9.6 Prediction Interval

The point and interval estimations of the mean in Sections 9.4 and 9.5 provide good information on the unknown parameter μ of a normal distribution, or a non-normal distribution from which a large sample is drawn. Sometimes, other than the population mean, the experimenter may also be interested in predicting the possible **value of a future observation.** For instance, in a quality control case, the experimenter may need to use the observed data to predict a new observation. A process that produces a metal part may be evaluated on the basis of whether a part meets process specifications on tensile strength. On certain occasions a customer may be interested in purchasing a **single part.** In this case a confidence interval on the mean tensile strength does not capture the requirement. The customer requires a statement regarding the uncertainty of one **single observation.** The type of requirement is nicely fulfilled by the construction of a **prediction interval.**

It is quite simple to obtain a prediction interval for the situations we have considered so far. Assume that the random sample comes from a normal population with unknown mean μ and known variance σ^2. A natural point estimator of a new observation is $\overline{X}$. It is known, from Section 8.5 that the variance of $\overline{X}$ is σ^2/n. However, to predict a new observation, not only do we need to account for the variation due to estimating the mean, but also should we account for the variation of the future observation. From the assumption, we know that the variance of the random error in a new observation is σ^2. The development of a prediction interval is best displayed by beginning with a normal random variable $x_0 - \bar{x}$, where x_0 is the new observation and $\bar{x}$ comes from the sample. Since x_0 and $\bar{x}$ are independent we know that

$$z = \frac{x_0 - \bar{x}}{\sqrt{\sigma^2 + \sigma^2/n}} = \frac{x_0 - \bar{x}}{\sigma\sqrt{1 + \frac{1}{n}}}$$

is $n(z;0,1)$. As a result, if we use the probability statement

$$\Pr[-z_{\alpha/2} < Z < z_{\alpha/2}]$$

with the z statistic above, and place x_0 in the center of the probability statement, we have the following occurring with probability $1 - \alpha$.

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}}.$$

As a result, the computed prediction interval is formalized as follows:

Prediction interval of a future observation: σ known

For a normal distribution of measurements with unknown mean μ and known variance σ^2, a $(1 - \alpha)/100\%$ **prediction interval** of a future observation, x_0, is

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n},$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

Example 9.5

Due to the decreasing of interest rates, the First Citizens Bank received a lot of mortgage applications. A recent sample of 50 mortgage loans resulted in an average of \$128,300. Assume a population standard deviation of \$15,000. If a next customer called in for a mortgage loan application, find a 95% prediction interval on this customer's loan amount.

Solution

The point prediction of the next customer's loan amount is $\bar{x} = \$128{,}300$. The z-value here is $z_{0.025} = 1.96$. Hence a 95% prediction interval for a future loan is

$$128300 - (1.96)(15000)\sqrt{1 + \frac{1}{50}} < x < 128300 + (1.96)(15000)\sqrt{1 + \frac{1}{50}},$$

which gives the interval (\$98,607.46, \$157,992.54).

The prediction interval provides a good estimate of the location of a future observation, which is quite different from the estimation of the sample mean value. It should be noted, that the variation of this prediction is the sum of the variation due to an estimation of the mean and the variation of a single observation. However, as in the past we consider the known variance case first. It is, therefore, important to deal with the prediction interval of a future observation in the situation that the variance is unknown. Indeed a Student's t-distribution may be used in this case as described in the following result. Here the normal distribution is merely replaced by the t-distribution.

Prediction interval of a future observation: σ unknown

For a normal distribution of measurements with unknown mean μ and unknown variance σ^2, a $(1 - \alpha)\,100\%$ **prediction interval** of a future observation, x_0, is

$$\bar{x} - t_{\alpha/2}s\sqrt{1 + 1/n} < x_0 < \bar{x} + t_{\alpha/2}s\sqrt{1 + 1/n},$$

where $t_{\alpha/2}$ is the t-value with $v = n - 1$ degrees-of-freedom, leaving an area of $\alpha/2$ to the right.

Example 9.6

A meat inspector has randomly measured 30 packs of acclaimed 95% lean beef. The sample resulted in the mean 96.2% with the sample standard deviation of 0.8%. Find a 99% prediction interval for a new pack. Assume normality.

Solution For $v = 29$ degrees-of-freedom $t_{0.005} = 2.756$. Hence a 99% prediction interval for a new observation x_0 is

$$96.2 - (2.756)(0.8)\sqrt{1 + \frac{1}{30}} < x_0 < 96.2 + (2.756)(0.8)\sqrt{1 + \frac{1}{30}},$$

which reduces to (93.96, 98.44). ⌟

Use of Prediction Limits For Outlier Detection

To this point in the text very little attention has been paid to the concept of **outliers** or abberant observations. The majority of scientific investigators are keenly sensitive to the existance of outlying observations or so called faulty or "bad data". We deal with the concept to a large extent in Chapter 12 where outlier detection in regression analysis is illustrated. However, it is certainly of interest to consider here since there is an important relationship between outlier detection and prediction intervals.

It is convenient for our purposes to view an outlying observation as one in which that observation comes from a population with a mean that is different from that which governs the rest of the sample of size n being studied. The prediction interval produces a bound that "covers" a future single observation with probability $1 - \alpha$ if it comes from the population from which the sample was drawn. As a result, a methodology for outlier detection involves the rule that an **observation is an outlier if it falls outside the prediction interval computed without inclusion of the questionable observation in the sample.** As a result, for the prediction interval of Example 9.6, if a new pack is observed and contains a percent fat content outside the interval [93.96 98.44] it can be viewed as an outlier.