Predicting a Future Observation

We have seen how to use a sample from a population to estimate the population mean μ, and how to quantify the uncertainty in that estimate by use of a confidence interval. Suppose instead that our interest is in using that sample to predict the value of a new observation to be taken from that population and to quantify the uncertainty in that prediction. In this presentation we will develop the statistical prediction tools for a population that follows a normal distribution with mean μ and variance σ^2 (i.e. a $N(\mu, \sigma^2)$ distribution).

Suppose to begin with, that we know μ, and that we use μ to predict the future observation. Call this future observation y_{new}. Since it comes from the same population, y_{new} has a $N(\mu, \sigma^2)$ distribution as well. The error in using μ to predict y_{new} will be $y_{new} - \mu = \epsilon_{new}$, which follows a $N(0, \sigma^2)$ distribution. Note that even if we know μ, there is still uncertainty in predicting y_{new}.

Of course, in reality we don't know μ, so we estimate it using $\hat{\mu} = \overline{y}$. When using $\hat{\mu}$ to predict a future observation, we'll call it $\hat{y}_{new}$. This is the **point predictor**.

We would also like a measure of how precise the prediction is. We measure this using the **standard error of prediction**, $\sigma^(y_{new} - \hat{y}_{new})$, which is computed as follows.

When using $\hat{y}_{new}$ to predict a new observation, the prediction error is

$$y_{new} - \hat{y}_{new} = (\mu + \epsilon_{new}) - \hat{y}_{new} = (\mu - \hat{y}_{new}) + \epsilon_{new}. \tag{1}$$

The term $(\mu - \hat{y}_{new})$ in the rightmost expression of (1) equals $(\mu - \hat{\mu})$ and is the error due to using $\hat{\mu}$ to estimate μ. Its variance is the variance of $\hat{\mu} = \overline{y}$, which, as we know, is σ^2/n. The second term, ϵ_{new}, is the random error inherent in y_{new}. Its variance, as we saw above, is σ^2. Since $\hat{y}_{new}$, being computed from the current data, is independent of the new observation y_{new}, the variance of $y_{new} - \hat{y}_{new}$ is the sum of the variances of $(\mu - \hat{y}_{new})$ and ϵ_{new}. That is,

$$\sigma^2(y_{new} - \hat{y}_{new}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2 \left[1 + \frac{1}{n}\right]. \tag{2}$$

Now, of course, we rarely know σ^2, so we use the sample variance, s^2, to estimate it, which gives the estimated standard error of prediction

$$\hat{\sigma}(y_{new} - \hat{y}_{new}) = \sqrt{s^2 \left[1 + \frac{1}{n}\right]} = s\sqrt{1 + \frac{1}{n}}.$$

It turns out that

$$t = \frac{y_{new} - \hat{y}_{new}}{\hat{\sigma}(y_{new} - \hat{y}_{new})}$$

has a t_{n-1} distribution. Therefore,

$$\begin{aligned} L &= P(-t_{n-1,\frac{1+L}{2}} < t < t_{n-1,\frac{1+L}{2}}) \\ &= P(\hat{y}_{new} - \hat{\sigma}(y_{new} - \hat{y}_{new})t_{n-1,\frac{1+L}{2}} < y_{new} < \hat{y}_{new} + \hat{\sigma}(y_{new} - \hat{y}_{new})t_{n-1,\frac{1+L}{2}}). \end{aligned}$$

It follows that a level L prediction interval for a future observation is

$$(\hat{y}_{new} - \hat{\sigma}(y_{new} - \hat{y}_{new})t_{n-1,\frac{1+L}{2}},\ \hat{y}_{new} + \hat{\sigma}(y_{new} - \hat{y}_{new})t_{n-1,\frac{1+L}{2}}).$$

As with confidence intervals for a model parameter, the interpretation of a prediction interval is based on repeated samples from the given population. Suppose for each sample a level L prediction interval is computed, and that a new observation is then taken from the population. A proportion L of all constructed intervals will contain their corresponding new observation.

A Caution

As we have previously seen, the Central Limit Theorem helps insure that, even if the population does not follow a normal distribution, in practice confidence intervals for the mean will still be valid if the sample size is not small. That is no longer true for prediction intervals, since the distribution of a single future observation will not be affected no matter how large a sample is used for prediction.

Example

A researcher at a biotechnology company is testing an artificial pancreas on laboratory rats. She gives four diabetic rats, who have had this pancreas implanted, an initial dose of glucose in solution and then measures their blood-sugar levels (serum/plasma glucose, mg/100ml) after one hour. The data are

266 149 161 220

We will use the artificial pancreas data to construct a 95% prediction interval for a future observation. From the data, we obtain the

$$\hat{y}_{new} = \overline{y} = 199, \quad s^2 = 2958$$

From this, we compute the estimated standard error of prediction:

$$\hat{\sigma}(y_{new} - \hat{y}_{new}) = \sqrt{2958\left[1 + \frac{1}{4}\right]} = \sqrt{3697.5} = 60.8.$$

Since

$$t_{n-1,\frac{1+L}{2}} = t_{3,0.975} = 3.18,$$

the desired interval is

$$(199 - (60.8)(3.18), 199 + (60.8)(3.18)) = (5.6, 392.4).$$

Notice that this 95% prediction interval for a future observation is much wider than the classical 95% confidence interval for μ. The extra width reflects the extra uncertainty involved in obtaining a completely new observation from the population.

In interpreting this interval, the researcher can say: "Suppose I obtain another rat (i.e., a rat not in the sample of four rats) with the artificial pancreas. Suppose that I measure the blood sugar level of this rat one hour after the ingestion of the glucose solution. I predict with 95% confidence that the blood sugar level of this new rat will fall between 5.6 and 392.4. My confidence rests in my use of a method to compute this interval which will produce an interval containing the blood sugar level of a new rat in 95% of all possible identically-run experiments."