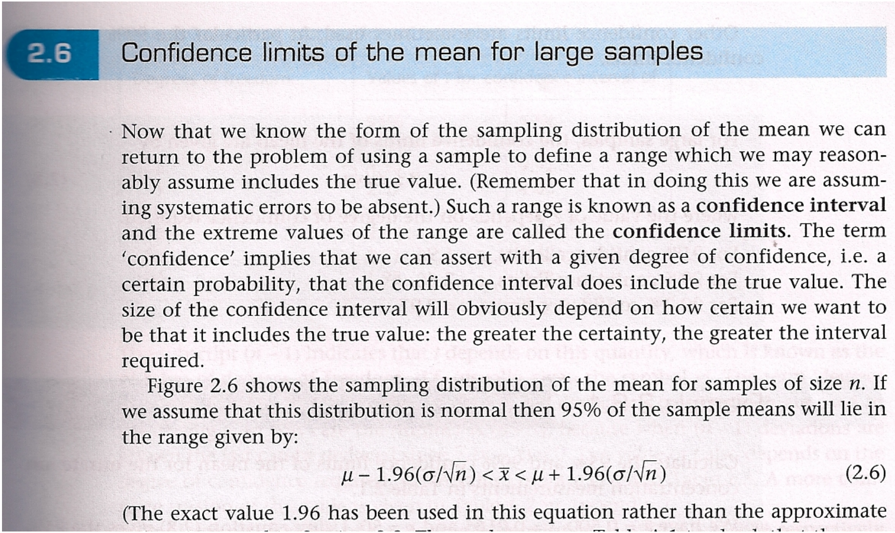
Confidence Intervals

Krzysztof Podgórski
Department of Mathematics and Statistics
University of Limerick

September 21, 2009

Large sample distribution of sample mean



2.6 Confidence limits of the mean for large samples

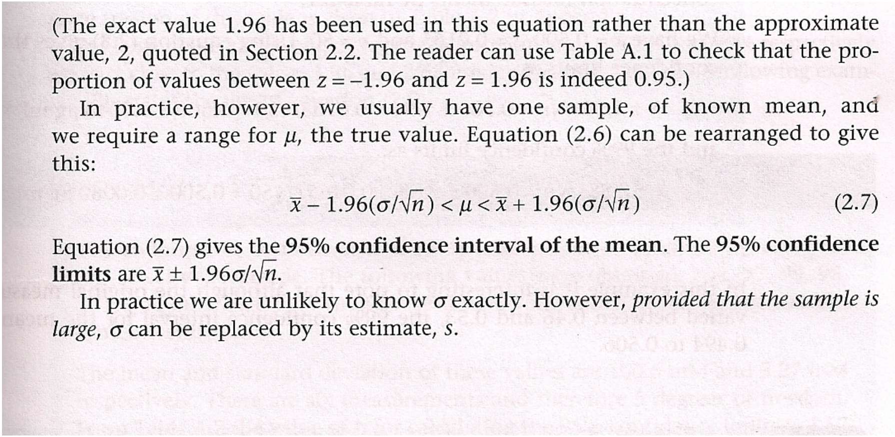
Now that we know the form of the sampling distribution of the mean we can return to the problem of using a sample to define a range which we may reasonably assume includes the true value. (Remember that in doing this we are assuming systematic errors to be absent.) Such a range is known as a **confidence interval** and the extreme values of the range are called the **confidence limits**. The term 'confidence' implies that we can assert with a given degree of confidence, i.e. a certain probability, that the confidence interval does include the true value. The size of the confidence interval will obviously depend on how certain we want to be that it includes the true value: the greater the certainty, the greater the interval required.

Figure 2.6 shows the sampling distribution of the mean for samples of size n. If we assume that this distribution is normal then 95% of the sample means will lie in the range given by:

$$\mu - 1.96(\sigma/\sqrt{n}) < \bar{x} < \mu + 1.96(\sigma/\sqrt{n}) \qquad (2.6)$$

(The exact value 1.96 has been used in this equation rather than the approximate

Large sample confidence interval based on sample mean



(The exact value 1.96 has been used in this equation rather than the approximate value, 2, quoted in Section 2.2. The reader can use Table A.1 to check that the proportion of values between $z = -1.96$ and $z = 1.96$ is indeed 0.95.)

In practice, however, we usually have one sample, of known mean, and we require a range for μ, the true value. Equation (2.6) can be rearranged to give this:

$$\bar{x} - 1.96(\sigma/\sqrt{n}) < \mu < \bar{x} + 1.96(\sigma/\sqrt{n}) \qquad (2.7)$$

Equation (2.7) gives the **95% confidence interval of the mean**. The **95% confidence limits** are $\bar{x} \pm 1.96\sigma/\sqrt{n}$.

In practice we are unlikely to know σ exactly. However, *provided that the sample is large*, σ can be replaced by its estimate, s.

Example of computations using **R**

Finding confidence intervals for the mean for the nitrate ion concentrations in Table 2.1.

```
#reading data
x=scan("Table2_1.txt")
#setting the confidence level
CL=0.95
#computing confidence interval
n=length(x)
pm=sd(x)*c(qnorm(0.025),qnorm(0.975))/sqrt(n)
CI=mean(x)+pm
```

Small-Sample Case ($n < 30$)

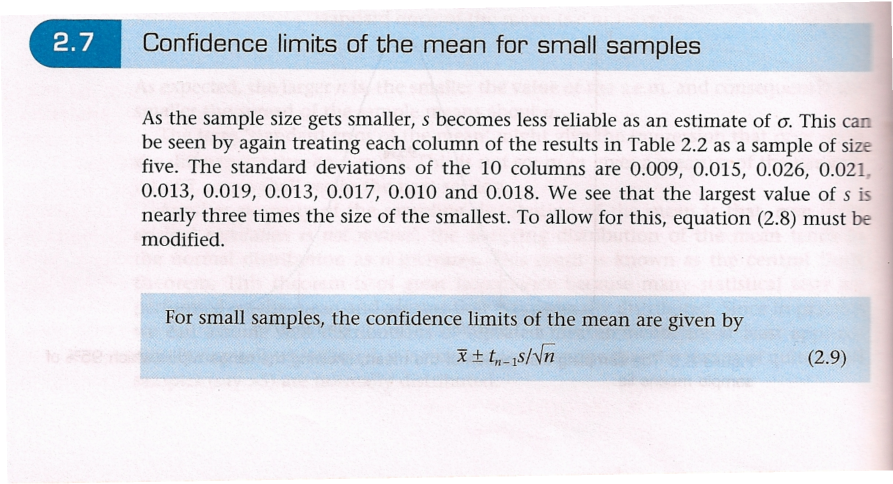
Small-Sample Case ($n < 30$)

If the data have a normal probability distribution and the sample standard deviation s is used to estimate the population standard deviation s, the interval estimate is given by:

$$\bar{X} \pm t_{\alpha/2} s/\sqrt{n},$$

where $t_a/2$ is the value providing an area of $a/2$ in the upper tail of a Student's t distribution with $n - 1$ degrees of freedom.

Small sample confidence intervals



2.7 Confidence limits of the mean for small samples

As the sample size gets smaller, s becomes less reliable as an estimate of σ. This can be seen by again treating each column of the results in Table 2.2 as a sample of size five. The standard deviations of the 10 columns are 0.009, 0.015, 0.026, 0.021, 0.013, 0.019, 0.013, 0.017, 0.010 and 0.018. We see that the largest value of s is nearly three times the size of the smallest. To allow for this, equation (2.8) must be modified.

For small samples, the confidence limits of the mean are given by

$$\bar{x} \pm t_{n-1} s/\sqrt{n} \qquad (2.9)$$

Student *t*-distribution

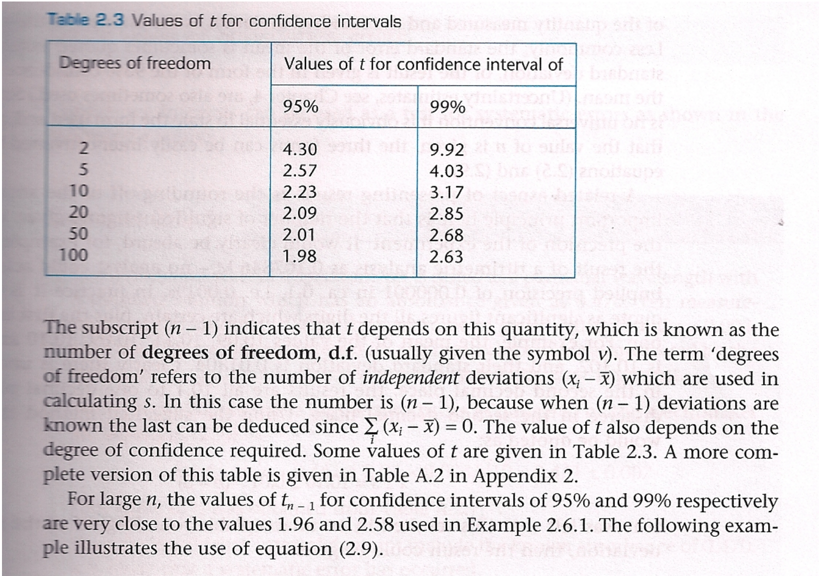


Table 2.3 Values of *t* for confidence intervals

Degrees of freedom	Values of *t* for confidence interval of	
	95%	99%
2	4.30	9.92
5	2.57	4.03
10	2.23	3.17
20	2.09	2.85
50	2.01	2.68
100	1.98	2.63

The subscript $(n - 1)$ indicates that t depends on this quantity, which is known as the number of **degrees of freedom, d.f.** (usually given the symbol ν). The term 'degrees of freedom' refers to the number of *independent* deviations $(x_i - \bar{x})$ which are used in calculating s. In this case the number is $(n - 1)$, because when $(n - 1)$ deviations are known the last can be deduced since $\sum_i (x_i - \bar{x}) = 0$. The value of t also depends on the degree of confidence required. Some values of t are given in Table 2.3. A more complete version of this table is given in Table A.2 in Appendix 2.

For large n, the values of t_{n-1} for confidence intervals of 95% and 99% respectively are very close to the values 1.96 and 2.58 used in Example 2.6.1. The following example illustrates the use of equation (2.9).

Why ”Student”?

William S. Gosset was a statistician employed by the Guinness brewing company which had stipulated that he not publish under his own name. He therefore wrote under the pen name “Student.” His main contribution was published in 1908.

Example of computations using **R**

Finding confidence intervals for the mean for the nitrate ion concentrations in Example 2.7.1.

```
#Typing data in
x=c(102,97,99,98,101,106)
mean(x)
sd(x)
n=length(x)
#setting the confidence level
CL=0.95
#computing confidence interval
pm=sd(x)*c(qt(0.025,n-1),qt(0.975,n-1))/sqrt(n)
CI=mean(x)+pm
```