MTHS2007 Advanced Mathematics and Statistics for Mechanical Engineers

Chapter 8: statistics

School of Mathematical Sciences

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Statistics: Introduction

Statistics concerns decision making in the presence of uncertainties and involves dealing with information from given data. Subsequent analysis is informed by probability theory.

In dealing with data the whole 'population' (i.e. all possible values) would ideally be evaluated but this is usually impractical for reasons such as

Expense The population may be too large or testing each item may be expensive.

Destructiveness Testing may require dismantling or running to destruction.

The starting point is usually data about a *sample* of the population, collected to represent the whole population.

In Statistics, we aim to draw conclusions about the whole population based on the sample(s).

Since the data samples are subject to random variation, we need to use probability models to quantify this variation and make informed, rational decisions based on probabilities.

Therefore, we suppose that the individual data points are random observations from some underlying probability distribution.

In realistic situations, we do not know this distribution exactly. However, the data will often approximately follow standard distributions such as those we have met previously. For instance:

- Continuous data, such as lengths/weights/strengths are often well modelled by a *normal* distribution;
- Discrete data, such as counting the number of times some event of interest occurs, are often well modelled by a *binomial* or *Poisson* distribution.

Population vs Sample

Population characteristics are properties of the entire population of interest.

Sample characteristics are properties of the sample we have observed. We use the sample to estimate unknown population characteristics.

For example, suppose we assume our sample values are random observations from a wider population, which we assume to be normally distributed with unknown mean μ and variance 1.

Then each data point is an observation of a random variable X, with

$$X \sim N(\mu, 1).$$

We might *estimate* the true, *unknown*, population mean μ using the mean of our observed sample.

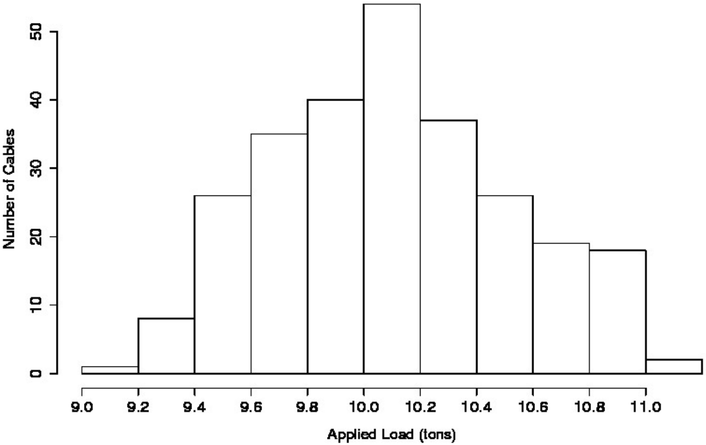
Example: Two hundred and sixty six lengths of cable were chosen at random from the production run of a cable manufacturer and the load necessary to break each of these cables was determined. The data are tabulated on the following slide.

A statistical analysis might proceed as follows:

- What is a suitable probability model for the observed data?
- What characteristic of the population are we interested in, and how can we estimate it?
- How can we quantify how certain we are about any conclusions we make?

Applied Load (tons)	Number of Cables
$<$ 9.0	0
9.0 - 9.2	1
9.2 - 9.4	8
9.4 - 9.6	26
9.6 - 9.8	35
9.8 - 10.0	40
10.0 - 10.2	54
10.2 - 10.4	37
10.4 - 10.6	26
10.6 - 10.8	19
10.8 - 11.0	18
$>$ 11.0	2

Often it is useful to show the sample results graphically, e.g. as a histogram. This can help to check if an assumed probability distribution for the data is reasonable.



Summary statistics (e.g sample mean, sample variance / standard deviation) can also be calculated. These are the sample versions of the theoretical quantities derived from probability distributions, such as $E[X]$ and $V[X]$.

These theoretical quantities often correspond to *parameters* of the probability distribution, and therefore represent the true population value.

For example, $E[X] = \mu$, the population mean (or equivalently, the mean of the underlying probability distribution we have assumed for the data).

The sample mean is calculated from the observed data, and would hopefully be a ‘good estimate’ of the true population mean μ.

Sample Mean and Variance

Suppose we collect a sample of n data points and label them $x_1, x_2, \dots, x_n$.

Sample Mean

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is a statistic to estimate the population mean;

Sample Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

is a statistic to estimate the population variance.

Example: Data values for concentration (in %) were obtained from measures on 20 samples of a chemical solution:

87	86	85	87	86	89	81	77	85	88
86	84	83	83	82	84	83	79	82	73

Sample Mean: $\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{87+86+\cdots+73}{20} = 83.5$

Sample Variance: $\sum_{i=1}^{20} x_i^2 = 87^2 + 86^2 + \ldots + 73^2 = 139728$

$$s^2 = \frac{1}{n-1}\left(\sum_{1=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{19}(139728 - 20 \times 83.5^2) \approx 14.8947$$

$$s = \sqrt{14.8947\ldots} \approx 3.86$$

Sampling

A key concept is that we are usually dealing with only a sample selected from an underlying population.

If $X_1, \ldots, X_n$ are random variables measuring a quantity of interest then we usually obtain sample values $x_1, ..., x_n$ to try and **infer** information about the population from which $X_1, \ldots, X_n$ are drawn.

The general method of approach is to define a Statistic Y (a *NEW* random variable) based on $X_1, \ldots, X_n$, and use the observed sample values $x_1, ..., x_n$ to calculate the *sample* statistic y. This is used to estimate quantities associated with the population.

The distribution of Y and the accuracy of the estimate will depend on the sample size, n, and the distribution of each X_i.

Statistical Inference

Statistical inference is concerned with using probability concepts to quantitatively deal with the uncertainty arising due to using representative samples in making decisions.

The basis is to obtain **samples** (from a population) to analyze and infer properties of the whole population.

For example, to obtain the 'true' (i.e. population) average concentration of a contaminant in a lake, one would need to test all the water!

Clearly this is not desirable or feasible, so an alternative is to take a number of **random samples** and obtain an estimate of the contaminant level from the samples.

This raises a number of questions, such as:

- How should we estimate the true population value?
- How does the sample size affect the accuracy of the estimate?
- How different might a sample estimate be from the true population value?
- How sure can we be that the population value lies within an acceptable range of the sample value?

The process of answering such questions is known as *statistical inference*, which is broadly divided into three sections:

1. **Sampling Distributions**
   Identifying a suitable probability distribution which captures the features of the population (and samples drawn from it).
2. **Estimation**
   To use sample values to infer, or estimate, a value of a parameter of the population. This may involve giving a likely range of values called a confidence interval.
3. **Hypothesis Testing**
   To make decisions, and to assign a probability of error when accepting or rejecting a given hypothesis.

A fundamental concept is that any collection of random variables X_i will form a statistic Y, say, given by $Y = g(X_1, X_2, \ldots, X_n)$ for some function g. Each of the quantities X_i will have its own distribution but also Y will have its own probability distribution.

Example:

5 measurements of contaminant concentration in a lake were taken, giving a sample of

$$57.4 \quad 59.5 \quad 62.1 \quad 56.6 \quad 58.2$$

Then an estimate of the true mean concentration might be

$$\tfrac{1}{5}(57.4 + 59.5 + 62.1 + 56.6 + 58.2) = 58.76 \quad \text{(sample mean)}$$

In this case, $n = 5$, $X_1, \dots, X_5$ are the random variables representing the measurements to be taken, and $x_1, \dots, x_5$ are the corresponding observed sample values.

The quantity $Y = \bar{X}$, the mean of $X_1, \dots, X_5$, is also a random variable (it is a function of other random variables) and so will have its own distribution. The corresponding sample statistic y is the observed sample mean, 58.76.

Decisions/conclusions will be based on comparing the observed statistic y with the probability distribution of Y.

Sampling distribution of the mean:

We focus on the case where data $X_i \sim N(\mu, \sigma^2)$ are independent.

Then $Y = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean.

We can show that $E(\bar{X}) = \mu$ and $V(\bar{X}) = \sigma^2/n$.

In fact $\bar{X} \sim N(\mu, \sigma^2/n)$.

The Central Limit Theorem says that if the X_i are independent but not Normal then so long as n is large we still have $\bar{X}$ approximately $N(\mu, \sigma^2/n)$.

In either case we have $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

Confidence Intervals

Given values $X_1, X_2, \ldots, X_n$, then a suitable statistic can be constructed to estimate a population value.

However, this gives no information about the accuracy of the estimate.

From knowledge of the distribution of the sample statistic one can proceed further to determine an **interval** within which the population value might lie with a specific probability.

Such a prescribed probability is called the **confidence level** and the resulting interval the **confidence interval**.

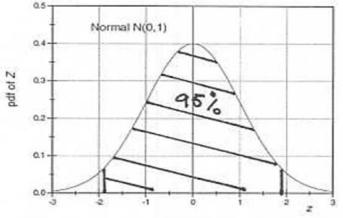
The confidence level may be expressed as a probability, e.g. 0.95, but is often given as a percentage, e.g. 95% confidence level.

As an example of the method, consider the case of independent samples X_i, each with a Normal distribution $N(\mu, \sigma^2)$ with σ assumed known.

Then we know $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

From the tables of $N(0,1)$, we can determine that $|Z| < 1.96$ with confidence level of 95%.

so the interval $-1.96 < Z < 1.96$ has an associated probability of 0.95.



Thus $-1.96 < \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96$ with confidence level 95%.

Rearranging for μ, we obtain

$$\bar{X} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\,\frac{\sigma}{\sqrt{n}}.$$

When a value for $\bar{X}$ is observed from a sample of data (i.e. the sample mean $\bar{x}$), a confidence interval can be obtained within which the population value is expected to lie with the specified confidence level. (*Interpretation requires care.*)

Substituting $\bar{x}$ in place of $\bar{X}$ in the above formula, we obtain our observed 95% confidence interval for the unknown μ:

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

So long as n is 'large', the same argument works even if σ^2 is not known (estimated from the data) and/or the sample data do not follow a Normal distribution.

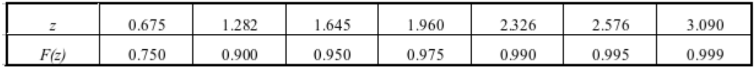
So the endpoints of the confidence interval are $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$.

We write this as $\bar{x} \pm C \frac{\sigma}{\sqrt{n}}$, where C satisfying $P(|Z| > C) = 0.05$.

For a general confidence level $100(1-\alpha)\%$ the same formula applies, with $C = z_{\alpha/2}$ satisfying

$$P(|Z| > C) = \alpha \quad \text{or} \quad P(Z > z_{\alpha/2}) = \alpha/2.$$

Table for finding C:



z	0.675	1.282	1.645	1.960	2.326	2.576	3.090
$F(z)$	0.750	0.900	0.950	0.975	0.990	0.995	0.999

Examples

In a sample of 10 components from a manufacturing process the measured sizes were

$$4.32, 4.35, 4.34, 4.30, 4.37, 4.39, 4.35, 4.35, 4.30, 4.33.$$

Assuming the distribution of component sizes is normal and has standard deviation 0.03, obtain the 95% confidence interval for the mean size produced in this process.

Solution: Write $X_i \sim N(\mu, \sigma^2)$ for the sizes.

Endpoints of the CI are $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$.

Here $\sigma = 0.03$, $n = 10$ and $\bar{x} = \frac{1}{10} \sum_i x_i = 4.34$.

So the endpoints are $4.34 \pm 1.96 \frac{0.03}{\sqrt{10}} \approx (4.32140\ldots\,, 4.35859\ldots)$.

Conclude that $\mu \in (4.32\,, 4.36)$ with 95% confidence.

Example: Based on a survey of 140 employees in a firm, the mean and standard deviation of the commuting distances between home and the place of work are found to be 8.6 miles and 4.3 miles, respectively.

Determine a 90% confidence interval for the mean commuting distance for the population of all employees of the firm.

Solution: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ (Standardized Normal variate)

From Normal table,

$$P(|Z| < C) = 0.90 \text{ for } C = 1.645.$$

Thus, $|Z| < C = 1.645$ with 90% probability.

So a 90% confidence interval for μ is

$$\bar{X} - 1.645\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.645\,\frac{\sigma}{\sqrt{n}}.$$

Replacing the unknown σ with its sample estimate s (the sample size n is 140, definitely 'large'), the confidence interval is

$$8.6 - 1.645\,\frac{4.3}{\sqrt{140}} < \mu < 8.6 + 1.645\,\frac{4.3}{\sqrt{140}}.$$

So the 90% CI for the mean commuting distance μ is $(8.0, 9.2)$.

Use of Confidence Intervals

- Confidence intervals for the mean give a range of values of the unknown parameter μ which are consistent with the observed data. (A range of plausible values for the true mean.)

  $\Longrightarrow$ Learning about the process/system being studied.

- Sometimes it is useful to interpret this in relation to a particular question relating to the process/data being investigated/collected.

  $\Longrightarrow$ Comparing this learnt information with the past / standards / specified requirements. (cf. Hypothesis testing.)

Possible conclusions

- The process/system is operating above/below the standard/historical level.

  What to do next is usually fairly clear.

- No conclusive evidence that the process/system is operating above/below the standard/historical level.

  What to do next is context-dependent. And also dependent on how wide the CI is.)

  Possibilities include: need to collect more data, take some action since some undesirable situation is plausible, take no particular action since some desirable situation is plausible, do nothing, ...

Example

- Suppose that in the manufacturing process example the specified size of the components is 4.3.
- The 95% CI for the mean size produced is $(4.32, 4.36)$.
- This suggests that the mean component size has moved away from the specification, so some intervention/recalibration is necessary.

What about the following situations?

- Data relating to a safety standard.
- Data relating to performance of a new product.

Confidence Level and Precision of Estimation

Our choice of confidence level was essentially arbitrary.

What if we had chosen a higher level of confidence, say 99%?

(It seems reasonable that we would want the higher level of confidence.)

At $\alpha = 0.01$, we obtain $C = 2.58$, while for $\alpha = 0.05$, $C = 1.96$.
Thus, the width of a 95% confidence interval is

$$2 \times 1.96 \frac{\sigma}{\sqrt{n}} = 3.92 \frac{\sigma}{\sqrt{n}}.$$

However, the length of the 99% confidence interval is

$$2 \times 2.58 \frac{\sigma}{\sqrt{n}} = 5.16 \frac{\sigma}{\sqrt{n}}.$$

The 99% confidence interval is wider than 95% interval.

The wider the confidence interval, the more confident we are that the interval actually contains the true value of μ. On the other hand, the wider the interval, the less information we have about the true value of μ.

Note that the larger the sample size, the narrower the confidence interval is (for fixed α and σ). So for known σ (or s), we could fix α and then calculate the sample size needed for a desired width of confidence interval (level of precision).

Example: If $\sigma = 10$, what sample size is needed to achieve a 95% confidence interval of width 8 or less?

We need $2 \times 1.96 \times \dfrac{10}{\sqrt{n}} \leq 8$.

This implies that $n \geq 24.01\ldots$.

So a sample size of $n = 25$ is needed to obtain the desired precision.

More examples

Example: Testing of 40 samples from a reservoir gave the following measurements for a contaminant concentration, in μg / l (micrograms per litre).

$$\sum_i x_i = 14,042 \qquad \sum_i x_i^2 = 4,934,319$$

Determine the sample mean and sample standard deviation of the concentration of the contaminant.

Calculate 95% and 90% confidence intervals for the mean contaminant level in the reservoir.

What do the results tell us about the level of contamination in the lake compared to environmental standards that specify a maximum safe level of 350 μg / l?

Example: Studies of CO concentration near a motorways gave the following measurements x_i in *ppm* (parts per million).

$$\sum_i x_i = 3,726 \qquad \sum_i x_i^2 = 393,355,$$

based on 36 samples.

Compute a 99% confidence intervals for the mean CO concentration along the motorway.

What does this data suggest about the mean CO concentration in comparison to a quality standard which says that the concentration should not exceed 70 ppm?