MAT 2379, Introduction to Biostatistics

Chapter 10. Confidence Intervals

The goal of the statistical inference is to draw conclusions about an unknown parameter which describes a feature of the population (given by a measurement X). Examples of parameters include: (a) the population mean μ; (b) the population variance σ^2; (c) the proportion p of individuals with a certain characteristic. In this chapter, we introduce the method of *estimation by confidence intervals*, for estimating the unknown parameter. This method produces an interval of possible values that includes the unknown parameter with a high probability.

Our conclusions will be based on the observed values $x_1, \ldots, x_n$ of the measurement X for a particular sample drawn from the population. Note that $x_1, \ldots, x_n$ can be regarded as the observed values which correspond to some random measurements $X_1, \ldots, X_n$. The measurements $X_1, \ldots, X_n$ are independent and have the same distribution as X.

A *point estimator* for the population mean μ is the theoretical quantity:

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}.$$

The sample mean $\bar{x}$ is the observed value of this estimator for our particular sample, and is called an *estimate of* μ.

A *point estimator* for the population variance σ^2 is the theoretical quantity:

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

The sample variance s^2 is the observed value of this estimator for our particular sample, and is called an *estimate of* σ^2.

In summary, we have the following table:

Parameter	Estimator	Estimate
μ	$\bar{X}$	$\bar{x}$
σ^2	S^2	s^2

Example 1. (Example 10.1 in the book) A research project supported in part by the Canadian Wildlife Federation, shows that in the recent years, an increased number of polar bears in the Beauford Sea are eating less, possibly due to a decrease in the number of ringed seals (the bear's main food source), during a critical spring feeding period. Further indication that the bears are fasting are smaller weights of their cubs at birth. The following data gives the weights at birth (in grams), for a sample of $n = 5$ cubs:

$$x_1 = 785, \quad x_2 = 825 \quad x_3 = 671 \quad x_4 = 981 \quad x_5 = 732.$$

The average weight for these 5 cubs is:

$$\bar{x} = \frac{785 + 825 + 671 + 981 + 732}{5} = 798.8\text{g}.$$

The sample variance for this data is:

$$s^2 = \frac{(785-798.8)^2+(825-798.8)^2+(671-798.8)^2+(981-798.8)^2+(732-798.8)^2}{5-1} = 13717.2.$$

The sample standard deviation is

$$s = \sqrt{13717.2} = 117.1205\text{g}.$$

In this example, X represents the weight of a (randomly chosen) cub at birth. We are interested in estimating the mean μ of X. From the practical point of view, μ can be interpreted as the average cub weight at birth for the entire population of polar bears in the Beauford Sea. The variance σ^2 gives an indication about the amount of variability of the cub weights at birth. An estimate of μ is $\bar{x} = 798.8$g, whereas an estimate for σ is $s = 117.1205$g.

Section 10.1 Confidence Interval for the Mean μ: σ^2 known

The statistic $\bar{X}$ is called a *point estimator* because it gives a single numeric value. This value is not exactly the value of the parameter μ of interest. The method of *estimation by confidence intervals* gives an interval $[L_1, L_2]$ of values which has a high probability of containing the unknown parameter μ. For instance

$$P(L_1 \leq \mu \leq L_2) = 0.95.$$

To obtain a confidence interval in this case, one uses the fact that: (Theorem 9.5 in the book)

Central Limit Theorem: If n is large, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has approximately a standard normal distribution.

If we know that the random sample is drawn from a population which is normally distributed with mean μ and variance σ^2, then the distribution of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is *exactly* standard normal (Theorem 9.3 in the book). In this case, we do not need n to be large.

From Table 17.3, we know that $P(-1.96 \leq Z \leq 1.96) = 0.95$. By replacing Z with $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ we get

$$P(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

After some calculations (see p. 112 of your book), this becomes

$$P\left(\bar{X} - (1.96)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + (1.96)\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

The interval $[\bar{X} - (1.96)(\sigma/\sqrt{n}),\ \ \bar{X} + (1.96)(\sigma/\sqrt{n})]$ is called a 95% *confidence interval* for the mean μ because it includes the mean μ, with probability 95%.

The value $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is called **the standard error of the mean**.

Similarly, we have $P(-2.575 \leq Z \leq 2.575) = 0.99$, and hence the 99% confidence interval is

$$\left[\bar{X} - (2.575)\frac{\sigma}{\sqrt{n}},\ \ \bar{X} + (2.575)\frac{\sigma}{\sqrt{n}}\right]$$

Also $P(-1.645 \leq Z \leq 1.645) = 0.90$ and the 90% confidence interval is:

$$\left[\bar{X} - (1.645)\frac{\sigma}{\sqrt{n}}, \quad \bar{X} + (1.645)\frac{\sigma}{\sqrt{n}}\right]$$

Example 1. (continued) Suppose that the cub weight X at birth is a normal random variable with standard deviation $\sigma = 115$g. The general formula for the 95% confidence interval is

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}.$$

Since X has a normal distribution, the construction of the interval is justified by Theorem 9.3 mentioned above. Note that the sample size is small. We know that $n = 5$, $\bar{x} = 798.8$g and $\sigma = 115$g. The 95% confidence interval for the average cub weight μ at birth is

$$798.8 \pm 1.96\frac{115}{\sqrt{5}}, \quad \text{or} \quad 798.8 \pm 100.8, \quad \text{that is} \quad [688.0;\ 899.6].$$

Interpretation: with probability 95%, the average cub weight is between 688.0g and 899.6g.

10.2. Confidence intervals for the mean: σ^2 unknown

The fact that σ^2 is known is not a realistic assumption. We replace σ with its estimator S. *Suppose that X has a normal distribution.* Using techniques which are beyond the level of this course, one can prove that: (see Theorem 9.4 in the book)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ has a } T \text{ distribution with } n - 1 \text{ degrees of freedom (df).}$$

Table 17.4 (or the back inner cover of the book) gives some values for the T distribution with the number of degrees of freedom between 1 and 30.

Using the same technique as in the case when σ^2 is known, we conclude that a 95% confidence interval for μ is

$$\bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right) \tag{1}$$

where t is the value that we read in Table 17.4 at level $\nu = n - 1$, such that $P(-t \leq T \leq t) = 0.95$, i.e $P(T \leq t) = 0.975$.

The value $s\{\bar{X}\} = s/\sqrt{n}$ is called **the estimated standard error of the mean**.

Example 2. Let X be the amount of butterfat in pounds produced by a cow during a 305-day milk production period during her first and second calves. Suppose that X has a normal distribution. The data for a sample of $n = 20$ cows is:

481 537 513 583 453 510 570 500 457 555

618 327 350 643 499 421 505 637 599 392

We want to find a 90% confidence interval for the average amount μ of butterfat.

From the data, we get the estimate $\bar{x} = 507.5$ for the mean μ, the estimate $s = 89.75$ for the standard deviation σ, and the estimate $s^2 = 8055.0625$ for the variance σ^2. A 90% confidence interval for μ is based on the T distribution with $20 - 1 = 19$ degrees of freedom. For this level of confidence, the probability at the right of the point t is 0.05 (half of $1 - 0.90 = 0.10$). Table 17.4 gives the value $t = 1.729$. Therefore, the 90% confidence interval for μ is

$$507.5 \pm 1.729 \left(\frac{89.75}{\sqrt{20}} \right), \text{ which is } 507.5 \pm 34.7, \text{ or equivalently } [472.8;\ 542.2]$$

In this example, the estimated standard error of the mean is:

$$s\{\bar{X}\} = \frac{89.75}{\sqrt{20}} = 20.07$$

Remark: If the sample size is large (i.e. $n > 30$), one may replace in (1) the value t by the value z, corresponding to the same confidence level. The formula for the confidence interval becomes:

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$

This is called a *large sample interval* and is justified by Theorem 9.5 in the book, which says that when n is large, the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ is approximately standard normal. This is why for $n > 30$, the values t given in Table 17.4 coincide with the values z corresponding to the standard normal distribution.

10.3. Confidence Intervals for the Proportion

In this section we introduce some statistical methods that can be applied for drawing conclusions about the proportion p of individuals with a certain characteristic in a given population.

Examples: proportion of voters favorable to a certain political candidate, proportion of people affected by diabetes, proportion of students who are in a favor of a multiple-choice final exam, proportion of brain cancer cases, proportion of left-handed individuals

If we have a sample of size n from the population, and we denote by Y the total number of individuals (in the sample) who possess the desired characteristic, then a *point estimator* for p is: $\hat{p} = Y/n$.

Example 3. The last 30 years have seen a rise in childhood obesity in Canada. It was found that in a group of 975 randomly selected children aged 2-13, 78 are obese. Based on this data, the estimate for the proportion of obese children in Canada is:

$$\hat{p} = \frac{78}{975} = 0.08 \quad (\text{or} \quad 8\%)$$

To construct a confidence interval for p, we will use the following fact:

$$\text{if the sample size } n \text{ is large, then the distribution of } \hat{p} = \frac{Y}{n} \text{ is approximately normal}$$

$$\text{with mean } p \text{ and variance } p(1-p)/n$$

(This is a consequence of the Central Limit Theorem.) Hence, by standardization,

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \text{ is approximately a standard normal random variable}$$

Using the same technique as for μ, we conclude that a 95% confidence interval for p is:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For other confidence levels, 1.96 has to be replaced by a different value from Table 17.3.

Example 3. (continued) A 90% confidence interval for the proportion p of obese children in Canada is:

$$0.08 \pm 1.645\sqrt{\frac{(0.08)\cdot(0.92)}{975}}, \quad \text{or} \quad 0.08 \pm 0.0143, \quad \text{or} \quad [0.0657;\ 0.0943]$$

Interpretation: with a probability of 90%, the proportion of obese children in Canada is between 6.57% and 9.43%.