LECTURE 06

CONFIDENCE INTERVALS

Outline of today's lecture:

I. Introduction

Confidence intervals (CI) is an important part of statistical inference.

It refers to obtaining statements such as

$$P\left[a(X_1,...,X_n) \le \theta \le b(X_1,...,X_n)\right] = 1-\alpha$$

where θ is the parameter of interest and a, b are quantities computed based on the *iid* sample $X_1, . . . ,X_n$. The probability $1-\alpha$ is called the *confidence coefficient*. It is generally taken to be 0.9, 0.95 or 0.99.

In contrast to point estimators $\hat{\theta}$ which give us a specific guess for θ, CIs provide an interval - which is less accurate than a specific number. The advantage of confidence intervals is that we can characterize the confidence of our statement $\theta \in [a,b]$. CIs of the form $[-\infty,b]$ or $[a,\infty]$ are called one-sided CIs (lower or upper).

- In general, to construct a CI, we need to know some partial information concerning the unknown distribution - for example that it is a normal distribution.
  - Such CIs are called *small sample confidence intervals*.

- If we can not make such an assumption we can still construct CIs by appealing to the *central limit theorem*. However, in this case, the CI will be only approximately correct - with the approximation improving in its quality as the sample size increases $n \to \infty$.
  - Such CIs are called *large sample confidence intervals*.

II. Large Sample Confidence Intervals

If the target parameter θ is μ or μ_1-μ_2, then for large samples

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

possesses approximately a standard normal distribution. Consequently, $Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$ forms (at least approximately) a pivotal quantity, and hence the pivotal method can be employed to develop intervals for the target parameter θ.

Example 1 Let $\hat{\theta}$ be a statistic that is normally distributed with mean θ and standard error $\sigma_{\hat{\theta}}$. Find a confidence interval for θ that possesses a confidence coefficient equal to (1-α).

<u>*Solution*</u>

We know that $Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$ has a standard normal distribution.

Now, we select two values in the tails of this distribution, namely, $z_{\alpha/2}$ and $-z_{\alpha/2}$, such that (see figure below)

$$P\left[-z_{\alpha/2} \le Z \le z_{\alpha/2}\right] = 1 - \alpha$$

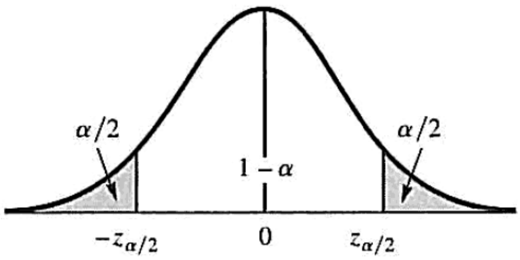


Figure 1 Location of -$z_{\alpha/2}$ and $z_{\alpha/2}$

Substituting for Z in the probability statement yields

$$P\left[-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}\right] = 1-\alpha$$

Multiplying by $\sigma_{\hat{\theta}}$, we obtain

$$P\left[-z_{\alpha/2}.\sigma_{\hat{\theta}} \le \hat{\theta}-\theta \le z_{\alpha/2}.\sigma_{\hat{\theta}}\right] = 1-\alpha$$

Subtracting $\hat{\theta}$ from each term of the inequality produces

$$P\left[-z_{\alpha/2}.\sigma_{\hat{\theta}} - \hat{\theta} \le -\theta \le z_{\alpha/2}.\sigma_{\hat{\theta}} - \hat{\theta}\right] = 1-\alpha$$

Finally, multiplying each term by -1 (which would change the direction of inequality signs) yields

$$P\left[\hat{\theta} + z_{\alpha/2}.\sigma_{\hat{\theta}} \ge \theta \ge \hat{\theta} - z_{\alpha/2}.\sigma_{\hat{\theta}}\right] = 1-\alpha$$

rearranging, we have

$$P\left[\hat{\theta} - z_{\alpha/2}.\sigma_{\hat{\theta}} \le \theta \le \hat{\theta} + z_{\alpha/2}.\sigma_{\hat{\theta}}\right] = 1 - \alpha$$

Thus the endpoints for 100(1-α)% confidence interval for θ are given by:

$$\hat{\theta}_L = \hat{\theta} - z_{\alpha/2}.\sigma_{\hat{\theta}} \text{ and } \hat{\theta}_U = \hat{\theta} + z_{\alpha/2}.\sigma_{\hat{\theta}}$$

Attention Similarly, we determine that 100(1-α)% <u>*one-sided*</u> confidence limits, often called upper and lower bounds, respectively, are given by:

100(1-α)% lower bound for $\theta \rightarrow$ $\hat{\theta}_L = \hat{\theta} - z_{\alpha}.\sigma_{\hat{\theta}}$

100(1-α)% upper bound for $\theta \rightarrow$ $\hat{\theta}_U = \hat{\theta} + z_{\alpha}.\sigma_{\hat{\theta}}$

Example 2 The shopping times of n=64 randomly selected customers at a local supermarket were recorded. The *average* and *variance* of the 64 shopping times were 33 minutes and 256, respectively. Estimate μ, the true average shopping time per customer, with a confidence coefficient of 1-α = 0.90.

<u>*Solution*</u>

- In this case we are interested in the parameter $\theta=\mu$.
- Thus, $\hat{\theta} = \bar{Y} =$ 33 and $s^2=256$ for a sample of $n=64$ shopping times.

- The confidence interval $\hat{\theta} \pm z_{\alpha/2}.\sigma_{\hat{\theta}}$ has the form:

$$\bar{Y} \pm z_{\alpha/2}.\left(\frac{\sigma}{\sqrt{n}}\right)$$

- The population variance σ^2 is *unknown*, from CLT we know that <u>*in the large sample case*</u> it can be replaced (with no serious loss in accuracy) by the sample estimate s^2. Hence:

$$\bar{Y} \pm z_{\alpha/2}.\left(\frac{\sigma}{\sqrt{n}}\right) \approx \bar{Y} \pm z_{\alpha/2}.\left(\frac{s}{\sqrt{n}}\right)$$

Consulting Standard Normal Distribution table, we see that $z_{\alpha/2}=z_{0.05}=1.645$; hence, the confidence limits are given by

$$\bar{Y} - z_{\alpha/2}.\left(\frac{s}{\sqrt{n}}\right) = 33-1.645.\left(\frac{16}{\sqrt{64}}\right) = 33-3.29 = 29.71$$

$$\bar{Y} + z_{\alpha/2}.\left(\frac{s}{\sqrt{n}}\right) = 33 + 1.645.\left(\frac{16}{\sqrt{64}}\right) = 33 + 3.29 = 36.29$$

z	$P(-z < Z < z)$
1.65	0.9
1.96	0.95
2.58	0.99

Figure 2 Common Values of Z

Attention Thus, our confidence interval for μ is (29.71, 36.29). In <u>*repeated sampling*</u>, approximately 90% of all intervals of the form $\bar{Y} \pm \underbrace{1.645}_{z_{\alpha/2}}.\left(\frac{s}{\sqrt{n}}\right)$ include μ, the true mean shopping time per customer.

- Although we do not know whether the particular interval (29.71,36.29) contains μ, the procedure generating it yields intervals that do capture the true mean in approximately 90% of all instances where the procedure is used.

A. Interpretation of (1-α) Confidence Intervals

- *Before we generate* a confidence interval, we can say that the interval we will produce has probability 1-α of containing the true mean μ.

- Once we generate a specific interval, it either contains the true mean μ or doesn't. *It's a very common mistake* to say that the specific interval we obtain contains the mean with probability 0.95. From our viewpoint, the truth (hence, a *constant*) is not a random variable, so this interpretation is not valid.
- What we can say is that if *we repeated the experiment a large number of times (repeated sampling)* and generate a $(1-\alpha)$ confidence interval for each one, then approximately $1-\alpha$ of these intervals will contain μ.
- It seems a bit confusing at first, so let's draw an analogy to simple probability calculations.
  - For instance, we know that the probability of rolling a six from a fair die is 1/6. However, once we roll the die and look at the result, it either shows a six or does not. This tells us that probabilities are useful for thinking about the likelihood of a certain event before the experiment takes place, but once the results are in, there is no longer anything random to compute the probability of. Similarly, *confidence intervals are random variables* that have a $1-\alpha$ chance of containing μ, but once we've collected the data we use to compute them, then the interval either contains μ or does not.

B. Simulation to Show that a Confidence Interval is a Random Variable

See file: *confidence_interval.xlsx or confidence_interval.xls*

III. Selecting the Sample Size

- Above, we assumed that based on fixed α and n we calculated the resulting confidence interval. One could reverse the reasoning as follows.
    - We may ask what is the sample size n that will provide a specific confidence interval $\theta \in \left[\bar{Y} - a, \bar{Y} + a\right]$ at a specific confidence level *1*-α.

- In this case we should take

$$a = z_{\alpha/2} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$n \geq \left[z_{\alpha/2} \cdot \left(\frac{\sigma}{a} \right) \right]^2$$

where we use inequality since n has to be integer.

- Since the value of σ is usually not known, <u>for large samples,</u> we can estimate its value by the standard deviation s of a sample:

$$n \geq \left[z_{\alpha/2} \cdot \left(\frac{s}{a} \right) \right]^2$$

- As an alternative, we can estimate the range R of observations in the population and use it to estimate, $\sigma \approx R/4$.
  - Why divide the range by 4?
    - The range covers the entire distribution and ± 2 (or 4) standard deviations cover 95% of the area under the normal curve. Since we are estimating one standard deviation, we divide the range by 4.

Example Management of a firm wants to know customers' level of satisfaction with their service. They propose conducting a survey and asking for satisfaction on a scale from 1 to 10. (since there are 10 possible answers, the range=10).

- Management wants to be 95% confident in the results and they do not want the allowed error to be more than ±0.5 scale points.
- What should be the sample size of the survey for this purpose?

Solution

s=10/4=2.5

$z_{\alpha/2}$=1.96 (95% confidence)

a= 0.5

Hence $n = \left[z_{\alpha/2} . \left(\frac{s}{a} \right) \right]^2 = \left[1.96 * \frac{2.5}{0.5} \right]^2 = 96.04$. Then n should be 97.

Note: Solve the Example 8.10 from your book about the differences of means.

IV. Small Sample Confidence Intervals

If we know the distribution of the data we can do better than the large sample approximations based on the Central Limit Theorem.

A. Confidence Interval for μ

- Suppose that Y is a random variable from a normally distributed population, $Y \sim N(\mu, \sigma^2)$, $S^2 = \dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$

- Then, in case of small samples, the pivot $\dfrac{\bar{Y} - \mu}{S / \sqrt{n}}$ has a *<u>t distribution</u>* with *n-1* degrees of freedom (df).

It leads to the confidence interval:

$$P\left[-t_{\alpha/2} \le \frac{\bar{Y} - \mu}{S / \sqrt{n}} \le t_{\alpha/2}\right] = 1 - \alpha$$

Then, after simple manipulations we get:

$$P\left[\overline{Y} - t_{\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{Y} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right] = 1-\alpha$$

We can also obtain 100(1 - a)% <u>*one-sided*</u> confidence limits for μ.

- 100(1-α)% lower confidence interval is given by

  $$P\left[\overline{Y} - t_{\alpha}\frac{S}{\sqrt{n}} \le \mu\right] = 1-\alpha$$

- 100(1-α)% upper confidence interval is given by

  $$P\left[\overline{Y} + t_{\alpha}\frac{S}{\sqrt{n}} \le \mu\right] = 1-\alpha$$

B. Confidence Interval for μ_1-μ_2

- Suppose that we are interested in comparing the means of two normal populations, one with mean μ_1 and variance σ_1^2 and the other with mean μ_2 and variance σ_2^2.

In this section we assume that Y_1 and Y_2 are independent random samples from normal populations:

$$Y_1 \sim N(\mu_1, \sigma_1^2),\ S_1^2 = \frac{\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y_1})^2}{n_1 - 1}$$

$$Y_2 \sim N(\mu_2, \sigma_2^2),\; S_1^2 = \frac{\sum_{i=1}^{n_2}(Y_{2i} - \overline{Y_2})^2}{n_2 - 1}$$

If $\overline{Y_1}$ and $\overline{Y_2}$ are the respective sample means from independent random samples from normal populations, the large sample confidence interval can be developed using:

$$Z = \frac{(\overline{Y_1} - \overline{Y_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

as a pivotal quantity.

If we assume that the two populations have a common, but unknown variance, $\sigma_1^2 = \sigma_2^2 = \sigma$ (unknown), the pivotal quantity Z can be rewritten as:

$$Z = \frac{(\overline{Y_1} - \overline{Y_2}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Since, we have assumed that the common variance is unknown, we need to find an estimator for it.

The usual unbiased estimator of the common variance σ^2 is obtained by *pooling* the sample data to obtain the pooled estimator, S_p :

$$S_p^2 = \frac{\sum_{i=1}^{n_1}\left(Y_{1i} - \overline{Y}_1\right)^2 + \sum_{i=1}^{n_2}\left(Y_{2i} - \overline{Y}_2\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where S_i^2 is the sample variance from the i^{th} sample, i=1, 2.

i. When $n_1 = n_2$

If $n_1 = n_2 = n$, S_p^2 is simply the average of of S_1^2 and S_2^2 :

$$S_p^2 = \frac{(n-1)S_1^2 + (n-1)S_2^2}{n + n - 2} = \frac{(n-1)S_1^2 + (n-1)S_2^2}{2(n-1)} = \frac{S_1^2 + S_2^2}{2}$$

ii. When $n_1 \neq n_2$

However, if $n_1 \neq n_2$, S_p^2 is the weighted average of S_1^2 and S_2^2.

Recall that for $Z \sim N(0,1)$ and $W \sim \chi_v^2$, we have $\frac{Z}{\sqrt{W/v}} \sim t_v$. Hence, we will use the random variable $\frac{Z}{\sqrt{W/v}} \sim t_v$ as a pivot with

$$Z = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

For W in the pivot, we use

$$W = \frac{\sum_{i=1}^{n_1}\left(Y_{1i} - \overline{Y}_1\right)^2}{\sigma^2} + \frac{\sum_{i=1}^{n_2}\left(Y_{2i} - \overline{Y}_2\right)^2}{\sigma^2}$$

$$W = \frac{\sum_{i=1}^{n_1}\left(Y_{1i} - \overline{Y}_1\right)^2 + \sum_{i=1}^{n_2}\left(Y_{2i} - \overline{Y}_2\right)^2}{\sigma^2}$$

We know that :

$$S_p^2 = \frac{\sum_{i=1}^{n_1}\left(Y_{1i} - \overline{Y}_1\right)^2 + \sum_{i=1}^{n_2}\left(Y_{2i} - \overline{Y}_2\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Hence, $\sum_{i=1}^{n_1}\left(Y_{1i} - \overline{Y}_1\right)^2 + \sum_{i=1}^{n_2}\left(Y_{2i} - \overline{Y}_2\right)^2 = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2$.

Thus, W becomes:

$$W = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}$$

$$W = \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2_{n_1-1+n_2-1} = \chi^2_{n_1+n_2-2}$$

Substituting Z and W in the pivot $\frac{Z}{\sqrt{W/v}} \sim t_v$ yields:

$$\frac{Z}{\sqrt{\frac{W}{v}}} = \frac{\left(\frac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)}{\sqrt{\frac{\left(\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\right)}{n_1+n_2-2}}} = \frac{\left(\frac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)}{\sqrt{\frac{1}{\sigma^2}\left[\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}\right]}}$$

$$\frac{Z}{\sqrt{\frac{W}{v}}} = \frac{\left(\frac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)}{\sqrt{\frac{S_p^2}{\sigma^2}}} = \frac{\frac{1}{\cancel{\sigma}}\left(\frac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)}{\frac{S_p}{\cancel{\sigma}}} \rightarrow$$

$$\frac{Z}{\sqrt{\frac{W}{v}}} = \frac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_v$$

Hence:

$$P\left[-t_{\alpha/2} \leq \frac{(\overline{Y_1}-\overline{Y_2})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq t_{\alpha/2}\right]=1-\alpha$$

$$P\left[-t_{\alpha/2}.S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq (\overline{Y_1}-\overline{Y_2})-(\mu_1-\mu_2) \leq t_{\alpha/2}.S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]=1-\alpha$$

$$P\left[-t_{\alpha/2}.S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}-(\overline{Y_1}-\overline{Y_2}) \leq -(\mu_1-\mu_2) \leq t_{\alpha/2}.S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}-(\overline{Y_1}-\overline{Y_2})\right]=1-\alpha$$

$$P\left[t_{\alpha/2}.S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}+(\overline{Y_1}-\overline{Y_2}) \geq (\mu_1-\mu_2) \geq -t_{\alpha/2}.S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}+(\overline{Y_1}-\overline{Y_2})\right]=1-\alpha$$

Hence the confidence interval is:

$$P\left[(\overline{Y_1}-\overline{Y_2})-t_{\alpha/2}.S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq (\mu_1-\mu_2) \leq (\overline{Y_1}-\overline{Y_2})+t_{\alpha/2}.S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]=1-\alpha$$

where $t_{\alpha/2}$ is determined from the *t distribution* with *(n_1+n_2-2)* degrees of freedom.

Solve Example 8.12 from your textbook.

V. Confidence Intervals for σ^2

We use the pivot $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ to obtain the confidence interval for σ^2.

$$P\left[\chi^2_L \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_U\right] = 1-\alpha$$

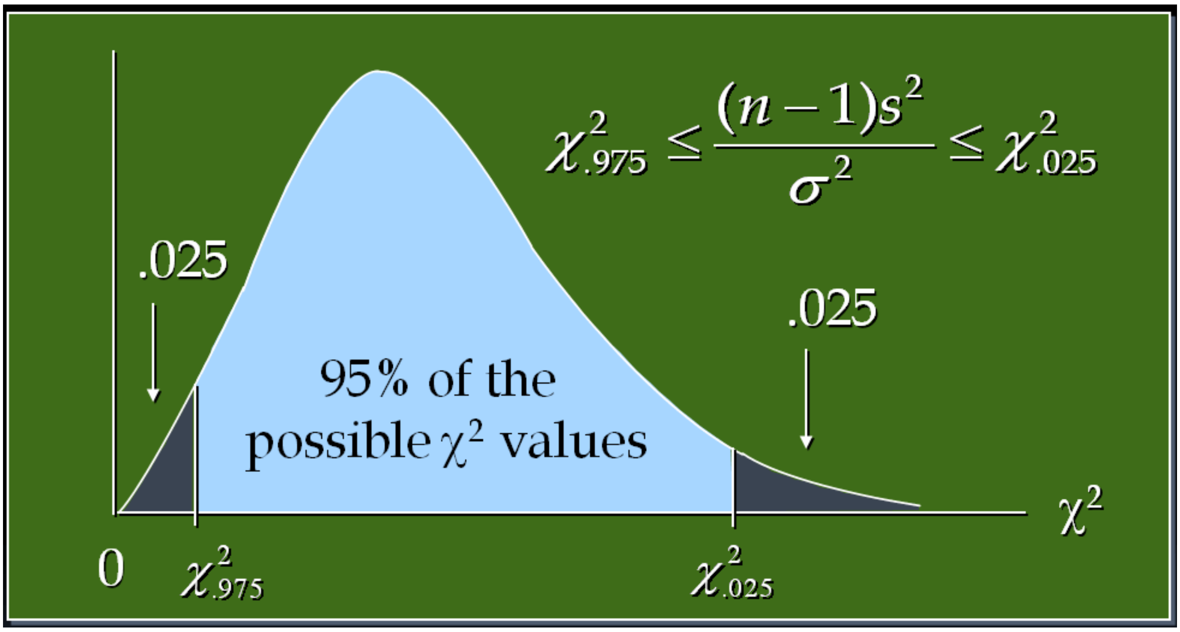


Figure 3 Chi-Square Distribution

Hence, there is a (1–α) probability of obtaining a χ^2 value such that:

$$P\left[\chi^2_{1-(\alpha/2)} \le \chi^2 \le \chi^2_{\alpha/2}\right] = 1-\alpha$$

Substituting $\frac{(n-1)S^2}{\sigma^2}$ for the χ^2 we get

$$P\left[\chi^2_{1-(\alpha/2)} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2}\right] = 1-\alpha$$

$$P\left[\frac{\chi^2_{1-(\alpha/2)}}{(n-1)S^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{\alpha/2}}{(n-1)S^2}\right] = 1-\alpha$$

Hence, the confidence interval is as follows:

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2)}}\right] = 1-\alpha$$

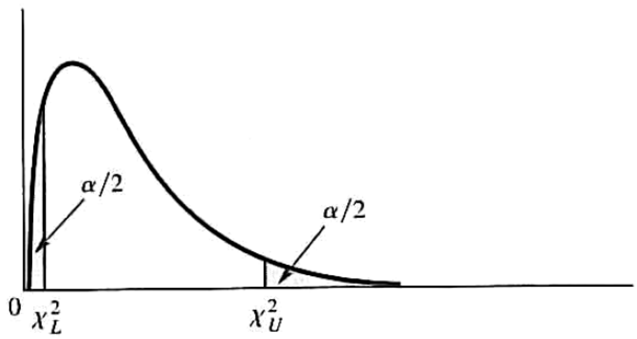
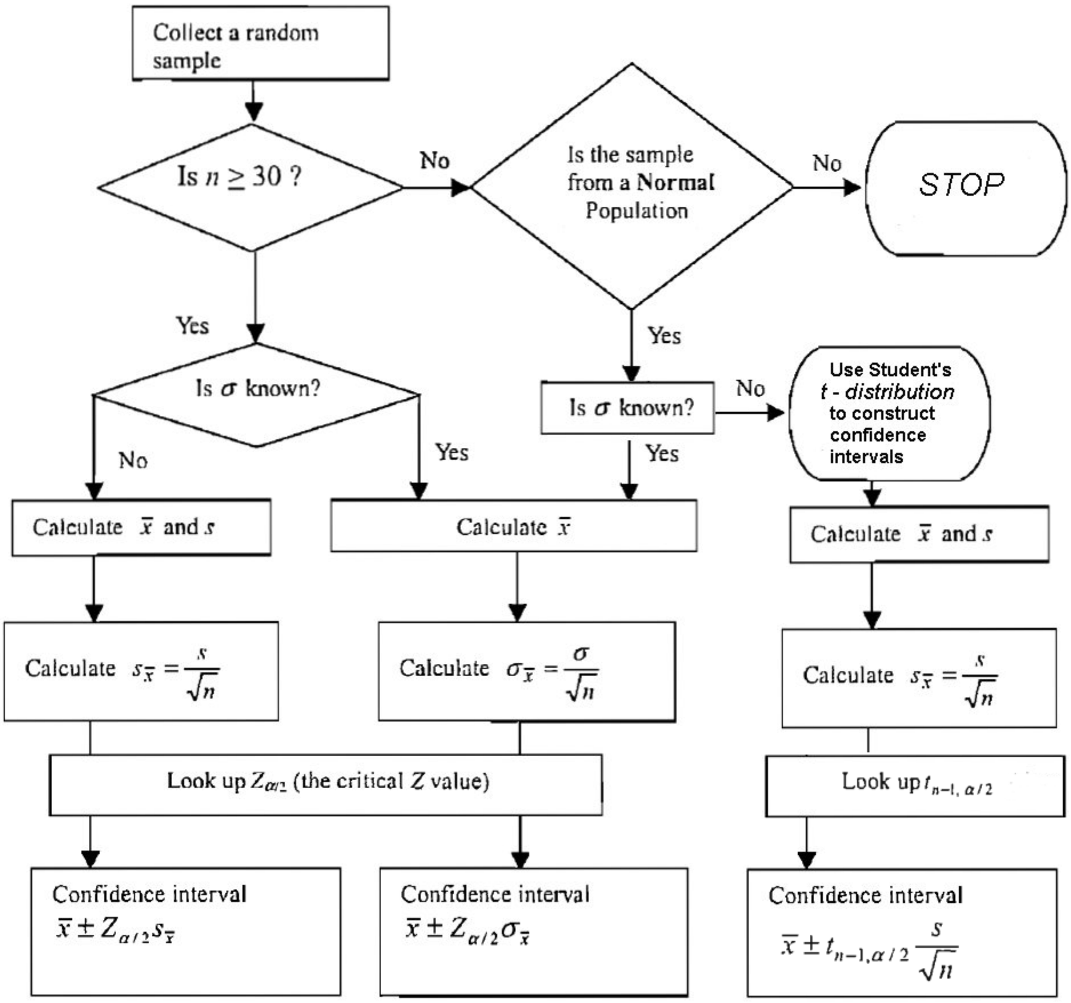


Figure 4 Location of $\chi^2_L = \chi^2_{1-(\alpha/2)}$ and $\chi^2_U = \chi^2_{\alpha/2}$

Solve Example 8.13 from your textbook.

Appendix Constructing Confidence interval for the *mean*



The $(1-\alpha)100$ % confidence interval for the population mean μ