1 Random Samples from Normal Distributions

1.1 Estimators and confidence intervals

Motivation

Assume for example that we are given a set of data, which we regard as plausibly normal, and we might wish to find a point estimate of the mean μ. The previous lectures suggest that $\overline{X}$ is an obvious candidate. We also need to know what is the likely error range. What If we had a different set of data? How reliable is our estimate, can we trust it? To within what error bounds? We need some theory, making use of the previous lectures.

Definition An *estimator* of a parameter θ is a statistic, say a function $A\left(X_1,\, X_2\,, \ldots,\, X_n\right)$ of the random sample, which does not depend on any unknown parameters in the model and which we use to give a point estimate of the parameter from the data.
□

An example of this is the way we use $\overline{X}$ to estimate the mean of a distribution.
If the estimator is to have any use at all, it should have some nice properties. For example, we know that $\overline{X} \xrightarrow{P} \mu$ by the weak law of large numbers, ensuring that $\overline{X}$ is a sensible estimator for μ.

A starting point for considering the likely error using the normal distribution is given by

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N\left(0, 1\right).$$

Definition A *Z-statistic* is a statistic with a standard normal distribution (as above).
$\square$

The main use of Z-statistics stems from the facts that, for a general distribution, the Central Limit Theorem implies asymptotically that

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N\left(0, 1\right),$$

and that the standard normal distribution involves no unknown parameters: it can be (and is) tabulated.

We can use the Z-statistic to calculate a range of plausible values for μ, under the assumption that σ^2 is known.

Definition *Confidence interval*

Let $\mathbf{X}$ represent a vector of random variables with entries X_i. If $(a(\mathbf{X}), b(\mathbf{X}))$ is a random interval such that

$$P\left(a(\mathbf{X}) < \mu < b(\mathbf{X})\right) = 1 - \alpha,$$

then a realisation of that interval, $(a(\mathbf{x}), b(\mathbf{x}))$ is said to be a $100\,(1-\alpha)\,\%$ confidence interval for μ.
$\square$

It is not easy to get to grips with what is meant by a confidence interval. Clearly one cannot say that the parameter μ has probability $(1-\alpha)$ of lying within the calculated interval $(a(\mathbf{x}), b(\mathbf{x}))$ because the ends of the interval are fixed numbers, as is μ, and without random variables being present, probability statements cannot be made: either μ lies between the two numbers or it doesn't, and we have no way of knowing which. The only viable interpretation is to say that we have used a procedure which, if repeated over and over again, would give an interval containing the parameter $100\,(1-\alpha)\,\%$ of the time: the rest of the time we will be unlucky.

Central $100(1-\alpha)\%$ confidence intervals using Z-statistics are found as follows. Remembering that $Z \sim N(0, 1)$, choose $z_{\alpha/2}$ such that

$$\begin{aligned} P\left(Z \le z_{\alpha/2}\right) &= 1 - \frac{\alpha}{2} \\ \Longrightarrow P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) &= 1 - \alpha. \end{aligned}$$

If $Z = \dfrac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}$ as above, then

$$\begin{aligned} P\left(-z_{\alpha/2} \le \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \le z_{\alpha/2}\right) &= 1 - \alpha \\ \Longrightarrow \quad P\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2} \le \mu \le \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\right) &= 1 - \alpha. \end{aligned}$$

Hence the appropriate random interval is

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

and the $100\,(1-\alpha)\,\%$ confidence interval is

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

The most common value of α in use is 0.05, in which case $z_{\alpha/2} = z_{0.025} = 1.960$.

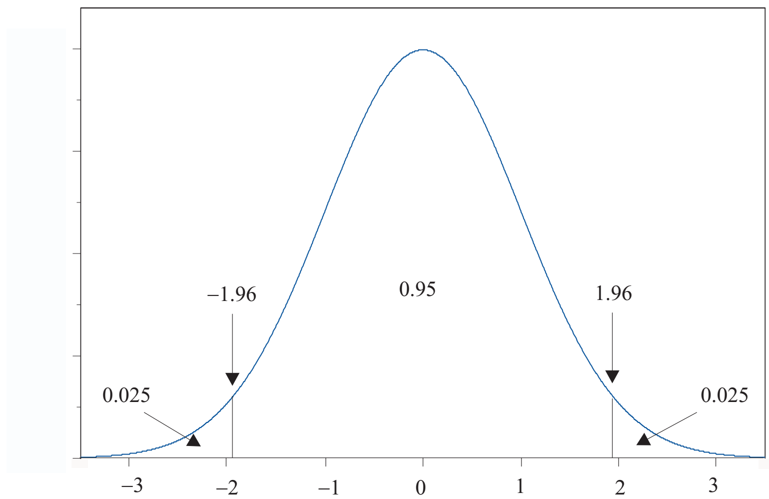

Figure 3.3 95% interval for $N(0,1)$

Example 1 *Radioactive-carbon dating*

Assume the samople mean is $\overline{x} = 2505.86$. In order to estimate the age of the site, we need to take the following steps.

(i) Check that the data are plausibly normal. We can use a normal probability plot.

(ii) Estimate the mean of the distribution by the sample mean and write $\widehat{\mu} = \overline{x} = 2505.86$.

(iii) Use a Z-statistic to find a 95% confidence interval which gives a range of plausible values for the mean age. This is

$$\left(\overline{x} - \frac{1.96\sigma}{\sqrt{n}}\ ,\ \overline{x} + \frac{1.96\sigma}{\sqrt{n}}\right),$$

and, putting in $n = 7$ and $\overline{x} = 2505.86$, we find a central 95% confidence interval

$$\left(2505.86 - \frac{1.96\sigma}{\sqrt{7}}\ ,\ 2505.86 + \frac{1.96\sigma}{\sqrt{n}}\right),$$

Unfortunately we are no better off. We cannot obtain the confidence interval because we do not know σ, so what should we do? We would like to replace σ by s, the sample standard deviation, but can we? $\left[\text{Recall that } S^2 = \dfrac{1}{n-1}\sum\left(x_i - \overline{x}\right)^2.\right]$

■

We know that, if $X_1,\ X_2,\ldots,\ X_n$ is a random sample from a normal distribution $N(\mu,\sigma^2)$, then

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} \sim t\left(n-1\right)$$

We can now look for a confidence interval by replacing the Z-statistic with the t-statistic. Writing $t_{\alpha/2}\left(n-1\right)$ for the $1-\frac{\alpha}{2}$ quantile from the distribution $t(n-1)$,

$$P\left(-t_{\alpha/2}\left(n-1\right) < \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} < t_{\alpha/2}\left(n-1\right)\right) = 1-\alpha.$$

Re-arranging gives the random interval

$$\left(\overline{X} - \frac{t_{\alpha/2}}{\sqrt{n}}\,S,\ \overline{X} + \frac{t_{\alpha/2}}{\sqrt{n}}\,S\right),$$

and the $100\left(1-\alpha\right)\%$ confidence interval is the realisation of this interval.

Example 2 *Radioactive-carbon dating*

For the carbon-dating example, $n = 7$ and $t_{0.025}(6) = 2.447$, from a t-distribution with 6 degrees of freedom, $s = 56.44$. Plugging these values into the formula results in a 95% confidence interval of (2453.5, 2558.3), thereby giving a range of plausible values for μ.

■

1.2 Application of Central Limit Theorem

The Central Limit Theorem states that, for any random sample $X_1,\ X_2,\ldots,\ X_n$ such that the sample size n is sufficiently large, we have

$$\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \mathrel{\dot\sim} N\left(0,1\right).$$

Notation: $\dot\sim$ means 'approximately distributed as'. Provided we are dealing with moderate to large sample sizes we can therefore use the approximate normality to find confidence intervals, using approximate Z-statistics.

Example 3 *Binomial Proportion*

In an opinion poll prior to a Staffordshire South East by-election, of 688 constituents chosen at random 368 said they would vote Labour (53.5%). The newspapers are perfectly happy to use these data to estimate p, the probability that a constituent selected at random would vote Labour, but they rarely, if ever, give any idea of the quality of the estimate. Let us see how to obtain a 95% confidence interval for p.

First identify the random sample. Constituents questioned are labelled 1,..., 688. Let

$$X_i = \begin{cases} 1, & \text{if } i^{\text{th}} \text{ constituent says "I will vote Labour"}, \\ 0, & \text{otherwise.} \end{cases}$$

Then X_i has a Bernoulli distribution $B(1, p)$, the sample size n is 688, and $E(X_i) = p$, $V(X_i) = p(1-p)$. We know that p can be estimated by the sample mean $\overline{x} = \dfrac{368}{688} =$ 0.535. We can also apply the Central Limit Theorem to find an approximate confidence interval using the asymptotic normality with $\mu = p$, $\sigma^2 = p(1-p)$. Thus

$$\frac{\sqrt{n}\left(\overline{X} - p\right)}{\sqrt{p\,(1-p)}} \dot\sim N(0, 1) \, .$$

The 95% random interval is of the form

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{0.025},\ \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{0.025}\right)$$

but unfortunately σ is a function of p. We could solve a quadratic inequality for p, but, since $n = 688$ is large, we will replace σ by its estimator $\sqrt{\overline{x}(1-\overline{x})}$. This gives (0.498, 0.572) as a 95% confidence interval for p, with point estimate 0.535.

If we required a 99% confidence interval we would use $z_{0.005} = 2.576$ to replace 1.960, and get a wider interval (0.486, 0.584) about which we are slightly more confident.

■