Some Theory behind Statistical Analysis

In statistics, we begin by collecting data from an experiment. Based on descriptive statistics, we make a "model" for our experiment. Typically, this model is a distribution function with some unknown parameters. Our next task is to estimate these parameters based on the data collected.

To help us do this, we write the likelihood function of these parameters based on the data collected. This allows us to compare one choice of the parameter with other choices of the parameter using the *ratio* of the likelihoods. How does one interpret the ratio of the likelihoods? One way is to compare these likelihood ratios with some "standard" (or *reference*) experiment. If we have two coins, one unbiased and another completely biased, the likelihood of n successive heads with the first is 2^n times smaller than the likelihood in the second case. So a likelihood ratio of 8 is like using three coin flips to say a coin is biased, while a likelihood ratio of 1000 is like using 10 coin flips to decide a coin is biased.

On the one hand this suggests that we estimate the parameter as the one that leads to the maximum likelihood; this is the method of maximum likelihood estimation. This suggests *one* reasonable way to write the estimator function.

On the other hand, if someone else uses an estimator which leads to a likelihood of (say) 1/4-th of the maximum, then a parameter choice based on MLE is "better" than this second method as much as deciding a coin is biased based on two successive heads! This does not sound like much and it is *not*.

So, even if the model completely specified (with no free parameters), we still *need* to ask "How good is this model?" In other words, "How well does the model fit the data?"

Bayesian Inference

Even though we may feel happy with 95% or even more so with 99% confidence in our estimation procedures, we should know that this is not always adequate. To see why let us consider the following example (taken from Lavine's book on Statistics).

Suppose that 1 out of every thousand persons has a certain disease.

Suppose that there is test for the presence of the disease which is 95% accurate.

We formulate this information as follows. Let D be the random variable representing the *existence* of the disease in a randomly chosen member of the population. We have $P(D) = 1/1000$.

Let T be the random variable representing the *result of the* "test* on a randomly chosen member of the population. We are given that the test is 95% successful. In other words:

- Given that the person has the disease ($D = 1$) the probability of the test showing the disease ($T = 1$) is 0.95; in terms of probability theory we write this as $P(T = 1|D = 1) = 0.95$.
- Given that the person does not have the disease ($D = 0$) the probability thet the test showing the disease ($T = 1$) is 0.05; in terms of probability theory we write this as $P(T = 1|D = 0) = 0.05$.

We want to calculate the probability that the person has the disease *given* that the test shows its presence. In other words, we want $P(D = 1|T = 1)$. We calculate

$$P(D = 1|T = 1) = \frac{P(D = 1 \cap T = 1)}{P(T = 1} = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1|D = 1)P(D = 1) + P(T = 1|D = 0)P(D = 0)} = \frac{0.9}{0.95 \times 0.00}$$

In other words, even though the test is 95% accurate, there is only a 2% chance of a person having the disease when the test says that the disease is present!

The relation with statistics is as follows. Let us assume that the parameter c of our distribution follows a certain probability distribution g. The data that we get from our experiments is the distribution of $(X_1, \ldots, X_n|c)$; in other words, the probability of the results given the value of the parameter c. This is what we called the likelihood $l(c|X_1, \ldots, X_n)$.

We then calculate the joint distribution $p(X_1, \ldots, X_n, c)$ by combining these two. Finally, we get the probability of $(c|X_1, \ldots, X_n)$ as the probability that c has the given value given the results of the experiment. We check easily that $p(c|X_1, \ldots, X_n)$ is proportional to $p(c)l(c|X_1, \ldots, X_n)$.

The crucial problem is that we may not have a good idea about $p(c)$. If all parameters are equally likely, then the probability of obtaining c correctly from the data is proportional to the likelihood. However, all values of the parameter are (in general) *not* equally likely!

The idea of Bayesian inference is to use the experiment with some *a priori* (before the fact) value for $p(c)$ and use this to calculate $p(c|X_1, ;X_n)$ as above. By feeding in the actual results of the experiments we have values for $X_i = x_i$ that have already happened. Hence, we can get an *a posteriori* $p'(c) = p(c|X_1 = x_1, \ldots, X_n = x_n)$. We can now use this *improved* estimate for the probability distribution of c in later experiments.

In conclusion, we need to understand the distribution of the parameter better in order to correctly interpret the results of the experiment.

Distribution of an Estimator

A function of the data is called a *statistic.* Since the (collected) data is a (finite) sequence of random variables, this function of the data is itself a random variable. Hence, it also has a distribution function.

In particular, an estimator is a function of the data, hence it has a distribution. This distribution should be a guide to the accuracy of the estimator which is what we want to understand.

For many distributions that we have studied, the primary parameter is equal or proportional to the mathematical expectation. Moreover, the maximum likelihood estimator for the mathematical expectation is (very often) the sample mean. (Note that counting the number of Heads and dividing by the the total number of experiments is also the sample mean).

Recall that we said that an estimator was *unbiased* if *its* mathematical expectation m is the same as the value of the parameter (of the distribution) that it is trying to estimate. This is certainly the case with the sample mean.

The standard deviation s of the estimator is sometimes called its standard error. In the case of the sample mean, we see that this is $s = \sigma/sqrtn$ where σ is the standard deviation of an individual random variable X_i. This s *decreases* with sample size, which is always a good thing!

The sample mean is even more of a "Good Thing" since we have the Central Limit Theorem. This tells us that for a large sample size (of independent identically distributed random variables), the distribution of the sample mean is approximately the same as the normal distribution $N(m, s)$; where, m and s are as above.

Since we know the normal distribution well we can use this as a guide to understanding more about the sample mean as an estimator.

Interval Estimates

We would like to replace the "point" estimates that we have given earlier with "interval" estimates. In other words, we would like to use the data to calculate an interval $[a, b]$ which we *estimate* will contain the parameter that we are trying to calculate. One reason for doing this is that limit distributions are generally continuous and there is no probability associated to individual values for such distributions.

Returning to the case of the sample mean, we have seen above that its distribution is well approximated by the normal distribution for large sample size. We understand the normal distribution well. In particular, we know that if Y is normally distributed with mean m and standard deviation s, then

$$P(m - 2s \leq Y \leq m + 2s) \simeq 0.95 \text{ and } P(m - 3s \leq Y \leq m + 3s) \simeq 0.99$$

and so on. We can see this as the area under the Guassian curve lying over the different intervals.

This allows us to conclude that for large enough sample size, the probability of the sample mean ending up in the interval $[m - 2s, m + 2s]$ is approximately

.95. Since s is also small for a large enough sample size, we this interval is quite small.

The sample variance Z is another random variable and its expectation is σ^2. For large sample sizes we have (for reasons similar to those above) high probability that Z is close to σ^2.

This leads to the procedure for determination of an interval associated with the sample. For a sample of size n, let Y be the sample mean and Z be the sample variance, the interval $[Y - 2\sqrt{Z/n}, Y + 2\sqrt{Z/n}]$ is our estimated interval within which the parameter m is to be located. This is called the 95% *confidence interval*; we can similarly define the 99% confidence interval and other confidence intervals using calculations of the area under the Gaussian lying over these intervals.

Some cautionary statements are in order (see Wikipedia on Confidence Intervals).

- It is not correct to say that m lies in the interval $I = [Y - 2\sqrt{Z/n}, Y + 2\sqrt{Z/n}]$ with probability 0.95! Instead, one can say (from a frequentist interpretation of probability) that for 95% of all (large) samples looked at the estimated interval will contain m. In other words, the probability is associated with the end-points of the interval and not with m.
- It is not necessary that 95% of the data will lie in this interval.
- If one experiment results in an interval I as the 95% confidence interval, then it is not correct to say that there is 95% probability that later experiments will have sample mean in this interval.
- The above description works well only for really large sizes of n. In other cases, one must use Student's t distribution for the correct multiple k_a of $\sqrt{Z/n}$ to be used for a certain confidence level a. We will not worry about this correction in this course.

In other words, one can call the interval obtained as the 95% confidence interval *mathematically* as long as one does not try to interpret this meaning in terms of the "English" language!

One other point which is worth noting is that the ratio between the *value* of $e^{-t^2/2}$ and $t = 0$ and at $t = 2$ (or $t = 3$) is about 7.4 (respectively 20). This is another indication that likelihood ratios of these sizes should not be considered too large.