Statistical Estimation

Random Sampling

Conduct an opinion poll about voting intentions. How reliable are the results?

Measure the actual volume of every 50^{th} 750mL bottle of vodka. What can we say about the distribution of volume in all bottles?

A chemical is delivered to a lab in batches of 1000 1L bottles. How can we reliably test the purity of the batch?

These problems all involve *statistical inference*, that is, obtaining information from a sample of data about the population from which the sample is drawn and setting up a mathematical model to describe this population. In general, we cannot test/measure every member of the population because:

it would take too long;

it would be too expensive;

may require testing to destruction;

the total population may be unknown.

We will focus on two aspects of statistical inference:

1. Estimation of the unknown parameters of the mathematical model, e.g. expected value and variance.
2. Testing a hypothesis about the mathematical model, e.g. “FG/Labour will win the election".

1 Point estimates and confidence intervals

Point Estimates and confidence intervals

Recall that *measurements* tend to follow a normal distribution. To describe the normal distribution and answer useful questions (as in the previous chapter), we need to know two numbers; the expectation or mean μ and the standard deviation (square root of the variance) σ. Then the quantity we measure X follows the normal distribution

$$X \sim N(\mu, \sigma^2).$$

To calculate these two numbers - or parameters - precisely, we would need to measure every member of the population. Instead, we use a statistic which is an estimate of a given parameter.

Parameter	Sample	Population
Mean	$\overline{x}$	$\mu = E(X)$
Standard deviation	s	σ

Example
From a batch of 5000, 8 samples of plastic granules were tested for fire resistance and the combustion temperatures in ^oC were as follows:

Sample.	1	2	3	4	5	6	7	8
Temp.	510	535	498	450	491	505	487	500

Calculate the mean and standard deviation for the samples. We use these definitions. For a sample of size n, the *sample mean* is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Recall that the variance (σ^2) is the average distance of each value from the mean. So

$$s_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}.$$

For technical reasons, we use the following slightly different formula for the *sample standard deviation*:

$$s_{n-1} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}.$$

Sample.	1	2	3	4	5	6	7	8
Temp.	510	535	498	450	491	505	487	500

We find that

$$\overline{x} = 497^oC, \qquad s_{n-1} = 24^oC.$$

Then we expect that

$$\mu \simeq 497^oC, \qquad \sigma \simeq 24^oC.$$

How confident can we be about these estimates?
If the population follows a normal distribution, so that if X is the quantity being measured, we have $X \sim N(\mu, \sigma^2)$, then we can use the following results:

> Fact One: If random samples of size n are taken from a distribution $N(\mu, \sigma^2)$, then the sample mean forms a distribution having the same mean μ but with a smaller standard deviation given by
>
> $$STEM = \frac{\sigma}{\sqrt{n}}.$$
>
> (STEM comes from STandard Error of the Mean.)

If the population does not follow a normal distribution, we still have a quite powerful result.

> Fact Two - the Central Limit Theorem: If random samples of size n are taken from a population with mean μ and standard deviation σ, then the sample mean $\overline{x}$ is approximately normally distributed with mean μ and standard deviation $STEM$. The approximation improves as n increases.

Simplified version
When we measure $\overline{x}$ of a sample from a population, we may assume that

$$\overline{x} \sim N(\mu, \frac{\sigma^2}{n}),$$

where μ, σ are the mean and standard deviation of the population.

Confidence Intervals I- when σ is known.
For the data above giving the fire resistance of plastic granules, the sample mean temperature was 497^oC. This is an estimate for the mean of the whole batch of 5000. We expect this mean to be close to 497. How close? We can find a range of values of temperature within which we can be 95% (or 90%, or 99%,...) certain that the true means lies.
Let us assume that the estimated value 24^oC is actually the true value of the standard deviation σ. We have just seen that

$$\overline{x} \sim N(\mu, \frac{\sigma^2}{n}).$$

Translate to a standard normal variate:

$$z \sim N(0, 1), \quad \text{where} \quad z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}.$$

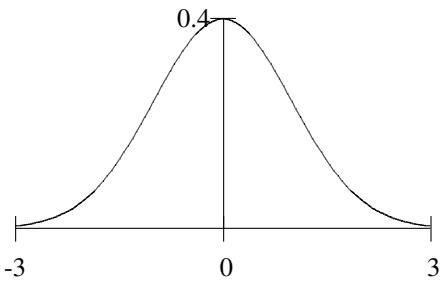

Figure 1: The standard normal curve.

Confidence Intervals
From the tables, we can work out that

95% of z lies between -1.96 and $+1.96$. That is,

$$P(-1.96 < z < 1.96) = 0.95.$$

99% of z lies between -2.58 and $+2.58$. That is,

$$P(-2.58 < z < 2.58) = 0.99.$$

90% of z lies between -1.64 and $+1.64$. That is,

$$P(-1.64 < z < 1.64) = 0.90.$$

We can show that $-1.96 < z < 1.96$ is the same as

$$\overline{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Thus the interval

$$I_{95} = \left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is called the 95% *confidence interval for the mean*.

Significance of I_{95}.
To calculate I_{95}, we need to *know* the standard deviation σ, but we only need an *estimate* $\overline{x}$ of the true mean. We can say with 95% certainty that the true mean lies in the interval I_{95}.

Example

Calculate the 90% confidence interval for the mean combustion temperature of the batch of plastic granules.

$$I_{90} = \left(\overline{x} - 1.64\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.64\frac{\sigma}{\sqrt{n}}\right)$$

$$\begin{aligned} I_{90} &= n\left(497 - \frac{1.64 \times 24}{\sqrt{8}}, 497 + \frac{1.64 \times 24}{\sqrt{8}}\right) \\ &= (483.08, 510.96). \end{aligned}$$

Confidence Intervals II - when σ is not known.
Things are different here because we need to use estimates for *both* the standard deviation and the mean. We can't use the normal distribution. The appropriate distribution was discovered in 1908 by William Gosset who worked as a statistician in Guinness. It is known as Student's $t-$distribution; see p.38 of the log tables.
Given a sample of size n, calculate the sample mean $\overline{x}$ and the sample standard deviation $s = s_{n-1}$. The confidence intervals are given by

$$I_{\%} = \left(\overline{x} - t_{\%,r}\frac{s}{\sqrt{n}}, \overline{x} + t_{\%,r}\frac{s}{\sqrt{n}}\right),$$

where $r = n - 1$ gives the *degrees of freedom* of the sample.

Using the tables

$t_{\%,r}$ is determined from the tables as follows.

The left hand column gives the value of r and tells us which row to look in.

The top row gives the value of $100 - P$ for the $P\%$ confidence interval, and tells us which column to look in.

Example

The lengths of 10 fossilized skulls of an extinct species of bird are measured and found to have mean $5.68cm$ and standard deviation $0.29cm$. Determine 95% and 99% confidence intervals for the mean.

$$I_{95\%} = \left(5.68 - t_{95\%,9}\frac{0.29}{\sqrt{10}}, 5.68 + t_{95\%,9}\frac{0.29}{\sqrt{10}}\right).$$

$$I_{95\%} = \left(5.68 - 2.262\frac{0.29}{\sqrt{10}}, 5.68 + 2.262\frac{0.29}{\sqrt{10}}\right) = (5.473, 5.887).$$

$$I_{99\%} = \left(5.68 - t_{99\%,9}\frac{0.29}{\sqrt{10}}, 5.68 + t_{99\%,9}\frac{0.29}{\sqrt{10}}\right).$$

$$I_{99\%} = \left(5.68 - 3.25\frac{0.29}{\sqrt{10}}, 5.68 + 3.25\frac{0.29}{\sqrt{10}}\right) = (5.382, 5.978).$$

Notes

The $t-$distribution and normal distribution are similar when the sample size n is large ($n > 30$ is usually large enough). This is because the sample standard deviation gives a better and better estimate of the true standard deviation for large values of n.

Notice that

$$s = s_{n-1} = \sqrt{\frac{n}{n-1}}s_n,$$

so that $s > s_n$. This correction factor is included in the definition of s because s_n is usually too small to give a good estimate for σ.

The reason for this is that if we take a sample of only a few values ($n < 30$), there is a very good chance that we will only get values near the mean; the chance of finding extreme values is low. If all our sample values are close to the mean, our estimate s_n would be artificially low. Remember that σ is the average distance of values from the mean.