Introduction - What is "Statistics"?

"The revolutionary idea behind modern science...inspecting the facts, the scientist perceives a theory that might work. With this first guess, the process of deduction, experimental test, induction and then making a further hypothesis continues, and science is built up, moving forward in this zig-zag way..."

- David Lindley, "The End of Physics".

To a scientist, "Statistics" has nothing to do with baseball records, the national census, the theory of computers and data storage, and so on. To understand what the word "Statistics" means to a scientist, we have to examine how science works, and we start with the big picture, and consider the scientific processes of deduction and induction. (The above quotation also describes these two processes and their interaction.)

Deduction and Induction

These two words are often used interchangeably in ordinary conversation, but they have completely opposite meanings, especially in science. We now discuss what these are. The following diagram summarizes the discussion below.

```
  deductions
  
Theories about the (unknown) real world

  observations

  inductions
```

Science progresses through an interplay between theory and observations - see above quote. Perhaps the best known, and most famous, interplay arose in astronomy. First, having discarded geocentric theories, Copernicus noted that the planets move around the sun (observation). We speculate that gravity keeps them in their orbits (theory). We next observe (Kepler) that the orbits are ellipses. Newton shows (theory) that the inverse square law of gravitation exactly implies elliptical orbits. The inverse square law theory thus looks pretty good, and in fact reigns supreme until we observe (late 19th century) that the orbits are not quite ellipses. Einstein (1916, theory) shows that general relativity implies slightly non-elliptical orbits, exactly as observed. So far, then, general relativity is our "best" theory, since it explains best what we observe: but we admit
that future observations might lead us to discard it and find a better theory. And so on....

This interplay between theory and observation extends to all levels of science, not just the central paradigm work of people like Newton. An example is given below about the genetic basis of cancer, which is clearly of great current importance.

Any result coming from a theory, for example that under the inverse square law of gravitation, planets will move in elliptical orbits around the sun observe, is called a deduction. Often mathematics is used in this deductive process (as in the inverse square law example).

Any statement coming from observations, for example that planets do not quite move in ellipses, which could change our theories about how nature works, is called an induction.

Thus science consists of making progress through the continual interplay of deductive and inductive arguments, that is of theory and observation.

In all of the above, there has been no mention of randomness in the observations. In the biological and medical sciences in particular, but also in economics, politics and similar areas, random data are the norm. For example, we cannot predict exactly how a given drug will affect a patient, since we know little of the complex physiology of that patient. Nor can we tell exactly how a given fertilizer will affect the growth of plants, because of conditions unknown to us like soil fertility and rainfall. We do not know in advance the result of Gallup poll before we conduct it, because of the randomness in the choice of people interviewed in the Gallup poll. Thus data from experiments concerning new drugs, fertilizers, Gallup polls, etc involve randomness.

Learning how to deal with this randomness is one of the major features of present-day science. This brings us to....

Probability and Statistics

Suppose we throw a thumbtack in the air 1000 times, and note that it came down "point up" 541 times and "point down" 459 times. What does this observation tell us about the hypothesis that a thumbtack is equally likely to land "point up" and "point down"?

Here we are involved with data in whose generation randomness is involved. We cannot tell in advance whether the thumbtack will land up or down. This brings us to the definition of Statistics. In doing statistics we attempt to make an inductive statement about the real world on the basis of observations which unavoidably come from a situation involving randomness. Thus in the thumbtack example, in attempting to say something useful about the hypothesis that the thumbtack is equally likely to land point up or point down, based on the observation of 541 ups and 459 downs, we would be doing statistics.

We often use the word data to mean observations coming from a random process. Statistics is
thus an *inductive* procedure, based on (random) data. The data in the above example is the 541 ups and the 459 downs.

We cannot answer the thumbtack question unless we first carry out a corresponding *deductive* calculation. Since we are dealing with (random) data, this deductive process uses probability theory. Any probability calculation starts with some theory (or hypothesis, or assumption) about the properties of the real world, and then finds, by mathematical operations, various consequences of this theory. So we can extend the diagram above to the following:-

```
[diagram]
```

For the thumbtack example, our initial theory might be that the thumbtack is equally likely to land “point up” and “point down”. A competing theory might be that it is more likely to land “point up” than “point down”. In broad terms, we would calculate the probability of getting 541 or more cases of "point up", *assuming* that the thumbtack is equally likely to land point up or point down. Deductive probability theory shows that this probability is very small, about .0052. Because the probability of the event we saw (or one more extreme) is so small under the hypothesis that a thumbtack is equally likely to land point up and point down, we might feel inclined to claim we have significant evidence against this hypothesis. This claim is our statistical statement.

It is crucial to note that this example demonstrates that (i) we cannot make any statistical statement, that is, a statement based on observed data, unless we first have some probability calculation to base the statement on, and (ii) no statistical statement can be one of complete certainty, (because it is based on random data).

In science we have to deal with problems that are far more important than asking about thumbtacks. But the principles outlined above for our approach to the thumbtack question still apply. For example, we do not know the causes of cancer. We might have a theory (for example that there is a cancer-disposing gene on chromosome 14). To test this theory, we take observations of various kinds. These observations involve considerable randomness (mainly in the random choice of the individuals studied in the experiment), and any statement we make from the data (that is, from our observations in this random experiment) is thus a *statistical* statement. As in the thumbtack example, this statistical statement will be based on some appropriate probability calculation, in this case involving probability calculations assuming there is indeed a cancer-disposing gene on chromosome 14. (The probability theory for this is very complex.) Speaking loosely, if the probability of what we observe is sufficiently small under this
theory, we will reject the theory, and we might try a new theory (for example that there is a cancer-causing gene on chromosome 17). If it is not small, we might make a more refined theory, for example that there is a cancer causing gene on some specified part of chromosome 14. And so on.....

To summarize:-

(1) Statistics is an *inferential* science involving observations of *random* phenomena, from which we attempt to say something about the state of the real world.

(2) These statements, however, cannot be made unless we first do some corresponding *deductive* probability theory calculation on which to base our statistical inference.

(3) Statistics is central in biological and medical research because in these areas, randomness in our observations is essentially unavoidable.

(4) No statement of statistics can be 100% certain of being correct, because it is based on (random) data.

(5) Because statistics is based on probability theory, we must first learn some of that theory. In this course, we only learn those parts of the theory that are relevant to the statistical operations we consider later.

*Relevance to STAT 512/432.*

Much of STAT 512/432 is involved with deriving the theory necessary to carry out tests of hypotheses - that is, essentially with probability theory calculations. One simple test of hypothesis is a *t* test, so let's see how deductive and inductive arguments apply for *t* tests.

A *t* test is carried out using a *t* statistic, whose definition is here assumed to be known. The probability theory, or deductive, part of the procedure is finding the probability distribution of this statistic when some given (null) hypothesis is true. This involves some serious math, which we shall do. Once this is done, the inductive, or statistical, part is almost trivial.

What this implies is that the difficult part in any statistical procedure is in the probability theory and calculations involved before the experiment of interest is even carried out. This is why probability courses are done before statistics courses, and why there is a lot of probability theory in statistics courses. STAT 512/432 considers, in great part, the probability theory necessary for various statistical procedures.

*Final note: Much of STAT 512/432 is concerned with the question of how we do our statistical procedures optimally. In fact the optimality problem is the key aspect of the entire course.*