Queen Mary
University of London

MTH4106 Introduction to Statistics

Notes 13 **Spring 2012**

Interval estimation

So far, all our estimates have been *single numbers*: these are called *point estimates*. Sometimes it is better to give *interval estimates*.

Definition If c is a number associated with a population, and L and U are statistics (that is, numerical functions of samples), and $P\%$ is a percentage, then the open interval

$$(L,U) = \{x \in \mathbb{R} : L < x < U\}$$

is a *$P\%$ confidence interval for c* if

$$\mathbb{P}(L < c < U) = \frac{P}{100}.$$

Note that L and U are the random variables, not c. So the left-hand-side of the above equation could be written as

$$\mathbb{P}(L < c \text{ and } U > c).$$

If we estimate that c lies in the interval (L,U), we shall be right $P\%$ of the time.

Example The time required by workers to complete an assembly job usually has a mean of 50 minutes and a standard deviation of 8 minutes. On a snowy day, the supervisor wonders if they are working at the same speed as usual. He intends to record the times that 60 workers take to complete one assembly job apiece.

Let μ be the unknown mean assembly time on a snowy day, and let $\bar{X}$ be the mean of a sample of 60. The population is known to be (approximately) normal with $\sigma = 8$, so $\bar{X}$ is normal, with

$$\mathbb{E}(\bar{X}) = \mu \quad \text{and} \quad \text{standard error } = \frac{8}{\sqrt{60}}.$$

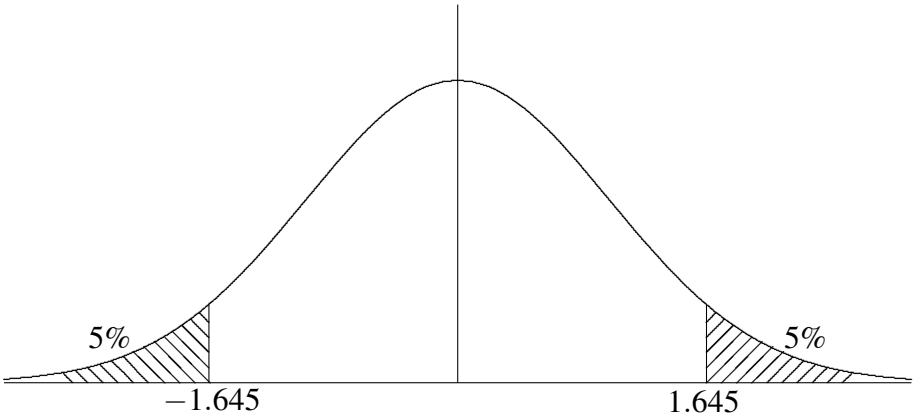
To find a 90% confidence interval for μ, we need to choose L and U so that

$$\mathbb{P}(L < \mu < U) = \frac{90}{100}.$$

As the normal distribution is symmetric, it makes sense to do this in such a way that

$$\mathbb{P}(\mu \leqslant L) = \mathbb{P}(\mu \geqslant U) = \frac{1}{2}\left(1 - \frac{90}{100}\right) = \frac{5}{100} = 5\%.$$

For a standard normal random variable Z, Table 5 of the *New Cambridge Statistical Tables* [1] shows that $\mathbb{P}(Z \geqslant 1.645) = 0.05$.



So

$$\mathbb{P}\left(\frac{\bar{X} - \mu}{\text{S.E.}} \geqslant 1.645\right) = 0.05,$$

so

$$\mathbb{P}(\bar{X} - \mu \geqslant 1.645 \times \text{S.E.}) = 0.05,$$

so

$$\mathbb{P}(\bar{X} - 1.645 \times \text{S.E.} \geqslant \mu) = 0.05.$$

So we put $L = \bar{X} - 1.645 \times \text{S.E.}$. Similarly, $U = \bar{X} + 1.645 \times \text{S.E.}$.

Since S.E. $= 8/\sqrt{60}$, we get $1.645 \times \text{S.E.} = 1.70$ to 2 decimal places, so a 90% confidence interval for μ is $(\bar{X} - 1.70, \bar{X} + 1.70)$. This means that, for 9 out of 10 samples, we shall be correct in claiming that μ lies in the interval $(\bar{x} - 1.70, \bar{x} + 1.70)$.

If the supervisor measures $\bar{x} = 52.57$, then a 90% confidence interval for μ is

$$(52.67 - 1.70, 52.67 + 1.70) = (50.97, 54.37).$$

Note that μ is *not* a random variable, so μ either is or is not in our calculated confidence interval, and we do not know whether it is or not.

Comment on notation

Let Φ be the cumulative distribution function for the standard normal distribution. Then Φ is a bijection from $\mathbb{R}$ to the open interval $(0,1)$, so it has an inverse function Φ^{-1}. In the preceding example, we wanted to find the value of u such that $1-\Phi(u)=0.05$. Rearranging gives $\Phi(u)=0.95$, so $u=\Phi^{-1}(0.95)$. Then Table 5 tells us that $\Phi^{-1}(0.95)=1.645$.

If you think about the cdf, then it is natural to write $1.645=\Phi^{-1}(0.95)$. Indeed, if you use Minitab, then you need to enter 0.95 to get the value 1.645. However, most textbooks concentrate on the area to the *right* of u instead of the area to the left, so they use the notation $u=z_{0.05}$.

Your lecturers are just as likely as you are to get confused about whether we should write this notation with '0.95' or '0.05'.

General procedure to find a $P\%$ confidence interval for the mean μ when σ is known and $\bar{X}$ may be assumed normal

Put

$$\alpha = 1 - \frac{P}{100}.$$

Use Table 5 of NCST [1] to find the number, called $z_{\alpha/2}$, such that $\mathbb{P}(Z \geqslant z_{\alpha/2}) = \alpha/2$, where $Z \sim N(0,1)$. The confidence interval is

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

General procedure to find a $P\%$ confidence interval for the mean μ when σ is unknown, $\bar{X}$ may be assumed normal, and $n \geqslant 30$

If $n \geqslant 30$, then s is a very good estimate for σ, so

$$\frac{\bar{X}}{S}$$

is also approximately normal. So we construct the confidence interval as before, but replacing $\sigma/\sqrt{n}$ by the estimated standard error, which is

$$\begin{aligned}
&\frac{s}{\sqrt{n}} && \text{for a mean;}\\
&\sqrt{\frac{\hat{p}\hat{q}}{n}} && \text{for a proportion;}\\
&N\sqrt{\frac{\hat{p}\hat{q}}{n}} && \text{for a total count.}
\end{aligned}$$

Example Let us find a 95% confidence interval for the number of hurt children. Then

$$P = \frac{95}{100} \quad \text{so} \quad \alpha = \frac{5}{100} \quad \text{so} \quad \frac{\alpha}{2} = \frac{2.5}{100} = 0.025,$$

and Table 5 of NCST [1] gives $z_{0.025} = 1.96$ (which is the number which we have been approximating by 2 up until now).

We have already seen that

$$\begin{aligned} \text{the point estimate} &= 311968, \\ \text{the estimated standard error} &= 34497, \end{aligned}$$

so the 95% confidence interval is

$$\begin{aligned} & (311968 - 1.96 \times 34497,\ 311968 + 1.96 \times 34497) \\ &= (311968 - 67614,\ 311968 + 67614) \\ &= (244354,\ 379582), \end{aligned}$$

which it is sensible to present as $(245000, 380000)$, given our previous remarks about spurious accuracy.

Note that 41% of this interval is below $300,000$, but the newspaper reports all claimed that the number is "more than $300,000$".

Confidence interval for a Poisson mean

If $X_i \sim \text{Poisson}(\lambda)$ for $i = 1, 2, \ldots, n$ then X_i has the same mean and variance, so we can use a slightly more sophisticated procedure. Now

$$\mathbb{E}(\bar{X} = \lambda) \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\lambda}{n},$$

so

$$\frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$

is approximately $N(0, 1)$.

Here we will just consider a 95% confidence interval. The calculations are similar for P% confidence intervals for other values of P.

For a 95% confidence interval, we want

$$-1.96 < \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} < 1.96.$$

Squaring gives

$$(\bar{X}-\lambda)^2 < (1.96)^2\,\frac{\lambda}{n},$$

or

$$n(\bar{X}-\lambda)^2 < (1.96)^2\,\lambda.$$

Given the actual sample mean $\bar{x}$, we want to find the values of λ which satisfy the quadratic inequality

$$n(\bar{x}-\lambda)^2 < (1.96)^2\,\lambda,$$

which we can rewrite as

$$n\lambda^2 - (2n\bar{x}+(1.96)^2)\lambda + n\bar{x}^2 < 0.$$

Since the coefficient of λ^2 is positive, the parabola of the corresponding quadratic funtion is "the right way up", and so the values of λ for which the function values are less than zero are those lying between the two roots of the quadratic. So the endpoints of the confidence interval are

$$\begin{aligned}
&\frac{2n\bar{x}+(1.96)^2 \pm \sqrt{[2n\bar{x}+(1.96)^2]^2 - 4n^2\bar{x}^2}}{2n} \\
= \;& \bar{x}+\frac{(1.96)^2}{2n} \pm \frac{\sqrt{4n^2\bar{x}^2+4n\bar{x}(1.96)^2+(1.96)^4-4n^2\bar{x}^2}}{2n} \\
= \;& \bar{x}+\frac{(1.96)^2}{2n} \pm \frac{\sqrt{4n\bar{x}(1.96)^2+(1.96)^4}}{2n} \\
= \;& \bar{x}+\frac{(1.96)^2}{2n} \pm \frac{1.96}{2n}\sqrt{4n\bar{x}+(1.96)^2}.
\end{aligned}$$

---

[1] D. V. Lindley and W. F. Scott, *New Cambridge Statistical Tables*, Cambridge University Press.