Confidence Intervals for Means - Unknown Variance

Gene Quinn

Unknown Variance

We have seen that when $\overline{x}$ is the mean of a random sample $x = \{x_1, x_2, \ldots, x_n\}$ from a normal population $N(\mu, \sigma^2)$ with **known** variance σ^2, the quantity

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution, $N(0, 1)$.

Unknown Variance

We have seen that when $\overline{x}$ is the mean of a random sample $x = \{x_1, x_2, \ldots, x_n\}$ from a normal population $N(\mu, \sigma^2)$ with **known** variance σ^2, the quantity

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution, $N(0, 1)$.

Based on this fact, a $100(1-\alpha)\%$ confidence interval for the parameter μ is:

$$\left[\overline{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \overline{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right]$$

Unknown Variance

In practice, the population variance σ^2 is seldom known, and usually must be estimated from the sample.

Unknown Variance

In practice, the population variance σ^2 is seldom known, and usually must be estimated from the sample.

We will see that the following is an unbiased estimate of the population variance σ^2:

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

Unknown Variance

In practice, the population variance σ^2 is seldom known, and usually must be estimated from the sample.

We will see that the following is an unbiased estimate of the population variance σ^2:

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

The statistic S^2 is called the **sample variance**.

Unknown Variance

In practice, the population variance σ^2 is seldom known, and usually must be estimated from the sample.

We will see that the following is an unbiased estimate of the population variance σ^2:

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

The statistic S^2 is called the **sample variance**.

The square root of S^2 is denoted by S and called the **sample standard deviation**

Unknown Variance

Just after the turn of the twentieth century, an employee of the Guiness brewery in Dublin named William S. Gosset derived the distribution of the statistic

$$t = \frac{\overline{x} - \mu}{S/\sqrt{n}}$$

for a random sample from a normal population.

Unknown Variance

Just after the turn of the twentieth century, an employee of the Guiness brewery in Dublin named William S. Gosset derived the distribution of the statistic

$$t = \frac{\overline{x} - \mu}{S/\sqrt{n}}$$

for a random sample from a normal population.

In so doing, Gosset proved that unlike its counterpart

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}},$$

t does **not** have a normal distribution.

Unknown Variance

Forbidden by his employer to publish research, Gosset published the result under the psuedonym "Student" and the result became known as the "Students-t" distribution.

Unknown Variance

Forbidden by his employer to publish research, Gosset published the result under the psuedonym "Student" and the result became known as the "Students-t" distribution.

The t distribution is actually a family of distributions with a parameter called the *degrees of freedom* that is one less than the sample size n.

Unknown Variance

Forbidden by his employer to publish research, Gosset published the result under the psuedonym "Student" and the result became known as the "Students-t" distribution.

The t distribution is actually a family of distributions with a parameter called the *degrees of freedom* that is one less than the sample size n.

Like the normal distribution, the t distribution is bell-shaped and centered at zero.

Unknown Variance

Forbidden by his employer to publish research, Gosset published the result under the psuedonym "Student" and the result became known as the "Students-t" distribution.

The t distribution is actually a family of distributions with a parameter called the *degrees of freedom* that is one less than the sample size n.

Like the normal distribution, the t distribution is bell-shaped and centered at zero.

As the sample size n and degrees of freedom $n - 1$ become large, the t curve becomes indistinguishable from the normal curve.

Unknown Variance

Confidence intervals for the population mean μ are obtained from the t distribution in a manner similar to those used when the variance σ^2 is known, except:

- σ is replaced by S
- $z_{\alpha/2}$ is replaced by $t_{\alpha/2,n-1}$

Unknown Variance

Confidence intervals for the population mean μ are obtained from the t distribution in a manner similar to those used when the variance σ^2 is known, except:

- σ is replaced by S
- $z_{\alpha/2}$ is replaced by $t_{\alpha/2,n-1}$

Values of $t_{\alpha/2,n-1}$ can be obtained from a table of the Student-t distribution or from a spreadsheet using

$$t_{\alpha/2,n-1} = \quad \text{=TINV}(\alpha, n-1)$$

where n is the sample size and α is the desired alpha level.

Unknown Variance

Theorem: If

$$x_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n$$

with μ and σ^2 unknown, then $100(1-\alpha)\%$ of the time the interval

$$\left[\overline{x} - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}}, \quad \overline{x} + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}}\right]$$

will contain the (unknown) population mean μ.

Unknown Variance

Example: A sample of 12 fifth-grade students found an average weight $\overline{x}$ of 83.4 pounds with a sample variance of 17.0.

Unknown Variance

Example: A sample of 12 fifth-grade students found an average weight $\overline{x}$ of 83.4 pounds with a sample variance of 17.0.

In this case $S^2 = 17.0$ and $t_{.025,11} = 2.20$ so the 95% confidence interval for μ is

$$\left[\overline{x} - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}}, \quad \overline{x} + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}}\right]$$

$$= \left[83.4 - 2.20\sqrt{\frac{17}{12}}, \quad 83.4 + 2.20\sqrt{\frac{17}{12}}\right] = [80.7, \ 86.0]$$