BEREA COLLEGE

1 Estimation Theory: Some Additional LECTURE Notes

****

****

****

****

As we have mentioned earlier, the 2 main areas of statistical estimations are *point estimation* and *interval estimation*. In point estimation, we are concerned with estimating a single value of a parameter or parameters. But, announcing a single value as one's estimate of a parameter θ calls for some auxiliary indication of how much that value can be trusted- a measure of reliability. The notion of standard error of an estimator is a commonly used measure , and sometimes estimates are given given as a point estimate (single value) plus or minus a standard error. The standard error is not a maximum error and so one could not expect the real value of the parameter to be within one standard error of the point estimate in every instance. Sometimes it will, and sometimes it won't.

Given a sample, a confidence interval provides a range of values that estimates an unknown population parameter, such as the mean, proportion, or variance.

Confidence intervals : provide a range of values that contain the unknown parameter with a certain degree of confidence.

- The **length** of a confidence interval conveys information about the precision of the estimate .
- NARROW confidence intervals suggest precise answers.
- WIDE confidence intervals indicate that we know little about the population.

We consider several cases of confidence interval estimation.

1.1 Confidence Interval for Means

Suppose we are willing to accept as a fact the numerical outcome X of a random experiment is a random variable that has a normal distribution with known variance σ^2 but unknown mean μ. That is, μ is some constant, but its value is unknown. To elicit some information about the parameter μ, we repeat the random experiment under identical conditions n independent times, n being a fixed positive integer, and obtain a random sample $X_1, X_2, \cdots, X_n$, so that $X_1, X_2, \cdots, X_n$ denote the outcomes obtained on these n repetitions of the experiments.

Consider a point estimate of μ, such as the m.l.e. $\hat{\mu} = \bar{X}$.

We want to investigate the closeness of $\bar{X}$ to the unknown parameter μ. Since we are sampling from a normal population, we know that $\bar{X} \sim N(\mu, \sigma^2/n)$.

Case 1: Confidence Interval for μ with KNOWN σ

To construct a confidence interval for the unknown parameter μ when the variance σ^2 is known. For illustration, consider the following situation

$$\mathbb{P}\left(-2 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 2\right) = 0.954$$

Notice that the following events

$$-2 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 2$$

$$\frac{-2\sigma}{\sqrt{n}} < \bar{X} - \mu < \frac{2\sigma}{\sqrt{n}}$$

and

$$\bar{X} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{2\sigma}{\sqrt{n}}$$

are equivalent. Thus these events have the same probability.That is,

$$\mathbb{P}\left(\bar{X} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.954$$

Since σ is a known number, each of the random variables $\bar{X} - \frac{2\sigma}{\sqrt{n}}$ and $\bar{X} + \frac{2\sigma}{\sqrt{n}}$ is a statistic. The interval

$$\left(\bar{X} - \frac{2\sigma}{\sqrt{n}}, \bar{X} + \frac{2\sigma}{\sqrt{n}}\right)$$

is a random interval. The last probability statement can be interpreted as follows: Prior to the repeated independent performances of the random experiment, the probability is 0.954 that the random interval $\left(\bar{X} - \frac{2\sigma}{\sqrt{n}}, \bar{X} + \frac{2\sigma}{\sqrt{n}}\right)$ includes the the unknown fixed parameter μ.

Suppose the experiment yields $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$. Then the sample value of $\bar{X}$ is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

a known number.We cannot say that 0.954 is the probability that the particular interval $\left(\bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}\right)$ includes the parameter μ, for μ, although unknown, is some constant, and this particular interval either does or does not include μ. But, we can say that the known interval $\left(\bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}\right)$ is a *95.4% confidence interval for* μ. The confidence coefficient is equal to the probability that the random interval includes the parameter.

We can obtain an 80, a 90, or a 99 percent confidence interval for μ or more generally a $(1-\alpha)\%$ confidence interval for μ as follows.

Write

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

For example, if $\alpha = 0.05$, we have $z_{\alpha/2} = z_{0.025} = 1.96$, and if $\alpha = 0.10$, then $z_{\alpha/2} = z_{0.05} = 1.645$. The following events are equivalent

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

$$-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \bar{X} - \mu \leq z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Because the events are equivalent, we have

$$\mathbb{P}\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1 - \alpha$$

So *the probability that the random interval*

$$\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

includes the unknown mean μ is $1 - \alpha$.

Once the sample is observed and the sample mean computed to equal $\bar{x}$, the interval

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

becomes known. Since the probability that the random interval covers μ before the sample is drawn is equal to $1-\alpha$, we call the computed interval $\left[\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$ a $100(1-\alpha)\%$ **confidence interval** for the unknown parameter μ. The term $1 - \alpha$ is called the **confidence coefficient**.

Notice that as n increases, $z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ decreases, resulting in a shorter confidence interval with the same confidence coefficient $1 - \alpha$.

- *A shorter confidence interval indicates that we have more credence in $\bar{x}$ as an estimate of μ.*

For a fixed sample size, the length of the confidence interval can also be shortened by decreasing the confidence coefficient $1 - \alpha$. But if this is done, we achieve a shorter confidence interval at the expense of losing some confidence.

Example 1

Let X be the length of life of a 60-watt light bulb marketed by a certain manufacturer. Assume that the distribution of X is $N(\mu, 1296)$. If a random sample of size n=27 bulbs is tested until they burn out, yielding a sample mean of $\bar{x} = 1478$ hours, find a 95% confidence interval for μ.

Solution

A 95% confidence interval for μ is given by

$$\begin{aligned}\left[\bar{x} - z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right)\right] &= \left[1478 - 1.96\left(\frac{36}{\sqrt{27}}\right), 1478 + 1.96\left(\frac{36}{\sqrt{27}}\right)\right] \\ &= [1478 - 13.58, 1478 + 13.58] \\ &= [1464.42, 1491.58]\end{aligned}$$

CASE 2: Confidence Interval for Means with σ UNKNOWN

We now turn to the problem of finding a confidence interval for the mean μ of a normal distribution when we do not know the variance σ^2.

Since we are sampling from a normal population, we know that the random variable

$$T = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{(n-1)S^2/[\sigma^2(n-1)]}} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution with n-1 degrees of freedom, whatever the value of $\sigma^2 > 0$, where S^2 is the usual unbiased estimator of σ^2. Select $t_{\alpha/2}(n-1)$ so that

$$\mathbb{P}[T \geq t_{\alpha/2}(n-1)] = \alpha/2$$

Then

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\left[-t_{\alpha/2}(n-1) \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right] \\
&= \mathbb{P}\left[-t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right) \leq \bar{X} - \mu \leq t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right] \\
&= \mathbb{P}\left[\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right) \leq -\mu \leq \bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right] \\
&= \mathbb{P}\left[-\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right) \leq -\mu \leq -\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right] \\
&= \mathbb{P}\left[\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right]
\end{aligned}$$

Thus, the observations of a random sample provide $\bar{x}$ and s^2, and

$$\left[\bar{x} - t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right)\right]$$

is a $100(1-\alpha)\%$ confidence interval for μ.

****

Summary

Let's summarize the findings about a confidence interval for the mean of a normal population.

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a normal population $N(\mu, \sigma^2)$

• If σ^2 is KNOWN, then the random variable $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is N(0,1) and

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

is a $100(1-\alpha)\%$ **confidence interval** for the unknown parameter μ. The term $1-\alpha$ is called the **confidence coefficient**.

• • If σ^2 is UNKNOWN , then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with n-1 degrees of freedom. and

$$\left[\bar{x} - t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right)\right]$$

is a $100(1-\alpha)\%$ confidence interval for μ.

****

Example 1

Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her firt and second calves. Assume that the distribution of X is $N(\mu, \sigma^2)$. To estimate μ , a farmer measures the butterfat production for n=20 cows and obtained the following data

$$481, 537, 513, 583, 453, 510, 570, 500, 457, 555$$

$$618, 327, 350, 643, 499, 421, 505, 637, 599, 392$$

For these data, $\bar{x} = 507.50$, and s=89.75. Therefore a point estimate for μ is $\bar{x} = 507.50$.

Since $t_{0.05}(19) = 1.729$, a 90% confidence interval for μ is

$$507.50 \pm 1.729\left(\frac{89.75}{\sqrt{20}}\right) = 507.50 \pm 34.70 = [472.80, 542.20]\square$$

If we are not able to assume that the underlying population distribution is normal, but μ and σ are both unknown, **approximate confidence intervals for** μ can still be constructed using the random test statistic

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

which now only has an approximate t-distribution. Generally, this approximation is quite good for many nonnormal distributions; in particular, it works well if the underlying distribution is SYMMETRIC, UNIMODAL, and of the CONTINUOUS TYPE. However, if the distribution is highly skewed , there is a great danger in using that approximation. In such a situation, it would be safer to use certain **nonparametric methods** for finding a confidence interval for the median of the distribution.

There is one other aspect of confidence intervals that should be mentioned. So far, we have created what are called **two-sided confidence intervals** for the men μ. Sometimes, however, we might want only a LOWER (or UPPER) bound for μ.

Again, we assume sampling from a normal population with distribution $N(]mu, \sigma^2)$ and $\bar{X}$ is the sample mean.

• If σ^2 is KNOWN , then

$$\mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq z_\alpha\right) = \alpha$$

or equivalently

$$\mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_\alpha\right) = 1-\alpha$$

or equivalently

$$\mathbb{P}\left[\bar{X} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu\right] = 1-\alpha$$

Once $\bar{X}$ is observed to be equal to $\bar{x}$, it follows that

$$\left[[\bar{x} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right), \infty\right)$$

is a $100(1-\alpha)\%$ **ONE-SIDED confidence interval for** μ.

That is , with the confidence coefficient $1-\alpha$, $\bar{x} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$ is a *lower bound* for μ with confidence coefficient (or confidence level) $1-\alpha$.

Similarly,

$$\left(-\infty, \bar{x} + z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

is an *Upper bound* for μ with confidence coefficient $1-\mu$.

• When σ^2 is UNKNOWN, we we use the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ to find the corresponding lower and upper bounds for μ, namely, $\bar{x} - t_\alpha(n-1)\left(\frac{s}{\sqrt{n}}\right)$ and

$\bar{x} + t_\alpha(n-1)\left(\frac{s}{\sqrt{n}}\right)$

EXERCISES

Exercise 1.1.1.

Ler the observed value of the mean $\bar{X}$ of a random sample of size 20 from a distribution that is $N(\mu, 80)$ be 81.2. Find a 95 percent confidence interval for $]mu$.

ANS: (77.28,85.12)

Exercise 1.1.2.

Let $\bar{X}$ be the mean of a random sample of size n from a distribution that is $N(\mu, 9)$. Find n such that

$$\mathbb{P}(\bar{X} - 1 < \mu < \bar{X} + 1) = 0.90$$

ANS: 24 or 25

Exercise 1.1.3.

Let a random sample of size 17 from the normal distribution $N(\mu, \sigma^2)$ yield $\bar{x} = 4.7$ and $s^2 = 5.76$. Determine a 90 percent confidence interval for μ.

ANS: (3.7, 5.7)

Exercise 1.1.4.

Let $\bar{X}$ be the mean of a random sample of size n from a distribution that has mean μ and variance $\sigma^2 = 10$. Find n so that the probability is approximately 0.954 that the random interval $(\bar{X} - \frac{1}{2}, \bar{x} + \frac{1}{2})$ includes μ

ANS: 160

Exercise 1.1.5.

Let $\bar{x}$ be the observed mean of a random sample of size n from a distribution having mean μ and known variance σ^2. Find n so that $\bar{x} - \sigma/4$ to $\bar{x} + \sigma/4$ is an approximate 95 percent confidence interval for μ.

Exercise 1.1.6.

To determine the effect of 100% nitrate on the growth of pea plants , several specimens were planted and then watered with 100% nitrate every day. At the end of teo weeks, the plants were measured. Here are data on seven of them

$$17.5, 14.5, 15.2, 14.0, 17.3, 18.0, 13.8$$

Assume that the data are observations from a normal distribution $N(\mu, \sigma^2)$.

a. Find the value of a point estimate of μ.

b. Find the value of a point estimate of σ.

c. Give the endpoints for a 90% confidence interval for μ.

Exercise 1.1.7.

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from the normal distribution $N(\mu, \sigma^2)$. Calculate the expected length of a 95% confidence interval for μ , assuming that n=5 and the variance is

a. known,

b. unknown.

Hint:
To find $\mathbb{E}(S)$, first determine $\mathbb{E}\left[\sqrt{(n-1)S^2/\sigma^2}\right]$, recalling that $\sqrt{(n-1)S^2/\sigma^2}$ is $\chi^2(n-1)$.

1.2 Confidence Intervals for Difference of Two Means

Suppose that we are interested in comparing the means of two normal distributions.

Let $X_1, X_2, \cdots, X_n$ and Let $Y_1, Y_2, \cdots, Y_m$ be, respectively, two independent random samples of sizes n and m from the 2 normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$.

Case 1: σ_X^2 and σ_Y^2 are KNOWN.

The random samples are independent; thus, the respective sample means $\bar{X}$ and $\bar{Y}$ are independent random variables and have distributions $N(\mu_X, \sigma_X^2/n)$ and $N(\mu_Y, \sigma_Y^2/m)$. Consequently, the random variable $W = \bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m)$. Hence,

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

After arrangement, we obtain

$$\mathbb{P}\left[(\bar{X} - \bar{Y}) - z_{\alpha/2}\sqrt{\sigma_X^2/n + \sigma_Y^2/m} \leq \mu_X - \mu_Y \leq (\bar{X} - \bar{Y}) + z_{\alpha/2}\sqrt{\sigma_X^2/n + \sigma_Y^2/m}\right] = 1 - \alpha$$

Once the experiments have been performed and the means $\bar{x}$ and $\bar{y}$ computed, the interval

$$\left[(\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{\sigma_X^2/n + \sigma_Y^2/m}, (\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{\sigma_X^2/n + \sigma_Y^2/m}\right]$$

provides a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$.

Example

Suppose that n=15, m=8, $\bar{x} = 70$, $\bar{y} = 75.3$, $\sigma_X^2 = 60$, $\sigma_Y^2 = 40$, and $1 - \alpha = 0.90$.
Hence, a 90% confidence interval for $\mu_X - \mu_Y$ is

$$70 - 75.3 \pm z_{0.05}\sqrt{60/15 + 40/8} = [-5.2 - 4.935, -5.2 + 4.935] = [-10.135, -0.265]$$

$\therefore$ Because the confidence interval does not include 0, we suspect that μ_Y is greater than μ_Y. $\square$

Case 2: Large Sample Confidence Interval: σ_X and σ_Y are UNKNOWN but with Large Samples

If the sample sizes are large enough and σ_X and σ_Y are unknown, we can replace σ_X^2 and σ_Y^2 by s_X^2 and s_Y^2 , where s_X^2 and s_Y^2 are the respective unbiased estimators of the variances. This means that

$$\left[(\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{s_X^2/n + s_Y^2/m}, (\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{s_X^2/n + s_Y^2/m}\right]$$

provides an approximate $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$.

Case 3: σ_X and σ_Y are UNKNOWN but Small Samples

We now consider the problem of constructing confidence intervals for the difference of the means of two normal distributions when the variances are unknown but the sample sizes are small.

Let $X_1, X_2, \cdots, X_n$ and $Y_1, Y_2, \cdots, Y_m$ be two independent random samples from two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively. If the sample sizes are not large(say, considerably smaller than 30), this problem can be a difficult one.However, if we can assume common, but unknown, variances, say , $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, we can solve the problem.

We know that

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim N(0, 1)$$

Moreover, since the random samples are independent, the random variable

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2}$$

is the sum of 2 independent chi-square random variables, with n-1 and m-1 degrees of freedom respectively. Hence,

$$U \sim \chi^2(n + m - 2)$$

According to the definition of the Student random variable,

$$T = \frac{Z}{\sqrt{U/(n+m-2)}}$$

has a t-distribution with n+m-2 degrees of freedom. That is,

$$\begin{aligned} T &= \frac{\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\sigma^2/n+\sigma^2/m}}}{\sqrt{\left[\frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2}\right]/(n+m-2)}} \\ &= \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\left[\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\right]\left[\frac{1}{n} + \frac{1}{m}\right]}} \end{aligned}$$

has a t-distribution with r=n+m-2 degrees of freedom. Thus, for a $100(1-\alpha)\%$ confident level , we have

$$\mathbb{P}(-t_{\alpha/2}(n+m-2) \leq T \leq t_{\alpha/2}(n+m-2)) = 1 - \alpha$$

Solving for $\mu_X - \mu_Y$ yields

$$\mathbb{P}\left(\bar{X} - \bar{Y} - t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \leq \mu_X - \mu_Y \leq \bar{X} - \bar{Y} + t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right) = 1-\alpha$$

where the **pooled estimator** of the common standard deviation is

$$S_p = \sqrt{\left[\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\right]}$$

If $\bar{x}$, $\bar{y}$, and s_p are the observed values of $\bar{X}$, $\bar{Y}$, and S_p, then a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\left[\bar{x} - \bar{y} - t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n}+\frac{1}{m}}, \bar{x} - \bar{y} + t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right]$$

Remark

The assumption of equal variances, namely, $\sigma_X^2 = \sigma_Y^2$, can be modified somewhat so that we are still able to find a confidence interval for $\mu_X - \mu_Y$. That is, if we know the ratio $\frac{\sigma_X^2}{\sigma_Y^2}$ of the variances, we can still make this type of statistical inference by using a random variable with a t-distribution.However, if we do not know the ratio of the variances and yet suspect that the unknown σ_X^2 and σ_Y^2 differ by a great deal , it is safest to return to

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}$$

for the inference about $\mu_X - \mu_Y$. Just replace the population variances by their sample variances and consider

$$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}}$$

Q: What is the distribution of W?

As before, if if n and m are large enough and the underlying distributions are close to normal, then W has approximately a normal distribution and confidence interval for $\mu_X - \mu_Y$ can be found by setting

$$\mathbb{P}(-z_{\alpha/2} \leq W \leq z_{\alpha/2}) \approx 1 - \alpha$$

For small n and m, it is common to use Welch's approximation (B.L. Welch) which we do not discussed here.

Exercise 1

Let $\bar{X}$ and $\bar{Y}$ be the means of two independent random samples, each of size n, from the respective distributions $N(\mu_1, \sigma^2)$, and $N(\mu_2, \sigma^2)$, where the common variance is known. Find n such that

$$\mathbb{P}(\bar{X} - \bar{Y} - \sigma/5 < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + \sigma/5) = 0.90$$

ANS: 135 or 136

Exercise 2

Let two independent random samples, each of size 10 , from two normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ yield $\bar{x} = 4.8$, $s_1^2 = 8.64$, $\bar{y} = 5.6$, $s_2^2 = 7.88$. Find a 95 percent confidence interval for $\mu_1 - \mu_2$.

ANS: [-3.6, 2.0]

1.3 Confidence Interval for Variances

In this section, we want to find confidence intervals for he variance of a normal distribution and for the ratio of two normal distributions.

Consider a random sample $X_1, X_2, \cdots, X_n$ from a normal population with a normal $N(\mu, \sigma)$ distribution.

The confidence interval for the variance σ^2 is based on the sample variance

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

Since we are sampling from a normal population, we know that $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$. We can select constants a and b from a chi-square with n-1 degrees of freedom such that

$$\mathbb{P}\left(a \leq (n-1)S^2/\sigma^2 \leq b\right) = 1 - \alpha$$

One way to do this is by selecting a and b so that $a = \chi^2_{1-\alpha/2}(n-1)$ and $b = \chi^2_{\alpha/2}(n-1)$. In order words, we can select a and b so that the probabilities in the 2 tails are equal. We have

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\left(\frac{a}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{b}{(n-1)S^2}\right) \\ &= \mathbb{P}\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) \end{aligned}$$

Thus, the probability that the random interval

$$\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right]$$

contains the unknown variance σ^2 is $1 - \alpha$.

Once the values of $X_1, X_2, \cdots, X_n$ are observed to be $x_1, x_2, \cdots, x_n$ and s^2 computed, the interval

$$\left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2. Accordingly, a $100(1-\alpha)\%$ confidence interval for σ, the standard deviation, is

$$\left[\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right] = \left[\sqrt{\frac{n-1}{b}}s, \sqrt{\frac{n-1}{a}}s\right]$$

Example

Assume that the time in days required for maturation of seeds of a species of Guardiola, a flowering plant found in Mexico , is $N(\mu, \sigma^2)$. A random sample of size n= 13 seeds , both parents having narrow leaves, yielded $\bar{x} = 18.97$ days and

$$12s^2 = \sum_{i=1}^{13}(x_i - \bar{x})^2 = 128.41$$

Find a 90% confidence interval for σ^2.

Solution

Because $\chi^2_{0.95}(12) = 5.226$ and $\chi^2_{0.05}(12) = 21.036$, a 90% confidence interval for σ^2 is

$$\left[\frac{128.41}{21.03}, \frac{128.41}{5.226}\right] = [6.11, 24.57]$$

The corresponding 90% confidence interval for the standard deviation is

$$[\sqrt{6.11}, \sqrt{24.57}] = [2.47, 4.96]\square$$

There are occasions when it is of interest to *compare the variances of two normal distributions.*The procedure is to find a confidence interval for the ratio of variances $\frac{\sigma_X^2}{\sigma_Y^2}$, using the ratio $\frac{S_X^2}{\sigma_X^2}$ to $\frac{S_Y^2}{\sigma_Y^2}$, where S_X^2 and S_Y^2 are the two unbiased sample variances based on two independent samples of sizes n and m from two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively.

Now , write

$$\frac{\frac{S_Y^2}{\sigma_Y^2}}{\frac{S_X^2}{\sigma_X^2}} = \frac{\left[\frac{(m-1)S_Y^2}{\sigma_Y^2}\right]/(m-1)}{\left[\frac{(n-1)S_X^2}{\sigma_X^2}\right]/(n-1)}$$

Now, since $\frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(m-1)$ and $\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1)$ and that they are independent random variables, it follows that the above equation is the ratio of two independent chi-square random variables divided by their corresponding degrees of freedom, and therefore we obtain an F(m-1,n-1) distribution. That is, focusing on the left hand side of the above result, we obtain

$$F = \frac{\frac{S_Y^2}{\sigma_Y^2}}{\frac{S_X^2}{\sigma_X^2}}$$

which has an F distribution with m-1, n-1 degrees of freedom.

To form the desired confidence interval, we select constants c and d from the F-table or software so that

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(c \le \frac{\frac{S_Y^2}{\sigma_Y^2}}{\frac{S_X^2}{\sigma_X^2}} \le d\right) \\ &= \mathbb{P}\left(c\frac{S_X^2}{S_Y^2} \le \frac{\sigma_X^2}{\sigma_Y^2} \le d\frac{S_X^2}{S_Y^2}\right). \end{aligned}$$

c and d are usually chosen by setting

$$c = F_{1-\alpha/2}(m-1, n-1) = 1/F_{\alpha/2}(n-1, m-1)$$

and

$$d = F_{\alpha/2}(m-1, n-1)$$

If s_X^2 and s_Y^2 are the observed values of S_X^2 and S_Y^2, respectively, then a $100(1-\alpha)\%$ confidence interval for the ratio $\frac{\sigma_X^2}{\sigma_Y^2}$ is given by

$$\left[\frac{1}{F_{\alpha/2}(n-1, m-1)}\frac{s_X^2}{s_Y^2}, F_{\alpha/2}(m-1, n-1)\frac{s_X^2}{s_Y^2}\right]$$

Accordingly, we can find a $100(1-\alpha)\%$ confidence interval for the ratio $\frac{\sigma_X}{\sigma_Y}$ by taking the square root of the endpoints in the above formula.

EXERCISES

Exercise #1

Let X equal the length (in centimeters) of a certain species of fish caught in the springtime. A random sample of n=13 observations of X is

$$13.1, 5.1, 18.0, 8.7, 16.5, 9.8, 6.8$$

$$12.0, 17.0, 25.4, 19.2, 15.8, 23.0$$

a. Give a point estimate of the standard deviation σ of this species of fish.

b. Find a 95% confidence interval for σ.

ANS: a. s=6.144; b. [4.406,10.142] or [4.107,9.521]

Exercise #2

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from $N(\mu, \sigma^2)$ with known μ. Describe how you would construct a confidence interval for the unknown variance σ^2.

HINT: Use the fact that $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.

ANS:

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{\alpha/2}(n)}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right]$$

Exercise #3

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from an exponential distribution with unknown mean θ.

a. Show that the distribution of the random variable $W = (2/\theta)\sum_{i=1}^{n} X_i$ is $\chi^2(2n)$.

b. Use W to construct a $100(1-\alpha)\%$ confidence interval for θ.

c. If n=7 and $\bar{x} = 93.6$, give the endpoints for a 90% confidence interval for the mean θ.

1.4 Confidence Intervals for Proportions

The histogram is a good description of how the observations of a random sample are distributed. We shall now consider the problem of determining a confidence interval for the unknown parameter p of a binomial distribution with parameter p.

Let Y be the number of successes in these n Bernoulli trials. Since $\mathbb{E}(Y) = np$, we see that $\mathbb{E}(Y/n) = p$ and therefore the random variable Y/n is an unbiased point estimator of p.

In general when observing n Bernoulli trials with probability p of success on each trial, we shall find confidence interval for the parameter p based on the random variable Y/n, where Y is the number of successes in n independent Bernoulli trials.

Provided that n is large enough, we know that the random variable

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{p(1-p)/n}}$$

has an approximate normal distribution N(0,1). Hence, for a given probability $1 - \alpha$, we can find $z_{\alpha/2}$ such that

$$\mathbb{P}\left[-z_{\alpha/2} \leq \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

Solving for p gives

$$\mathbb{P}\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] \approx 1 - \alpha$$

Unfortunately, the unknown parameter p appears in the endpoints of the inequalities. There are two possible ways out.

First, we could make an additional approximation, namely replacing p by its estimator Y/n in the expression $p(1-p)/n$ in the endpoints. That is, if the sample size n is large enough, it is still true that

$$\mathbb{P}\left[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}} \leq p \leq \frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}}\right] \approx 1 - \alpha$$

Thus, for large n, if the observed value of the random variable Y is y, then a $100(1-\alpha)\%$ confidence interval for p is provided by

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}, \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}\right]$$

We omit the second method of solving for p in the endpoints (Ref. HT, p.308-309)

Example

In a certain political campaign, one candidate has a poll taken at random among the voting population. The result are y=185 out n=351 voters favor this candidate. Even though y/n=185/351=0.527, or 52.7 support, should the candidate feel very confident of winning?

Solution

Compute an approximate 95% confidence interval for the fraction p of the voting population who favor the candidate.

The confidence interval is given by

$$0.527 \pm 1.96\sqrt{\frac{(0.527)(0.473)}{351}} = [0.475, 0.579]$$

From this confidence interval, there is a good possibility that p is less than 50%, and the candidate should certainly take this possibility into account in campaigning. □.

1.5 One-Sided Confidence Interval for Proportions

Sometimes, it is appropriate to use one-sided confidence interval for p. For example, we may be interested in an UPPER BOUND on the proportion of defectives in manufacturing some item. Alternatively, we may be interested in in a LOWER BOUND on the proportion of voters who favor a particular candidate.

An **upper bound** for p is given by the one-sided confidence interval

$$\left[0, \frac{y}{n} + z_\alpha \sqrt{\frac{(y/n)(1-y/n)}{n}}\right]$$

A **Lower bound** for p is given by the interval

$$\left[\frac{y}{n} - z_\alpha \sqrt{\frac{(y/n)(1-y/n)}{n}}, 1\right]$$

Remark

The fact that the variance of Y/n s a function of p caused us some difficulty in finding a confidence interval for p. Another way of handling the problem is to use the DELTA METHOD by trying to find a function u(Y/n) of Y/n whose variance is independent of p. We have seen previously that the function

$$u\left(\frac{Y}{n}\right) = \arcsin\sqrt{\frac{Y}{n}}$$

has an approximate normal distribution with mean $\sqrt{p}$ and variance

$$[u'(\mu)]^2\sigma^2/n = \left(\frac{1}{2\sqrt{p}}\frac{1}{\sqrt{1-p}}\right)^2 \frac{p(1-p)}{n} = \frac{1}{4n}$$

Hence we could find , for example, an approximate 95.4% confidence interval for p by using

$$\mathbb{P}\left(-2 < \frac{\arcsin\sqrt{Y/n} - \arcsin\sqrt{p}}{\sqrt{1/4n}} < 2\right) = 0.954$$

and solving the inequalities for p.

Example

Suppose that we sample from a distribution with unknown mean μ and variance σ^2=225. Find a sample size n so that $\bar{x} \pm 1$ serves as a 95 percent confidence interval for μ.

Solution

Using the fact that the sample of the observations,$\bar{X}$, is approximately normal, the interval $\bar{x} \pm 1.96(15/\sqrt{n})$ will serve as an approximate confidence interval for μ. That is, we want

$$1.96\left(\frac{15}{\sqrt{n}}\right) = 1$$

That is

$$\sqrt{n} = 29.4, \text{and hence n} \approx 864.36$$

or n=855 because n must be an integer.

Common Confusions in Interpreting Confidence Intervals

KEY POINT: **A confidence level gives a range of values for a population parameter with a certain level of confidence**

To avoid some common mistakes in interpreting confidence intervals. Confidence intervals describe the population parameter , not the the data in a sample

Take a sample of 140 customer balances at a bank .

1. *" 99% of all customers keep a balance of* $\$1,520$ *to* $\$2,460$ *."*

This statement is not accurate because the C.I. gives a range of values for the population μ, not the balance of an individual.

2. *" The mean balance of 95 % of samples of this size will fall between* $\$1,420$ *and* $\$2,460$*."*

The confidence interval describes μ, an unknown constant, not the means of other samples.

3. *" The mean balance* μ *is* $\$1,420$ *and* $\$2,460$*."*

Closer, but still incorrect. The average balance in the population does not have to fall between $\$1,420$ and $\$2,460$. This is a 95% confidence interval. It might not contain μ.

Here is the right way to interpret the confidence interval:

" I am 95% confident that the mean monthly balance for the population of customers who accept an application lies between $\$1,420$ *and* $\$2,460$*."*

1.6 A General Large Sample Confidence interval

Suppose that $\hat{\theta}$ is an estimator of a population parameter θ satisfying the following properties

1. It has approximately a normal distribution;

2. it is (at least approximately)n unbiased; and

3. an expression for $\sigma_{\hat{\theta}}$, the standard error of $\hat{\theta}$, is available. Hence we have

$$\mathbb{P}\left[-z_{\alpha/2} < \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right] \approx 1-\alpha$$

• Assume that $\sigma_{\hat{\theta}}$ does not involve any unknown parameters. The an approximate $100(1-\alpha)\%$ confidence interval for $\hat{\theta}$ is given by

$$\hat{\theta} \pm z_{\alpha/2}.\sigma_{\hat{\theta}}$$

• Now suppose that $\sigma_{\hat{\theta}}$ does not involve θ but does involve at least one other unknown parameter. Let $s_{\hat{\theta}}$ be the estimate of $\sigma_{\hat{\theta}}$ obtained by using estimates in place of the unknown parameters . Under certain conditions (essentially that $s_{\hat{\theta}}$ is close to $\sigma_{\hat{\theta}}$ for most samples) , a valid confidence

interval is $\hat{\theta} \pm z_{\alpha/2}.s_{\hat{\theta}}$

• Finally, assume that $\sigma_{\hat{\theta}}$ does involve the unknown parameter θ, as in the case of a population proportion. An APPROXIMATE SOLUTION can be obtained by replacing θ in $\sigma_{\hat{\theta}}$ by its estimate $\hat{\theta}$. This provides an estimated standard error $s_{\hat{\theta}}$.

In sum, this confidence interval is of the form

point estimate of $\theta\pm$ (z critical value)(estimated standard error of the estimator)

Let's apply this strategy to the problem of finding a confidence interval for a *population Proportion.*

1.7 Confidence Interval for a Population Proportion Revisited

Let p denote the the proportion of " successes " in a population, where "success" identifies an individual or object that has a specified property. Select a random sample of size n and let X be the number of successes in the sample. Provided that n is small compared to the population size, X can be regarded as a binomial random variable with mean $\mathbb{E}(X) = np$ and variance $\mathbb{V}ar(X) = np(1-p)$. Furthermore, if $np \geq 10$ and $nq \geq 10$, X has approximately a normal distribution.

An estimator for p is $\hat{p} = X/n$, the sample fraction of successes.We have

$\mathbb{E}(\hat{p}) = p$ and $\sigma_{\hat{p}} = \sqrt{p(1-p)}$. The standard error involves the unknown parameter p. We have

$$\mathbb{P}\left[-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right] \approx 1-\alpha$$

We have found previously an approximate or LARGE SAMPLE confidence interval by replacing p by $\hat{p}$ in the standard error formula for $\hat{p}$ given by

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right)$$

An improvement over the normal approximation CI was developed by **Edwin Bidwell Wilson (1927)**. The procedure consists by setting

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} = z_{\alpha/2} \Longrightarrow (p-\hat{p})^2 = z_{\alpha/2}\frac{p(1-p)}{n}$$

Solving this quadratic equation gives the solutions

$$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + \frac{z^2_{\alpha/2}}{4n^2}}}{1+z^2_{\alpha/2}/n}$$

where

$$\tilde{p} = \frac{\hat{p} + z^2_{\alpha/2}/2n}{1+z^2_{\alpha/2}/n}$$

We summarize these findings in the following proposition

****

Proposition(Score Confidence Interval)

Let $\tilde{p} = \frac{\hat{p}+z^2_{\alpha/2}/2n}{1+z^2_{\alpha/2}/n}$. Then a **confidence interval for a population p** with confidence level approximately $100(1-\alpha)\%$ is given by

$$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}\hat{q}/n + \frac{z^2_{\alpha/2}}{4n^2}}}{1+z^2_{\alpha/2}/n}$$

where

$\hat{q} = 1 - \hat{p}$.

This is often referred to as the **Score confidence interval for p**.

****

If the sample size n is very large, the dominant term in the second portion of the formula is $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$ and the score interval is approximately

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$$

1.8 Confidence Level,Precision, and Sample Size

Q: Why settle for a confidence level of 95% when a level of 99% is achievable?

A: Because the price paid for the higher confidence level is a WIDER interval. For example, for a 95% confidence interval for the mean of a normal population which extends $1.96\sigma/\sqrt{n}$ to each side of $\bar{x}$, the width of the interval is $2(\ 1.96\sigma/\sqrt{n}) = 3.92 \times \sigma/\sqrt{n}$. Similarly, the width of a 99% confidence interval for μ is

$2(\ 2.58\sigma/\sqrt{n}) = 5.16 \times \sigma/\sqrt{n}$. Hence, we have more confidence in the 99% interval precisely because it is wider.

$\therefore$ **The HIGHER the desired degree of confidence, the WIDER the resulting interval will be**.

If we consider the width of the confidence interval as a measure of its accuracy or precision, then the confidence level (or reliability) of the interval is inversely related to its precision. A highly reliable interval may be imprecise in that the endpoints of the interval may be far apart, whereas a precise interval may entail relatively low reliability. Thus it cannot be said that a 99% interval is to be preferred to a 95% interval; *the gains in reliability entails a loss in precision.*

Higher Confidence level 1-α $\Longrightarrow$ Wider Width of Confidence interval $\Longrightarrow$ Lower Precision

An appealing strategy is to specify both the desired confidence level AND interval width and then determine the necessary sample size.

In statistical consulting, the first question frequently asked is " How large should the sample size be to estimate a mean?"

One answer could be " Only one observation is needed, provided that the standard deviation of the distribution is zero." That is, if $\sigma = 0$, then the value of that one observation would necessarily equal the unknown mean of the distribution. This case is extreme and is not met in practice.

Key point: *The smaller the variance, the smaller is the sample size needed to achieve a given degree of accuracy.*

First, we consider the concept of margin of error.

Margin of Error

Recall.

• A statistic intended for estimating a population parameter is called a *point estimator* or simply an *estimator*

• The standard deviation of an estimator is called its *standard error: S.E.*

Without an assessment of accuracy, a single number quoted as an estimate may not serve a very useful purpose. We must indicate the extent of variability in the distribution of the estimator. The standard deviation, alternatively called *standard error* of the estimator, provides information about its variability. For example, for the population mean μ, we have the following

****

Point Estimate of the mean

• PARAMETER : Population mean μ.

• DATA : $X_1, X_2, \cdots, X_n$

• ESTIMATOR :$\bar{X}$, sample mean

•
$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}, \text{estimated S.E.}(\bar{X}) = \frac{S}{\sqrt{n}}$$

• For large n, the $100(1-\alpha)\%$ ERROR MARGIN is $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

If σ is unknown, we use S in place of σ.

• *Margin of Error:*
$$ME = \frac{2\sigma}{\sqrt{n}}$$

If σ is unknown, use and
$$\text{Margin of Error} = \frac{2s}{\sqrt{n}}$$

Caution

1. Standard error should not be interpreted as the "typical" error in a problem of estimation. For instance, if we obtain S.E. $(\bar{X})$=0.3,, we should not think that the error $|\bar{X} - 3|$ is likely to be 0.3, but rather, prior to observing the data, the probability is approximately .954 that the error will be within $\pm 2(S.E.) = \pm .6$.

2. An estimate and its variability are often reported in either of the forms : $ESTIMATE \pm S.E.$, or $ESTIMATE \pm 2(S.E.)$

Determining the Sample Size

How large a sample size do we need?.

Data collection costs time and money. Before collecting data, the investigator needs to know beforehand the sample size required to give the desired precision or to know whether the sample we can afford is adequate for what we want to learn.

In order to determine how large a sample is needed for estimating a population mean μ, we MUST SPECIFY

d= the desired error margin

and

$1 - \alpha =$ the probability associated with the error margin

Referring to the expression for a $100(1 - \alpha)\%$ error margin, we then set

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = d$$

Solving for n gives

$$n = \left[\frac{z_{\alpha/2}\sigma}{d}\right]^2$$

This determination of sample size is valid provided $n > 30$, so that the normal approximation is satisfactory.

****

To be $100(1-\alpha)\%$ **sure that the error of estimation** $|\bar{X}-\mu|$ **does not exceed d, the *required sample size* is**

$$n = \left[\frac{z_{\alpha/2}\sigma}{d}\right]^2$$

****

If σ is completely unknown , a small -scale preliminary sampling is necessary to obtain an estimate of σ to be used in the formula to compute n.

The expression for the desired margin of error , $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, is sometimes called the **maximum error of the estimate**

Example: *Determining a Sample Size for Collecting Water Samples*

A limnologist wishes to estimate the mean phosphate content er unit volume of lake water. It is known from studies in previous years that the standard deviation has fairly stable value $\sigma = 4$. How many water samples must the limnologist analyze to be 90% confident the error of estimation

does not exceed 0.8 milligrams?

Solution

We are given $\sigma = 4$, and $1 - \alpha = 0.90 \Longrightarrow \alpha/2 = 0.05$ with $z_{0.05} = 1.645$. We obtain

$$n = \left[\frac{1.645 \times 4}{0.8}\right]^2 = 67.65$$

Hence, the required sample size is n=68

The type of statistic we see most often in newspapers and magazines is an estimate of a proportion p. We might, for example, want to know the *percentage of the labor force that is unemployed* or the *percentage of voters favoring a certain candidate*. Sometimes, very important decisions are made on the basis on these estimates. Then we would desire *short confidence intervals* for p with *large confidence coefficients*. These conditions require a large sample size.

Example

Suppose we know that the unemployment rate has been about 8%. However, we wish to update our estimate in order to make an important decision about the national economic policy. Accordingly, let us say we wish to be 99% confident that the new estimate of p is within 0.001 of the true p. If we assume Bernoulli trials(an assumption that might be questioned), the relative frequency y/n , based upon a large sample size, provides the approximates 99% confidence interval

$$\frac{y}{n} \pm 2.576\sqrt{\frac{(y/n)(1 - y/n)}{n}}.$$

Although we do not know y/n exactly before sampling, since y/n will be near 0.08 we have

$$\frac{y}{n} \pm 2.576\sqrt{\frac{(y/n)(1 - y/n)}{n}} = 2.576\sqrt{\frac{(0.08)(1 - 0.08)}{n}}$$

This represents the margin of error. Now we can set

$$2.576\sqrt{\frac{(0.08)(1 - 0.08)}{n}} = 0.001 \Longrightarrow n \approx 488,394$$

Thus, under the assumptions, such a sample size is needed in order to achieve the reliability and accuracy desired. Because n is so large, we would probably be willing to increase the error , say, 0.01 and perhaps reduce the confidence level. to 98%. In such a case

$$\sqrt{n} = (2.326/0.010\sqrt{0.0736} \Longrightarrow n \approx 3,982$$

which is a more reasonable sample size.

In general, to find the required sample size to estimate p, we make use of the point estimate of p,$\hat{p} = y/n$ and then choose an approximate $100(1 - \alpha)\%$ confidence interval for p, which is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Now, suppose we want an estimate of p that is within d of the unknown p with $100(1-\alpha)\%$ confidence, where $d = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the MAXIMUM ERROR OF THE POINT ESTIMATE $\hat{p}$. Since $\hat{p}$ is unknown before the experiment is run, we CANNOT use the value of $\hat{p}$ in our determination of n. However, if it known that p is about equal to $p^\star$, then the necessary sample size n is the solution of the equation

$$d = z_{\alpha/2}\sqrt{\frac{p^\star(1-p^\star)}{n}}$$

Whose solution is

$$n = \frac{z^2_{\alpha/2}p^\star(1-p^\star)}{d^2} \leq \frac{z^2_{\alpha/2}}{4d^2}$$

Thus, if we want a $100(1-\alpha)\%$ confidence interval for p to be no longer than $y/n \pm d$, a solution for n that provides this protection is

$$n = \frac{z^2_{\alpha/2}}{4d^2} = \left[\frac{z_{\alpha/2}}{2d}\right]^2$$

Example

A possible gubernatorial candidate wants to assess initial support among the voters before making an announcement about her candidacy. If the fraction p of voters who are favorable, without any advance publicity, is around 0.15, the candidate will enter the race. From a poll of n voters selected at random, the candidate would like the estimate y/n to be within 0.03 of p. That is, the decision will be based on a 95 % confidence interval of the form $y/n \pm 0.03$. Since the candidate has no idea about the magnitude of p, a consulting statistician formulates the equation

$$n = \frac{(1.96)^2}{4(0.03)^2} = 1067.11$$

Thus, the sample size should be around 1068 to achieve the desired reliability and accuracy.

Suppose that 1068 voters around the state were selected at random and interviewed and y=214 express support for the candidate. Then $\hat{p} = 214/1068 = 0.20$ is a point estimate of p, and an approximate 95% confidence interval for p is

$$0.20 \pm 1.96\sqrt{(0.20)(0.80)/n} = [0.176, 0.224]\square$$

That is, we are 95% confident that p belongs to the interval[0.176,0.224] . On the basis of this sample, the candidate decided to run for office.

Suppose that you want to estimate the proportion p of a student body that favors a new policy.

Q: *How large should the sample size be*?

If p is close to 1,2 and you want to be 95% confident that the maximum error of the estimate is d=0.02, then

$$n = \frac{(1.96)^2}{4(.02)^2} = 2401$$

Such a sample size makes sense at large academic institutions. However, if you have a small student body , the entire enrollment can be much less than 2401. Here , a procedure is presented

that can be used to determine the sample size when the population is small relative to the desired sample size.

• Let N be the size of a given population. Assume that N_1 individuals in the population have a certain characteristic C (for example, favor a new policy).

• Let $p = N_1/N$, the proportion of the population with this characteristic. Then,

$1 - p = 1 - N_1/N.$

• Take a sample size n without replacement and let X be the number of observations from the sample with the characteristic C.

X is random variable having the *hypergeometric distribution*

$$\mu = n\left(\frac{N_1}{N}\right) = np$$

and variance

$$\sigma^2 = n\left(\frac{N_1}{N}\right)\left(1 - \frac{N_1}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right)$$

The mean and variance of X/n are given by

$$\mathbb{E}\left(\frac{X}{n}\right) = \frac{\mu}{n} = p$$

and

$$\mathbb{V}ar\left(\frac{X}{n}\right) = \frac{\sigma^2}{n^2} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$$

We can use the normal approximation to find a $100(1-\alpha)\%$ confidence interval for the proportion p:

$$\mathbb{P}\left[-z_{\alpha/2} \leq \frac{(X/n) - p}{\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}} \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

Solving the inequalities for p, we obtain

$$1 - \alpha \approx \mathbb{P}\left[\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \leq p \leq \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}\right]$$

Replacing p in the endpoints of the inequalities by the realized value $\hat{p} = x/n$, we obtain an approximate $100(1-\alpha)\%$ confidence interval for the parameter of interest, p :

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}$$

Observe that if N is large relative to n, then

$$\frac{N-n}{N-1} = \frac{1 - n/N}{1 - 1/N} \approx 1$$

Hence, in this case, the confidence interval for the binomial Bin(n,p) distribution is equivalent to that of the hypergeometric distribution.

Suppose now that we are interested in determining the sample size n that is required to have a $100(1-\alpha)\%$ confidence interval for p where the maximum error (the margin of error) of the estimate is d, where

$$d = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$$

Now, solve for n to obtain

$$\begin{aligned} n &= \frac{Nz_{\alpha/2}^2 p(1-p)}{(N-1)d^2 + z_{\alpha/2}^2 p(1-p)} \\ &= \frac{z_{\alpha/2}^2 p(1-p)/d^2}{\frac{N-1}{N} + \frac{z_{\alpha/2}^2 p(1-p)/d^2}{N}} \end{aligned}$$

Letting

$$m = z_{\alpha/2}^2 p(1-p)/d^2$$

the sample size is given by

$$n = \frac{m}{1 + \frac{m-1}{N}}$$

If we know nothing about p, then we set p=1/2 to determine m.

Example

Suppose that a college of N=3000 students is interested in assessing student support for a new form for teacher evaluation. To estimate the proportion p in favor of the new form, how large a sample is required so that the maximum error of the estimate of p is d=0.03 with 95 % confidence?

Solution

Assume that p is completely unknown. Use p=1/2 to obtain

$$m = \frac{(1.96)^2}{4(0.03)^2} = 1068$$

Thus the desired sample size is given by

$$n = \frac{m}{1 + \frac{m-1}{N}} = \frac{1068}{1 + 1067/3000} = 788$$

Exercises: Check your Understanding

Exercise #1

Consider a normal population distribution with the value of σ known.

a. What is the confidence level for the interval $\bar{x} \pm 2.81\sigma/\sqrt{n}$?

b. What is the confidence level for the interval $\bar{x} \pm 1.44\sigma/\sqrt{n}$?

Exercise #2

With a random sample of size n=81, someone proposes

$$(\bar{X} - 0.2S, \bar{X} + 0.2S)$$

to be a confidence interval fro μ. What then is the level of confidence?

ANS: .9282 or about 93%.

Exercise #3

A credit company randomly selected 50 contested items and recorded the dollar amount being contested. These contested items had a sample mean of $\bar{x} = 75.43$ dollars and s=24.73 dollars. Construct a 95% confidence interval for the mean amount contested, μ.

ANS: (68.58, 82.28)

EXERCISES

Exercise #1

A quality control engineer wanted to be 98% confident that the maximum error of the estimate of the mean length ,μ, of the left hinge on a vanity cover molded by a machine is o.25. A preliminary sample of size n=32 parts yielded a sample mean of $\bar{x} = 35.68$ and a standard deviation of s=1.723.

a. How large a sample is required ?

b. Does this seem to be a reasonable sample size?

ANS: a. 257 ; b. yes.

Exercise #2

For a public opinion poll for a close presidential election, let p denote the proportion of voters who favor candidate A. How large a sample should be taken if we want the maximum error of the estimate p to be equal to

a. 0.03 with 95% confidence?

b. 0.02 with 90% confidence?

c. 0.03 with 90% confidence?

ANS: a. 1068 ; b. 2401; c. 752

Exercise #3

Some dentists were interested in studying the fusion of embryonic rat palates by a standard transplantation technique . When no treatment is used , the probability of fusion equals approximately 0.89. The dentists would like to estimate p, the probability of fusion, when vitamin A is lacking.

a. How large a sample n of rat embryos is needed for $y/n \pm 0.10$ to be a 95% confidence interval for p?

b. If y=44 out of n=60 palates showed fusion, give a 95% confidence interval for p.

ANS: a. 38 ; b. [0.621,0.845]

Exercise #4

Let p equal the proportion of college students who favor a new policy for alcohol consumption on campus. How large a sample size is required to estimate p so that the maximum error of the estimate of p is 0.04 with 95% confidence when the size of the student body is

a. N= 1500?

b. N=15,000?

c. N=25,000?