Estimation, confidence intervals and tests

1 Statistics

A statistic is defined as follows:
If $X_1, X_2, X_3, ..., X_n$ is a random sample of size n from some population, then a statistic T is a random variable consisting of the X_i than involves no other quantities.

Since it is possible to repeat the process of taking a sample, if all samples are taken, they will form a probability distribution called the **sampling distribution of T**. This will usually depend on the distribution of the population X.

A random sample is defined as follows:
A random sample of size n consists of the observations $X_1, X_2, X_3, ..., X_n$ from a population where the X_i

- Are independent random variables
- Have the same distribution as the population

2 Estimators

A statistic that is used to estimate a population parameter is an **estimator**.
A particular individual value is an **estimate**.
Bias is how far the estimator is from the true value

If a statistic T is used as an estimator for a population parameter θ then the bias is:

$$E(T) - \theta$$

If $E(T) = \theta$ then the statistic is unbiased

$\hat{\theta}$ represents an estimator of θ, this is true for all population parameters, such as $\hat{\mu}$

2.1 Proving $\mathbf{E(\overline{X})} = \mu$

Prove that $\overline{X}$ is an unbiased estimator for μ when the population is normally distributed

A random sample $X_1, X_2, X_3...X_N$ is taken for a population with $X \sim N(\mu, \sigma^2)$

$$\overline{X} = \frac{1}{n}\sum X$$

$$E(\overline{X}) = E\Big(\frac{1}{n}\sum X\Big)$$

$$E(\overline{X}) = \frac{1}{n}\big(E(X_1) + E(X_2) + E(X_3)... + E(X_N)\big)$$

$$E(\overline{X}) = \frac{1}{n}(\mu + \mu + \mu... + \mu)$$

$$E(\overline{X}) = \frac{1}{n}(n\mu)$$

$$E(\overline{X}) = \mu$$

2.2 Variance

$$Var(\overline{X}) = \frac{\sigma^2}{\text{Sample size}}$$

We use S^2 as an estimator for σ^2

$$S^2 = \frac{1}{n-1}\left(\Sigma X^2 - n\overline{X}^2\right)$$

In formula book

$$S^2 = \frac{\Sigma(X_i - \overline{X})^2}{n-1}$$

Replace the top of the fraction with $S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$

$$S^2 = \frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1}$$

2.3 Uniform Distributions

Random variable X is continuously uniform $[0, \alpha]$. A sample $X_1, X_2..X_N$ is taken
Show that $\overline{X}$ is a biased estimate and state the bias

$$\overline{X} = \frac{0+\alpha}{2} = \frac{\alpha}{2}$$

$$E(\overline{X}) = E(\frac{\alpha}{2}) = \frac{1}{2}E(\alpha) = \frac{\alpha}{2}$$

$$\text{Bias} = \frac{\alpha}{2} - \alpha = -\frac{\alpha}{2}$$

3 Standard error

$$\text{Standard error} = \frac{\sigma}{\sqrt{n}} \text{ or } \frac{s}{\sqrt{n}}$$

4 Central limit theorem

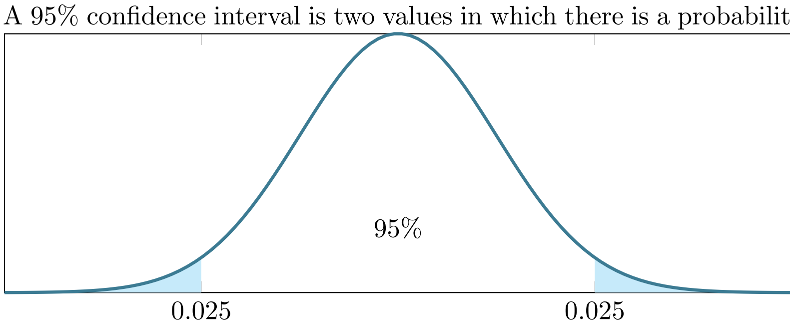
The central limit theorem (C.L.T.) states that if $X_1, X_2, X_3 ...$ is a random sample, from any distribution with mean μ and variance σ^2. Then the sample means, $\overline{X}$ are distributed with a normal distribution.

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

For this, the sample size must be greater than 50

5 Confidence intervals

A 95% confidence interval is two values in which there is a probability of 0.95 that it will contain the population mean.



Look up the value of one of the tails on the percentage points table and substitute into the formula.
For example:
At a 95% confidence interval each of the tails is 0.025 and the corresponding z value is 1.96.

$$\overline{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

6 Hypothesis testing

The test statistic for the population mean μ is $Z = \frac{X - \mu}{\frac{\sigma}{\sqrt{n}}}$

A hypothesis test can be done quicker by finding the critical values from the percentage points table.

6.1 Example

$\mu = 0.58$ $\sigma = 0.015$
Measured
$n = 50$ $\mu = 0.577$ $\sigma = 0.015$

$H_0 : \mu = 0.50$
$H_1 : \mu \neq 0.58$

$$\overline{D} \sim N(0.58, \frac{0.015^2}{50})$$

$$Z = \frac{0.577 - 0.58}{\sqrt{\frac{0.015^2}{50}}} = -1.41$$

At 0.5% per level Z=2.5758

$$-1.41 > -2.5758$$

Not in critical region so not significant, no evidence to reject H_0, the diameter of the bolts is not changed.

6.2 Hypothesis test for the difference between means

If $\overline{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$ and $\overline{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n}\right)$

$$\overline{X} - \overline{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}\right)$$

$$z = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}}}$$

6.2.1 Example

The weight of boys and girls in a certain school are known to be normally distributed with standard deviations of 5kg and 8kg respectively. A random sample of 25 boys had a mean weight of 48kg and a random sample of 30 girls had a mean weight of 45kg.
Stating your hypothesis clearly test, at the 5% level of significance, whether or not there is evidence that the mean weight of boys in the school is greater than the mean weight of girls.

Write down the two distributions

$$\overline{X_B} \sim N\left(\mu_B, \frac{5^2}{25}\right) \qquad \overline{X_G} \sim N\left(\mu_G, \frac{8^2}{30}\right)$$

State the two hypotheses

$$H_0: \ \mu_B = \mu_G$$

$$H_1: \ \mu_B > \mu_G$$

Find the z value

$$z = \frac{48 - 45 - (0)}{\sqrt{\frac{5^2}{25} + \frac{8^2}{30}}} = 1.6947$$

Find the critical value from the percentage points table

$$C.V. = 1.6449$$

Compare the critical value to the z value

$$1.6449 < 1.6947$$

Write conclusion As z is not in the critical region there is not enough evidence to reject H_0, there is not enough evidence to suggest that the mean weight of boys is greater than the mean weight of girls.