Lies, Damned Lies and Statistics

The science of statistics is all about interpreting data. Now that people are talking about "big data" this is all the more interesting. The reasoning behind the quote above (from Mark Twain or Benjamin Disraeli depending on whom you believe) is that statistics can be used in a very convincing way to "prove" something—even things you "know" are false!

When we model the real world using mathematics, we decide through observation (or wishful thinking!) which propositions are True/False (e.g. Maxwell's equations hold in all intertial frames) and make deductions using these propositions as axioms. We then test our theories by verifying the collection of propositions so obtained.

If we throw probability into the mix, we could assign probabilities to various propositions. As seen earlier in this course, this would allow us to make statements like, with a probabilty at least 60%, the measurement of (a random variable) X will result in a value that lies in the interval $[9.80, 9.82]$. We would then like to test these theories, but we cannot expect the same level of certainity as we found with the earlier deductive system. In fact, even the assignment of probabilities based on observation needs to be examined as we cannot be certain that such assignments are correct.

Statistics provides us with a systematic way of converting observational data into probabilities for various propositions. Statistics also provides us with a systematic approch to design experiments to verify the deductions based on theories involving probabilities.

A Simple Experiment

We conduct a simple experiment of flipping a coin a large number of times.

```
> load('coin.RData.gz')
> tapply(coinf,coinf,length)
H   T
511 489
```

Does this mean that the coin is biased? Assuming that the coin is unbiased, what is the probability that we would get a result such as the one above?

```
> dbinom(511,1000,0.5)
[1] 0.01980746
```

In other words, there is only a 2% chance of getting the result such as the one above! Does this mean that the coin is biased? No! The reasoning is flawed, after all we can imagine that with an unbiased coin we should get 500 Heads and Tails; what is the probability of that?

```
> dbinom(500,1000,0.5)
[1] 0.02522502
```

Just about 0.5% more! So this was not the correct way to interpret the experiment! Instead, we should calculate the probability of a *deviation* of at least 10 from 500.

```
> 1-sum(dbinom(490:510,1000,0.5))
[1] 0.50666
```

Thus there is a probability of almost 51% that there will be such a deviation!

In summary, all we can conclude is that the results are *consistent* with the hypothesis that the coin is unbiased. This does not *prove* that the coin is unbiased. All we can state is that we *have not* proved that the coin is *biased*!

Suppose that the results were different. For example, we calculate:

```
> 1-sum(dbinom(486:514,1000,0.5))
[1] 0.3591173
```

So, if we had got 515 Heads, then the probability of that happening seems to be significantly less than 50%. Does that mean that the coin *is* biased? How confident can we be of making such an assertion?

Conversely, suppose we have only 503 heads, can we now assert with greater confidence that the coin is, in fact *unbiased*?

Are there any other tests (other than merely counting Heads and Tails) that we can think of? Sure. For example, we could count lengths of “runs” and compare with the negative binomial distribution.

```
> runs <- rle(as.vector(coinf))
> runf <- factor(runs$length)
> tapply(runf,runf,length)
  1   2   3   4   5   6   7   8
242 136  72  20  15  12   5   1
```

We compare that with the expected numbers calculated using the exponential distribution

```
> sapply(1:8, function(k) { trunc(1000*2**(-1-k)); })
[1] 250 125  62  31  15   7   3   1
```

The agreement is not bad, but does this increase our confidence?

The above discussion should give a flavour of what statistics is about.

Tag Cloud for Statistics

Let us discuss some terms that are relevant to statistics.

Descriptive Statistics

This is where we find coarse descriptions of the data, through simple calculations of quantities like the *mean*, *variance*, *median* and *mode*. Through plotting the data or use of other forms of *data visualisation*.

Many of the above techniques require the data to be numerical. However, we can also try to study completely unstructured data with an attempt to infer structure. When this process is automated, it is called *machine learning*. For unstructured multidimensional data, *tag clouds* are a way of visualising data too!

Statistical Inference

After examining the data (or from some prior knowledge) we deduce that the data follows a certain “class” of distributions. For example, for many physics experiments to determine “constants of nature”, we may expect the distribution to be centred around the 'real" value m uniformly distributed in a small interval (determined by the precision of the experiment).

How do we *estimate* the value of m based on the experiments and what is the *likelihood* that the inferred value differs from the “true” (or real) value by an error less than some number e.

More generally, can we use experimental data to estimate the distribution or at least some key aspects of the distribution?

Experimentation and Testing

Once we have made inferences, the next step is to *test* the resulting *hypothesis*. Often, these are statements about the expectations of random variables over a large population, We thus need to *sample* the population carefully.

More generally, when there are a number of competing possibilities, we need to *design* the experiments so that we can clearly differentiate between them.

Finally, we must measure the *significance* of our results. If the results are indeed significant, we can go forward and build the theory further, else we must begin our analysis again!