Correctness of Estimation

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?

Using a Reference Experiment

One way is to compare these likelihood ratios with some "standard" (or *reference*) experiment. If we have two coins, one fair and another always giving heads, the likelihood of n successive heads with the first is 2^n times smaller than the likelihood in the second case.

Comparing with this reference experiment, if $L(c_1)/L(c_2) = 2^n$, and we pick c_1 as the more likely parameter than c_2, then it is like deciding a coin is biased if we get n successive heads with it.

So a likelihood ratio of 8 is like deciding on the basis of 3 successive heads that a coin is biased. Similarly, a likelihood ratio of 1000 is like using 10 succesive heads to decide that 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*.

In each context we must decide what likelihood ratio $L(c_1)/L(c_2)$ is large enough for us to consider that c_1 is a better estimate than c_2. Statisticians suggest that in many cases 10 is reasonable; in other words, the likelihood is "one order of magnitude" different.

Probability distribution of the estimator

As seen earlier, the (maximum likelihood) estimator is a *function* of the values of the random variables. Hence, it too can be thought of as a random variable with a distribution! What is the probability distribution of the estimator?

The justification for studying this distribution is as follows. If the estimator is unbiased, then the mathematical expectation of the estimator *is* the parameter being estimated. Thus, the nature of the probability distribution of the estimator gives us a good idea of how "tight" our estimate is; if the distribution is sharply peaked around the expectation, then we can say, with high probability, that our estimate will be close to the parameter.

As seen in many examples earlier, the parameter of the distribution can be taken to be the expection. In many cases, the *sample mean* can be shown to be the maximum likelihood estimator for the expection of a distribution. In any case, it is a good unbiased estimator of the expectation.

Let us model our data as a sequence of idependent random variables $X_1, X_2, \ldots, X_n$, that follow a distribution f. Assume that $E(X_i) = m$ and $\sigma^2(X_i) = s^2$. (Note that m and s are determined by the distribution f.)

The Central Limit Theorem also tells us that (for standard distributions) the probability distribution of the variable $Y = \sum_i (X_i - m)/s\sqrt{n}$ is well approximated by the normal distribution $N(0, 1)$, where m and s are as above.

Putting $\overline{X} = (\sum_i X_i)/n$, we see that $\overline{X} = (Y(s/\sqrt{n}) + m$, so that the probability distribution for the estimator $\overline{X}$ is well approximated by $N(m, (s/\sqrt{n})^2)$. (This is a bit misleading since the n appears *inside* the limiting distribution!) This can be used to justify the use of this normal distribution as the probability distribution of the estimator $\overline{X}$ for the parameter $m = E(X_i)$ of the distribution f.

Note that, since X_i's are independent random variables, we have

$$\sigma^2(\overline{X}) = \sum_i \sigma^2(X_i)/n^2 = s^2/n$$

Now, since we do not *know* the distribution f beforehand, we do not know m or s either. (Which is why we are using data to estimate them!) We thus often use the *sample variance* $(1/n-1)\sum_i (X_i - \overline{X})^2$ as an unbiased estimator for the variance s. (We note that for large n there is little difference between dividing by n and dividing by $n-1$.)

In summary, we use the *sample mean* $\overline{X}$ as an unbiased estimator for the principal parameter m of the distribution. Further, we approximate the probability distribution of this estimator by the Normal distribution whose variance is given by the *sample variance*. For large n this gives us a useful description of the (probabilistic) behaviour of the estimate.

In particular, if M is the sample mean and S is the sample variance for a large data set, then we can use $[M - 2S/\sqrt{n}, M + 2S/\sqrt{n}]$ as an interval for the value of m with a high (95%) probability. (We can replace 2 by 3 or 1.5 or another value as dictated by our requirements.)

The following two sections are provided for complete-ness of exposition. We will not be doing any problems dealing with them. Both of

these are used to improve the above approximations.

Student's t-distribution

For not too large values n, the above argument does not work well. However, there are many cases where we can feel confident that the distribution f of the variables X_i is in fact a normal distribution of type $N(m, s^2)$. In this case, it was shown by W. H. Gosset (who published under the pseudonym "Student") that we can correct for the fact that n is not too large.

As before we take the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We also take the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

The random variable

$$M = \frac{X - m}{S/\sqrt{n}}$$

can be shown to follow a distribution called "Student's t-distribution with $n-1$ degrees of freedom". Statistical tests that make use of this distribution are called Student's t-tests.

Chi-squared distribution

As seen above, the sample variance is an unbiased estimator for the variance of the distribution. Hence, it too has a probability distribution! When we assume (as above) that the results X_i of our experiments follow a normal distribution $N(m, s^2)$, one can show (Cochran's theorem) that the sample variance (random variable S defined above) follows a scaled Chi-squared distribution. Put differently, one can show that $(n-1)S^2/s^2$ follows the distribution χ^2_{n-1} (where the latter is a distribution that we have not defined!).

This theorem leads to statistical tests. Such as a test of the hypothesis that a certain population does actually follow a normal distribution or that the data does in fact have $n-1$ degrees of freedom and other such tests. All such tests are lumped together under the broad banner of Chi-squared tests.