}} \equiv \frac{s}{\sqrt{N}}. \tag{2.7}$$

Equation 2.7 is the single most important equation to be able to use in designing and analyzing experiments. It is not easy to derive, but conceptually it is not very complicated. What it says is this: if you are taking a series of measurements of some quantity, the individual measurements will continually fluctuate about some mean value. You can compute the mean from your measurements, and you can also compute the standard deviation s. Once you have computed the mean and standard deviation, you divide the standard deviation by the square root of your number of measurements, and that is how well you have determined the mean. The more times you measure it, the more accurately you know the mean, and the smaller the tolerances you can declare on how well you know it. You can also turn the equation around and use it to estimate how many times you will need to measure something in order to obtain desired accuracy.

So, continuing with the coin flip example, the standard error for the coin flips is $3.17/\sqrt{20} = .7$. The best compact summary of everything learned so far is that the coin comes up heads 12.45 ± 0.7 times out of 24.

2.2.9 Properties of Histograms

There are several points here that are difficult to understand on a first pass. So we return to them with additional activities:

Activity 2.2: Construct a histogram for 1000 students, each of whom flips 24 fair coins

You may think this is a joke. If the work had to be done by hand, it would be. However, there is a netlogo applet that makes it possible to see this histogram in just a few seconds

http:uteach.utexas.edu/ResearchMethods/coinflip1.nlogo

Experimenting with these computer models, you should convince yourself that no matter how many students flip 24 fair coins, the histogram describing this process never becomes narrow. However, the more students flip, and the more their results are combined into a histogram, the more and more closely the histogram converges upon an expected theoretical result, which is indicated in blue.

Activity 2.3: Now construct histograms for very large numbers of students flipping $N = 1$ fair coin, $N = 10$ fair coins, $N = 100$ fair coins, $N = 1000$ fair coins, and $N = 10000$ fair coins. Again, there is a web site to help with this: *http://stp.clarku.edu/simulations/cointoss/.* Once again, you may also make use of the netlogo model *coinflip1.nlogo* at

http:uteach.utexas.edu/ResearchMethods/coinflip1.nlogo

You can run the models until the histogram converges to a stable final shape whose values you trust.

Answer the following questions:

1. What is the mean number of heads after flipping each of these five numbers of times?
2. What is the mean *fraction* of heads in each case.
3. What is the standard deviation for the number of heads in each case?
4. What is the standard deviation for the *fraction* of heads in each case?
5. Prepare four plots, showing how mean numbers of heads, mean fraction of heads, standard deviation of heads, and standard deviation for fraction of heads vary with the number of flips N.
6. By examining the plots, try to deduce the rule for how each quantity varies as a function of the number of flips N.

2.2.10 Evaluating probabilities from histograms

When a very large number M of students flips N identical coins and their results are assembled into a histogram, the histogram adopts a shape whose mean and width stop changing. The more students contribute coin flip data, the taller the bars on the histogram will become. However, the total area under the histogram always equals the number M of students who contributed data. Therefore, it the height of each bar is divided by M, as more and more students contribute data, the histogram approaches a shape that stops changing altogether. For our coin flip example, it is a shape centered around a mean of 12, with a standard deviation of 2.449. If you feel uncertain about this point, please return to

http://www.math.uah.edu/psol/applets/BinomialCoinExperiment.html

and allow the coins to flip until you see the histogram approach the theoretical ideal outlined in blue. The histogram in this applet always scales the heights of the bars so that the area underneath them adds up to 1.

Once scaled so that the area underneath is 1, the histogram can be understood as a *probability distribution*. For example, suppose you would like to know the probability of obtaining 22 heads out of 24 flips. The answer can be found by scrolling down to the bottom of the column entitled *Distribution* in the Binomial Coin Experiment applet, and reading the value to the right of 22. It is 0.000002.

Exercise 2.4: Obtaining 22 heads out of 24 flips is very unlikely. So what is likely? Find the probability of obtaining 2 heads out of four flips, 4 heads out of 8 flips, 8 heads out of 16, and 30 heads out of 60. Do you expect the probability of obtaining exactly 50% heads to go up or down as you flip more and more times? What do you see? How can you explain it?

Once you worry about the preceding exercise, you will realize that there is a problem in the interpretation of probability distributions to overcome. Here is the solution. The height of a probability distribution at a particular point has no great significance. What has significance is the area under a probability distribution over a finite region. For example, one might ask "What is the probability of getting 75% heads **or more**?" from a fair coin. For 24 flips, this means finding the probability of getting 18 flips or more, which is $0.00802 + .00253 + .00063 + .00012 + .00002 = .011 = 1.1\%$.

This point is important enough to deserve some more examples. First, consider the number of oxygen atoms in a human. For someone around 75 kg, the number might be 4,515,829,843,508,098,043,125; that is, an integer on the order of 4.5×10^{21}. This number was just made up, and was chosen from around 10^{19} possibilities. The number 10^{19} is much larger than the number of humans on the planet. The odds that any human has this precise number of oxygen atoms in them right now is very small. But so what. What we normally want to know is not the answer to such an absurdly precise question, but the answer to a less precise question, such as "How many people are there right now whose weight is between 75.0 and 75.1 kg?" Rather than specifying a very precise numerical value, one specifies a range of values.

The same idea applies to coin flips. The odds of flipping a coin 1,000,000 times and getting heads exactly $500,000$ times are very small. Even worse, this probability is different from the probability of getting 500 heads in 1000 flips. It is much more meaningful to ask for the probability of getting heads between 50% and 51% of the time. The answer to this problem becomes independent of the number of coin flips, as the number becomes very large.

Exercise 2.5: Find the probability of obtaining between 50% and 60% heads for 15 flips, 30 flips, and 60 flips of a fair coin.

Finally, here is how probability distributions can be used. Suppose you believe you have a fair coin and you flip it 24 times, obtaining 21 heads. The probability of obtaining 21 heads (87.5% heads) or more from 24 flips of a fair coin is 0.014%. You can choose to believe that your coin is fair and came up 22 times by chance. However, the odds against this happening are so large that most people would conclude that the coin is not fair. This possibility of changing ones mind according to a careful account of likelihood is the most important outcome of statistics.

2.2.11 Limiting Distributions for Coin Flips

We return to the blue curve shown in

http://www.math.uah.edu/psol/applets/BinomialCoinExperiment.html.

A similar theoretical blue curve accompanies the model in *cointoss1.nlogo*

It is not an accident that even before taking any data, it was possible to describe the histogram that emerges after taking vast quantities of data.

The probability distribution to which all coin flipping experiments tend is the *binomial distribution*. Excel knows how to compute this distribution. To find any value on the blue histogram, one can bring up Excel, and type, for example

```
=BINOMDIST(10,24,0.5,0)
```

The = means "compute a function for me." The first argument, 10, means. "I just had 10 heads come up..." The second argument, 24, means "... out of 24 flips." The third argument, 0.5, means "The coin is fair, and has a 50–50 chance of coming up heads.” The final argument, 0, means “Tell me the probability of obtaining this number of heads,” while if it were 1 it would mean “Tell me the cumulative probability of obtaining this number of heads or less.”

Exercise 2.6: Use Excel to compute the probability of obtaining between 8 and 16 heads out of 24 flips of a fair coin.

Just knowing that Excel can compute this function gives no sense of where it comes from. The basic idea, from which all else can be derived, is that all individual strings of flips of a fair coin are equally likely. So, if many people flip a fair coin once, theoretically, one expects it to come up heads just as often as tails. The theoretical distribution in this case is a histogram where 0 and 1 heads both have probability .5.

Activity 2.4:

1. Now consider theoretical expectations for many people flipping two fair coins. How many different possible distinct outcomes are there when someone flips two fair coins?
2. For how many of these possible outcomes does one have 0, 1, and 2 heads?
3. What is the theoretical probability distribution for flipping a two fair coins?
4. Next three fair coins. Now what are all the possible outcomes? Again, from all the possible outcomes, construct the probability distribution for obtaining 0, 1, 2, or 3 heads when flipping a coin three times.

There is a systematic way to generate all possible outcomes for sequences of heads and tails. To generate all possible sequences of heads and tails in N coin flips, consider the polynomial

$$(\mathrm{H}+\mathrm{T})^N \tag{2.8}$$

There are 2^N distinct terms. Grouping all the terms with equal numbers of heads (H) and tails (T) together, the coefficients give the numbers of times that a certain number of heads and tails arose out of 2^N flips. These numbers are called *binomial coefficients*. Dividing each coefficient by 2^N gives the probability. The number of distinct sequences in which m heads arises during N flips is

$$\frac{N!}{m!(N-m)!} \tag{2.9}$$

The coefficients are also given by Pascal’s triangle.

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & .
\end{array}$$

For example, flipping a coin 5 times, the probability of getting 5 heads is $1/2^5$, the probability of getting 4 heads is $5/2^5$, 3 heads $10/2^5$, 1 head $5/2^5$ and 0 heads $1/2^5$.

2.2.12 Limiting Distributions for Everything

It may not surprise you that it was possible to work out a theory for coin flips in detail. The really surprising result is that *there is a theoretical prediction for the statistics of everything.* The statistical theory of everything applies to cases where you take a lot of data and average them. It tells you how *sample averages* fluctuate around the *true average*.

Example You have just spent three days in the lab, finding how long it takes a wet paper towel to break with a 1 kg weight clamped to the bottom. You have done the experiment 50 times, and find an average time of $\bar{x}$= 73 seconds, with a sample standard deviation s =5 seconds. You *imagine* that huge numbers of other students were stationed in labs around the country doing the same experiment. The statistical theory of everything provides you a histogram. To build this histogram, you ask "How many students got an answer between 73 and 73.1 seconds?" "How many students got an answer between 73.1 and 73.2 seconds?" "How many students got an answer between 73.2 and 73.3 seconds?"..."How many students got an answer between 72.9 and 73.0 seconds?".... And so on. To pose these questions, we had to choose a length of time in which to group data together. You can make any choice you want for this time interval, and the theory will give you an answer. However, if the time interval is too short (one millionth of a second), then you would need really huge numbers of students for any of their data to fall into it. If the time interval is too long (100 seconds) so much data will fall into it that you learn nothing.

The *Central Limit Theorem* tells you this:

- Suppose you are taking data from some experimental setup with true mean μ and true standard deviation σ. Usually when you are doing experiments, you do not know what these are, and must estimate them. But the mathematical *theory* assumes that *the mean and standard deviation exist* and that *you know them exactly*.

- You take N data points from this experimental setup. The statements of the Central Limit Theorem

$Z = (\bar{x} - \mu)/\sigma_\mu$	$\phi(\bar{x};\mu,\sigma_\mu)$
-1.5	0.129518
-1.4	0.149727
-1.3	0.171369
-1.2	0.194186
-1.1	0.217852
-1	0.241971
-0.9	0.266085
-0.8	0.289692
-0.7	0.312254
-0.6	0.333225
-0.5	0.352065
-0.4	0.368270
-0.3	0.381388
-0.2	0.391043
-0.1	0.396953
0	0.398942
0.1	0.396953
0.2	0.391043
0.3	0.381388
0.4	0.368270
0.5	0.352065
0.6	0.333225
0.7	0.312254
0.8	0.289692
0.9	0.266085
1	0.241971
1.1	0.217852
1.2	0.194186
1.3	0.171369
1.4	0.149727
1.5	0.129518

Table 2.1. Values of the normal distribution ϕ.

become exact as N becomes large.

- You find the average or sample mean of your data. The Theorem says nothing about any single measurement you will take. It tells you about what happens after you have averaged over many measurements.

- Your sample mean $\bar{x}$ will *never* equal the true mean μ *exactly.*

- The Central Limit Theorem tells you the precise probability that your sample mean differs by a certain amount from the true mean.

- The *characteristic amount* that your averaged data differ from the true mean is the *standard error*,

$$\sigma_\mu \equiv \sigma/\sqrt{N} \tag{2.10}$$

This equation is important, so I will write it again.

$$\sigma_\mu \equiv \sigma/\sqrt{N} \tag{2.11}$$

One more time for good luck:

$$\sigma_\mu \equiv \sigma/\sqrt{N} \tag{2.12}$$

- If you know how to *compute* the standard error, and if you know how to *interpret* what it tells you, then you know the central lesson of statistics for gathering experimental data. You will have moved from gathering data because of a vague belief that it is the right thing to do, to a precise understanding of how much data you *need* in order to obtain the certainty you *want*.

- The *larger* the standard deviation σ of the random data produced by your setup, the *larger* your standard error will be. The *larger* the number of data points N you take, the *smaller* your standard error will be.

- If you could have repeated your experiment vast numbers of times, your sample means $\bar{x}$ would have jiggled randomly around μ. They would have jiggled above and below it with a characteristic amount σ_μ. Therefore, define

$$Z \equiv \frac{\bar{x} - \mu}{\sigma_\mu}. \tag{2.13}$$

In vast numbers of repetitions of your experiment, the variable Z will fluctuate randomly around *zero*, and the characteristic distance by which it fluctuates above and below zero will be *one*.

- The probability that your sample mean $\bar{x}$ differs from the true mean μ by a certain amount is given by the *normal distribution*. I provide four separate representations of the normal distribution.

  1. Two graphical representations: the first shows the probability of obtaining a particular value of $\bar{x}$, using σ_μ as the unit of measure on the horizontal axis. The second shows the probability of obtaining a particular value of Z. The curves are the *same.*

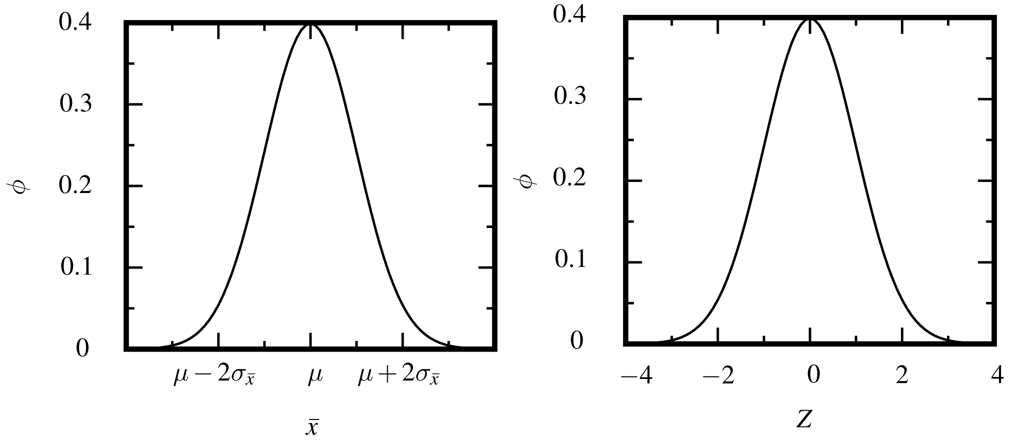


Figure 2.2. The normal distribution

2. The equation:

$$\phi(\bar{x};\mu,\sigma_\mu)) = \frac{e^{-(\bar{x}-\mu)^2/(2\sigma_\mu^2)}}{\sqrt{2\pi}\sigma_{\bar{x}}}. \tag{2.14}$$

.

3. A chart of values, Table 2.1.

- To interpret the normal distribution, ask the following question: "What is the probability of obtaining a sample mean that lies in the interval $[\mu-\sigma_\mu,\mu+\sigma_\mu]$?" The answer to this question is given by finding areas under the curve ϕ. Once again, I provide several representations:

1. Two graphical representations.

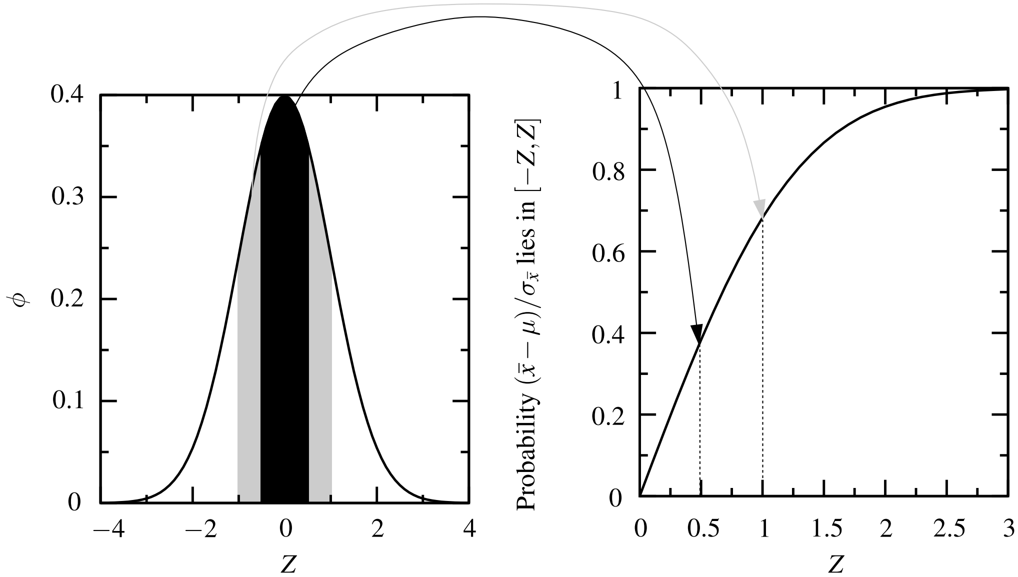


Figure 2.3. On the left, the probability that $(\bar{x}-\mu)/\sigma_\mu$ lies in the interval $[-0.5,0.5]$ or $[-1,1]$ indicated by shading in the relevant parts of ϕ. On the right, a graph of the probability $(\bar{x}-\mu)/\sigma_\mu$ lies in the interval $[-Z,Z]$ as a function of Z.

2. The equation:

$$\mathrm{P}\left(\bar{x}\in[\mu-Z\sigma_\mu,\mu+Z\sigma_\mu]\right) = \int_{-Z}^{Z} d\bar{x}\;\phi(\bar{x};\mu,\sigma_\mu)) = \int_{-Z}^{Z} d\bar{x}\,\frac{e^{-(\bar{x}-\mu)^2/(2\sigma_\mu^2)}}{\sqrt{2\pi}\sigma_\mu}. \tag{2.15}$$

.

3. A chart of values, Table 2.2

Interval size Z; Z is deviation from the mean in units of the standard error: $Z = (\bar{x} - \mu)/\sigma_\mu$.	Probability that your measured average lies between $-Z$ and Z.
0	0.000000
0.1	0.079656
0.2	0.158519
0.3	0.235823
0.4	0.310843
0.5	0.382925
0.6	0.451494
0.7	0.516073
0.8	0.576289
0.9	0.631880
1	0.682689
1.1	0.728668
1.2	0.769861
1.3	0.806399
1.4	0.838487
1.5	0.866386
1.6	0.890401
1.7	0.910869
1.8	0.928139
1.9	0.942567
2	0.954500
2.1	0.964271
2.2	0.972193
2.3	0.978552
2.4	0.983605
2.5	0.987581
2.6	0.990678
2.7	0.993066
2.8	0.994890
2.9	0.996268
3	0.997300

Table 2.2. Probability that your data fall in the interval $[\mu - \sigma_\mu Z, \mu + \sigma_\mu Z]$.

2.2.13 What to do

The statistical theory of everything tells you that *if* you know the mean and standard deviation of the experimental system from which you take your data, and *if* you repeat taking all your data an extremely large number of times *then* you know the probability with which you will obtain various sample averages.

However, in general you do *not* know the mean exactly, and you do *not* know the standard deviation exactly. It is because you do not know them that you are doing the experiment. Furthermore, you take your N data points precisely one time because that is the most data you have time to take.

Statistics seems useless because it assumes you know exactly the things you are trying to find out, and then it gives you a function that describes the outcome for doing an infinite number of times something that you can only do once.

So, what you do is to *estimate* the quantities you are trying to find out, and use statistics to describe the *confidence* with which you know them.

1. You take N data points $x_1 \ldots x_N$. You are intending to measure the *same thing* each time, and you are *repeating* the experiment so as to obtain *accurate results* by averaging.

2. You compute the sample mean $\bar{x}$.

3. You compute the sample standard deviation s.

4. You compute the sample standard error

$$s_{\bar{x}} = \frac{s}{\sqrt{N}}. \qquad (2.16)$$

   You observe that *if* you had a distribution with $\sigma = s$ and *if* you had a distribution with $\mu = \bar{x}$, *then* the standard error would be $\sigma_\mu = s_{\bar{x}}$. You define a *confidence interval* about the mean $[\mu - \sigma_\mu, \mu + \sigma_\mu]$. *If* the true mean is μ and if the true standard error is σ_μ, then 68% of attempts to perform the experiment you have just performed would lie in the range $[\mu - \sigma_\mu, \mu + \sigma_\mu]$. We conclude the value is 68% by consulting Table 2.2 and finding that the probability of measured averages falling within a range of $\pm\sigma_\mu$ about the mean is 0.68.

5. You enter the point you have just measured on a graph:

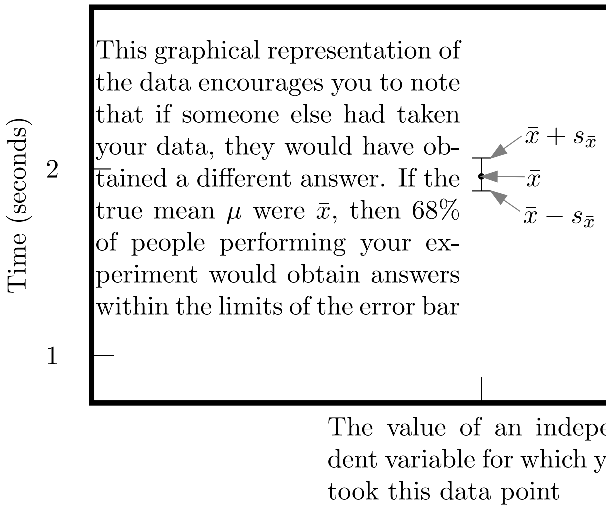

This graphical representation of the data encourages you to note that if someone else had taken your data, they would have obtained a different answer. If the true mean μ were $\bar{x}$, then 68% of people performing your experiment would obtain answers within the limits of the error bar
$\bar{x} + s_{\bar{x}}$
$\bar{x}$
$\bar{x} - s_{\bar{x}}$
Time (seconds)
2
1
The value of an independent variable for which you took this data point

Figure 2.4.

6. The point on your graph represents your best information on the value of the quantity you have measured, and the error bars indicate the range within which you would expect the point to fall 68% of the time, if you were able to repeat the experiment from the beginning many times.

7. You vary precisely one thing in your experimental setup, and again take N measurements. Once again you compute sample mean, sample standard deviation, and sample standard error, and plot them on the graph. To keep the measurements distinct, you label your first mean $\bar{x}_1$ and your second mean $\bar{x}_2$. You add similar subscripts to your standard error. You vary something yet again in your experiment, and plot point $\bar{x}_3$ on the graph. Here is the new graph, with measurements $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$ plotted as functions of Groups 1, 2, and 3.

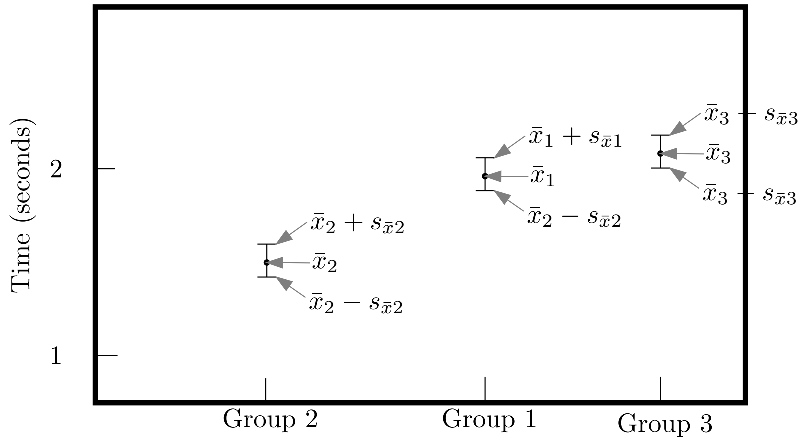


Figure 2.5.

8. Analyze the data by eye. There was something different in your opinion between Group 1 and Group 2. That is why you put them in different groups. However, maybe they are not really different. Maybe chance alone produced the difference between Groups 1 and 2. Making this guess will often be called the *null hypothesis*. By eye, you see that Group 1 and Group 2 are separated by around $3\sigma_\mu$. How likely is it that the average of *N* measurements from a system with true mean $\mu = \bar{x}_1$ would give you a value *outside* the range $[\mu - (\mu - \bar{x}_2), \mu + (\mu - \bar{x}_2)]$ = $[\bar{x}_2, \mu + (\mu - \bar{x}_2)]$? Turning to Table 2.2, you see that the probability of measurements falling *within* this interval is 0.997, so the chance of falling *outside* this interval is 0.003. That is, the probability of differing by $3\sigma_\mu$ or more from $\bar{x}_1$ due to *chance alone* is 0.3%. Most people faced with this information would conclude that chance alone was unlikely to explain the different averages you obtained for Groups 1 and 2. The *causal changes* you made in the experiment produced an *effect*. By contrast, the measurements for Groups 2 and 3 differ by only around σ_μ. If the true mean is $\mu = \bar{x}_1$, then chance alone would give the value $\bar{x}_3$ 32% of the time. Most scientists agree that one cannot conclude with adequate certainty that the averages obtained for Groups 1 and 2 are actually different.

9. The conventional point at which one says that averages $\bar{x}_1$ and $\bar{x}_3$ are different comes when $\bar{x}_3$ differs from $\bar{x}_1$ by $2\sigma_\mu$. Under these conditions, the confidence interval $[\mu + (\mu - \bar{x}_3), \bar{x}_3]$ contains 95% of the measurements around μ, and the odds of measuring something that differs from $\bar{x}_1$ as much as $\bar{x}_3$ have dropped to 5%. The *graphical* representation of this idea is that the top error bar of Group 1 just lines up with the lower error bar of Group 3, as shown below

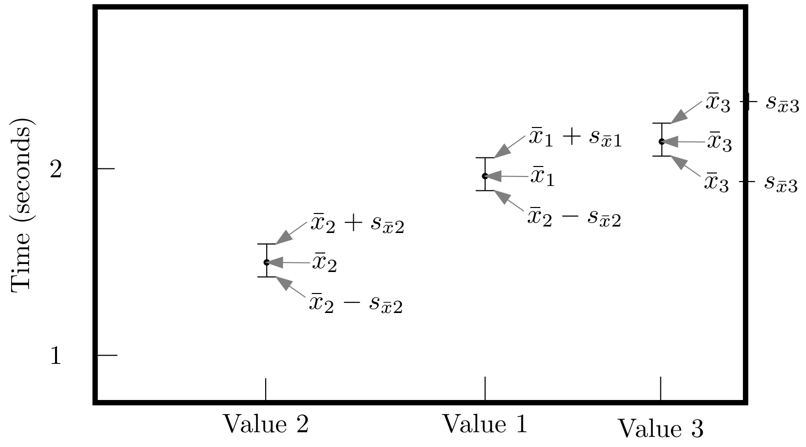


Figure 2.6.

Thus, simply checking by eye whether error bars overlap or not, one forms conclusions about whether data points can be said to be the same or different. Two points whose separation is much greater than either of their error bars are very likely to record some real change in experimental conditions. Two points whose error bars overlap cannot be distinguished from measurements of identical systems that differ only because of chance variation.

10. Suppose you are doing an experiment and find that two averaged measurements you expected to be different come out too close to distinguish, as in Figure 2.5. Now what? You have two choices. The first is to **live with it**, and admit that you found no effect. The second is to **take more data**. The more data you take, the smaller your error bars will become, and eventually you may be able to announce that you have found an effect.

11. **Warning:** It is *not* acceptable to take one data point at a time, monitoring your means and error bars and stopping the experiment as soon as they differ by a satisfying amount. This procedure will *always* tell you that you have an effect, whether or not there really is one. Instead, what you must do is to choose a target size for your error bars, calculate the number of data points N' that will enable you to obtain it, take N' measurements, and then perform a new statistical analysis.

12. **Warning:** The size of your error bars does not go down in proportion to the number of data points you take. The error bars decrease as the *square root* of the number of data points you take. Therefore, to drop the size of your confidence interval down by two, you must take *four* times as much data. If the original data took you one day, you will need to spend three more days taking data. If you set yourself a target of reducing the confidence interval by a factor of 10, you will need to take 100 times as much data. If the original data took you one day, you will need to spend over *three months* taking additional data. All real experiments are constrained by these sorts of considerations.

2.2.14 Preconceptions, and the Null Hypothesis

With these various exercises as background, we now lay out thought processes you should travel through every time you take data. There is a new idea about to be introduced, one where you always imagine that you have taken data in just one of a vast number of classrooms very similar to yours, where students similar but not identical to you are performing the same experiment. We have modeled this imaginative process through activities where you take data, as do all other students in the class, and we combine all the data into a class result. From now on you will be alone in taking data. However, we ask you to imagine that vast numbers of other students, in rooms you cannot see, are performing the same experiment. Based upon your understanding of probability, and making use of your data, we ask you to construct the histogram that would result from assembling the data of all those students. The key to using statistics is this process of constructing a histogram from vast numbers of experiments that have not actually been performed, making use of all the information and preconceptions you have available. If the histogram you construct reports to you that the data you have obtained is terribly unlikely based upon those preconceptions, you have rational grounds for questioning the preconceptions and maybe even changing them.

Often, when you start working on a problem, there is a natural result to expect, one which either common sense or expert knowledge would single out as special. If such a result exists, it is called the *null hypothesis.* For example, if you are flipping a coin, the null hypothesis is that the coin is fair and comes up heads 50% of the time. If you are testing a new medicine for trial, the null hypothesis would be that it is no better than the old medicine currently used. One might say that the null hypothesis is the belief held by your most skeptical and critical colleague.

Sometime there is no null hypothesis. For example, if someone gives you a rock and tells you to measure its weight, there is no natural and accepted value to check. You just have to measure and find out what its weight is. There is nothing to disprove.

2.2.15 Thought processes in taking data with a null hypothesis

1. Before taking any data, write down the null hypothesis. For example, if you are flipping a coin you expect to be fair, write that you expect the coin is fair, and that according to the null hypothesis it should come up heads 50% of the time.

2. Take data. Let us say that you flip the coin 1000 times.

3. Imagine vast numbers of classrooms where students have performed the same experiment. In each of them, someone has flipped a coin 1000 times. Making use of the null hypothesis, construct the histogram for vast numbers of students flipping a coin 1000 times.

4. Compare your actual data with the histogram. If your result lies near the peak of the histogram, your data are consistent with the statistical hypothesis with which you began, and there is no more to say. However, if your result lies far out in the wings, and if your preconceptions say that in vast numbers of classrooms performing your experiment it would be almost unheard of to see what you saw, you must begin to consider rejecting the null hypothesis, and changing the minds of all those who believe in the null hypothesis.

2.2.16 Thought processes in taking data without a null hypothesis

1. Before taking any data, record for yourself the statistical models that you expect to govern what you find. For example, if you are measuring the time it takes something to fall, your model would be that there is some time t you are setting out to measure, although particular measurements will deviate from it, and that the standard deviation of the measurements will have some value σ which you will try to determine as well.

2. Once you have taken data, compute the mean, standard deviation, and standard error. If you have nothing to compare your value with you are done, and you record the standard error as your uncertainty in the mean.

2.3 Reference Pages for Statistics

2.3.1 Questions Statistics Answers

Statistics appears when one makes repeated measurements, obtains many numbers, and wants to make sense of them.

- Measurements with a small, finite number of outcomes. *Examples:* Flipping a coin, obtaining heads or tails. Taking a survey, receiving answers "Yes" or "No."
    - Descriptive statistics summarize results of measurements. *Examples:* Record fraction of times coin gives heads or tails, or the fraction of people who respond "Yes" or "No" to survey. *What to use:* Descriptive statistics, averages, probability distributions.
    - Given an *hypothesis* about the likelihood of the outcomes, probability theory describes how probable it is that the data are consistent with the hypothesis. *Examples:* You might have the hypothesis that a coin should gives heads or tails with equal probability. After flipping a coin, you want to be able to state how confident you are your coin is fair. 60% of the respondents to your survey might have said "Yes." Do they represent the population as a whole, or are your results just due to chance? *What to use:* Cumulative binomial distribution.
    - You might want to know how much data to take in order to establish a certain degree of certainty. *Examples:* How many times should you flip your coin to know it is fair? How many people should you survey to have confidence in your results? *What to use:* Standard error for binomial distribution.
    - You might want to know how your outcomes depend upon discrete factors. *Examples:* Do quarters end up heads more often than pennies? Do men and women respond differently to your survey? *What to use: t* test (crude), χ^2 test, Poisson statistics for rare events.
- Measurements with continuous outcomes. *Examples:* Measure the weight of a Coke can. Measure the time it takes a penny to drop 2 feet.
    - Descriptive statistics summarize results of measurements. *Examples:* Record mean weight of coke can and standard deviation. Record mean time to drop and standard deviation. Construct histograms of weight of coke can or times for penny to fall.*What to use:* Mean, standard deviation, histogram.
    - The following tools of statistics only work if errors are distributed in a certain special way, according to *normal statistics*. You might want to check if your data fit the assumptions of the mathematical apparatus you are supposed to use. *Examples:* Errors in weighing Coke cans and penny drops are are certainly normally distributed. Magnitudes of earthquakes and sizes of islands are examples of things that are not. *What to use:* Histogram, normal or Gaussian distribution.
    - You might want to know how much data to take in order to establish a certain degree of certainty. *Examples:* How often should you record the final fluctuating digit given by the scale weighing the coke can to pin its weight down to that last digit? How often should you drop the penny to know the time it takes to fall with an uncertainty of .1%? *What to use:* Standard error.
    - You might want to know how your outcomes depend upon discrete factors. *Examples:* Do Coke cans from Atlanta weigh the same as Coke cans from Dallas? Does your penny fall at the same rate if it is wet? *What to use:* z test to compare with known mean, t test to compare two data sets.
    - You might want to know how your outcomes depend upon continuous factors. *Examples:* You open the Coke can, let the carbon dioxide out, and measure its weight as a function of time. You measure the time the penny takes to fall as a function of height. *What to use:* Error analysis, error propagation, regression analysis, χ^2 test to compare measurements with model.

2.3.2 Mean, Variance, Standard Deviation

Definitions. Suppose one has a quantity x being measured, and that the probability of finding a value of x is $\mathcal{P}(x)$. The *mean* of x or the *expected value* of x is

$$\langle x\rangle \equiv \mu = \equiv \int_{-\infty}^{\infty} \mathcal{P}(x)\, x\, dx. \tag{2.17}$$

The variance is defined similarly by

$$\left\langle (x-\mu)^2 \right\rangle \equiv \sigma^2 \equiv \int_{-\infty}^{\infty} \mathcal{P}(x)\,(x-\mu)^2\, dx. \tag{2.18}$$

and the standard deviation is defined by

$$\sigma = \sqrt{\langle (x-\mu)^2\rangle}. \tag{2.19}$$

Sometimes a variable x can take only discrete values $x_1, x_2 \ldots$. In this case, integrals over probability distributions are replaced by sums over discrete probabilities. More specifically, if p_i is the probability that x takes value x_i, then

$$\langle x\rangle \equiv \mu \equiv \sum p_i x_i \tag{2.20}$$

$$\left\langle (x-\mu)^2 \right\rangle \equiv \sigma^2 \equiv p_i (x_i - \mu)^2 \tag{2.21}$$

and once again the standard deviation is defined by

$$\sigma = \sqrt{\langle (x-\mu)^2\rangle}. \tag{2.22}$$

Example:. Suppose one has a coin for which the probability of coming up heads is p, and the probability of coming up tails is $1-p$. Denote heads by 1 and tails by 0. Then the expected number of heads is

$$\mu = (1-p)\times 0 + p\times 1 = p. \tag{2.23}$$

The variance in the number of heads is

$$\sigma^2 = (1-p)\times(0-\mu)^2 + p\times(1-\mu)^2 = (1-\mu)\mu^2 + \mu(1-\mu)^2 = \mu(1-\mu) = p(1-p) \tag{2.24}$$

and the standard deviation in the number of heads is

$$\sigma = \sqrt{p(1-p)} \tag{2.25}$$

2.3.3 Sample Mean, Sample Variance, Sample Standard Deviation

Idea. The sample mean, or sample average, of a group of numbers is a single number that can be used to represent the whole group. The sample standard deviation describes how much numbers in the group will typically differ from the average.
Given. A sequence of N numbers measuring the same quantity, $x_1, x_2 \dots x_N$. This is called a *sample.*

Questions. What is a single number that summarizes that many measurements in the sequence? What is a single number that summarizes how the measurements differ from one another?

Sample Mean. The *sample mean* or *sample average* of the measurements is

$$\bar{x} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i; \tag{2.26}$$

the sum of all the measurements divided by the number of measurements. The mean gives a single number to represent the many measurements.

Sample Variance. The *sample variance* of the measurements is

$$s^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2; \tag{2.27}$$

this is the average of the squared deviation of x_i from μ. The quantity $N-1$ is the number of independent *degrees of freedom* in the sum. See Section 2.3.8 for the justification for this formula, and see 2.3.16 for a discussion of degrees of freedom.

Sample Standard deviation. The *sample standard deviation* of the measurements is

$$s \equiv \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}; \tag{2.28}$$

this is the square root of the sample variance. The sample standard deviation gives a single number to represent how widely the measurements spread above and below the mean.

Example. Eleven people purchase a Ford Taurus on January 11 1987. The time in years before each of them buys a new car is

Person	1	2	3	4	5	6	7	8	9	10	11
Years to new car	2.5	3.1	2.2	6.3	4.1	2.3	3.1	4.5	10.4	12.3	8.4

The mean time to buy a new car is 5.38 years and the sample standard deviation about the sample mean is 3.52 years.

2.3.4 Sample Standard Error

Idea. The more often one measures a fluctuating quantity, the more accurately the sample average approaches the true mean. The sample standard error characterizes the degree of uncertainty that remains after making N measurements.

Given. A sequence of N numbers measuring the same quantity, $x_1, x_2 \ldots x_N$. The true mean of this random sequence is μ and the true standard deviation of the sequence is σ.

Questions. I average my numbers together to form the sample average $\bar{x}$. How close is $\bar{x}$ to the true mean μ?

Answer:. The likely difference between the measured sample mean $\bar{x}$ and the true mean μ is the standard error,

$$\sigma_\mu = \frac{\sigma}{\sqrt{N}} \tag{2.29}$$

Computation explaining why:. To be precise,

$$\sigma_\mu = \sqrt{\langle(\bar{x}-\mu)^2\rangle}. \tag{2.30}$$

That is, the standard error is the square root of the expected value of squared deviation between the sample mean and true mean.

Here is the computation showing that Equation 2.30 gives precisely Equation 2.29 for all values of N and independent of how the measurements x_i are distributed. There is just one condition: successive measurements of x_i must be *uncorrelated*. Intuitively, this means that knowing x_i gives one no special help at all in figuring out what x_j will be. Technically, saying x_i and x_j are uncorrelated is captured by the condition that $\langle(x_i-\mu)(x_j-\mu)\rangle = \langle x_i-\mu\rangle\langle x_j-\mu\rangle = 0$.

$$\begin{aligned}
& \left\langle \bar{(}x-\mu)^2 \right\rangle \\
&= \left\langle ([\sum_i \frac{x_i}{N}]-\mu)^2 \right\rangle \\
&= \left\langle \left(\sum_i \frac{x_i-\mu}{N} \right)^2 \right\rangle \\
&= \left\langle \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\frac{x_i-\mu}{N}\right)\left(\frac{x_j-\mu}{N}\right) \right\rangle \\
&= \sum_{i=1}^{N}\sum_{i=j}^{N} \frac{1}{N^2}\left\langle (x_i-\mu)(x_j-\mu) \right\rangle \\
&= \sum_i \frac{1}{N^2}\left\langle (x_i-\mu)(x_i-\mu) \right\rangle \quad \text{Using the fact that successive measurements are uncorrelated} \\
&= \sum_{i=1}^{N} \frac{1}{N^2}\left\langle (x_i-\mu)^2 \right\rangle = \sum_{i=1}^{N} \frac{1}{N^2}\sigma^2 \quad \text{This is just the definition of standard error: see Equation 2.19} \\
&= \frac{N}{N^2}\sigma^2 = \frac{\sigma^2}{N} \\
\Rightarrow \sigma_\mu &= \frac{\sigma}{\sqrt{N}}
\end{aligned}$$

2.3.5 Histograms and Probability Distributions

Idea. Histograms are pictures that represent measurements of supposedly a single quantity that gives a different answer every time one measures it. The higher the bar at a point on a histogram, the more likely it is to measure this value.

Given. A very large sequence of N numbers measuring the same quantity, $x_1, x_2 \dots x_N$.

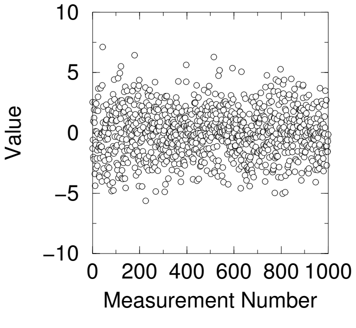
Questions. How can the measurements be displayed in a way that makes use of the fact that many of the measurements are essentially the same?

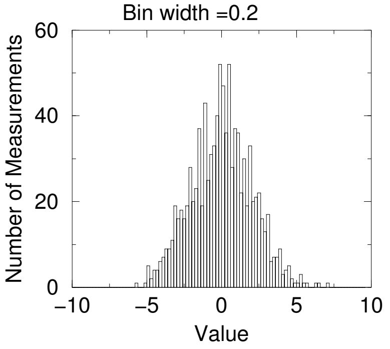
Histogram. Sort the measurements into groups. All measurements within some range are to be regarded as the same and put into the same group. On the horizontal axis place an index that varies over the groups. On the vertical axis show the number of measurements in each group.

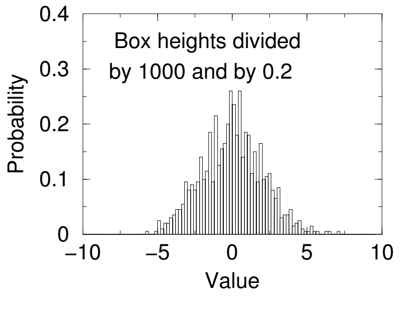
Probability Distribution. Plot a histogram, but change the vertical axis. Now divide the height of each bar by the total number of measurements N and by the bin width of each group. Once this is done, the y axis represents the *probability* of a measurement falling into a given group.

Example.

−1.2786 2.5442 −0.5886 0.9810 −2.0076 −2.1974 1.9349 1.2215 0.8505 −3.4185 2.5180 −4.3829 −0.7745 0.1316 −0.9861 −2.9297 2.4211 0.1107 1.8715 −2.1204 −2.7138 −1.4385 −0.0767
1.8333 2.9817 3.6656 −1.2161 −1.1891 −3.5463 −1.0217 −3.6671 −0.4403 −2.1058 −1.1934 3.7135 −1.4015 −0.1580 −1.1707 −0.0083 −1.1350 −4.0169 3.9440 −0.0429 7.1264 −2.0021 0.8525
2.1927 0.1770 −2.9562 −1.5594 −0.2478 0.1797 0.0763 −2.2953 0.4301 −3.3646 0.8052 −0.5388 1.0668 2.3460 1.0857 −3.9813 −0.3715 −3.4674 −2.8056 −0.5052 −1.7937 −0.0332 1.8878 −1.4054
−2.1752 −2.8963 1.5094 −4.2833 0.4809 0.8062 0.4143 −1.4486 2.6119 1.0385 0.7823 −4.8329 1.0476 1.5373 1.7955 −4.1104 0.3984 −1.1284 3.1901 −4.7823 −1.6988 1.7701 4.4316 2.7896 −1.2372
−0.2966 −1.5203 2.0613 1.4308 0.9694 1.6225 −1.1531 1.0217 −2.5166 −1.0349 −0.3024 −3.0851 −2.6653 4.5494 2.4869 −3.7372 2.0050 0.3269 2.2624 −0.4113 −3.4986 4.9675 2.6662 2.1550 1.1703
0.0269 5.5191 −0.8759 −3.7697 2.6931 −3.0932 −1.1530 −0.5290 2.4481 −3.8176 1.6876 −2.0655 −2.4828 −2.1435 1.7742 3.4000 −1.4895 2.1350 −3.7852 3.8191 −4.5570 −2.1773 0.3700 −0.5372
0.3058 0.9392 3.5107 −1.2697 −1.4349 0.4239 −2.4784 2.7887 1.8560 0.7799 1.4879 −0.0454 0.4789 −1.4395 2.7658 0.1164 1.5705 2.6590 1.0652 0.9155 0.8455 0.3217 −1.1445 −0.2003 −0.7119 −0.7506
3.4631 1.0801 −1.8429 1.3009 0.6133 −1.7390 3.9513 2.3306 0.1434 6.4459 −2.8861 −1.0111 1.5253 0.2066 −3.5093 0.7861 −0.0118 0.1000 −1.4478 2.8290 −0.0473 −1.5553 3.6632 3.0800 −2.2021
−3.2464 −4.4036 3.9264 −2.3448 4.3965 2.2660 −0.8148 −1.9048 0.2413 −2.5973 0.9976 1.6936 −0.5565 −0.6970 2.2261 1.5235 −4.4571 1.0935 −1.5195 −0.4586 0.4499 1.8109 −2.2080 −2.4450
1.8392 −1.9091 1.1926 −0.0367 1.4417 3.7576 −0.00018 −5.6311 −2.4226 −2.1358 2.5038 −2.8741 1.2066 −0.3661 1.3489 0.1777 3.2787 1.0951 3.0657 −3.0932 2.2782 0.4135 −1.9099 −0.6312 3.5701
1.8485 0.6197 −0.2644 −0.7789 −1.6320 0.9403 3.0068 0.3229 0.7408 −3.0912 0.8055 0.3835 −0.2160 −2.2847 −2.0863 −2.6133 0.7909 1.9111 1.7733 −2.8636 2.1552 4.1980 −1.1880 −2.4284 −4.8467
0.0672 2.1366 1.9924 0.1728 3.1196 −0.7401 1.3036 0.7957 1.9057 −3.1432 0.4721 2.5541 0.5288 −1.1061 0.2239 0.5628 −0.9009 0.9181 −0.7352 −1.2725 −2.0148 −1.3609 2.9414 −2.1701 −0.9448
−0.3568 1.4501 −1.8722 −1.0145 −1.3606 −0.6195 −2.0146 −0.0086 1.9998 0.2083 −2.5756 1.8724 2.9106 −4.8802 1.5053 −4.2771 −1.3729 3.0402 0.5940 −1.0394 −2.3579 3.0649 −1.8697
−0.8412 0.9050 1.2080 −0.8282 2.2117 −3.7834 0.8662 −0.0591 −1.6214 −0.0697 0.2853 −0.5315 2.5648 0.1338 1.8164 −0.3422 −0.8549 −0.3483 0.9790 −3.0767 −1.1361 −2.1876 −0.9322 1.0756
0.5110 −1.9921 −0.0467 0.8901 0.5092 −0.3098 −2.5837 1.5395 0.7203 0.0260 2.2114 −1.9351 −0.6787 −0.1482 2.9448 2.4833 1.7181 −1.1458 −1.4022 −3.1520 −0.8553 −1.1245 0.5162 4.2730
2.1075 −1.2607 −0.9456 −2.5753 −1.5597 0.3583 1.2521 1.1668 0.2471 −4.0580 −3.3781 −0.1001 −1.4318 0.8114 −1.8017 1.7593 −0.6614 1.0696 −0.1899 −1.0537 −2.6052 −2.2922 0.6788 0.7595
−2.4225 0.5425 1.2588 −2.6931 1.2373 −0.5175 2.2242 −0.1819 5.6359 4.2364 0.5060 0.5879 1.8254 −0.0090 0.2439 −1.7324 −1.1678 −0.0655 0.8646 0.5233 −1.6396 0.6715 0.4375 1.7107 0.3468
−0.2405 −0.6741 1.8548 0.7395 −0.9268 −0.1957 −0.6633 1.5470 −2.0462 −1.1394 −0.7916 −0.2607 −3.8538 −2.4225 −2.1595 0.5256 −0.1937 3.1824 2.4960 −1.3813 2.5052 2.4889 1.7575 1.9387
−1.3451 −0.8392 −0.0532 3.9447 0.4780 −3.9671 0.7060 0.9999 1.8009 0.1656 −0.0151 2.9379 −1.6059 −3.8062 1.5197 1.3992 −1.7002 −0.0442 0.9814 −2.7787 −0.5323 1.5330 −2.7083 −0.2127
−2.3941 −2.3783 1.9924 −1.8261 −0.8041 −0.6937 −1.3767 −1.5578 0.4416 1.2714 0.5764 3.5509 −1.7936 −2.4762 2.2231 1.1927 0.3957 1.8397 −0.0925 1.5119 2.4133 −0.2916 −1.3185 −3.1495
1.8512 0.7302 1.4868 −0.4909 2.1843 3.3576 −0.1679 0.3323 −1.2927 3.8851 2.6156 −0.6365 −0.0167 −1.8058 0.5392 1.5395 0.5674 −3.4116 −0.3588 1.9297 2.3780 0.2709 2.0315 0.7454 −2.1340
1.8119 1.4519 −1.2137 0.3827 6.3016 0.6146 −3.1369 −3.2734 1.8261 3.2227 0.0513 0.9060 1.6044 −2.3212 2.3527 1.1443 −0.2602 −1.1013 0.3685 −0.1187 −2.7509 −3.6235 −0.3137 4.7417 −2.2909
−0.3049 2.1598 0.4469 5.2058 2.2608 0.0303 0.0701 0.4593 1.4113 0.5205 −0.5412 3.1047 1.0982 −3.5577 0.0623 −0.6794 −2.8624 0.5832 2.4035 −4.0941 1.8786 0.3845 0.4970 2.5767 0.1174 0.9045 0.1909
2.6004 −0.1750 0.7509 1.0887 2.3862 −0.7806 −0.5347 −0.2104 1.5433 0.3155 −1.7695 −0.7301 2.2859 −4.2530 0.1581 0.5223 0.3216 0.5801 −1.5757 −1.1353 0.8801 1.3411 −1.4252 1.4370 3.1004
−1.1471 −2.2213 1.7786 2.5228 −1.6401 −0.4085 5.3686 0.7166 1.1111 2.7807 0.5058 −2.5553 0.5748 0.8340 −0.0075 1.5993 −1.8726 −1.0722 −1.0492 1.1729 0.7241 −0.4005 0.3517 1.0820 −1.5730
−1.6379 −0.4372 −0.2330 −0.3397 −0.4873 2.2681 −0.3392 −0.2431 −3.0192 −1.4327 −0.6977 −1.4366 1.7828 1.7334 1.0577 0.6951 −4.6004 −2.2413 5.0790 −2.7895 0.2570 −0.9461 0.7907
−0.2336 0.5434 1.4910 −1.7486 0.9488 0.8521 −0.7788 3.1784 −2.6609 −1.8428 −0.4485 −3.5233 0.2977 2.8366 −0.6166 0.3069 −0.9238 −3.9497 0.4345 −2.8711 −0.1538 0.2137 −3.0446 −0.5286
−1.9371 −3.0387 −2.3529 −0.5324 −2.9185 1.5775 −4.4005 −2.1798 −0.6940 −2.0087 −0.0294 −2.6743 −0.5231 −3.1477 2.6768 0.5897 −3.0824 2.9612 −3.3482 −1.0176 −0.1553 −0.4542
2.5115 −3.6759 −1.5779 −0.3477 −1.4915 1.1624 1.2236 2.1847 −0.2129 −0.1639 3.0321 −1.1708 −0.3593 1.3141 0.6896 1.6392 0.5500 −1.6071 −2.6792 −0.0091 −1.6633 0.1389 2.1274 0.8639 0.8222
−1.4617 −2.3747 2.4621 0.1541 1.1330 −0.3100 1.2260 3.3178 −1.8712 −1.4201 1.8389 −1.9291 −2.8069 −1.4744 −0.3607 2.0700 0.0661 1.4276 0.0860 −0.9942 2.1804 −1.6419 1.3313 −1.2622 1.1278
−4.0304 −2.6691 3.7853 −1.0460 −3.2546 0.5403 −1.0645 4.7774 −0.1382 −0.7746 0.2652 0.4713 0.1758 −1.4027 −0.7999 0.2052 0.0189 3.0629 2.4909 −2.0349 −0.9239 0.1478 1.8373 0.4830 3.8837
0.8133 1.3104 2.6069 −0.6519 0.1967 0.9508 −1.9090 −3.2601 −3.0904 −0.3292 −2.9871 3.4489 −2.7269 1.2628 1.1972 −0.8266 −4.9746 −2.0894 1.6176 −1.6682 −0.0729 4.4096 −1.1438 −2.0052
−1.9314 −0.6224 1.1346 −1.1683 −0.9121 −1.9539 3.4352 −3.6811 4.2731 2.3429 1.4118 −2.8424 3.1394 0.3338 0.3991 −0.5226 −1.0969 5.2833 −0.0558 −3.7018 0.8387 0.3197 0.1885 2.6102 2.8613
−5.0512 1.2975 4.0192 3.2592 0.0048 0.4776 −1.1339 4.4027 −0.1011 2.6302 −1.6480 −4.9243 1.9848 −3.5634 1.8727 −0.6557 −0.8234 1.4409 −1.0222 1.0162 −0.5781 −0.1141 −2.8537 0.1772
−2.6535 −1.6151 −3.2842 2.9084 3.1640 1.1167 −2.3328 0.8332 0.5257 −0.1487 0.0615 0.0022 −1.2455 4.4892 −0.5578 1.1513 −0.1160 0.1082 −1.5938 −1.0761 0.4265 2.3183 1.7732 −0.2844 0.8985
1.0883 2.8346 3.6508 3.1054 −3.2404 −0.5157 −0.2396 −2.1729 −0.0777 −0.1220 −2.3166 2.1420 0.7982 −0.3356 0.5750 2.3919 −1.5277 −2.4368 1.2734 0.0562 4.0330 1.9181 1.2732 0.9551 0.4446
1.9342 1.4582 0.4193 1.8610 −0.6842 2.0248 −3.9520 −3.3411 −1.7644 −1.4743 −1.0750 3.8830 0.0363 3.9734 2.9009 1.8933 −0.8715 0.7010 −2.0149 −0.2749 −0.7893 −1.6058 −1.5878 0.1577 1.5041
−0.2027 2.3083 1.6134 2.9510 −1.0659 1.1149 0.0019 −3.1054 0.9316 −1.2004 2.1319 −3.3372 −2.9651 3.1304 −1.1002 0.5016 0.6210 0.0389 0.1383 1.0955 0.9184 −0.4260 −0.9396 −2.1202 −0.1676
0.2439 2.0732 0.0324 −2.3473 −0.3996 −1.3516 −0.4693 0.3832 0.1033 2.4062 −0.5043 1.1745 −0.2387 −1.0588 −0.1687 −3.0588 −4.0922 −0.0863 −2.6245 −1.4040 0.4400 2.5722 −1.8666 1.1953
−3.0879 −0.4067 −1.5899 −2.0720 0.0177 1.5775 2.1379 −1.0894 −2.9720 −0.9590 3.6982 0.0701 −2.4379 0.5395 1.0121 −0.0028 −2.3659 2.6489 0.6706 −1.5142 −1.6650 0.0527 2.4051 0.4774
−2.1305 0.7333 −3.1763 −0.1271 −2.6083 −0.2274 1.3476 −0.0622 0.4246 2.2813 3.4840 −1.5977 −0.9616 −1.1260 2.7315 −1.0861 −2.6991 −0.4531 −1.5202 −1.5570 −0.6853 −0.1418 −1.8475







2.3.6 Normal Distribution

Idea. Many of the quantities one wants to study are averages. If one uses N measurements to obtain a sample average $\bar{x}$, it will almost never be the same as the true mean μ; instead, it will be distributed about the true mean. However, there is a miracle, and in the limit that N becomes large, the probability distribution for $\bar{x}$ becomes a universal function, whose shape is always the same, and whose width and center point can be described by just two numbers.

Given. N measurements $x_1 \dots x_N$ of some quantity, in the limit where N is very large. The measurements are floating point numbers.

Assume. One computes the sample average $\bar{x}$, and the values going into the average do not fluctuate too wildly.

Normal Distribution. The *Central Limit Theorem* states that the probability distribution for $\bar{x}$ is a *normal distribution*, which has the form

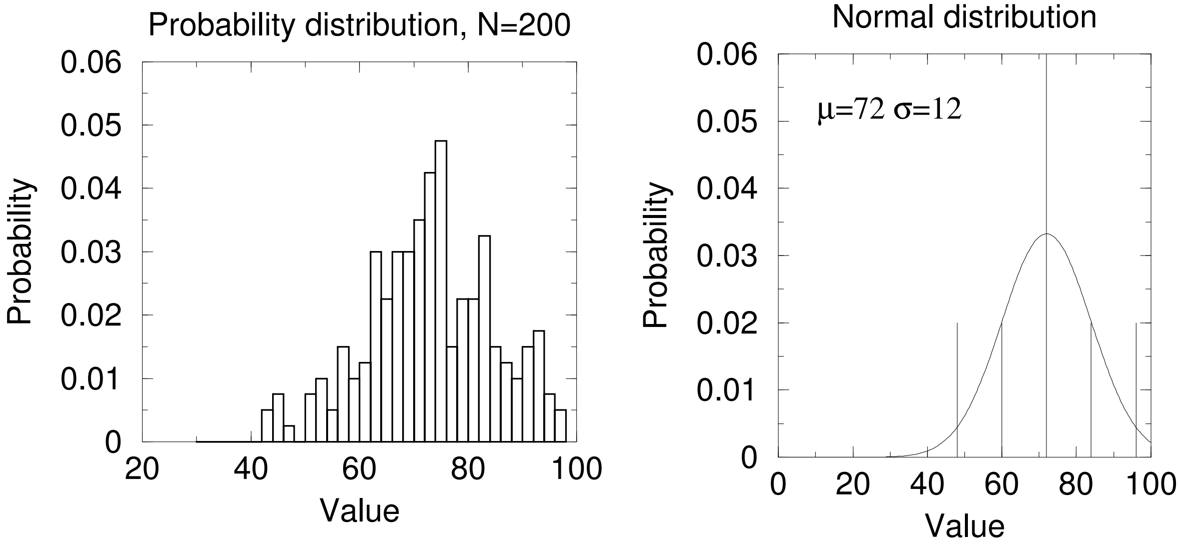
$$\phi(\bar{x}) = \frac{e^{-(\bar{x}-\mu)^2/(2\sigma_\mu^2)}}{\sqrt{2\pi}\sigma_\mu}. \tag{2.31}$$

Normal distributions are described by the two numbers μ and σ_μ, which give the point where they are centered and their width. Otherwise, all normal distributions are the same.

Writing the normal distribution out explicitly with the expression for standard error and sample size N gives

$$\phi(\bar{x}) = \frac{e^{-(\bar{x}-\mu)^2/(2\sigma^2/N)}}{\sqrt{2\pi}(\sigma/\sqrt{N})}. \tag{2.32}$$

Example. In the picture below, the left panel shows a histogram of 200 measurements drawn randomly from a normal distribution with mean 72 and standard deviation 12, while the right panel shows the normal distribution with mean 72 and standard deviation 12.



What is **not** true is that all distributed quantities in the world are normal. The heights of people are not normally distributed. Test scores of kids are not normally distributed. For example, scores often cannot be larger than 100 or smaller than 0, but the normal distribution would predict a nonzero probability. A case where real tails are larger than normal tails arises with earthquakes and many other natural phenomena. According to the normal distribution, the probability of earthquakes larger than about 6 on the Richter scale would be negligible, but such earthquakes do occur with a measurable frequency.

2.3.7 Sampling

Idea. If you are measuring something, and your measurements are distorted by error, the more often you measure it, the more accurately your average value will approach the true value you are trying to measure.

Given. A sample consisting of N measurements $x_1 \dots x_N$ of some quantity. The measurements are floating point numbers.

Assume. The measurements are independent, and come from an underlying distribution which has mean μ and standard deviation σ.

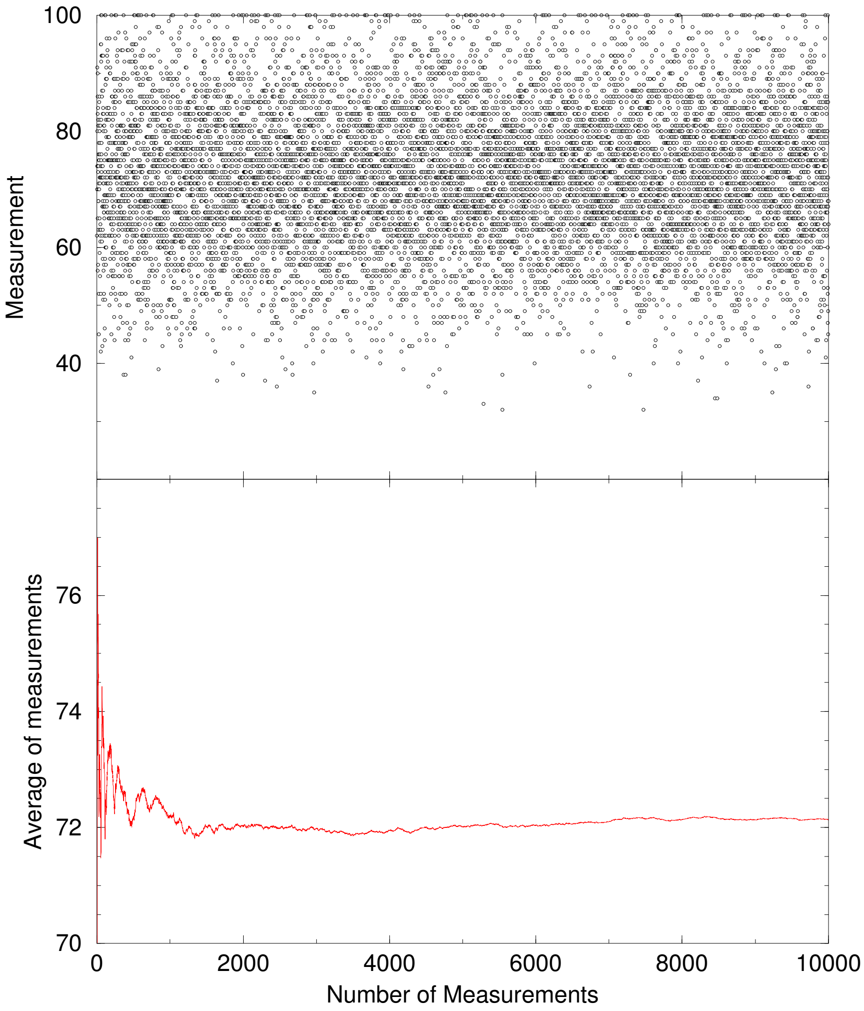
Normal Distribution. As N becomes large, the sample mean $\bar{x}$ will approach the mean μ, and the sample standard deviation s will approach σ.

Averages. The average $\bar{x}$ after N measurements is normally distributed around μ. The fluctuations of $\bar{x}$ around μ are described by the standard error, which one can estimate from the *sample standard error*, $s_{\bar{x}}$ whose value is

$$s_{\bar{x}} = \frac{s}{\sqrt{N}}. \tag{2.33}$$

Example.

In the example below, measurements are drawn from a normal distribution with mean 72 and standard deviation 12 (with however an upper bound of 100). The average of the measurements is plotted as a function of the number of samples entering in the average. So when the horizontal axis has 100, the first 100 measurements were averaged, and so on. Notice that the average converges toward the expected mean value of 72.



2.3.8 Unbiased Estimators

Idea:. When one takes a limited number of measurements to estimate some quantity, one should perform the estimation so that the expected value of the estimate gives the right answer.
Computation:. Consider trying to estimate the standard deviation of a quantity x by taking a sample of N measurements. One might guess that the best procedure would be to compute

$$\sigma^2 \approx^? \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2. \tag{2.34}$$

However, computing the expected value of this quantity shows that it does not provide the best possible estimate of σ^2. First, recall

$$\sqrt{\langle(\bar{x} - \mu)^2\rangle} = \sigma_\mu = \frac{\sigma}{\sqrt{N}}. \tag{2.35}$$

This relation is exact for all N so long as the distribution $\mathcal{P}(x)$ is well behaved and successive measurements are not correlated.

Using this relation, and the fact that the expected value of a sum is the sum of expected values ($\langle A + B\rangle = \langle A\rangle + \langle B\rangle$) gives

$$\left\langle \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \right\rangle \tag{2.36}$$

$$= \left\langle \frac{1}{N}\sum_{i=1}^{N}([x_i - \mu] - [\bar{x} - \mu])^2 \right\rangle \tag{2.37}$$

$$= \left\langle \frac{1}{N}\sum_{i=1}^{N}\left\{(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2\right\} \right\rangle \tag{2.38}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{ \left\langle(x_i - \mu)^2\right\rangle - 2\left\langle(x_i - \mu)(\bar{x} - \mu)\right\rangle + \left\langle(\bar{x} - \mu)^2\right\rangle \right\} \tag{2.39}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{ \left\langle(x_i - \mu)^2\right\rangle + \left\langle(\bar{x} - \mu)^2\right\rangle \right\} - 2\left\langle (\frac{1}{N}\sum_{i=1}^{N}x_i - \mu)(\bar{x} - \mu) \right\rangle \tag{2.40}$$

$$= \left\{\frac{1}{N}\sum_{i=1}^{N}\left\langle(x_i - \mu)^2\right\rangle\right\} + \left\langle(\bar{x} - \mu)^2\right\rangle - 2\left\langle(\bar{x} - \mu)(\bar{x} - \mu)\right\rangle \tag{2.41}$$

$$= \left\{\frac{1}{N}\sum_{i=1}^{N}\left\langle(x_i - \mu)^2\right\rangle\right\} - \left\langle(\bar{x} - \mu)^2\right\rangle \tag{2.42}$$

$$= \left\{\frac{1}{N}\sum_{i=1}^{N}\sigma^2\right\} - \left\langle(\bar{x} - \mu)^2\right\rangle \tag{2.43}$$

$$= \sigma^2 - \frac{\sigma^2}{N} \tag{2.44}$$

$$= \sigma^2\frac{N-1}{N}. \tag{2.45}$$

What this computation shows is that Eq. (2.34) does not on average give the correct value of the variance; it gives $(N-1)/N$ times the variance. For this reason, the sample variance is defined instead as

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2; \tag{2.46}$$

repeating the calculation above with sample variance defined in this way gives the happy result that

$$\left\langle s^2 \right\rangle = \sigma^2. \tag{2.47}$$

2.3.9 *Z* Test

Idea. The Z test helps you decide whether some values you have measured could reasonably have come from a distribution with known mean μ and known standard deviation σ.

Given.

1. N measurements $x_1 \dots x_N$ of some quantity.
2. A claim about the mean value μ one would get after an infinite number of measurements.
3. A claim about the standard deviation σ one would get after an infinite number of measurements.

Assume. The measurements are drawn from a normally distributed population with with standard deviation σ.

Questions. How likely is it for me to obtain the value of $\bar{x}$ I see if I assume that the underlying distribution has mean μ and standard deviation σ?

Z. Define Z by

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \tag{2.48}$$

where $\bar{x}$ is the average of the measurements. So Z is the ratio between [how far the mean you measure is from the mean you expect] and [the standard error]. If this ratio is large, your data may be telling you that you were expecting the wrong mean.

95% Confidence Interval. The 95% confidence interval for μ is $[-1.96 < Z < 1.96]$. If Z lies outside of this interval, say "I reject the *null hypothesis* $\bar{x} = \mu$ and accept the alternate hypothesis $\bar{x} \neq \mu$ with a confidence of 95%."

p-value (two-sided). The two-sided *p-value* gives the probability of drawing N measurements from a distribution with mean μ and standard deviation σ and obtaining a value of z whose absolute value is greater than what was actually obtained.

$$p = \Pr(z < |\mathrm{Z}|) = 1 - 2\int_0^{|} \mathrm{Z}| dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} \tag{2.49}$$

The smaller is p, the greater the evidence we have to to accept the alternative hypothesis that $\bar{x} \neq \mu$.

p-value (one-sided). The one-sided *p-value* for $\bar{x} > \mu$ gives the probability of drawing N measurements from a distribution with mean μ and standard deviation σ and obtaining a value of z that is greater than what was actually obtained.

$$p = \Pr(z > \mathrm{Z}) = \int_Z^{\infty} dz \frac{e^{-z^2/(2)}}{\sqrt{2\pi}} \tag{2.50}$$

The smaller is p, the greater the confidence to accept the alternative hypothesis that $\bar{x} > \mu$.

p-value (one-sided). The one-sided *p-value* for $\bar{x} < \mu$ gives the probability of drawing N measurements from a distribution with mean μ and standard deviation σ and obtaining a value of z that is less than what was actually obtained.

$$p = \Pr(z < \mathrm{Z}) = \int_{-\infty}^{Z} dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} \tag{2.51}$$

The smaller is p, the greater the stronger the evidence to accept the alternative hypothesis that $\bar{x} < \mu$.

2.3.10 *t* Test

Idea. The t test tells you the liklihood that an average $\bar{x}$ you have measured comes from a distribution with mean μ ; you no longer assume that you know the standard deviation in advance.

Given.

1. N measurements $x_1 \dots x_N$ of some quantity.
2. A claim about the mean value μ one would get after an infinite number of measurements.

Assume. The measurements are drawn from a normally distributed population.

Questions. How likely is it for me to obtain the value of $\bar{x}$ I see if I assume that the underlying distribution has mean μ and standard deviation s?

t. Define t by

$$t = \frac{\bar{x}-\mu}{s/\sqrt{N}} = \frac{\bar{x}-\mu}{s_{\bar{x}}}, \tag{2.52}$$

where $\bar{x}$ is the sample mean, and the sample standard deviation is

$$s \equiv \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}; \tag{2.53}$$

So t is the ratio between [how far the sample mean lies from the mean you expect] and [the standard error]. If this ratio is large, your data may be telling you that you were expecting the wrong mean. It differs from Z only because you don't know the standard deviation σ in advance, but have to estimate it from the data.

95% Confidence Interval. The 95% confidence interval for t is $[-1.96 < t < 1.96]$. If t lies outside of this interval, say "I reject the *null hypothesis* $\bar{x} = \mu$ and accept the alternate hypothesis $\bar{x} \neq \mu$ with a confidence of 95%."

p-value (two-sided). The two-sided p-value gives the probability of drawing N measurements from a distribution with mean μ and standard deviation σ and obtaining a value of z whose absolute value is greater than what was actually obtained. The correct formula for p is complicated, depends upon the number of measurements N and we will not record it. Excel performs the calculation with *TTEST(t,N,1).* For large sample sizes it is *approximately*

$$p \approx \Pr(|x| > |t|) = 1 - 2\int_0^t dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} \tag{2.54}$$

The smaller is p, the greater the evidence to accept the alternative hypothesis that $\bar{x} \neq \mu$.

p-value (one-sided). The one-sided p-value for $\bar{x} > \mu$ gives the probability of drawing N measurements from a distribution with mean μ and standard deviation σ and obtaining a value of t that is greater than what was actually obtained. The smaller is p, the greater the evidence to accept the alternative hypothesis that $\bar{x} > \mu$.

p-value (one-sided). The one-sided p-value for $\bar{x} < \mu$ gives the probability of drawing N measurements from a distribution with mean μ and standard deviation σ and obtaining a value of t that is less than what was actually obtained. The smaller is p, the greater the evidence to accept the alternative hypothesis that $\bar{x} < \mu$.

2.3.11 2-sample t Test

Idea. The 2-sample t test tells you whether two quantities you have measured are greater than, less than, or different from each other.

Given.

1. N_x measurements $x_1 \ldots x_{N_x}$ of some quantity, and N_y measurements $y_1 \ldots y_{N_y}$ of a comparable but potentially different quantity.

Assume. The measurements are drawn from a normally distributed populations, but the first and second sets of measurements may have different means. The measurements come from normal distributions with the same standard deviation (pooled data).

Question. How likely is it for me to obtain the two values $\bar{x}$ and $\bar{y}$ I see if all my measurements come from the same underlying probability distribution?

t. Let $\bar{x}$ and $\bar{y}$ be the averages of the N_x and N_y measurements of x and y. If there is doubt that the standard deviations of the two sequences are the same, one can define

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{(s_{\bar{x}}/\sqrt{N_x})^2 + (s_{\bar{y}}/\sqrt{N_y})^2}}. \tag{2.55}$$

For conservative probability estimates, use this quantity with tables of t distributions using the smaller of N_x and N_y for the number of degrees of freedom.

If there is confidence that the standard deviations of x and y are the same, one may define the *pooled standard error* s_D

$$s_D = \sqrt{\frac{\sum_{i=1}^{N_x}(x_i - \bar{x})^2 + \sum_{i=1}^{N_y}(y_i - \bar{y})^2}{N_x + N_y - 2}\left(\frac{1}{N_x} + \frac{1}{N_y}\right)} \tag{2.56}$$

and t by

$$t = \frac{\bar{x} - \bar{y}}{s_D}. \tag{2.57}$$

and employ $N_x + N_y - 2$ degrees of freedom.

95% Confidence Interval. The 95% confidence interval for t is $[-1.96 < t < 1.96]$. If t lies outside of this interval, say "I reject the *null hypothesis* $\mu_x = \mu_y$ and accept the alternate hypothesis $\mu_x \neq \mu_y$ with a confidence of 95%."

p-value (two-sided). The two-sided p-*value* gives the probability of drawing $N_x + N_y$ measurements from a distribution with mean 0 and standard deviation s_D, and obtaining a value of t whose absolute value is greater than what was actually obtained. The precise formula for p is again too complicated to record here. The smaller is p, the greater the evidence to accept the alternative hypothesis that $\mu_x - \mu_y \neq 0$.

p-value (one-sided). The one-sided p-*value* for $\mu_x > \mu_y$ gives the probability of drawing $N_x + N_y$ measurements from a distribution of common mean value and standard deviation s_D and obtaining the observed value of $\mu_x - \mu_y$ (or a greater one) by chance. The smaller is p, the greater the confidence to accept the alternative hypothesis that $\mu_x > \mu_y$.

p-value (one-sided). The one-sided p-*value* for $\mu_x < \mu_y$ gives the probability of drawing $N_x + N_y$ measurements from a distribution of common mean value and standard deviation s_D and obtaining the observed value of $\mu_x - \mu_y$ (or a lesser one) by chance. The smaller is p, the greater the confidence to accept the alternative hypothesis that $\mu_x < \mu_y$.

2.3.12 Binomial Distribution

Idea. The binomial distribution tells you the probability of series of events that are chosen randomly from two possible outcomes; for example, it gives the probability of obtaining 38 heads when flipping a fair coin 100 times.
Given. Measurements have two possible outcomes, **yes** or **no**. In an infinite series of measurements, the fraction of times **yes** appears is f and the fraction of times **no** appears is $1-f$.
Questions. One makes n measurements, and obtains **yes** m times and **no** $n-m$ times. What were the odds according to pure chance of obtaining these measurements?
Answer:. Suppose N samples are drawn from a collection where the fractions of **yes** and **no** are f and $1-f$. Treat all cases where m samples of **yes** and $n-m$ samples of **no** arise as equivalent, no matter what order the **yes** and **no** appear. Then the probability of making this measurement is

$$p_m = \frac{N!}{m!(N-m)!} f^m (1-f)^{N-m}. \tag{2.58}$$

As N becomes large, all the probabilities p_m tend to zero, so it can be more helpful to consider the cumulative binomial distribution, which gives the probability of drawing m or fewer samples of **yes**. The cumulative distribution is obtained by summing up all values of $p_{m'}$ where $m' \leq m$.
Explanation:. Each time **yes** appears, the probability of that happening was f, so if there are m instances of **yes** and $N-m$ instances of **no**, the factor of $f^m(1-f)^{N-m}$ should not be too surprising. But why the additional nonsense with the factorials? That factor comes from asking what were all the possible different ways to get the same measurement. For example, suppose one makes 10 measurements, and there are two cases of **yes** and eight cases of **no**. The ways to get this are:

1. Getting the first **yes**: The first measurement could be yes. Or the second measurement could be yes. Or the third could be yes. All in all, there are ten possibilities.

2. Getting the second **yes**: If the first measurement was yes (from the first step), then the second could also be yes, the third could also be yes, the fourth could also be yes...nine possibilities. If the second measurement was yes (from the first step), then the first measurement could also be yes, the third measurement could also be yes...nine possibilities. Altogether, $10 \times 9 = 90$ possibilities. However, this method of counting double counts, since "Measurement 1 is **yes** and measurement 2 is **yes**" shows up twice, and so does every other possibility. So have to divide by 2. Final answer is

$$p_8 = \frac{10 \times 9}{2} f^2 (1-f)^8 = \frac{10!}{2!8!} f^2 (1-f)^8. \tag{2.59}$$

Relation to normal distribution:. If one asks a **yes/no** question with underlying probability f a large number N times, then the probability of seeing a fraction F of **yes** answers is approximately normally distributed. The mean of F is f, and the standard deviation σ_N of F is

$$\sigma_N = \sqrt{f(1-f)/N}. \tag{2.60}$$

If one is designing a survey with **yes/no** answers, this expression tells how many people to sample for a desired accuracy.

Sample usage:.

> In a trial comparing whole versus partial breast irradiation, 708 patients were randomized following BCS to receive whole or partial breast irradiation. At a median follow-up of 7 years, local relapse rates were 11% for the whole breast irradiation group and 19.6% for the partial breast irradiation group ($p < 0.008$). There was no difference in survival. [Cancer Prevention and Control 1997; 1(3):228-240.]

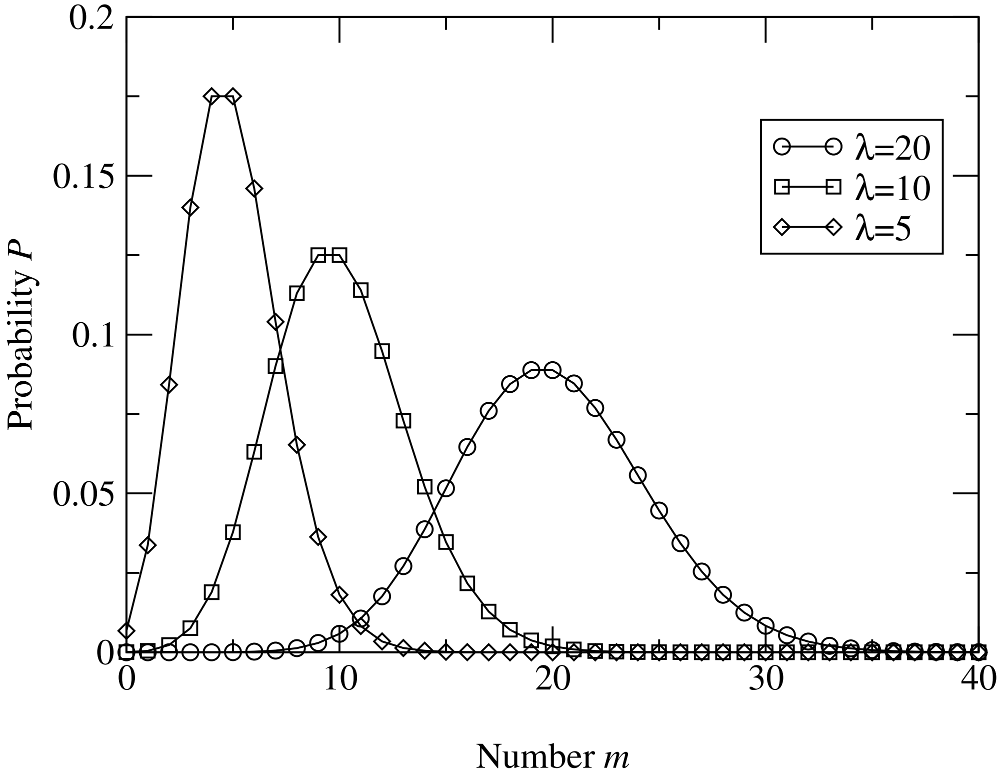
Simple analysis: 350 patients are in partial irradiation group. 69 have a relapse, 281 do not. 350 patients are in full irradiation group. 39 have relapse, 311 do not. Lumping all patients together, there is a 0.15 relapse rate. Taking this to be the true probability of relapse f, the probability of having 39 relapse cases from a sample of 350 is $p = 7.4 \times 10^{-3}$. ([2cd]Distr→ binompdf(350,.15,39)). However, the cumulative probability of 39 relapses or less is the meaningful measure, and is larger, 0.022.

2.3.13 Poisson Distribution

Idea. The Poisson distribution tells how likely it is for a certain number of rare events to happen when the rate at which they happen is known; for example, it gives the probability that six people will come to a supermarket counter in 15 minutes when the overall rate at which they come to the store is known.
Given. In every small time interval dt there is a very small probability of an event occurring.
Questions:. After a finite amount of time T, how many times will the event occur?
Answer:. The probability of m occurrences of the event in time T is of the form

$$P(m) = \frac{\lambda^m e^{-\lambda}}{m!}. \tag{2.61}$$

λ depends upon the probability of the seeing the event in the small time dt, and on the size of the large time interval T; if r is the rate at which the event occurs per time, then $\lambda \approx rT$. Poisson statistics can be obtained as an approximation to the binomial distribution where the number m of expected observations of **yes** is much smaller than the total number n of observations, and the fraction $f \ll 1$.
Examples. One famous case described by Poisson statistics is radioactive decay; the probability that a certain number of radioactive nuclei will have radiated in one hour. Another case where it arises is in queuing theory, which describes the number of people one should expect to arrive at a supermarket checkout counter in an hour, or the number of packets of information one should expect to arrive at a certain point on the Internet in a minute.



Note that when λ is small the distribution is asymmetric, while when λ gets larger the distribution becomes more symmetric and normal. A final comment is that the mean value of the Poisson distribution is just λ, so if a computer program asks for the mean, it means λ.

2.3.14 χ^2 in general

Idea. The χ^2 distribution is a tool that answers the question "Could my data have come from a theoretical distribution that I imagine produced them?"
Given. Measurements of m quantities, with mean values x_1, $x_2 \ldots x_m$, standard *errors* or uncertainties for each of those mean values, which we will here denote by σ_1, σ_2 … σ_m, and theoretical predictions for each quantity, which we will call μ_1, $mu_2 \ldots \mu_m$.
Questions:. Could chance alone have allowed the measurements x_i I made to differ from the values μ_i I expect?
Answer:.

Compute

$$\chi^2 \equiv \sum_{i=1}^{m} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{2.62}$$

Roughly speaking, if χ^2 is much larger than m, it is unlikely that chance alone can explain the deviation between x_i and μ_i, while otherwise, chance alone could well be operating. Calculating the probability of a particular value of χ^2 is sufficiently complicated that we are not going to record the formulas here; every statistics book includes tables enabling you to find the probability that a value of χ^2 arose by chance. Excel has a function *CHIDIST* that calculates the liklihood of obtaining a value of χ^2 for an experiment with m measurements.
Example: .

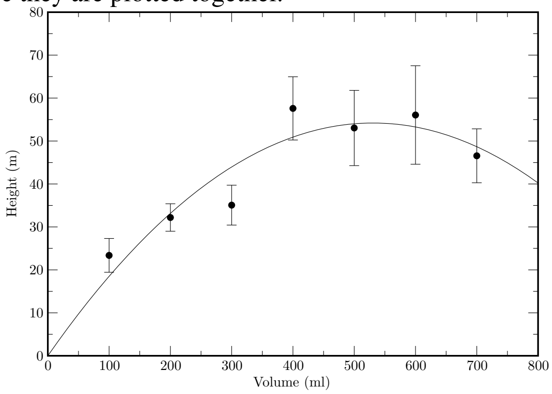
The height of a bottle rocket launch in meters is obtained as function of volume of water in the rocket in ml; here are the data.

Volume (ml)	Height (m)	Standard Error (m)
100	23.37	3.93
200	32.19	3.2
300	35.08	4.64
400	57.6	7.36
500	53.04	8.76
600	56.06	11.47
700	46.56	6.27

One comes up with a fitting function for height (y in meters) versus volume (V in ml).

$$y = .204V - 1.925 \times 10^{-4} V^2; \tag{2.63}$$

here they are plotted together.



Computing χ^2, one finds $\chi^2 = 6.25$. The Excel function *CHIDIST(6.25,7)* gives a probability of .51 that chance alone could explain the difference between the model and the observations. So the model is perfectly consistent with the observations.

2.3.15 χ^2 for categories

Idea. The χ^2 distribution is a tool that answers the question “Could my data have come from a theoretical distribution that I imagine produced them?” Here is a special case where the tool can be used to analyze data that are divided into a number of categories.

Given. A collection of m categories, measurements of N_1, $N_2 \ldots N_m$ data points in each of these categories, and an expectation that n_1, $n_2 \ldots n_m$ data points should theoretically have been observed in each category.

Questions:. Could chance alone have allowed the measurements N_i I made to differ from the values n_i I expect?

Answer:.

Compute

$$\chi^2 \equiv \sum_{i=1}^{m} \frac{(N_i - n_i)^2}{n_i} \tag{2.64}$$

Roughly speaking, if χ^2 is much larger than m, it is unlikely that chance alone can explain the deviation between N_i and n_i, while otherwise, chance alone could well be operating. Calculating the probability of a particular value of χ^2 is sufficiently complicated that we are not going to record the formulas here; every statistics book includes tables enabling you to find the probability that a value of χ^2 arose by chance, and Excel allows you to compute this probability directly from the sequences of numbers $N_1 \ldots N_m$ and $n_1 \ldots n_m$.

WARNING: *The mathematics lurking behind the χ^2 test is of questionable validity unless the number of counts in each cell is 5 or more. Sometimes one can respond to this problem by grouping categories together. However, since it is common in this class to conduct surveys where the results violate this guideline, the instructors have prepared a computer program,* `Survey.nlogo` *that correctly computes probabilities even when there are fewer than 5 entries in cells.*

Example: Favorite math course

Group	Number Asked	Calc 1	Calc 2	Diff EQ	Lin. Alg.
Engineering Majors	300	10%	15%	25%	50%
CNS Majors	500	15%	15%	25%	45%

Table 2.3.

To ask whether Natural Science students or Engineering students have significantly different preferences in their favorite math course, chosen from a group of four classes all respondents have taken, begin by constructing a null hypothesis. This can be done by assuming that all the responses come from a single random sample and can sensibly be grouped together. In this case the total number and fractions of preferences are

	Calc 1	Calc 2	Diff EQ	Lin. Alg.
Total Number	105	120	200	375
Fraction	0.13	0.15	0.25	0.47

Table 2.4.

Based upon the preferences in the table above, one can calculate the expected responses in the two groups:

Group	Calc 1	Calc 2	Diff EQ	Lin. Alg.
Observed in Engineering	30	45	75	150
Expected for Engineering	39.38	45	75	140.63
Observed in CNS	75	75	125	225
Expected for CNS	65.63	75	125	234.38

Table 2.5.

Using *CHITEST* in Excel, the probabilities of the observed results arising as a consequence of the expected distribution are 41% for Engineering, and 63% for Natural Sciences, if one compares predictions and observations for the two groups separately, or around 20% if one groups all the predictions and observations together. There are no grounds to reject the null hypothesis, and no reason to think that people in Natural Sciences and Enginering feel differently about their favorite math course.

2.3.16 Degrees of Freedom

Many statistical tests require one to specify the number of degrees of freedom. There are rules for how to compute degrees of freedom that cover most common cases, although it is challenging explain in words exactly what this quantity means.

Definition:. The parameter "degrees of freedom" in a statistical test refers to the number of completely independent random variables required by the statistical model behind the test.

Examples:. The first nontrivial instance of degrees of freedom arose in the definition of sample variance and sample standard deviation (Section 2.3.3, where one divides by $N-1$ than N. Suppose one has measurements $x_1 \dots x_N$ that vary randomly around a known mean μ. Then one could compute

$$\sum_{i=1} \frac{(x_i-\mu)^2}{N}, \tag{2.65}$$

and because of the N independent measurements x_i, one would say that N degrees of freedom are involved. However, this is not the definition of sample variance; sample variance is constructed instead from the N values

$$x_1-\bar{x}, x_2-\bar{x} \dots x_N-\bar{x}. \tag{2.66}$$

These are not completely independent, because

$$\sum_{i=1}^{N}(x_i-\bar{x}) = N\bar{x} - N\bar{x} = 0. \tag{2.67}$$

One says that there is *one constraint.* Therefore, if one knows

$$x_1-\bar{x}, x_2-\bar{x} \dots x_{N-1}-\bar{x}, \tag{2.68}$$

the final value in the series can be determined from

$$\sum_{i=1}^{N-1}(x_i-\bar{x}) + (x_N-\bar{x}) = 0 \Rightarrow (x_N-\bar{x}) = -\sum_{i=1}^{N-1}(x_i-\bar{x}). \tag{2.69}$$

That is, once the sample average $\bar{x}$ has been computed, there are only $N-1$ truly independent quantitities appearing in the computation of the variance, and for this reason the variance is defined by

$$s^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2. \tag{2.70}$$

This explanation is incomplete; there are still N independent quantities showing up in the sum, once one includes $\bar{x}$, and without proving theorems about the distribution of s^2, one cannot see for sure that the $N-1$ degrees of freedom are most appropriate to divide by. However, it should provide some motivation for using $N-1$ rather than N.

Another common example of degrees of freedom occurs in applying χ^2 tests to survey data, as in Tables 2.3, 2.4 and 2.5. If one computes χ^2 from Eq. (2.64) and then wants to compute the liklihood that this value of χ^2 arose by chance, from a table or with the Excel function `CHIDIST`, one must know the number of degrees of freedom. For charts such as these, the number of degrees of freedom is given by

$$\text{dof} = (\text{Number of rows} - 1) \times (\text{Number of columns} - 1). \tag{2.71}$$

The reason is similar to the reason that $N-1$ rather than N appears in the sample variance. To compute χ^2 one makes use of the total number of items sampled in each row; thus there is a constraint for each row. In computing expected values, one makes use of the sum over all items in each column, and there is one constraint for each column as well.

3. Experimental Design

Talking about the design of experiments means designing procedures for the first three research procedures from Chapter 1: Hypothesis, Experimental Correlations, or Experimental Value. Use of statistics is important both in designing the experiments and in interpreting the results.

An experiment is a set of actions designed to convince scientists of a conclusion. Together with analysis and written reports, the experiments construct scientific arguments. The more important and significant are the results the experiment demonstrates, the more skeptical will be the audience. Therefore, the experiments have to be designed to address as many critical questions as possible. There is no way to anticipate every possible criticism, but some criticisms are extremely common, and preparing for them provides a rational basis for experimental design.

Hypothesis

Experiments involving an hypothesis are common in medicine, agriculture, and psychology; they certainly may appear in other disciplines as well.

Ideas that enter in the design of these experiments:

Null Hypothesis In the spirit of maximum skepticism, it is customary to assume that a desired effect is **not** present and to try to disprove this hypothesis.

The formation of a null hypothesis is an art. One way to obtain a null hypothesis is to imagine objections of critics who believe that something you are searching to find is not actually there and try to prove them wrong. For example, suppose you are investigating the effectiveness of a three-week course of radiation therapy for breast cancer and comparing it to a six-week course of radiation that is standard. Your null hypothesis could be that the three-week course of treatment will lead to lower survival rates than the six-week treatment.

Control One decides upon a limited collection of factors to vary and responses to measure. The goal is to keep as many extraneous elements out as possible.

In investigating a null hypothesis, there will often need to be a *control group.* This is a group where, in some sense, nothing new is being done. If a six-week radiation treatment is standard, and a three-week treatment is being tested, then patients receiving six weeks of treatment are the control group.

Experimental technique plays a large role in obtaining control. For example, if one is measuring the time it takes for water to boil, careful measurement of the amount of water put in the pot would be part of control. Placement of the pot on the burner, care in setting the burner at constant level between trials, and many other things would enter.

One way that control can be improved is through the technique of *blocking.* Even if certain factors are not controlled, they can at least be recorded and accounted for. For example, suppose there is a study looking at the effects of Diet Coke on the rate of heart attacks. The aspirins could just be given to 20,000 people at random. However, the age and gender of patients may be important. If these are recorded, then the patients can later be grouped according to whether they are male or female, and according to age. It could turn out that men between the ages of 40 and 50 have fewer heart attacks if they drink a six-pack of Diet Coke a day, while women in this age bracket have more. If men and women are all grouped together, the effect might be lost completely.

When data can be separated according to various distinctions between cases, one says the data have been *disaggregated*. For example, it is fairly easy to find the total number of science teachers in Texas schools teaching on emergency permits, but harder to find the number of biology teachers in Texas schools teaching on emergency permits, because the public data have not been disaggregated in this way.

Keeping all but one thing constant is not nearly as simple as it seems. For example, consider the question of how fast water freezes as a function of the starting temperature. If a hot glass of water is put in the freezer, a noticeable amount of water evaporates as it cools. So, in comparing the freezing times of two glasses of water, what should really be held constant? The initial volume of water, or the volume of water present in the hot and cold glasses at the time that both hit, say, a temperature of 40°? There are no fixed answers, but one has to be ready to address such questions.

Repetition Experiment are usually repeated many times. The repetition has many goals. It convinces other researchers that an observed effect can be seen more than once. It reduces error by making possible averages over random extraneous influences. It makes possible an estimate of the size and nature of those extraneous influences.

The question of how many repetitions will be enough is a question to address through use of statistics. A basic mathematical result that is helpful in many cases is the observation about averages on page 42. Suppose one has a quantity that is normally distributed, with standard deviation σ. Then after taking N measurements, the standard deviation of the average is $\sigma/\sqrt{N}$. In settling on a number of repetitions, one has to think about the accuracy one wants to have, and the confidence one wants to have in reaching this accuracy. Return to the example of new teaching methods leading to a test where the district mean score is 75, and the standard deviation is 12. Suppose one wants to be able to detect improvements of 5 points. One might demand therefore that the standard deviation of the **average score** be less than 2.5 points. Then 5 points would correspond to two standard deviations, and the odds of being off that much would be less than 5%. That is, if one gets a certain average score, the odds of getting something more than 5 points higher, or less than 5 points lower, purely by chance, should be less than 5%. Since the standard deviation associated with each individual student test is 12, the standard deviation associated with the average over N students is $12/\sqrt{N}$, and to get this value down to 2.5 requires $N = (12/2.5)^2 \approx 24$. Now, if one is really certain the ones students will be better than average, and the only question is how much, one can choose to do a one-sided test. The probability of scores being 1.6 standard deviations below the mean or less is also 5%. In this case, the uncertainty would only have to be reduced to 3 points, rather than 2.5, leading to an estimate that only 15 students are needed to have 95% confidence. The reason that fewer student scores are leading to the same confidence is that one has made the firm decision (maybe misguided!) that his students are better than average, and given this information less extra testing is required.

More generally, with roughly 95% confidence, and a two-sided test,

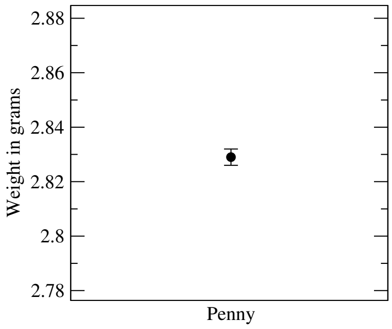
$$N \approx \left[2 \frac{\text{Standard Deviation}}{\text{Minimum Acceptable Uncertainty}} \right]^2 . \tag{3.1}$$

The factor of 2 in the numerator comes from thinking about a two-tailed test where errors would be eliminated with 95% confidence. Statistics texts write $2 = z_{\alpha/2}$ where $\alpha = 0.05$. Demanding higher confidence could be obtained by putting smaller factors in the denominator; the higher the confidence desired, the larger the number of trials needed. The number of trials also increases as the acceptable uncertainty diminishes, or as the standard deviation of the underlying distribution increases.

Doing the same thing over and over again is usually boring. It is tempting to perform just a few trials and leap to conclusions, especially if they are what one wants to see. But if one cannot answer hard questions about the influence of chance, the results are not believable.

Randomization Extraneous influences cannot completely be avoided. Therefore, as the experiment is repeated, it is done in such a way that they act as randomly as possible.

This experimental principle attempts to compensate for the influence of unknown factors. In testing a new drug, one group of patients might be given a new medicine, and another group a placebo. The patients receiving the placebo should be chosen completely at random, and neither they nor the doctor treating them should know that they are the ones on the placebo. After all, mental reactions to treatment can influence the effect of treatment, and doctors might alter their behavior depending on what they



know about the new drug. In testing cooling rates of hot and cold water in a freezer, both hot and cold glasses should be scattered randomly about at many locations in the freezer to compensate for possible temperature differences.

Experimental Value

There are some new things to think about when the goal of an experiment is to measure a number. It often does not help much to have an hypothesis about what the answer ought to be, and one thinks about the problem in a different way. Common experience shows that in carrying out such measurements, there is always some difference between successive attempts. Various contributors to this variation need to be considered.

Instrumental Error. It may be that the quantity one is trying to measure is fixed and absolute, but the measuring procedure is influenced by random factors from the environment. Suppose one is trying to measure the mass of a single penny. The last digit on the scale flips about randomly because of fluctuations in electrical current inside the scale, or because of small air currents that hit the plate where the penny sits. Factors such as this are called *noise* and contribute to *instrumental error*. The important point is that this sort of error can be eliminated completely by taking a sufficient number of independent measurements. Every particular measurement is influenced by noise, but the contribution of noise to the average of the measurements vanishes for sufficiently many measurements. The way averaging works is described by Equation 2.33 on page 43. If the standard deviation associated with measuring the weight of a penny is σ, then the standard error associated with averaging its weight over N trials is $\sigma/\sqrt{N}$. The standard error is often called the *uncertainty*.

Faced with uncertainty, there is conventional language to speak about it, a conventional way to describe it with numbers, and a conventional way to indicate it graphically.

Continuing with the example of the penny, imagine taking 10 measurements, which in grams are

2.832	2.825	2.826	2.831	2.834
2.827	2.825	2.828	2.829	2.831

Using Equations 2.26 through 2.28, the mean is 2.829, the variance is $s^2 = 8.51 \times 10^{-6}$, and the sample standard deviation is 2.92×10^{-3}. Rounding 0.00292 up to 0.003, one says that the weight of the coin is 2.829 ± 0.003 grams. There is little sense in retaining lots of decimal places to describe the standard error.

In drawing a graph, it is traditional in physics to indicate 2.829 ± 0.003 in this way:

Sometimes there is uncertainty in the values used both for x and y axes, and in this case there can be horizontal and vertical bars indicating the standard error in both directions.

Finally note that instead of drawing bars to display the standard error like physicists, statisticians are more likely to draw bars to indicate a 95% confidence interval. Since the standard error corresponds to one standard deviation, and the 95% confidence interval corresponds to two standard deviations, a statistician might have made the bar in Figure 3 twice as high.

Systematic error. Some sorts of error cannot be cured by repeated measurement. For example, if a scale is calibrated incorrectly it will give a wrong measurement of the weight of a penny, and using it over and over again simply will not help. The only way to catch systematic error is to have some point of reference for comparison, maybe a scale one knows is more accurate, a number of other scales thought to be equally inaccurate but statistically distributed, the known weight of a quarter, or even some internal check, such as seeing whether two new pennies weigh twice as much as one. Sometimes knowledge of physical theories can be used to estimate a systematic error.

Here is an example of systematic error in action: Reynolds conducted experiments on the speed water could be forced through a pipe before the flow becomes turbulent. He found a value of a dimensionless ratio $R = vd/\eta$ (v the velocity of the fluid, d the diameter of the pipe, η the viscosity of the fluid) where this happened. Turbulence started when $R = 25,000$. These experiments in the 1880's. Thirty years later *Check dates*, a new generation of researchers at his laboratory repeated the experiments. They found $R = 5000$. Nothing they tried would give any other result. Eventually they settled upon the guess that vibrations caused by trucks driving outside the laboratory were causing the discrepancy, and waited until late at night on Christmas eve when they could be guaranteed several hours with no trucks passing by. $R = 25,000$.

Hypothesis testing is also subject to systematic error. Another story. During the second world war the British wanted to know where to place armor on bombers to protect them against enemy fire. They subjected the bomber fleet to an exhaustive analysis to find the frequency with which bullets hit various parts of the planes. Question for reader: *Would it make sense to put extra armor on the parts of the planes where bullets had hit most frequently?*

Measurement of distributed quantities. The quantity one wants to measure does not always have a precise value that can be measured to infinite precision with sufficient effort. There is little reason to be interested in the weight of a particular penny. It makes more sense to be interested in the weight of pennies. But no two are precisely the same, and repeated measurement of the weight of many different pennies will find a distribution of values. The mathematics associated with this range of values is identical to the mathematics that describes instrumental error, and one can still say, for example, that the weight of pennies is 2.832 ± 0.002 grams. However there is a difference, which is that the uncertainty belongs to the object being measured, and cannot be eliminated through improved instrumentation. The weight of pennies really is distributed. There still is something that can be measured with precision, namely the distribution itself. That is, one can set out to measure with arbitrary accuracy the fraction of pennies that have a given weight as a function of the weight. Although it is customary to assume that distributions in the natural world are normal, any particular distributed quantity measured with enough accuracy will probably differ from the normal distribution.

Propagation of Error

Suppose one has measured one or more uncertain quantities and wants to compute functions of these uncertain things, and find how uncertain the functions are. Most tasks of this sort can be accomplished with a few rules.

When two quantities with uncertainty are added together, the uncertainty of the sum is **not** the sum of uncertainties. Instead, if one has $A_1 \pm \epsilon_1$ and $A_2 \pm \epsilon_2$ then the uncertainty in $A_1 + A_2$ is $\sqrt{\epsilon_1^2 + \epsilon_2^2}$, so one gets

$$A_1 + A_2 \pm \sqrt{\epsilon_1^2 + \epsilon_2^2}. \tag{3.2}$$

Using this expression repeatedly one obtains expressions already given for the uncertainties of averages.

If one

For a more complete discussion of errors and error propagation, see for example
http://www.physics.mun.ca/˜cdeacon/errornotes.html.

Experimental Correlations

In studying experimental correlations, one measures not just one number, but many, and searches for a relationship between one or more control variables, and one or more responses. For example, one might measure the distance a projectile travels as a function of the angle at which it is fired. An example of some data from my laboratory are shown in Figure 3.1

All the individual measurements from which the relationship is built are subject to the rules discussed above for measuring numbers. Individual measurements need to be repeated to avoid instrumental error, they can be plagued by systematic error, and the quantities one measures often turn out to be distributed, rather than having definite and precise values.

The new task that arises in taking a series of measurements as a function of a control variable is the task of making sense of the results. It is desirable to find a compact way of expressing all the information in the data without reading out a list with all the numbers. The way to do that is to find a mathematical function that passes through all the points.

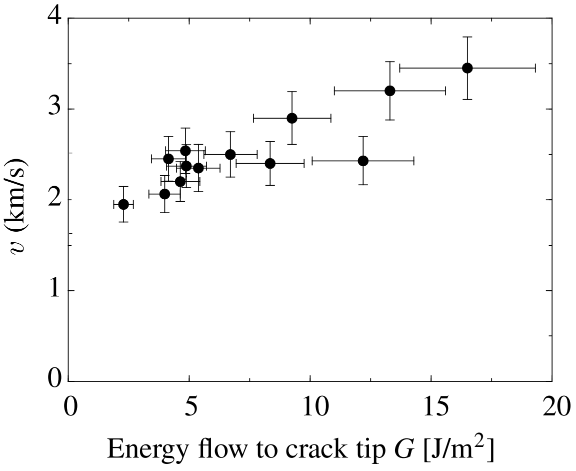


Figure 3.1. Data on speed of a crack in a single crystal of silicon as a function of the amount of energy sent to the crack tip. Both the velocity of the crack and the energy flow were measured independently, and each of these measurements had some error associated with it. The crack speed does seem to depend upon the energy flow, but the data points are very scattered, and it is difficult to tell whether the relationship can well be described by anything other than a straight line passing through most of the points.

The number of functions that passes through a finite set of points is infinite, and when all of those points are uncertain, the space of functions becomes even larger. So functions are constrained. They need to be as simple as possible, but no simpler.

The most popular function to fit to a collection of data points is a straight line. The *method of least squares* finds the unique line for which the sum of squares of vertical distances to the data points is as small as possible. Computer programs performing a least squares fit should also report uncertainties in the slope and intercept that result from uncertainties in the locations of the data points. The sum of squares of vertical distances from line to data points gives a measure of how good the fit is. If the line passed perfectly through all the data points, the sum would be zero, and the more the line deviates from the data points, the larger the sum becomes.

3.1 Cause and effect

One of the most important goals of science is to establish the links between cause and effect. The links are not simple. Causes and effects can be linked in many different ways:

- Correlation. Cause and effect are often said to be operating when two things are correlated while all other factors are held constant. (It is not always so easy to determine what "holding constant" really means.)

- Theories. In scientific theories, an underlying theory is said to cause the effects that result as consequences.

- Temporal ordering. If there is a temporal ordering of events related by cause and effect, the causes are the ones that come first, and the effects are the ones that come afterward.

- Human agency. Causes can relate to choices that humans must make, while effects are outcomes that humans want to change.

Famous statement: "Correlation is not causality." But is causality anything other than correlation?

Another common statement: "Cause and effect are never firmly established without a plausible mechanism." However, different fields have completely different views of what makes for a plausible mechanism. In physics, the mechanism usually needs to be accompanied by a detailed calculation starting with accepted equations that accounts for the results. In biology, the mechanism may be a believable story that physicists would dismiss in their own field as "hand-waving." The biologists meanwhile view the physicists as cluttering the field with arrogant and irrelevant detail.

Examples:

- Gravity of sun causes planets to move in circular orbits (no human influence, laboratory and observational tests are possible, there is no temporal ordering).

- Asteroid impact caused extinction of dinosaurs (no human influence, cannot be repeated, there is temporal ordering).

- A light switch causes lights to go on (human influence, prediction, temporal ordering hard to establish).

- (A particular) Fertilizer makes (a particular) plant grow larger (human influence, prediction, temporal ordering, statistical connection).

- Example of correlation without causality: The number of Starbucks coffee houses in the United States has grown enormously since 1990. The number of personal computers in the United States has grown enormously since 1990. However, almost no one would think that the coffee houses caused people to by computers. This example may seem silly because no one can imagine any possible mechanism so that the first could affect the second. However, there are plenty of things people believe where I find an equal difficulty in establishing a mechanism. Many intelligent people believe that the positions of the planets affect their lives; I cannot see how this would be the case. Still, my mother tells the following story. I was around 4 years old, and she entered an elevator with me in Manhattan. A complete stranger was also there. After listening to me talk for a few seconds, he turned and said with complete confidence to my mother "He's Sagittarius, of course." A chance in twelve, one might say, but my mother can come up with other stories as well that make her unwilling to dismiss astrology. A skeptic will say that the success of astrology is due to the selective memories of its believers, who develop correlations between its predictions and real happenings by forgetting all the cases where the predictions failed.

- Health effects of smoking:

  > When smoking was first suspected of causing lung cancer and heart disease, Sir Ronald Fisher, then the world's greatest living statistician and a smoker, offered the "constitution hypothesis"; that people might genetically be disposed to develop these diseases and to smoke; that is, that genetics was confounding the association. [The Little Handbook of Statistical Practice Gerard E. Dallal, http://www.tufts.edu/%7Egdallal/LHSP.HTM]

3.2 Design of experiments involving apparatus

Modern experimental design begins with Galileo. His experiments to determine the laws governing falling objects contain all the ingredients that make up scientific experiments today.

From *Dialogues on Two New Sciences* [213] (179). Salvatore (Galileo) has just been challenged to justify the claim that falling objects accelerate uniformly.

> SALV. The request which you, as a man of science, make, is a very reasonable one; for this is the custom – and properly so — in those sciences where mathematical demonstrations are applied to natural phenomena, as is seen in the case of perspective, astronomy, mechanics, music, and others where the principles, once established by well-chosen experiments, become the foundations of the entire superstructure. I hope therefore it will not appear to be a waste of time if we discuss at considerable length this first and most fundamental question upon which hinge numerous consequences of which we have in this book only a small number, placed there by the Author, who has done so much to open a pathway hitherto closed to minds of speculative turn. So far as experiments go they have not been neglected by the Author; and often, in his company, I have attempted in the following manner to assure myself that the acceleration actually experienced by falling bodies is that above described.

> A piece of wooden moulding or scantling, about 12 cubits[1] long, half a cubit wide, and three finger-breadths thick, was taken; on its edge was cut a channel a little more than one finger in breadth; having made this groove very straight, smooth, and polished, and having lined it with parchment, also as smooth and polished as possible, we rolled along it a hard, smooth, and very round bronze ball. Having placed this board in a sloping position, by lifting one end some one or two cubits above the other, we rolled the ball, as I was just saying, along the channel, noting, in a manner presently to be described, the time required to make the descent. We repeated this experiment more than once in order to measure the time with an accuracy such that the deviation between two observations never exceeded one-tenth of a pulse-beat. Having performed this operation and having assured ourselves of its reliability, we now rolled the ball only one-quarter the length of the channel; and having measured the time of its descent, we found it precisely one-half of the former. Next we tried other distances, comparing the time for the whole length with that for the half, or with that for two-thirds, or three-fourths, or indeed for any fraction; in such experiments, repeated a full hundred times, we always found that the spaces traversed were to each other as the squares of the times, and this was true for all inclinations of the plane, i. e., of the channel, along which we rolled the ball. We also observed that the times of descent, for various inclinations of the plane, bore to one another precisely that ratio which, as we shall see later, the Author had predicted and demonstrated for them.
>
> For the measurement of time, we employed a large vessel of water placed in an elevated position; to the bottom of this vessel was soldered a pipe of small diameter giving a thin jet of water, which we collected in a small glass during the time of each descent, whether for the whole length of the channel or for a part of its length; the water thus collected was weighed, after each descent, on a very accurate balance; the differences and ratios of these weights gave us the differences and ratios of the times, and this with such accuracy that although the operation was repeated many, many times, there was no appreciable discrepancy in the results.
>
> SIMP. I would like to have been present at these experiments; but feeling confidence in the care with which you performed them, and in the fidelity with which you relate them, I am satisfied and accept them as true and valid.
>
> SALV. Then we can proceed without discussion.

The ingredients of this experiment are

1. Ingredients

   (a) **Controlled environment.** The experiment is performed under precisely controlled conditions. The channel for the ball is made as smooth as possible. The ball is as round as can be made. Channel and ball are kept clean. In short, part of experimental design is trying to manipulate the world so that it imitates mathematical perfection. There is an important preconception lying behind this effort. Physical science presumes that the real external world does have mathematical perfection, and if we only control it sufficiently, the perfection will be revealed. Galileo was challenged on this point. If his experiments only worked in a carefully controlled environment, how could one say they describe the real world?

   (b) **Precise description.** Part of an excellent scientific experiment is that it can be reproduced by other observers. Therefore the experiment must be characterized very carefully by the experimenter, with even seemingly irrelevant details being recorded carefully. Publication of an experiment should include enough of the details that others can return to it years later and duplicate the results.

   (c) **Control of limited set of variables.** Deciding what to vary and what to leave constant is one of the most difficult parts of designing an experiment. Vary too little and no laws or lessons emerge, just a single observation over and over. Vary too much, and again no laws or lessons emerge because it is impossible to distinguish random variation from the effects of the variables one controls. It is often helpful to think forward to a skeptical listener who will ask you questions about your results. Questions you are likely to face include

[1] 1 cubit=1.5 feet=45.72 cm. So the apparatus is 18 feet or 5.5 m long, and 9 inches=23 cm wide

- What laws or relations have you uncovered?
- If you simply measured some numbers, why are they significant?
- Anyone would imagine that the following things might have affected your results [temperature in the room, humidity, last Thursday's earthquake, the stomach flu epidemic, etc.] How do you know your results are not due to those effects, rather than what you claim?
- Have you measured everything your techniques make possible? Have you developed new techniques?

3.3 Design of experiments involving people

Experimentation with apparatus established an ideal. There is an external reality, which obeys perfect and unchanging laws. This reality can be uncovered by performing experiments, or by making careful observations. It is natural to hope that this ideal can be extended to the domain of people, that they can be studied in a similar way, and that laws of behavior can be determined similar to laws of nature.

However, people are more complicated than steel balls rolling down inclined planes. Much more complicated. Their behavior does not obey laws with the same precision. Grappling with this complexity can lead in many different directions. The directions span much of human learning, and include literature, philosophy, government, sociology, linguistics, marketing, economics, anthropology, history, psychology, education, and law. Each one of these disciplines has numerous distinctive ways to describe what its practitioners understand about people and their interactions.

Only some in these disciplines would call themselves scientists studying human behavior. Among those who do, there is a basic division between qualitative and quantitative researchers.

I will bring up qualitative research mainly to admit that I have no formal experience with it, and will not discuss it further. Here is a chart contrasting qualitative with quantitative research that I found at *http://www.iptv.org/FINELINK/publications/criteria.html.* It gives an idea what the differences are supposed to be.

I will mention just a few ideas concerning the quantitative study of people.

Quantitative studies involve numbers, but it is never straightforward to characterize complex human qualities such as belief or understanding by numbers. This class is not going to deal with issues of testing individuals, but will touch briefly on surveying populations. So some brief comments on surveying:

In performing a survey, the goal is to learn something about a large population of people by asking questions of a small subset of them. In order for the responses of the small subset to represent well the larger population, they may be chosen completely randomly. Obtaining an unbiased random sample is hard. For example, suppose one wants to know how long students at UT spend at the library. One strategy would be to go to the library and ask students how long they have been there. But this would give completely incorrect answers to the original question because students who never go to the library would never take part in the survey.

Therefore, before beginning any survey, one must decide upon the population one wishes to sample, and identify a strategy for obtaining a representative random sampling of that population. If the random sample cannot be obtained, one has to change the question.

Any activity involving human subjects creates a set of ethical concerns. You will need a certificate showing you have participated in training on how to treat human subjects: see *http://cme.cancer.gov/c01/.*

Characteristics	**Quantitative Research**	**Qualitative Research**
Phrases	experimental; numerical data; empirical; statistical	descriptive; naturalistic; word-oriented
Key Concepts	variables; operationalizing; reliability; hypotheses; validity; statistical significance; replication, validity	meaning; common-sense understanding; process; social construction; themes; trustworthiness
Designs	structured; predetermined; formal; specific; detailed plan of operation	evolving; flexible; general; negotiated; a hunch as to how to proceed
Sample	large; stratified; control groups; precise; random selection; control for extraneous variables; representative	small; theoretical sampling; purposive sampling; selected to take into account as much context as possible
Techniques or Methods	experiments; survey research; structured interviewing; quasi- experiments; structured observation; data sets	observation; participant observation; reviewing documents and artifacts; open-ended interviewing
Data	quantitative; operationalized variables; quantifiable coding; statistical; counts, measures	descriptive; people's own words; personal documents; field notes; artifacts; official documents; audiotapes, videotapes, transcripts
Instruments and Tools	inventories; questionnaires; scales; test scores; computers; indexes	tape recorder; transcriber; notes; researcher is often only instrument
Data Analysis	deductive; statistical; occurs at conclusion of data collection	inductive; ongoing; models; themes; concepts; constant comparative method
Problems in Using	controlling other variables; Approach obtrusiveness; validity	time consuming; procedures not standardized; reliability

4. Simple Mathematical Models

4.1 Introduction

One of the branches of mathematics is *applied mathematics* which roughly refers to the portion of mathematics that is most often useful in representing the physical world. Or, since so much of mathematics might be used to represent the physical world, it refers to mathematics employed with the intention of representing the world.

In too many courses, the subtle process of mathematical modeling is replaced by an ugly caricature that students describe by saying “You get the formula, plug in the numbers, and then you get the answer.” Learning modeling by example in this way is not too bad, so long as one remembers the following:

Every formula has a context.

There are conditions where the formula applies, often rather limited. For example, momentum of a system of particles is conserved **only** if the particles are not subject to any outside forces. Light absorption in a solution is proportional to the concentration of the solution only if it is sufficiently dilute. And so on.

To discuss the process of modeling, one has to mention the mathematical constructs employed in the models.

4.2 Ingredients of mathematical modeling

4.2.1 Maps

Maps provide an example of mathematical representation. Most of the features of a city are removed; a few features remain, indicated schematically by lines. The map is at a different scale from the original object. In many city maps all objects are in the same relation to each other as in life, but projected onto two dimensions and shrunk down in scale. In other cases, such as the New York City subway map, only the relative order of the subway stops is preserved, and the connections of the different lines. There is no attempt to indicate precise locations or distances. Maps do not use Arabic numerals, but they do constitute a symbolic representation of the world, which makes them mathematical. The ancient Greeks, incidentally, performed their mathematics using geometrical arguments, without either numerals or the abstract symbols such as x and y we now take for granted.

4.2.2 Numbers

Following maps and diagrams, the next mathematical objects to consider are numerals for counting, and to indicate quantity. Using numbers to represent the world proceeds in stages:

1. Positive integers. These represent increasing quantities of essentially identical objects. Take note of the basic idea, so simple and yet so abstract, that one object can be essentially the same as another, while not literally being identical. Even two symbols, X, X, are different if only because they are in different places, but everyone seems to understand what it means to say that they are the same, and there are two of them.

2. Zero. This is a number representing the absence of some quantity. It makes it possible to state that something is not present.

3. Negative integers. These numbers seem to have been conceived in India during our Medieval period, along with the Arabic numerals (brought by Arabs from India to Europe.) Long regarded as “imaginary” numbers since one cannot “really” have negative two sheep.

4. Rational fractions. Capture the idea of parts of a whole. Known to ancient mathematicians. Note the word used to describe these numbers.

5. Irrational numbers. Numbers that denote quantity but cannot be written as ratios of integers. Early examples include π and $\sqrt{2}$. Note again the word used to describe these numbers.

6. All the numbers mentioned until this point are now collectively called "real numbers." Real numbers stand in opposition to "imaginary numbers" which include multiples of $\sqrt{-1}$. Having gotten used to the idea that when someone takes two of your sheep, they have added negative two sheep to your flock, now it is $\sqrt{-1}$ that does not really exist. Except, unfortunately, that the basic mathematical theory of matter, quantum mechanics, cannot be written down in any fashion that does not involve these imaginary numbers. So like all the previous mathematical constructs, this one too acquires reality as one uses it to represent the world.

7. Vectors. Matrices. Quaternions. Groups. Fiber bundles. Many others. The process of creating new mathematical structures becomes easier once one is no longer bound by intuitive senses of what is real. The ones mentioned here all have a place in physical theories.

4.2.3 Operations

The basic operations underlying mathematical modeling are addition, subtraction, multiplication, and division. Performing well-defined problems with these operations is not hard, particularly if one gets to use a calculator. What is hard is being able to select the correct operation when faced with a question about something quantitative in the world.

Here are examples where each of these basic operations would be used:

Addition.

- Ellen was 13 years old at the start of her freshman year in high school. At the start of her senior year, how old is she?

- Twelve pennies sit in a jar. I put in another 15. How many are now in the jar?

Subtraction.

- There are two hundred and fifteen Daphnia swimming in a jar on Monday morning. On Tuesday morning, seven lie dead on the bottom of the jar. How many are left alive?

- On Monday, a mold colony occupies 1.3 square centimeters. On Tuesday, it occupies 2.2 square centimeters. How much did it grow between Monday and Tuesday?

Multiplication.

- There are 215 Daphnia in each of 8 jars. How many Daphnia do I have?

- My house has a rectangular shape 220 feet long and 80 feet wide. What is the floor space of my house?

- My car is traveling at 20 miles per hour. How far does it go after 4 hours?

- How many distinct numbers can I form with two digits, when one digit ranges from 1 to 4, and the other ranges from 5 to 9?

Division.

- You have 4 ounces of plant food, and want to put the same amount in 12 pots. How much goes in each pot?

- You have 4 ounces of plant food, and want to put .2 ounces in pots. How many pots should you get?

- My house has 2100 square feet, and is 70 feet long. How wide is it?

- I drove 210 miles in 3 1/2 hours. What was my average speed?

Many problems involve operations in combination:

- My car is traveling at 20 miles per hour. How many meters does it go after 10 seconds?

- Once he reached his 20th birthday, John started paying \$20 per year for physical exams. How old was he when he had paid \$180?
- An internship program pays students \$10/hour. Students work an average of 15 hours/week for 12 weeks, and typically 45 students participate per semester. How much money does the program cost per year?
- An internship program pays students \$10/hour. Students work an average of 15 hours/week for 12 weeks. You have \$100,000 for two semesters. How many students can you take into the program?
- On Monday my plant is 2 cm long. It grows .3 cm/day. How long is it after a week?
- My plant grows .3 cm/day. After a week it is 7 cm long. How long was it originally?
- On Monday my plant is 2 cm long. After a week it is 7 cm long. How fast does it grow?

4.2.4 Functions

Functions can be used as an alternative to pictures in order to describe shapes.

Linear Functions. The most basic function is the straight line, $y = Ax + b$. It can be used to represent any situation where something is proportional to something else. Think that that gas mileage of a car should go down the more heavily it is loaded? Try a linear fit.

Mileage = Mileage with one driver − (Number of passengers-1)×A, where A is a mileage loss coefficient that would need to be determined.

The idea behind linear functions is very simple. Do something once, get an effect. Do it twice, get twice the effect.

Linear functions can be generalized to functions of many variables,

$$y = b + \sum_{i=1}^{N} A_i x_i. \tag{4.1}$$

So, if gas mileage depends upon the weight of people in the car (x_1) and the wind speed (x_2) and the average slope of the hill the car is climbing (x_3), there might be a formula for mileage involving a linear sum of these three variables.

Products and Dimensional Analysis. Another useful class of functions is built just by taking products. Functions of this type arise naturally when one performs *dimensional analysis*.

In modern practice, numerals should usually be accompanied by *units* when they refer to objects in the world. That is, one should always accompany the number by a word which indicates the quantity to which the number refers. This practice is valuable for catching errors. But it provides an almost magical way to obtain relationships and construct physical theories on the fly. Example: speed of (transverse) wave on string. Slinky. Tests. Units of force. Formula!

More complicated example, worked out: Suppose one wants to estimate characteristic frequency for oscillation of a small drop of water due to surface tension. The most crucial step in the analysis is to decide upon a collection physical quantities that the result should depend upon. In this case, let's guess that the result depends upon the surface tension σ of the drop, the density of the drop ρ, and the radius of the drop r. We want to obtain a frequency, and so we write

$$[\nu] = [\sigma]^a [\rho]^b [r]^c,$$

where the square brackets mean that one wants to consider the dimensions of the quantities within them. In this case, indicating units in cgs, one has

Frequency	$[\nu]$	sec^{-1}
Surface Tension	$[\sigma]$	$\mathrm{gm\ cm^2/sec^2/cm^2}$
Density	$[\rho]$	$\mathrm{gm/cm^3}$
Radius of Drop	$[r]$	cm

Thus one has from Eq. (4.2.4) that

$$\sec^{-1} = (\mathrm{gm}\ /\sec^2)^{\mathrm{a}}\ (\mathrm{gm}/\mathrm{cm}^3)^{\mathrm{b}}\ \mathrm{cm}^{\mathrm{c}}.$$

The equation only holds if the units for time, mass, and distance balance separately, so it is actually three equations in three unknowns. Taking the logarithm of both sides of Eq. (4.2.4) one has

$$-1 = -2a\ (\sec)$$
$$0 = a + b\ (\mathrm{gm})$$
$$0 = -3b + c(\mathrm{cm})$$

which has the unique solution

$$a = \frac{1}{2}$$
$$b = -\frac{1}{2}$$
$$c = -\frac{3}{2}.$$

So one has that

$$\nu \sim \sqrt{\frac{\sigma}{\rho r^3}}$$

The remaining constants cannot be determined by this type of argument, but it is an article of faith that it is of order unity.

Other Functions.

There is no limit to functions of one variable, but a small number of them – the most common *special functions* – arise very often, and have special significance.

Exponentials and powers The exponential function

$$f(t) = e^{rt} \tag{4.2}$$

arises whenever the rate at which something grows is proportional to how much there is now. The constant r gives the rate of growth. The essential behavior of the function is unchanged if one substitutes another positive constant for e. One way to see this is to note that

$$e^{rt} = p^{r't} \quad \text{where} \quad p = e^{r/r'}. \tag{4.3}$$

This function describes population growth of bacteria (or people) in the absence of limits due to food supply or disease. If something is growing exponentially, then if it doubles in time t_0, after time $2t_0$ it will have increased by four, and so on. The function grows so fast that almost nothing in the real world can increase exponentially for too long. For example, suppose that one has ten rats, and that together they weigh 1 kg. Suppose that every year the rat population increases by 20%. Then after 312 years, the rats weigh a total of

$$1\mathrm{kg} \times 1.2^{312} \approx 6.1 \times 10^{24}\mathrm{kg}, \tag{4.4}$$

which is about equal to the entire weight of the earth.

Logarithm The logarithm is the inverse of the exponential:

$$\ln[e^t] = t. \tag{4.5}$$

It has a completely different feel to it. There are very few natural processes that increase logarithmically in time. The logarithm has two types of significance.

First, it provides a very handy calculational tool, because it converts multiplication to addition. That is

$$\ln[ab] = \ln[a] + \ln[b]. \tag{4.6}$$

Multiplication of a and b can therefore be carried out in the following way

1. Find the logarithms of $\ln a$ and $\ln b$ of a and b. This can be done by looking them up in tables.
2. Add the logarithms $\ln a + \ln b = \ln[ab]$.
3. Find the antilogarithm $ab = \exp[\ln ab]$.

This procedure was used for centuries, and underlies the construction of the slide rule, which was the mechanical device that preceded today's calculators. Why tabulate logarithms rather than just tabulate multiplication?

Second, the logarithm shows up in expressions for some exceptionally important physical quantities. Later we will see that the logarithm is used to define the amount of information carried by a collection of messages, and it is also used to define the entropy of a collection of atoms. (In fact, the information in messages and the entropy of atoms are different manifestations of the same idea.)

Sines and cosines Sines and cosines have two basic uses in modeling.

First, they are the functions to think about when faced with any periodic or recurring phenomenon. For this purpose, there is little difference between the two; the sine function starts at 0 ($\sin 0 = 0$) while the cosine function starts at 1 ($\cos 0 = 1$). The two functions are related to each other by a horizontal offset: $\sin(x) = \cos(x - \pi/2)$. Examples of when to think about these functions:

- Mean monthly temperature in Arizona as a function of month over 10 years.
- Air pressure measured every millisecond as a sound wave passes by.
- Sea level at a beach measured every ten minutes over several days.

Second, sines and cosines give *projections*. For example, if a two meter ladder leans against a wall at $70°$, and the sun is directly overhead, then the shadow cast by the ladder is $2\cos 70°$ m long and the ladder rises to a height on the wall of $2\sin 70°$ m.

4.2.5 Regression

Where dimensional analysis provides a basic example of the action of deduction in physical modeling, regression provides a basic example of induction. Regression begins with observations, and asks for a function that fits them. The simplest example is *linear regression:* a collection of points appears to fall on a line. The method of *least squares* provides a systematic way to find the best line passing through the points. This is the line that minimizes the sum of squared vertical distances from the line to the points. If the points lie perfectly on a line, then this sum of squares is zero.

The literature is full of optimistic lines drawn through swarms of seemingly random points. To make a fitting function legitimate, one can employ notions from statistics. One way to do this is

- Take one particular data point, and make repeated measurements. So, if one is measuring the velocity v of marbles rolling down a track as a function of the angle of incline θ of the track, make repeated measurements of v for fixed θ. Having done this, one can get an estimate of the standard deviation σ associated with measuring the particular quantity.
- Repeating this process for many different points, form the sum

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - (Ax_i + B)}{\sigma_i} \right)^2 . \tag{4.7}$$

The method of least squares simply asks for the values of A and B that make Eq. (4.7) as small as possible. Once these values of A and B have been found, there will still be some remaining, nonzero, value of χ^2. The plausibility of the linear fit can be determined by asking, "if y_i are normally distributed variables with standard deviation σ_i, what is the probability of finding the observed value χ^2, or something larger, by chance? In short, a one-sided statistical test. If the odds of finding a value of χ^2 this large by chance alone are near 1 then the linear fit is plausible. If chance alone would be unlikely to give such a value of χ^2, then the linear fit is unlikely to be right.

This discussion mentioned fitting to a linear function $Ax+B$, but exactly the same argument can be used with any other function, although if the function depends upon parameters, it may be difficult the values of the parameters leading to a best fit. Most calculators and plotting or statistical packages contain a variety of fitting functions that one can try to apply to data. Many of the routines try to estimate σ internally as they go along, and do not need to be given it as additional information. They all report a number that describes the goodness of fit. If this number is not near 1 (or in some cases -1) then the function at hand does not fit the data well.

4.2.6 Matching functions to data

- Multiplying a function $f(x)$ by a constant C makes it increase in scale in the *vertical* direction.

- Dividing the argument of a function $f(x)$ by a constant C, obtaining $f(x/C)$ makes it increase in scale by factor C in the *horizontal* direction.

- Adding a constant as in $f(x)+C$ slides a function *up* by C.

- Subtracting a constant as in $f(x-C)$ slides a function *to the right* by C.

In short, these operations capture the ideas of *scale* and *offset.*

The need to understand scale and offset arises so often during inquiries that it requires some more detailed explanation. Here is an example that is intended to give you an idea of what to do in general.

Suppose you have done an inquiry where you have a turning wheel and you want to know how fast it is turning. You decide to measure this by putting a small speaker near the edge of the wheel, hooking it to a tone generator so that it plays a constant tone, and mounting a microphone on the wheel. As the wheel turns, the microphone gets nearer and farther from the speaker, and the strength of the sound it measures goes up and down. A schematic diagram of the apparatus appears in Fig. 4.1 and a graph of sample data appears in Fig. 4.2. As you are designing your experiment, the teaching assistant comes by and says that the data will be "basically sinusoidal."

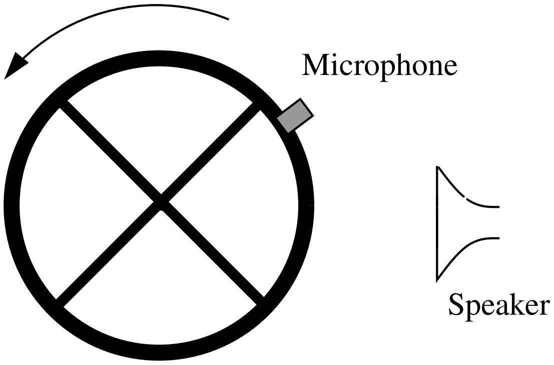


Figure 4.1. A microphone is attached to a rotating wheel, and catches sound from a speaker. The measured sound amplitude is used to find the rotation rate of the wheel.

You are not entirely clear on what this means, but the most logical guess is that the data are the same as $\sin(x)$. So you plot $\sin(x)$ (Figure 4.3). It looks nothing like the data. What was the TA talking about?

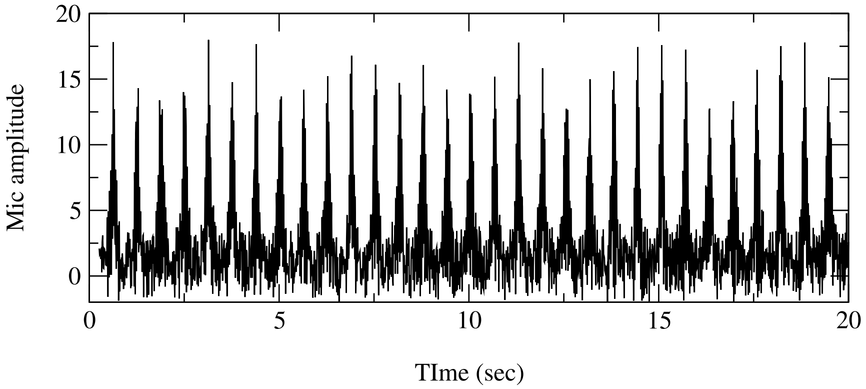


Figure 4.2. Here is a graph of the data collected by LoggerPro of output from microphone versus time.

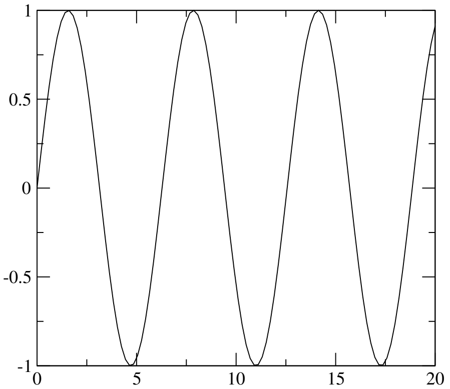


Figure 4.3. A plot of $\sin(x)$.

There many reasons why the sine function does not look exactly like the data. There is nothing that can be done to make the fit perfect, or even particularly close. However, the data and the function can be brought into much closer accord than seems apparent at first. In order to see how, it is valuable to have a computing environment that makes it possible to plot the data, type in a function, and see the function in comparison with the data. *Fathom* is well suited for this task, and one can also carry it out, although with greater difficulty in *Excel.* In my scientific work I usually make use of a free program called *xmgrace* that I find faster and easier to use than either *Fathom* or *Excel.* There are many other free and commercial programs to perform this task as well.

To begin the process of analyzing the data, zoom in on the region between 1 and 5 seconds. (Fig. 4.4). There are two oscillations visible. The slower oscillation takes about 1 second, which is more or less the distance between the highest peaks. This oscillation corresponds to the rotation of the wheel. How does one know? Well, from watching the wheel, that seems about right. The second oscillation is much more rapid, and comes from the rapid vibrations in air pressure measured by the microphone that our ears interpret as a tone. This assumption can be checked by collecting data from the microphone when the wheel is stationary, but the speaker is turned on, and seeing the fast oscillations without the slow ones.

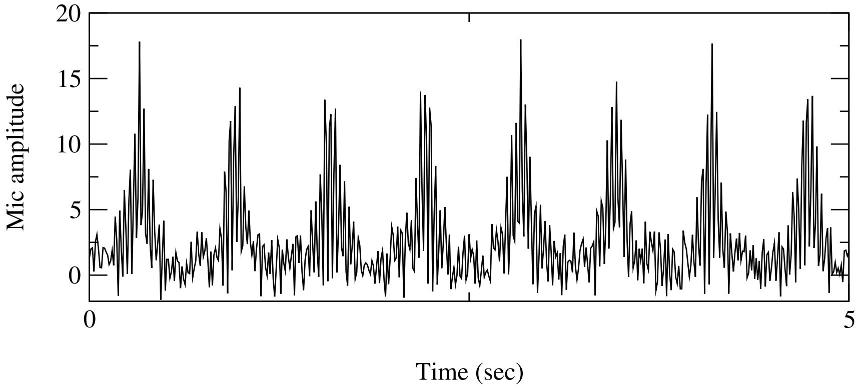


Figure 4.4. Zoom in on region between 1 and 5 seconds.

Since this project is aimed at measuring properties of the rotating wheel, and not the vibrations emitted by the speaker, further analysis focuses on the slow oscillation, neglecting the rapid one. It would be nice somehow to see the slow oscillation without the distracting rapid one getting in the way. This step is not necessary, and not every plotting program makes it easy, but the free program xmgrace contains a simple menu–driven command for this purpose. The operation is to take a *running average*, which means one moves along the data, takes the average of, say, the first 10 points in the data. Then one collects together points 2 through 11, and takes the average of those, then points 3 through 12, and so on until one has traversed the whole data set. Averages have the effect of smoothing out wiggles, and a running average of 10 on these data smooths out any wiggles that have a wavelength smaller than 10, while leaving intact the structure that is much larger than 10 data points in width. Having performed this running average, the data now look as in Fig 4.5.

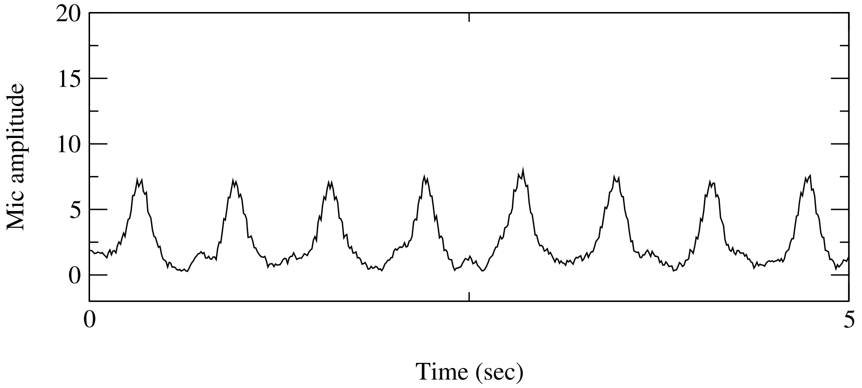


Figure 4.5. Take a running average over 10 consecutive points.

We are now at the point where we can ask the questions about scale and offset with which the section began. The data look roughly like a sine wave, but in many details they are wrong. A sine function has a maximum at 1 and a minimum at -1. These data have a maximum at around 7 and a minimum at around 0. A sine function goes through a complete period when its argument goes through $2\pi \approx 6.2$. The data go through a complete period when time goes through 0.6 seconds, as one can find by measuring the distance between peaks. The sine function is zero when its argument is zero, while the data start at a value of 0.3. These four discrepancies between the data and the function are what can be cured with the four constants the govern scale and offset.

The vertical offset of a function is determined by adding a constant, and the vertical scale is determined by multiplying it by a constant. So the sine function is offset and scaled by writing

$$A + B\sin(x) \tag{4.8}$$

Choose A and B to match the maximum and minimum of the data. The sine function maximum at 1 should correspond to the data maximum at 7. The sine function minimum at -1 should correspond to the data minimum at 0. Write down the two relations

$$7 = A + B \times 1 \tag{4.9}$$

$$0 = A + B \times (-1) \tag{4.10}$$

$$\Rightarrow A = B = 3.5 \tag{4.11}$$

At this point one has the function

$$3.5 + 3.5\sin(x). \tag{4.12}$$

Next, adjust the horizontal scale and offset. This is accomplished by applying addition and multiplication of constants to the argument of the sine function. That is, one considers the function

$$3.5 + 3.5\sin(t/C - D). \tag{4.13}$$

The constants have a more natural interpretation if one divides and subtracts rather than adding and multiplying. To set the function C, notice that when x travels through 0.6 sec, the sine function must travel through a full period, which means that x/C must change by 2π, or

$$2\pi = 0.6\text{sec}/C \Rightarrow C = 0.6\text{sec}/(2\pi) \approx 0.1\text{sec}. \tag{4.14}$$

One says that the period of the oscillation is 0.1 sec. The final constant to determine is D. The data have a first maximum when time equals .35 sec, while the sine function has a first maximum when its argument equals $\pi/2$. So

$$0.35\text{sec}/C - D = \pi/2 \Rightarrow D = .35\text{sec}/C - \pi/2 \approx 3.5 - 1.5 = 2. \tag{4.15}$$

Finally, one has the function

$$3.5 + 3.5\sin(t/(0.1\text{sec}) - 2). \tag{4.16}$$

A comparison of the scaled sine function with the data appears in Fig. 4.6. Whether this is considered a good fit or not depends completely upon the context. A pure mathematician would be likely to think that the functions have nothing to do with one another and not worth further discussion. An applied mathematician would want to find a way to quantify the difference between the data and the fit. A theoretical physicist would view the situation as an encouraging start, but would suggest 15 ways to find more accurate functions, all of them time consuming (and requiring extensive summer funding.) An experimental physicist would view the theoretical fit as acceptable, but want to improve the quality of the data (requiring extensive summer funding). And so on.

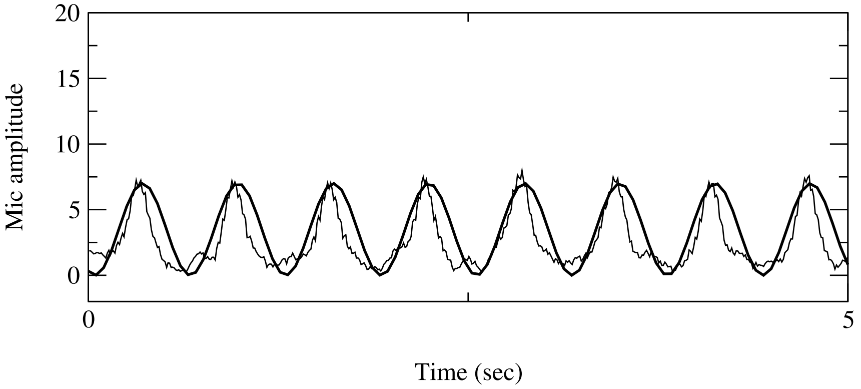


Figure 4.6. Data compared with properly scaled sine function.

5. Collecting and Distributing Scientific Information

5.1 Scientific Reasoning

There are two basic types of reasoning employed in scientific arguments, *deductive* and *inductive*. All branches of science and mathematics employ both of them. Roughly speaking, given premises, the role of deductive reasoning is to predict consequences. The role of inductive reasoning is to provide the premises. A turn-of-the-century mathematician and philosopher, Alfred George Whitehead wrote

> There is a tradition of opposition between adherents of induction and deduction. In my view it would be just as sensible for two ends of a worm to quarrel.

What is the source?

5.1.1 Logical and Deductive Reasoning

Few people think they are illogical. Few professions promote illogical thought. Yet different professions have different standards for logical arguments. The same is true within science and mathematics. Mathematicians have the tightest standard, and the rigor of argument falls off as one moves away from mathematics. The first question we must address is what different people mean by logical thought.

The importance of this question is that one of the reasons to teach science and mathematics to every student in the country is to promote logical thinking. In the past, addition and multiplication were survival skills in their own right. Now that the price ten Big Macs with tax can be determined by pushing on a picture of a hamburger, learning logical thinking is probably the most important goal. Precisely what form this logical thinking is supposed to take, and how learning science and mathematics are supposed to produce it is a question often ignored. Here is how the State of Texas has described the incorporation of logic in mathematics.

> (K.15) **Underlying processes and mathematical tools.** The student uses logical reasoning to make sense of his or her world. The student is expected to reason and support his or her thinking using objects, words, pictures, numbers, and technology. *[Texas Essential Knowledge and Skills, Kindergarten Mathematics]*
>
> (6.13) **Underlying processes and mathematical tools.** The student uses logical reasoning to make conjectures and verify conclusions. The student is expected to:
>
> A make conjectures from patterns or sets of examples and nonexamples; and
>
> B validate his/her conclusions using mathematical properties and relationships.
>
> *[Texas Essential Knowledge and Skills, Sixth-Grade Mathematics]*
>
> (Geometry.3) The student understands the importance of logical reasoning, justification, and proof in mathematics. Following are performance descriptions.
>
> A The student determines if the converse of a conditional statement is true or false.
>
> B The student constructs and justifies statements about geometric figures and their properties.
>
> C The student demonstrates what it means to prove mathematically that statements are true.
>
> D The student uses inductive reasoning to formulate a conjecture.
>
> E The student uses deductive reasoning to prove a statement.
>
> *[Texas Essential Knowledge and Skills, Geometry]*

What type of logical thinking is it that the State really wishes to promote? The goal is certainly not to have every student become a professional mathematician. A popular image of the logical thinker is Sherlock Holmes. Here is how this (purely fictional) detective works:

> The portly client puffed out his chest with an appearance of some little pride and pulled a dirty and wrinkled newspaper from the inside pocket of his greatcoat. As he glanced down the advertisement column, with his head thrust forward and the paper flattened out upon his knee, I took a good look at the man and endeavoured, after the fashion of my companion, to read the indications which might be presented by his dress or appearance.
>
> I did not gain very much, however, by my inspection. Our visitor bore every mark of being an average commonplace British tradesman, obese, pompous, and slow. He wore rather baggy gray shepherd's check trousers, a not over-clean black frock- coat, unbuttoned in the front, and a drab waistcoat with a heavy brassy Albert chain, and a square pierced bit of metal dangling down as an ornament. A frayed top-hat and a faded brown overcoat with a wrinkled velvet collar lay upon a chair beside him. Altogether, look as I would, there was nothing remarkable about the man save his blazing red head, and the expression of extreme chagrin and discontent upon his features.
>
> Sherlock Holmes's quick eye took in my occupation, and he shook his head with a smile as he noticed my questioning glances. "Beyond the obvious facts that he has at some time done manual labour, that he takes snuff, that he is a Freemason. that he has been in China, and that he has done a considerable amount of writing lately, I can deduce nothing else."
>
> Mr. Jabez Wilson started up in his chair, with his forefinger upon the paper, but his eyes upon my companion.
>
> "How, in the name of good-fortune, did you know all that, Mr. Holmes?" he asked. "How did you know, for example, that I did manual labour? It's as true as gospel, for I began as a ship's carpenter."
>
> "Your hands, my dear sir. Your right hand is quite a size larger than your left. You have worked with it, and the muscles are more developed."
>
> "Well, the snuff, then, and the Freemasonry?"
>
> "I won't insult your intelligence by telling you how I read that, especially as, rather against the strict rules of your order, you use an arc-and-compass breastpin."
>
> "Ah, of course, I forgot that. But the writing?"
>
> "What else can be indicated by that right cuff so very shiny for five inches, and the left one with the smooth patch near the elbow where you rest it upon the desk?"
>
> "Well, but China?"
>
> "The fish that you have tattooed immediately above your right wrist could only have been done in China. I have made a small study of tattoo marks and have even contributed to the literature of the subject. That trick of staining the fishes' scales of a delicate pink is quite peculiar to China. When, in addition, I see a Chinese coin hanging from your watch-chain, the matter be- comes even more simple."
>
> Mr. Jabez Wilson laughed heavily. "Well, I never!" said he. "I thought at first that you had done something clever, but I see that there was nothing in it, after all." *[Arthur Conan Doyle, "The Red-Headed League" (1892)]*

There are many examples of thinking of this type in Feynman's stories about himself. They are probably embellished, and probably essentially true. See "He Fixes Radios By Thinking," for example, which casts Feynman as a functioning Sherlock Holmes by the age of twelve. The ability to select a pertinent fact out of masses of irrelevant ones and to focus upon its significance is the hard part of the Holmes skill, more than the deductions.

Most people have mixed emotions about this sort of logical thinking. Logic is opposed to emotion. It is opposed to humanity. An excess of logic leads to an imbalanced personality. Sherlock Holmes has a talent that many people want someone to have, but are not sure they want for themselves:

> "Watson.... you have given prominence ... to those incidents which may have been trivial in themselves, but which have given room for those faculties of deduction and of logical synthesis which I have made my special province...." "You have erred, perhaps... perhaps in attempting to put colour and life into each of your statements instead of confining yourself to the task of placing upon record that severe reasoning from cause to effect which is really the only notable feature about the thing."
>
> "It seems to me that I have done you full justice in the matter," I remarked with some coldness, for I was repelled by the egotism which I had more than once observed to be a strong factor in my friend's singular character.
>
> "No, it is not selfishness or conceit," said he, answering, as was his wont, my thoughts rather than my words. "If I claim full justice for my art, it is because it is an impersonal thing – a thing beyond myself. Crime is common. Logic is rare. Therefore it is upon the logic rather than upon the crime that you should dwell. You have degraded what should have been a course of lectures into a series of tales." *[Arthur Conan Doyle, "The Adventure of the Copper Beeches" (1892)]*

Maybe the crime is the opposite. Schools and universities degrade what could be a series of tales into a course of lectures. The paradox, at any rate, is that everyone seems convinced we should be training young minds to proceed in a logical fashion, while at the same time there is a rather deep general suspicion of those whose minds most completely show the effect of the training. One explanation is the following argument:

- Successful doctors, lawyers, businessmen, and other professionals think logically in their professions.
- Mathematics involves logic.
- If my son or daughter studies mathematics it will help them become a successful professional

There may be scientific evidence for this argument, but it has no basis in pure logic.

Of Kurt Gödel, a foremost logician of the century, his brother Rudolf wrote:-

> My brother had a very individual and fixed opinion about everything and could hardly be convinced otherwise. Unfortunately he believed all his life that he was always right not only in mathematics but also in medicine, so he was a very difficult patient for doctors. After severe bleeding from a duodenal ulcer ... for the rest of his life he kept to an extremely strict (over strict?) diet which caused him slowly to lose weight.

> Towards the end of his life Gödel became convinced that he was being poisoned and, refusing to eat to avoid being poisoned, starved himself to death. *[J J O'Connor and E F Robertson]*

Propositional Logic.

Scientific reasoning is reasoning that convinces scientists. Mathematical reasoning is reasoning that convinces mathematicians. This is probably the only correct definition of mathematical proof, standards for which have varied over the centuries. The type of argument that constitutes proof is nevertheless not completely arbitrary, and has itself become part of the formal study of mathematics.

The simplest subject in formal logic is *propositional logic.* Propositional logic distinguishes between the *validity* of an argument and whether it is *true* or *false*. This form of reasoning is a series of statements of precisely three very specific types.

1. First, there are *premises*. These can be true or false. For example, "My car is green" is a premise (which for me happens to be false.)

2. Second, statements can be negated. Given the statement "My car is green," one can also consider the statement "My car is not green" and its truth value is the opposite of the original statement.

3. Third there are *if-then relations* such as "If [my car is green] then [I will sell it.]" An if-then relation is also said to be true or false. This part is tricky. It works like this. Suppose my car really is green. If I sell it, the if-then relation is true. If I don't actually sell it, the if-then relation is false. If my car is not really green, then the if-then relation is said to be true whether I sell the car or not. This rule can be summarized in a *truth table*.

A	*B*	If *A* then *B*
T	T	T
T	F	F
F	T	T
F	F	T

Finally adding a *rule of inference* makes it possible to play a game in which one starts with statements that are true or false and constructs valid arguments leading to other statements that are true or false. The rule of inference says: Given *A* followed by [If *A* then *B*], it is *valid* to state *B*. So the following is a valid argument.

- The man has a fish tatoo.
- If a man has a fish tattoo the man has been to China.
- Therefore, the man has been to China.

The ideas of *and* and *or* can be defined purely in terms of the structures defined until now, but I will not show that. The idea of "and" is contained in the truth table

A	*B*	*A* and *B*
T	T	T
T	F	F
F	T	F
F	F	F

Another valid mode of reasoning is to begin with certain premises, proceed logically to a conclusion that is false, and thereby show that at least one of the premises must be false. This is a theorem that can be proved based upon ideas already introduced, but again I will not show it, just provide an example:

- The economy is in recession when real GNP declines for two consecutive quarters.
- Last quarter, real GNP increased.
- Therefore, the economy is not in recession.

The statements here are compound, and beginning to approach natural language, but they still can be placed in the context of propositional logic. Notice that although the definition of recession refers to two quarters, the proof there is no recession only mentions one quarter. That is enough, since both quarters must have dropping indicators to satisfy the definition of recession.

An *invalid* statement is one that does not follow from the rules of propositional logic. Here is an example of an invalid argument:

- If my son is a mathematician then he knows mathematics.
- My son knows mathematics.
- Therefore my son is a mathematician.

Other Logics.

Propositional logic is a primitive system. It is rich enough to allow mathematicians to prove some interesting theorems, but not nearly complex enough to describe typical proofs in mathematics, let alone ordinary discussion in English. Adding a few additional entities makes it possible to construct much more complicated mathematical entities. A *first-order logic* must contain the idea of *For all*, and the idea of *There exists.* Now it is possible to construct arguments such as

- All Greeks are people.
- All people are mortal.
- Therefore, all Greeks are mortal. *[Aristotle, Prior Analytics]*

In terms of propositional logic, "All people are mortal" would have to be made as an infinite number of separate statements about all people who ever have been and ever will be. To the idea of moving from a simple premise to a conclusion has been added the idea of moving from the general to the specific.

Similarly one can construct somewhat more interesting invalid arguments:

- All doctors and lawyers think logically.

- All students who learn mathematics well learn to think logically.
- My son is a student.
- My son is learning mathematics well.
- Therefore my son will become a doctor or a lawyer. *[Invalid!]*

As formal logical systems become more and more complex, they become equivalent to more and more complicated contexts for proofs in mathematics.

I will not say any more about the details of formal logic — one can read a book on it (check out *http://www.trentu.ca/academic/math/sb/stefan.html*) or take a course. The full story of formal logic is both very long and very short. It is very long because the mathematical results fill many volumes. It is very short because the original aim of developing formal logic was unsuccessful. The idea was to reduce the whole activity of mathematical proof to a mathematical system. Mathematicians would learn a new symbolic language. It would be possible to make all mathematical statements in terms of this new language. There would be a small number of rules for proceeding from statement to statement. Constructing a proof would consist in nothing than taking an allowed set of steps in the language. Anyone who learned the rules could easily check the most complicated proof, needing little more than patience. This dream is just part of a larger dream. In the larger dream, there is a language that replaces English so that all human thoughts can be expressed in a completely logical fashion, truth can trivially be distinguished from falsehood, and misunderstanding becomes impossible.

Frege began an assault on the attempt to formalize mathematical proof at the end of the nineteenth century. As he was preparing to send the second volume of his life's work to the printer he received a letter from a young British philosopher, Betrand Russell, pointing out that two axioms with which Frege began were inconsistent. Frege added in proof

> Hardly anything more unwelcome can befall a scientific writer than that one of the foundations of his edifice be shaken after the work is finished. I have been placed in this position by a letter of Mr Bertrand Russell just as the printing of the second volume was nearing completion... [F. Frege, Grundgesetze der arithmetik (1903)]

Russell and Whitehead took up the challenge of correcting Frege's axioms, and set about deriving all of mathematics in a new symbolic language, published as *Principia Mathematica*. While this book was enormously influential, ultimately it did not succeed in two ways. First, Gödel proved in 1931 that the ultimate goal of reducing all of mathematics to a completely logical scheme could not succeed. Second, mathematicians continued to use the messy aid of their native languages in discussing mathematics, although formal symbols such as $\forall$ and $\exists$ do routinely appear as parts of proofs. In transforming mathematical argument, there has been partial success.

Theorem (Stoke's Theorem). If ω is a $k-1$–form on an open set $A \in \mathbf{R}^n$ and c is a k–chain in A then

$$\int_c d\omega = \int_{\partial c} \omega. \tag{5.1}$$

Proof: See Michael Spivak (1965), *Calculus on Manifolds* Benjamin, New York, p. 102.

Spivak continues,

> Stoke's theorem shares three important attributes with many fully evolved major theorems:
>
> 1. It is trivial.
> 2. It is trivial because the terms appearing in it have been properly defined.
> 3. It has important consequences.

While not rigidly locked into the constraints of formal proof, the values expressed by this passage show an influence from the formal impulse.

There is another sense in which the research program of formalizing proof has been wildly successful, surpassing original hopes. The idea of reducing mathematics to a sequence of trivial operations led directly to the computer. Many of the scientists (Turing, von Neumann) most responsible for conceiving computers in the 1930's and 1940's were mathematicians working to advance pure research in formal logic. Formal logic still forms the basis for computer science. Here at UT, every computer science major still has to take a formal logic course, which is taught in the philosophy department, and which acts like a mathematics credit.

So, some final thoughts on deductive reasoning:

1. Mathematics has developed the most rigorous standard of argument that exists.
2. Mathematical arguments provide the closest approach to certainty that we will ever know **but**
3. Efforts to place all of mathematics on a completely formal and logical base have failed.
4. It is not possible to have complete agreement on what constitutes a logical argument.
5. Deductive arguments play a role in all scientific discourse, but they are not sufficient for the progress of the sciences, or even the progress of mathematics.
6. The successful application of logic in one sphere of life does not guarantee its successful application in other spheres of life.

5.1.2 Inductive Reasoning

Even within mathematics itself, progress depends upon generalization from examples. For example, the *Goldbach conjecture* is that

> Every even integer greater than 2 can be represented as the sum of two primes.

This statement has no logical foundation. It has resisted proof since it was guessed in 1742. Most mathematicians, however, strongly suspect that it is true, since a vast number of examples has been tested and no exceptions have been found. The conjecture first appeared in a letter to Euler, who replied

> That every even number is a sum of two primes, I consider an entirely certain theorem in spite of that I am not able to demonstrate it.

The conjecture had been verified for all integers up to 10,000 by 1855 (A. Desboves) and has now been verified for all integers up to 4×10^{14} (J.-M. Deshouillers, H. J. J. te Riele, and Y. Saouter).

The reasoning that leads one to believe Goldbach's conjecture is *inductive*. There are 2×10^{14} distinct known cases where it is true, and none where it is false. Experience leads one to seek out patterns, and to predict their recurrence, even if pure logic could permit something different to happen. Most of mathematics probably originates from inductive arguments — some regularity that the mathematician first observes without explanation, and that he or she later manages to prove in deductive fashion.

The role of induction in the rest of science is even more important. All the different forms of experimental science consist in probing the world for regular phenomena that can then be captured in some generalization or model. The general law can never be deduced from the observations. That would be a logical fallacy. Nevertheless, it is an essential form of scientific reasoning.

- When I drop an object in the air it falls down.
- Leave bread out for two weeks and it goes moldy.
- Baking soda fizzes when mixed with vinegar.

These are all examples of knowledge gained inductively. This seems a simple sort of reasoning at first, but becomes extremely difficult when one attempts to work out complicated relationships, or to grapple with uncertainty. Much of the rest of the course will be spent learning how to do it.

Put cause and effect in here when doing this again.

5.2 Publishing

When I was an undergraduate, I remember sitting around with another physics major and agonizing about publication. It was hard enough to hand in weekly problem sets. What could it possibly be like to do scientific research, something original, and actually have it published in a real scientific journal? Things could hardly have seemed worse if I sat in a valley, looked up at a mountain peak, and imagined climbing there — barefoot.

But this sort of mystique is little different from what surrounds any profession whose members require a lot of training. Looking back over more than 50 papers and a book, the process now seems fairly routine. No, writing a scientific paper that truly changes the way that people view the world, founds a new discipline, or gives birth to new industries, is truly as difficult as it ever could have seemed. But simply performing a piece of publishable research, and getting it into print, is not terribly different from the inquiries and writeups in this class. The inquiries and writeups have all the ingredients of the professional counterpart: twists and turns, a constant feeling of uncertainty, the gap between original intentions and final accomplishments, simple errors that come back to haunt you after you have solved ten ten times more difficult problems along the way, the important result that comes out easily, almost by accident and is nearly overlooked, the impossibility of something that seems easy, the trivial solution of something that seems hard, and the terribly unexpected reactions of the outside world to ones results.

Really, the only difference between the inquiries in this class and publishable research is that for the inquiries we have been completely indifferent to whether the results were original or not, and the time frame for each inquiry has been limited to about three weeks. In the great world of science, the results of an inquiry should actually be new, and the time frame is a little longer, around one to five years, which is roughly the time that companies, funding agencies, or universities will grant before demanding to see results.

So how, then, does one determine what is new? Or more to the point, how does one find what has already been done? Some people think they don't need to. "Why, I'm so damn smart, anything I do is bound to be world-shatteringly original, and no one else could possibly ever have done it." One of the reason that other scientists love physicists so much is that so many physicists think like this, particularly when they decide to work on some field other than physics. A slightly more humble twist on this logic is "The world has such infinite variety, and inspires so many questions, that no one could possibly have come up before with the one that I am working on, and it would cramp my originality to check." The truth is, I rather like this particular line myself, although I shouldn't admit it. But to be responsible, one has to acknowledge that there is The Literature.

5.3 Searching

The problem is that The Literature has gotten to be very big. It is growing exponentially. Any number of examples will illustrate the problem. Here is one. The rectification of current — that is, the fact that electrical current sometimes flows more easily in one direction than another — was discovered in 1874 by Schuster and Braun. Writing in 1947 on the subject of metal rectifiers, Henisch wrote that after 80 years of work on the problem

> The student of rectifier problems is confronted with an extensive literature, including some 80 to 100 papers of major historical, practical, or theoretical importance.

Now, jump ahead to 1986, when a new phenomenon was discovered, demonstrably less important, of high–temperature superconductivity, by Bednorz and Müller. By 1988, the literature on high temperature superconductors already included 5200 articles, and by the year 2001, over 73,900 articles on the subject are in print.

The computer is inextricably bound up in the explosion of the research literature. On the one hand, it makes it possible for researchers to make multiple copies of their articles with minor changes and send them to multiple journals on multiple occasions. Researchers do this, since some professional rewards simply follow from the number of papers one publishes, whether or not they are different from one another, and whether or not anyone ever reads them. For example, in the physics department here at UT, yearly raises are largely determined by a point system, and every published article contributes a point. On the other hand, the computer makes it possible to search for scientific information and distribute it in completely new ways.

First, on the matter of searching. I do not know how many of you have carried out a literature search using the high-octane tools that libraries have constructed. So let's take my search on high-temperature superconductivity as an example. This is a physics topic, so I know where to look.

1. Go to *www.lib.utexas.edu.*
2. Click on *Indexes, Abstracts, and Full Text.*
3. Click on *Science/Technology/Health.*

If you don't know the name of the database containing the discipline you are searching for, do a word search on this page. The word "physics" shows up in the description of four databases. One of them is primarily for aerospace engineering, one is primarily for astronomy, one is primarily for electrical engineering, and finally comes INSPEC, which has physics as its first target. Like most of these databases, INSPEC is available to you when you are off campus so long as you have a student ID number and properly configure your computer to use the university proxy server. This service is only available to current students and staff, but this might be worth working to change.

Right now, access to INSPEC is through Ovid; that is, there are separate companies maintaining the database and providing the software interface. So going from one database to another one has to learn dozens of different interfaces, but they have enough family similarities that this is not a problem. To find the number of articles on high–temperature superconductivity, search on

```
high temperature and supercond*
```

The * is a wild-card character that finds anything containing supercond, including superconductor, superconductors, superconductivity. The "and" means to look for all the terms, but not demand that supercond$ be contiguous to the other terms. Searching straight out for "high temperature superconductivity" gives only 2805 matches. Going to another database, like IEEEXplore, that is not so specialized, gives only 2200 matches total. However, this one has the feature that one can click and immediately get the text of every article, something that INSPEC does not now provide.

When carrying out literature searches, one has to constantly zoom in and out until the scale is right. Make the search too general, as in "high temperature and superconduct$," and so many articles pop up that the result is overwhelming. Make it too specific, as in "high temperature and superconduct$ and power cables and Texas," and no hits come up at all. A query at the right scale gets a few hundred results, as in "high temperature and superconduct$ and power cable," which pulls up 113 articles describing the current effort to make power transmission lines out of these materials. Most of the recent articles are by Japanese and European authors, which is one reason why restricting the search to Texas was a mistake.

Something really quite new is the ascendancy of Google. This search engine is so good that it really can serve as a starting point for huge numbers of "why" and "how" questions. Thousands of academics have put sophisticated web pages on display, and these even can lead to the primary research literature. A Google search will never uncover 78,000 articles the way an INSPEC search does. It will, however, find pages by top researchers who provide clear statements of what is important. The ones I have found on this topic are rather superficial. In some fields however, a Google search may now compete with the library databases when it comes to finding important information fast.

5.4 Authorities

One of the most appealing hopes in science is that it provides a set of tools to make all people equal. All trained scientists faced with a certain collection of information and the time to puzzle matters out should arrive at the same conclusion. If this hope were realized, then scientific communities would always present a unified front to the rest of the world, and would only change their opinions when new data emerged that forced old conclusions to be discarded.

There is some truth to this picture of a happy scientific consensus. I do not believe you will find a single faculty in our physics department who argues that Maxwell's equations for electromagnetism are wrong (there is, however, someone in the math department who argues this loudly to anyone who will listen, and someone in the history department who points out that they are not due to Maxwell!)

On the other hand, this image of unanimous scientific authority makes it impossible to account for scientific controversies. Scientific controversies are absolutely commonplace. They do not have to involve anything so grand as overthrowing accepted notions of space and time. Much new scientific work involves explaining a previously mysterious phenomenon, or challenging an accepted explanation. Almost every single published article passes through a debate with the referees. Industrial patent disputes involve expert witnesses supporting both sides.

So, on the one hand, there is a very broad consensus on a wide range of scientific issues, while on the other hand the daily life of most scientists is taken up with disagreements with other scientists. This fact suggests asking how consensus establishes itself. At some point, on most questions, we all fall back on authorities. So, what are the main authorities, and how do they establish themselves? This question can be important for teachers, in trying to navigate through controversial issues.

One type of authority is a person. Specialists in any given field often agree that there is a small number of people whose opinion really matters, and must be reckoned with. Any scientist who wins a Nobel prize obtains a license to be taken seriously for life (even here there are exceptions, but it is not polite to mention any of them.) In theoretical condensed matter physics, P. W. Anderson and J. R. Schrieffer are two prize winners whose opinions and attitudes play a large role. Given the tens of thousands of papers regularly published in any hot field, the role of an authority, even more than deciding what is right, is to rule on what is interesting and important. Personal authority is clearly asserted at conferences, as well as in articles and books.

Some institutions carry a natural authority. The word of any professor from Harvard carries extra weight just because of where he or she works. Only a few other places in the US have similar effect; Princeton and Cal Tech are in the club. Oxford and Cambridge have a magical resonance. (No matter how good a department at the University of Illinois may be, it just doesn't sound quite as good to say "I come from Urbana–Champaign....") All scientists who work at research institutions have in fact survived an enormously daunting set of hurdles. In addition to teaching and administrative work, they must achieve national and international recognition in some area of research, as attested by letters of support and publications. Every university, every national lab, every industrial lab, has an elaborate mechanism for evaluating its scientists, ensuring that they are respected and productive. At the University of Texas, professors are reviewed every 6 years. The review is unlikely to lead to dismissal, except for assistant professors, but certainly affects raises. By which I mean to say, much as I may dislike the idea of authority, research scientists are subject to a great deal of review before they acquire any. It is not unreasonable to look at the institution where a scientist works as part of judging whether to trust his or her words.

Similarly, certain journals carry a lot of weight. The authority of journals is more evenhanded than the authority of individuals or of institutions, since anyone from anywhere has a chance to publish in the journal if only the ideas are good enough. Ever since the second world war, all the dominant scientific journals have been in English (the Russian literature is translated!) At the top of the heap are *Nature* and *Science*, which claim to be interdisciplinary, although they actually emphasize the biological sciences. Every discipline has its top journal. In physics, *Physical Review Letters* is the most prestigious.

The authority of journal rests on the quality of the peer reviews of articles. Reviewers are not paid; reviewing is understood to be part of the professional responsibility of every scientist. Reviewers automatically try to adjust their judgment scales as they move from working for one journal or another. I have personally lost a fair amount of faith in the quality of the reviewing system as I have proceeded through my professional career. The problem, I believe, is that the sheer number of submissions to the best journals overwhelms them, and good reviews are harder and harder to find. I get one or two articles per week to review. Right now I have pending reviews for the *Physical Review* (on a new derivation of Ohm's law), for the *Journal of Applied Mechanics* (on a new mathematical description of rapidly moving cracks), for the *Journal of Applied Physics* (on the growth of nanoclusters), plus a request to review several papers concerning the storage of nuclear waste from a research group at a national lab. I have been very discouraged to see papers that I think are obviously wrong showing up in prestigious journals, and I believe that the load on reviewers, together with publish–or–perish pressures, is largely to blame. Some—but by no means all— scientists feel that peer–reviewed journals are approaching the ends of their useful days. One possible modification might be to stop placing so much emphasis upon the decision of whether or not an article should be published, but whenever it is published, publish it together with its reviews.

Another source of authority lies in textbooks and handbooks. These are also peer reviewed, and now the reviewers get paid, so the reviews may be a bit more careful. Errors at different levels can certainly creep into university–level texts, but with thousands of students working through them, and hundreds of professors questioning them, and the possibility of multiple corrected printings, there is a natural correction mechanism.

High school and elementary texts are a different story. Large sums of state money are involved. State and national politics is involved. They are written by science and mathematics educators, and are rarely reviewed in detail by practicing scientists and mathematicians. Feynman got a chance to review textbooks in California, and writes about it in your book. If you haven't already, please read it. Even if practicing scientists and mathematicians are given a chance to review, their opportunities to influence are constrained. So, the College of Natural Sciences was given the opportunity, for this coming summer, to bid on the review of all science textbooks for the state. However, the charge was to check for factual errors, very narrowly construed. If a text said that Avagadro's number is 6.02×10^{22} we could have fixed that. If a text said that evolution and creation are two hypothesis describing the origin of life, that would be an opinion, and we would be instructed to leave it alone. So the message here is

You probably already know as much science or mathematics or computer science as the people who wrote the textbooks from which you will be asked to teach. Therefore, you are qualified to evaluate these books, to question their claims, to choose among them, and to improve upon their approaches.

Millions of web pages on every conceivable topic have also sprung up. There is no peer review mechanism at all for most of these. Their authority derives mainly from the credentials of who the authors is, or claims to be.

Regardless of the source of information, I honestly believe that everyone is capable of forming their own judgments about scientific claims in a wide range of situations, and that the essential step is simply the decision not to be persuaded unduly by authority that cannot back its claims up with logic and evidence. Here are some things I would look for in evaluating any source of information:

- Who is the author? Where does he or she work? Is there any evidence he or she is an accepted authority? If not, is there evidence that David is fighting a Goliath of conventional wisdom, but that David ought to win?
- Is the argument well written? Does it use unnecessary jargon? Is the work well motivated? Most great scientists have also been great communicators. This is not an accident.
- Passing beyond these superficial tests, one can plow into details of the argument, whatever it is, and see if it makes sense. With some effort and practice, it is possible to read, make sense from, and evaluate, articles from a wide range of disciplines. An excellent way to get practice is by trying to read *Nature* and *Science*. The technical articles in these journals are fairly well written, and in addition the journals publish commentary up front to help explain many of the more technical points. The one thing the commentary never raises, however, is the possibility that the technical article might be wrong. That is something that you, the reader, will have to bring to the experience.

Final comment: most parts of science involve communities. Single individuals are rarely granted too much power. Papers and grant proposals usually involve an absolute minimum of three parties other than the author who judge correctness. Very important results must be reproduced independetly before they are accepted. Most papers have more than one author, and sometimes much more than one author. Work is communal, and judgements are collective.

There is a major exception to this general rule, and that is the classroom. Traditionally, the teacher is the authority. The teacher judges correctness of all work, and assigns scores. There is no appeals process. The teachers' responsibilities are heavy, and teachers should at least be aware of the tension between the traditional authoritarian process for conveying scientific knowledge, and the spirit of freedom and democracy at the scientific frontier.

6. The Research Enterprise

6.1 Creation of today's institutions

At the turn of the last century, I doubt that the US had more than a few tens of thousands of scientists. The forefront of science was in Europe, where almost all the basic discoveries were being made. The US did excel in practical invention and manufacturing, but was not yet the world economic power we now take for granted.

The second world war changed our position. On the one hand, many of Europe's best scientists fled Hitler and occupation, so the U.S. acquired an extraordinary collection of talent without having to invest in education. The work done by these scientists and by the students they trained turned the U.S. into the world's dominant scientific force. The international language of science has remained English ever since. Before the war, all functioning scientists mainly needed to know German and French. On the other hand, science helped the U.S. win the war, and the government took note.

Here is the letter that set our set our current scientific enterprise on its path. From Franklin Roosevelt to Vannevar Bush:

> THE PRESIDENT OF THE UNITED STATES,
> The White House,
> Washington, D. C.
> DEAR DR. BUSH:
>
> The Office of Scientific Research and Development, of which you are the Director, represents a unique experiment of team-work and cooperation in coordinating scientific research and in applying existing scientific knowledge to the solution of the technical problems paramount in war. Its work has been conducted in the utmost secrecy and carried on without public recognition of any kind; but its tangible results can be found in the communiques coming in from the battlefronts all over the world. Some day the full story of its achievements can be told.
>
> There is, however, no reason why the lessons to be found in this experiment cannot be profitably employed in times of peace. The information, the techniques, and the research experience developed by the Office of Scientific Research and Development and by the thousands of scientists in the universities and in private industry, should be used in the days of peace ahead for the improvement of the national health, the creation of new enterprises bringing new jobs, and the betterment of the national standard of living.
>
> It is with that objective in mind that I would like to have your recommendations on the following four major points:
>
> First: What can be done, consistent with military security, and with the prior approval of the military authorities, to make known to the world as soon as possible the contributions which have been made during our war effort to scientific knowledge?
>
> The diffusion of such knowledge should help us stimulate new enterprises, provide jobs four our returning servicemen and other workers, and make possible great strides for the improvement of the national well-being.
>
> Second: With particular reference to the war of science against disease, what can be done now to organize a program for continuing in the future the work which has been done in medicine and related sciences?
>
> The fact that the annual deaths in this country from one or two diseases alone are far in excess of the total number of lives lost by us in battle during this war should make us conscious of the duty we owe future generations.
>
> Third: What can the Government do now and in the future to aid research activities by public and private organizations? The proper roles of public and of private research, and their

> interrelation, should be carefully considered.
>
> Fourth: Can an effective program be proposed for discovering and developing scientific talent in American youth so that the continuing future of scientific research in this country may be assured on a level comparable to what has been done during the war?
>
> New frontiers of the mind are before us, and if they are pioneered with the same vision, boldness, and drive with which we have waged this war we can create a fuller and more fruitful employment and a fuller and more fruitful life.
>
> I hope that, after such consultation as you may deem advisable with your associates and others, you can let me have your considered judgment on these matters as soon as convenient - reporting on each when you are ready, rather than waiting for completion of your studies in all.
>
> Very sincerely yours,
> (s) FRANKLIN D. ROOSEVELT

Vannevar Bush responded in 1945 with a report, *Science, the Endless Frontier* that determined US postwar science policy. [1] Just one brief quotation to understand the tenor of the report:

> The Government should accept new responsibilities for promoting the flow of new scientific knowledge and the development of scientific talent in our youth. These responsibilities are the proper concern of the Government, for they vitally affect our health, our jobs, and our national security. It is in keeping also with basic United States policy that the Government should foster the opening of new frontiers and this is the modern way to do it. For many years the Government has wisely supported research in the agricultural colleges and the benefits have been great. The time has come when such support should be extended to other fields.

Over the following five to ten years, the nation established the National Science Foundation, and put the national laboratories and nuclear energy under civilian control. Following the launching of Sputnik in 1957(?), new waves of energy and money went into increasing the scientific infrastructure. The types of textbooks we now use, the sequence and difficulty of courses, the sizes of the research universities and national laboratories, and the funding mechanisms to support it all, were put in place at that time.

6.2 The situation today

Today we have 1.5 million engineers, 170,000 life scientists, 1.5 million computer scientists and mathematicians (14,000 mathematicians), and 200,000 physical scientists.[2] While these positions are a tiny fraction of the 160 million US jobs, their significance in the economy is greater than the numbers suggest. The future of science and science education is the subject of a constant stream of public reports, usually warning that we have just entered a period of crisis.

There are many reports that could be used to illustrate the attitude of the government toward advanced science and engineering. Here are the summary recommendations from a recent one, *Unlocking our Future*, a report to Congress by the House Committee on Science, 1998. [3] Notice the many different types of expectations for scientific accomplishment, involving industrial development, international relations, education, and many other things.

6.3 Summary of *Unlocking our Future*

1. New ideas form the foundation of the research enterprise. It is in our interests for the Nation's scientists to continue pursuing fundamental, ground-breaking research. Our experience with 50 years of government investment in basic research has demonstrated the economic benefits of this investment. To maintain our Nation's economic strength and international competitiveness, Congress should make stable and substantial federal funding for fundamental scientific research a high priority.

[1] http://www.nsf.gov/od/lpa/nsf50/vbush1945.htm

[2] See the Bureau of Labor Statistics at *http://stats.bls.gov/emptab0.htm*, or for a more detailed study, Science and Engineering Indicators 2000, *http://www.nsf.gov/sbe/srs/seind00/start.htm.*

[3] *http://www.house.gov/science/science_policy_report.htm*

2. Notwithstanding the short-term projections of budget surpluses, the resources of the federal government are limited. This reality requires setting priorities for spending on science and engineering. Because the federal government has an irreplaceable role in funding basic research, priority for federal funding should be placed on fundamental research.

3. The primary channel by which the government stimulates knowledge-driven basic research is through research grants made to individual scientists and engineers. Direct funding of the individual researcher must continue to be a major component of the federal government's research investment. The federal government should continue to administer research grants that include funds for indirect costs and use a peer-reviewed selection process, to individual investigators. However, if limited funding and intense competition for grants causes researchers to seek funding only for ”safe” research, the R&D enterprise as a whole will suffer. Because innovation and creativity are essential to basic research, the federal government should consider allocating a certain fraction of these grant monies specifically for creative, groundbreaking research.

4. The practice of science is becoming increasingly interdisciplinary, and scientific progress in one discipline is often propelled by advances in other, seemingly unrelated, fields. It is important that the federal government fund basic research in a broad spectrum of scientific disciplines, mathematics, and engineering, and resist concentrating funds in a particular area.

5. Much of the research funded by the federal government is related to the mission of the agency or department that sponsors it. Although this research is typically basic in nature, it is nevertheless performed with overriding agency goals in mind. In general, research and development in federal agencies, departments, and the national laboratories should be highly relevant to, and tightly focused on, agency or department missions.

6. The national laboratories are a unique national resource within the research enterprise, but there are concerns that they are neither effective nor efficient in pursuing their missions. A new type of management structure for the federal labs may provide one solution and deserves exploration. To that end, a national laboratory not involved in defense missions should be selected to participate in a corporatization demonstration program in which a private contractor takes over day-to-day operations of the lab.

7. We also have the obligation to ensure that the money spent on basic research is invested well and that those who spend the taxpayers' money are accountable. The Government Performance and Results Act was designed to provide such accountability. Government agencies or laboratories pursuing mission-oriented research should employ the Results Act as a tool for setting priorities and getting the most out of their research programs. Moreover, in implementing the Results Act, grant-awarding agencies should define success in the aggregate, perhaps by using a research portfolio concept.

8. Partnerships in the research enterprise can be a valuable means of getting the most out of the federal government's investment. Cooperative Research and Development Agreements (CRADAs) are an effective form of partnership that leverages federal research funding and allows rapid commercialization of federal research. When the research effort involved in a CRADA fulfills a legitimate mission requirement or research need of the federal agency or national lab, these partnerships should be encouraged and facilitated. Partnerships between university researchers and industries also have become more prevalent as a way for universities to leverage federal money and industries to capture research results without building up in-house expertise. University-industry partnerships should, therefore, be encouraged so long as the independence of the institutions and their different missions are respected.

9. International scientific collaborations form another important aspect of the research enterprise. While most international collaborations occur between individuals or laboratories, the U.S. participates in a number of large-scale collaborations where the costs of large scale science projects can be shared among the participants. In general, U.S. participation in international science projects should be in the national interest. The U.S. should enter into international projects when it reduces the cost of science projects we would likely pursue unilaterally or would not pursue otherwise. Our experience with

international collaborations has not been uniformly successful, as our participation in Mir and the International Space Station demonstrate. Therefore, a clear set of criteria for U.S. entry into, participation in, and exit from an international scientific project should be developed.

10. Large-scale international projects often take place over many years, requiring stable funding over long periods. The annual appropriations cycle in Congress can lead to instability in the funding stream for these projects, affecting our ability to participate. The importance of stability of funding for large-scale, well-defined international science projects should be stressed in the budget resolution and appropriations processes.

11. It is also important that international science projects not appear to be simply foreign aid in the guise of research. To that end, when the U.S. is a major contributor of funds to projects with international participation, funding priority must be placed on the U.S.-based components.

12. America's pre-eminent position in the world suggests new roles for U.S. science policy in the international arena. To take advantage of these opportunities, the State Department must broaden its scientific staff expertise to help formulate scientific agreements that are in America's interest. The evidence suggests that the State Department is not fulfilling this role. Mechanisms that promote coordination between various Executive branch Departments for international scientific projects must be developed. The State Department should strengthen its contingent of science advisors within its Bureau of Oceans and International, Environmental, and Scientific Affairs and draw on expertise in other agencies.

13. A private sector capable of translating scientific discoveries into products, advances and other developments must be an active participant in the overall science enterprise. However, there is concern that companies are focusing their research efforts on technologies that are closest to market instead of on mid-level research requiring a more substantial investment. Capitalization of new technology based companies, especially those that are focused on more long-term, basic research, should be encouraged. In addition, the R&D tax credit should be extended permanently, and needlessly onerous regulations that inhibit corporate research should be eliminated.

14. Partnerships meant to bring about technology development also are important. Well-structured university-industry partnerships can create symbiotic relationships rewarding to both parties. These interactions and collaborations, which may or may not involve formal partnerships, are a critical element in the technology transfer process and should be encouraged.

15. Partnerships that tie together the efforts of State governments, industries, and academia also show great promise in stimulating research and economic development. Indeed, States appear far better suited than the federal government to foster economic development through technology-based industry. As the principal beneficiaries, the States should be encouraged to play a greater role in promoting the development of high-tech industries, both through their support of colleges and research universities and through interactions between these institutions and the private sector.

16. The university community, too, has a role in improving research capabilities throughout its ranks, especially in states or regions trying to attract more federal R&D funding and high-tech industries. Major research universities should cultivate working relationships with less well-established research universities and technical colleges in research areas where there is mutual interest and expertise and consider submitting, where appropriate, joint grant proposals. Less research-intensive colleges and universities should consider developing scientific or technological expertise in niche areas that complement local expertise and contribute to local economic development strategies.

17. To exploit the advances made in government laboratories and universities, companies must keep abreast of these developments. The RAND Corporation's RaDiUS database and the National Library of Medicine's PubMed database serve useful purposes in disseminating information. Consider expanding RaDiUS and PubMed databases to make them both comprehensive and as widely available as possible.

18. Intellectual property protections are critical to stimulating the private sector to develop scientific and engineering discoveries for the market. The Bayh-Dole Act of 1980, which granted the licensing

rights of new technologies to the researchers who discover them, has served both the university and commercial sectors reasonably well. A review of intellectual property issues may be necessary to ensure that an acceptable balance is struck between stimulating the development of scientific research into marketable technologies and maintaining effective dissemination of research results.

19. While the federal government may, in certain circumstances, fund applied research, there is a risk that using federal funds to bridge the mid-level research gap could lead to unwarranted market interventions and less funding for basic research. It is important, therefore, for companies to realize the contribution investments in mid-level research can make to their competitiveness. The private sector must recognize and take responsibility for the performance of research. The federal government may consider supplementary funding for private-sector research projects when the research is in the national interest. Congress should develop clear criteria, including peer review, to be used in determining which projects warrant federal funding.

20. Science and engineering also provide the basis for making decisions as a society, as corporations and as individuals. Science can inform policy issues, but it cannot decide them. In many cases science simply does not have the answer, or provides answers with varying degrees of uncertainty. If science is to inform policy, we must commit sufficient resources to get the answers regulators need to make good decisions. At the earliest possible stages of the regulatory process, Congress and the Executive branch must work together to identify future issues that will require scientific analysis. Sufficient funding to carry out these research agendas must be provided and should not be overly concentrated in regulatory agencies.

21. For science to play any real role in legal and policy decisions, the scientists performing the research need to be seen as honest brokers. One simple but important step in facilitating an atmosphere of trust between the scientific and the legal and regulatory communities is for scientists and engineers to engage in open disclosure regarding their professional background, affiliations and their means of support. Scientists and engineers should be required to divulge their credentials, provide a resume, and indicate their funding sources and affiliations when formally offering expert advice to decision-makers. The scientific opinions these experts offer also should stand up to challenges from the scientific community. To ensure that decision-makers are getting sound analysis, all federal government agencies pursuing scientific research, particularly regulatory agencies, should develop and use standardized peer review procedures.

22. Peer review constitutes the beginning, not the end, of the scientific process, as disagreement over peer-reviewed conclusions and data stimulate debates that are an integral part of the process of science. Eventually, scientists generate enough new data to bring light to previously uncertain findings. Decision-makers must recognize that uncertainty is a fundamental aspect of the scientific process. Regulatory decisions made in the context of rapidly changing areas of inquiry should be re-evaluated at appropriate times.

23. Aside from being based on a sound scientific foundation, regulatory decisions must also make practical sense. The importance of risk assessment has too often been overlooked in making policy. We must accept that we cannot reduce every risk in our lives to zero and must learn to deploy limited resources to the greatest effect. Comprehensive risk analysis should be standard practice in regulatory agencies. Moreover, a greater effort should be made to communicate various risks to the public in understandable terms, perhaps by using comparisons that place risks in the context of other, more recognizable ones.

24. The judicial branch of government increasingly requires access to sound scientific advice. Scientific discourse in a trial is usually highly contentious, but federal judges have recently been given the authority to act as gatekeepers to exclude unreliable science from the courtroom. More and more judges will seek out qualified scientists to assist them in addressing complex scientific questions. How these experts are selected promises to be an important step in the judicial process. Efforts designed to identify highly qualified, impartial experts to provide advice to the courts for scientific and technical decisions must be encouraged.

25. In Congress, science policy and funding remain scattered piecemeal over a broad range of committees and subcommittees. Similarly, in the Executive branch, science is spread out over numerous agencies and departments. These diffusive arrangements make effective oversight and timely decision making extremely difficult. Wherever possible, Congressional committees considering scientific issues should consider holding joint hearings and perhaps even writing joint authorization bills.

26. No factor is more important in maintaining a sound R&D enterprise than education. Yet, student performance on the recent Third International Math and Science Study highlights the shortcomings of current K-12 science and math education in the U.S. We must expect more from our Nation's educators and students if we are to build on the accomplishments of previous generations. New modes of teaching math and science are required. Curricula for all elementary and secondary years that are rigorous in content, emphasize the mastery of fundamental scientific and mathematical concepts as well as the modes of scientific inquiry, and encourage the natural curiosity of children must be developed.

27. Perhaps as important, it is necessary that a sufficient quantity of teachers well-versed in math and science be available. Programs that encourage recruitment of qualified math and science teachers, such as flexible credential programs, must be encouraged. In general, future math and science teachers should be expected to have had at least one college course in the type of science or math they teach, and, preferably, a minor. Ongoing professional development for existing teachers also is important. Another disincentive to entry into the teaching profession for those with a technical degree is the relatively low salaries K-12 teaching jobs offer compared to alternative opportunities. To attract qualified science and math teachers, salaries that make the profession competitive may need to be offered. School districts should consider merit pay or other incentives as a way to reward and retain good K-12 science and math teachers.

28. The revolution in information technology has brought with it exciting opportunities for innovative advances in education and learning. As promising as these new technologies are, however, their haphazard application has the potential to adversely affect learning. A greater fraction of the federal government's spending on education should be spent on research programs aimed at improving curricula and increasing the effectiveness of science and math teaching.

29. Graduate education in the sciences and engineering must strike a careful balance between continuing to produce the world's premier scientists and engineers and offering enough flexibility so that students with other ambitions are not discouraged from embarking on further education in math, science, or engineering. While continuing to train scientists and engineers of unsurpassed quality, higher education should also prepare students who plan to seek careers outside of academia by increasing flexibility in graduate training programs. Specifically, Ph.D. programs should allow students to pursue coursework and gain relevant experience outside their specific area of research.

30. The training of scientists and engineers in the U.S. occurs largely through an apprenticeship model in which a student learns how to perform research through hands-on experience under the guidance of the student's thesis advisor. A result of this link between education and research is that students and post-doctoral researchers are responsible for actually performing much of the federally-funded research done in universities. Mechanisms for direct federal funding of post-docs are already relatively common. Expansion of these programs to include greater numbers of graduate students in math, science and engineering should be explored.

31. Increased support for Masters programs would allow students to pursue an interest in science without making the long commitment to obtaining a Ph.D., and thus attract greater numbers of students to careers in science and technology. More university science programs should institute specially-designed Masters of Science degree programs as an option for allowing graduate study that does not entail a commitment to the Ph.D.

32. The length of time involved and the commensurate forfeiture of income and benefits in graduate training in the sciences and engineering is a clear disincentive to students deciding between graduate training in the sciences and other options. Universities should be encouraged to put controls on the length of

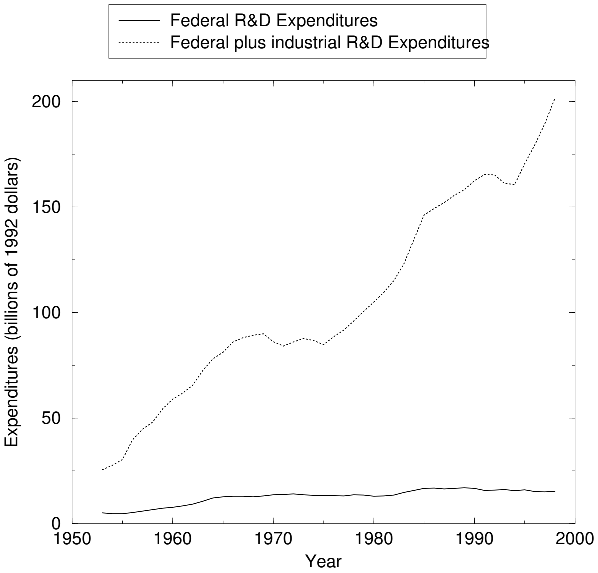


Figure 6.1. Federal and industrial research and development expenditures, expressed in billions of 1992 dollars (that is, compensating in some fashion for the change in value of a dollar with inflation) from 1952 to 1998. Source: *Science and Engineering Indicators 2000, http://www.nsf.gov/sbe/srs/seind00/start.htm.*

time spent in graduate school and post-doctoral study, and to recognize that they cannot attract talented young people without providing adequate compensation and benefits.

33. Educating the general public about the benefits and grandeur of science is also needed to promote an informed citizenry and maintain support for science. Both journalists and scientists have responsibilities in communicating the achievements of science. However, the evidence suggests that the gap between scientists and journalists is wide and may be getting wider. Closing it will require that scientists and journalists gain a greater appreciation for how the other operates. Universities should consider offering scientists, as part of their graduate training, the opportunity to take at least one course in journalism or communication. Journalism schools should also encourage journalists to take at least one course in scientific writing.

34. As important as bridging the gap between scientists and the media is, there is no substitute for scientists speaking directly to people about their work. In part because science must compete for discretionary funding with disparate interests, engaging the public's interest in science through direct interaction is crucial. All too often, however, scientists or engineers who decide to spend time talking to the media or the public pay a high price professionally, as such activities take precious time away from their work, and may thus imperil their ability to compete for grants or tenure. Scientists and engineers should be encouraged to take time away from their research to educate the public about the nature and importance of their work. Those who do so, including tenure-track university researchers, should not be penalized by their employers or peers.

35. The results of research sponsored by the Federal government also need to be more readily available to the general public, both to inform them and to demonstrate that they are getting value for the money the government spends on research. Government agencies have a responsibility to make the results of federally-funded research widely available. Plain English summaries of research describing its results and implications should be prepared and widely distributed, including posting on the Internet.

6.4 Science and Controversies

The National Academy Press has made nearly 2000 books available on the web, covering almost every scientific question that affects public policy. Ask whether powerlines cause cancer, whether marijuana has medical

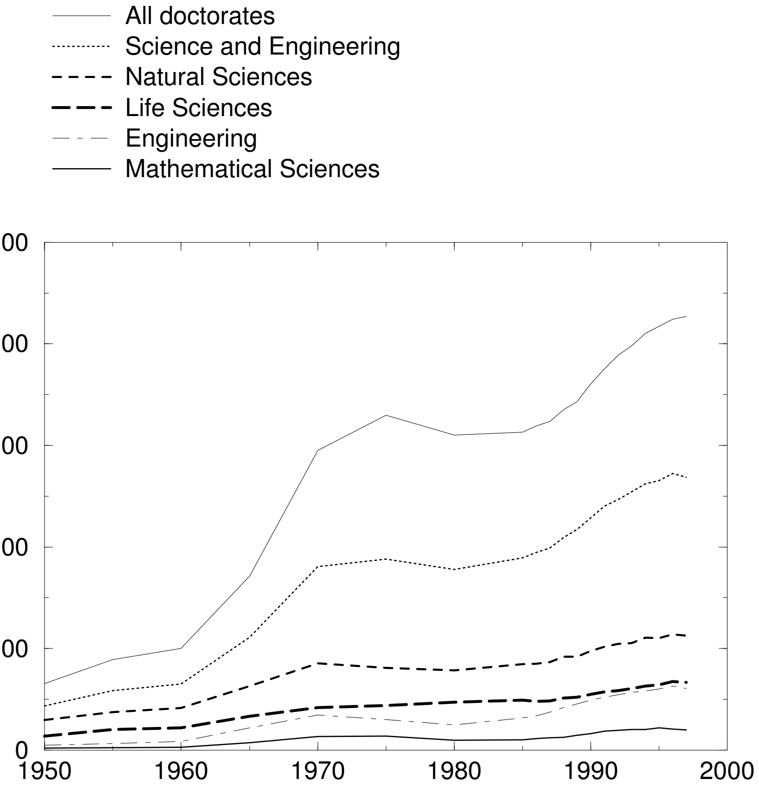


Figure 6.2. Numbers of doctorates granted by year, total, and broken down into science and engineering categories. Number of active researchers is roughly given by area under the curve. Source: *Science and Engineering Indicators 2000, http://www.nsf.gov/sbe/srs/seind00/start.htm.*

benefits, and whether creationsim is science, and you will find answers from the scientific establishment, *www.nap.edu.*

For example, searching on "Creationism" gives

1. Quantum Leaps in the Wrong Direction: Where Real Science Ends...and Pseudoscience Begins (2001, 200 pp.) Charles M. Wynn and Arthur W. Wiggins, With cartoons by Sidney Harris

2. Teaching About Evolution and the Nature of Science (1998, 150 pp.) Working Group on Teaching Evolution, National Academy of Sciences

3. Science and Creationism: A View from the National Academy of Sciences (1992, 28 pp.) Committee on Science and Creationism, National Academy of Sciences

4. Fulfilling the Promise: Biology Education in the Nation's Schools (1990, 168 pp.) Committee on High-School Biology Education, National Research Council

5. Headline News, Science Views (1991, 248 pp.) David Jarmul, Editor; Foreword by Frank Press, President, National Academy of Sciences; National Research Council

6. High-School Biology Today and Tomorrow (1989, 364 pp.) Committee on High-School Biology Education, National Research Council

7. Responsible Science, Volume I: Ensuring the Integrity of the Research Process (1992, 224 pp.) Panel on Scientific Responsibility and the Conduct of Research, National Academy of Sciences, National Academy of Engineering, Institute of Medicine

7. Excel Tutorial