Confidence Intervals generalities

If β is a population parameter and B is the sample statistic that is being used to estimate it, most of the time a $(1-\alpha)100\%$ confidence interval for β has the form
$B - \epsilon < \beta < B + \epsilon$
where ϵ is some expression involving a known probability distribution (such as the normal distribution), and it is called the **error of the estimate** or the **margin of error** (or some similar name).
$1-\alpha$ is called the **level of confidence**. Common values are 0.95, 0.98, 0.99 and they are usually given as percents, indicated by multiplying by 100% in the formula.
For example, if $\alpha = 0.05 = 5\%$, this would give a 95% confidence interval, as $(1-0.05)100\% = (0.95)100\% = 95\%$

In the following, $z_{\alpha/2}$ represents the **critical value** of z which cuts off a right-hand tail with area $\frac{\alpha}{2}$; that is, $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$

Confidence Intervals for the Mean: known standard deviation

Conditions: Random sampling from an infinite (or very large) population, where σ is known, and either
- X has normal distribution, or
- n is large (so we can invoke the Central Limit Theorem) to get an approximate confidence interval

Then $\bar{x} - z_{\alpha/2}(\frac{\sigma}{\sqrt{n}}) < \mu < \bar{x} + z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$
is a $(1-\alpha)100\%$ confidence interval for μ (the population mean)
Note: when n is large, we often use this formula with s replacing σ. The formula below involving the T distribution would be more correct (if the population distribution is not too far away from normal) but the critical values of t are practically equal to the critical values of z for large (≥ 30) values of n.

Finite population: for random sampling from a finite population, when the other conditions listed above hold, the formula for the confidence interval is
$\bar{x} - z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})\left(\sqrt{\frac{N-n}{N-1}}\right) < \mu < \bar{x} + z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})\left(\sqrt{\frac{N-n}{N-1}}\right)$

Example: (11.1 on p. 359)
We intend to use the mean of a random sample of size $n = 150$ to estimate the average mechanical aptitude of workers. If, based on experience, we can assume that $\sigma = 6.2$ for such data, what can we assert with 99% confidence about the maximum error of that estimate?

Example: (11.2 on p. 356)
If a random sample of size $n = 20$ from a normal population with $\sigma^2 = 225$ has mean $\bar{x} = 64.3$, construct a 95% confidence interval for the population mean μ.

Finding a sample size needed to control the size of the error in the estimate of μ

We want the error term to be no more than ϵ:

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \epsilon$$

Solve this for n to find the sample size needed.

Example: Suppose that we want to estimate the average mechanical aptitude of workers as in Example 11.1, assuming that $\sigma = 6.2$, with 99% confidence, and we want our estimate to be within 1.0 units of the true value. That means that we want the error term to be no more than 1.0:

$$2.58\left(\frac{6.2}{\sqrt{n}}\right) \leq 1.0$$

Solve this for n:

Confidence Intervals for a Population Proportion

The role of the Central Limit Theorem is played here by the following theorem:

Theorem (DeMoivre-Laplace): If X has binomial distribution $b(x; n, \theta)$, then as the number of Bernoulli trials n goes to infinity, the distribution of the statistic

$$Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

approaches the standard normal distribution.

To use this to find an interval estimate for a population proportion, we want to take the number of sucesses X and divide it by the number of trials n. This gives the proportion of successes in the n trials, and this should give an estimate of the proportion of successes in the population. Thus let $\hat{\theta} = \frac{X}{n}$; then $\hat{\theta}$ has mean θ and standard deviation $\sqrt{\frac{\theta(1-\theta)}{n}}$, and the above theorem tells us that the statistic $Z = \frac{\hat{\theta}-\theta}{\sqrt{\theta(1-\theta)/n}}$ has nearly standard normal distribution if n is large enough (depending on θ).

We estimate a population proportion θ by a sample proportion $\hat{\theta}$.

Conditions:

- n is large [rule of thumb: $n \geq 100$ to justify using $\hat{\theta}$ in place of θ in the error term]
- both $n\hat{\theta}$ and $n(1-\hat{\theta})$ are not too small [rule of thumb: both ≥ 5, but ≥ 10 or so is better, to justify estimating binomial by normal]
- and in the finite population case, also need N large compared with n [rule of thumb: $n \leq 0.05N$ to justify estimating hypergeometric by binomial]

Then $\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} < \theta < \hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
is a $(1-\alpha)100\%$ confidence interval for θ (the population proportion)

Example: (11.7 on p. 363)
In a random sample, 136 of 400 people given a flu vaccine experienced some discomfort. Construct a 95% confidence interval for the true proportion of people who will experience some discomfort from the vaccine.

Confidence Intervals for a Population Variance

The role of the Central Limit Theorem is played by the following:

Theorem: If $X_1, X_2, ..., X_n$ is a random sample of a random variable X with normal distribution $n(x; \mu, \sigma)$, then the sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ has a χ^2 distribution with $\nu = n - 1$ degrees of freedom.

We use this to get a confidence interval for the population variance:

Condition: X has normal distribution (cannot be relaxed much)
Then

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

is a $(1-\alpha)100\%$ confidence interval for σ^2 (the population variance)
Note: $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are critical values of χ^2 with $\nu = n - 1$ degrees of freedom.

Example: (11.10 on p. 367)
In 16 test runs the gasoline consumption of an experimental engine had a standard deviation of 2.2 gallons. Construct a 99% confidence interval for the true variance of the gasoline consumption of the engine.)