Chapter 8

Estimation

8.1 The estimation game

Let's play a game. I'll think of a distribution, and you have to guess what it is. We'll start out easy and work our way up.

I'm thinking of a distribution. I'll give you two hints; it's a normal distribution, and here's a random sample drawn from it:

$\{-0.441, 1.774, -0.101, -1.138, 2.975, -2.138\}$

What do you think is the mean parameter, μ, of this distribution?

One choice is to use the sample mean to estimate μ. Up until now we have used the symbol μ for both the sample mean and the mean parameter, but now to distinguish them I will use $\bar{x}$ for the sample mean. In this example, $\bar{x}$ is 0.155, so it would be reasonable to guess $\mu = 0.155$.

This process is called **estimation**, and the statistic we used (the sample mean) is called an **estimator**.

Using the sample mean to estimate μ is so obvious that it is hard to imagine a reasonable alternative. But suppose we change the game by introducing outliers.

I'm thinking of a distribution. It's a normal distribution, and here's a sample that was collected by an unreliable surveyor who occasionally puts the decimal point in the wrong place.

$\{-0.441, 1.774, -0.101, -1.138, 2.975, -213.8\}$

Now what's your estimate of μ? If you use the sample mean your guess is -35.12. Is that the best choice? What are the alternatives?

One option is to identify and discard outliers, then compute the sample mean of the rest. Another option is to use the median as an estimator.

Which estimator is the best depends on the circumstances (for example, whether there are outliers) and on what the goal is. Are you trying to minimize errors, or maximize your chance of getting the right answer?

If there are no outliers, the sample mean minimizes the **mean squared error** (MSE). If we play the game many times, and each time compute the error $\bar{x} - \mu$, the sample mean minimizes

$$MSE = \frac{1}{m} \sum (\bar{x} - \mu)^2$$

Where m is the number of times you play the estimation game (not to be confused with n, which is the size of the sample used to compute $\bar{x}$).

Minimizing MSE is a nice property, but it's not always the best strategy. For example, suppose we are estimating the distribution of wind speeds at a building site. If we guess too high, we might overbuild the structure, increasing its cost. But if we guess too low, the building might collapse. Because cost as a function of error is asymmetric, minimizing MSE is not the best strategy.

As another example, suppose I roll three six-sided dice and ask you to predict the total. If you get it exactly right, you get a prize; otherwise you get nothing. In this case the value that minimizes MSE is 10.5, but that would be a terrible guess. For this game, you want an estimator that has the highest chance of being right, which is a **maximum likelihood estimator** (MLE). If you pick 10 or 11, your chance of winning is 1 in 8, and that's the best you can do.

Exercise 8.1 Write a function that draws 6 values from a normal distribution with $\mu = 0$ and $\sigma = 1$. Use the sample mean to estimate μ and compute the error $\bar{x} - \mu$. Run the function 1000 times and compute MSE.

Now modify the program to use the median as an estimator. Compute MSE again and compare to the MSE for $\bar{x}$.

8.2 Guess the variance

I'm thinking of a distribution. It's a normal distribution, and here's a (familiar) sample:

$\{-0.441, 1.774, -0.101, -1.138, 2.975, -2.138\}$

What do you think is the variance, σ^2, of my distribution? Again, the obvious choice is to use the sample variance as an estimator. I will use $S^2$to denote the sample variance, to distinguish from the unknown parameter σ^2.

$$S^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

For large samples, $S^2$is an adequate estimator, but for small samples it tends to be too low. Because of this unfortunate property, it is called a **biased** estimator.

An estimator is **unbiased** if the expected total (or mean) error, after many iterations of the estimation game, is 0. Fortunately, there is another simple statistic that is an unbiased estimator of σ^2:

$$S_{n-1}^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

The biggest problem with this estimator is that its name and symbol are used inconsistently. The name "sample variance" can refer to either S^2or S_{n-1}^2, and the symbol $S^2$is used for either or both.

For an explanation of why $S^2$is biased, and a proof that $S_{n-1}^2$is unbiased, see `http://wikipedia.org/wiki/Bias_of_an_estimator`.

Exercise 8.2 Write a function that draws 6 values from a normal distribution with $\mu = 0$ and $\sigma = 1$. Use the sample variance to estimate σ^2 and compute the error $S^2 - \sigma^2$. Run the function 1000 times and compute mean error (not squared).

Now modify the program to use the unbiased estimator S_{n-1}^2. Compute the mean error again and see if it converges to zero as you increase the number of games.

8.3 Understanding errors

Before we go on, let's clear up a common source of confusion. Properties like MSE and bias are long-term expectations based on many iterations of the estimation game.

While you are playing the game, you don't know the errors. That is, if I give you a sample and ask you to estimate a parameter, you can compute the value of the estimator, but you can't compute the error. If you could, you wouldn't need the estimator!

The reason we talk about estimation error is to describe the behavior of different estimators in the long run. In this chapter we run experiments to examine those behaviors; these experiments are artificial in the sense that we know the actual values of the parameters, so we can compute errors. But when you work with real data, you don't, so you can't.

Now let's get back to the game.

8.4 Exponential distributions

I'm thinking of a distribution. It's an exponential distribution, and here's a sample:

{5.384, 4.493, 19.198, 2.790, 6.122, 12.844}

What do you think is the parameter, λ, of this distribution?

In general, the mean of an exponential distribution is $1/\lambda$, so working backwards, we might choose

$\hat{\lambda} = 1 \ / \ \bar{x}$

It is common to use hat notation for estimators, so $\hat{\lambda}$ is an estimator of λ. And not just any estimator; it is also the MLE estimator[1]. So if you want to maximize your chance of guessing λ exactly, $\hat{\lambda}$ is the way to go.

But we know that $\bar{x}$ is not robust in the presence of outliers, so we expect $\hat{\lambda}$ to have the same problem.

Maybe we can find an alternative based on the sample median. Remember that the median of an exponential distribution is $\ln(2) \ / \ \lambda$, so working backwards again, we can define an estimator

$$\hat{\lambda}_{1/2} = \ln(2)/\mu_{1/2}$$

where $\mu_{1/2}$ is the sample median.

Exercise 8.3 Run an experiment to see which of $\hat{\lambda}$ and $\hat{\lambda}_{1/2}$ yields lower MSE. Test whether either of them is biased.

[1]See http://wikipedia.org/wiki/Exponential_distribution#Maximum_likelihood.

8.5 Confidence intervals

So far we have looked at estimators that generate single values, known as **point estimates**. For many problems, we might prefer an interval that specifies an upper and lower bound on the unknown parameter.

Or, more generally, we might want that whole distribution; that is, the range of values the parameter could have, and for each value in the range, a notion of how likely it is.

Let's start with **confidence intervals**.

I'm thinking of a distribution. It's an exponential distribution, and here's a sample:

{5.384, 4.493, 19.198, 2.790, 6.122, 12.844}

I want you to give me a range of values that you think is likely to contain the unknown parameter λ. More specifically, I want a 90% confidence interval, which means that if we play this game over and over, your interval will contain λ 90% of the time.

It turns out that this version of the game is hard, so I'm going to tell you the answer, and all you have to do is test it.

Confidence intervals are usually described in terms of the miss rate, α, so a 90% confidence interval has miss rate $\alpha = 0.1$. The confidence interval for the λ parameter of an exponential distribution is

$$\left(\hat{\lambda} \frac{\chi^2(2n, 1-\alpha/2)}{2n}, \hat{\lambda} \frac{\chi^2(2n, \alpha/2)}{2n} \right)$$

where n is the sample size, $\hat{\lambda}$ is the mean-based estimator from the previous section, and $\chi^2(k, x)$ is the CDF of a chi-squared distribution with k degrees of freedom, evaluated at x (see `http://wikipedia.org/wiki/Chi-square_distribution`).

In general, confidence intervals are hard to compute analytically, but relatively easy to estimate using simulation. But first we need to talk about Bayesian estimation.

8.6 Bayesian estimation

If you collect a sample and compute a 90% confidence interval, it is tempting to say that the true value of the parameter has a 90% chance of falling in the