MAST7866 — Foundations of Data Science

Computing Session 7 — Point and Interval Estimation

Task 1 – Introducing Estimation

Estimation is a procedure by which a numerical value or values are assigned to a population parameter based on the information collected from a sample.

In inferential statistics, μ is called the **true population mean** and π is called the **true population proportion**.

The value assigned to a population parameter based on the value of a sample statistics is called an **estimate** of the population parameter.

Example 1

Suppose a manager takes a sample of 40 new employees and finds that the mean time, $\bar{x}$, taken to learn this job for these employees is 5.5 hours. If he or she assigns this value to the population mean, then 5.5 hours is called an estimate of μ.

The sample statistic used to estimate a population parameter is called an **estimator.** Thus, the sample mean, $\bar{x}$, is an estimator of the population mean, μ and the sample proportion, p, is an estimator of the population proportion π.

The estimation procedure involves the following steps:

1. Select a sample.
2. Collect the required information from the members of the sample.
3. Calculate the value of the sample statistic.
4. Assign value(s) to the corresponding population parameter.
5. If we select a sample and compute the value of a sample statistic for this sample, then this value gives the **point estimate** of the corresponding population parameter.

Each sample selected from a population is expected to yield a different value of the sample statistic. Thus, the value assigned to a population mean, μ, is based on a point estimate, depends on which of the samples is drawn. Consequently, the point estimate assigns a value to μ that almost always differs from the true value of the population mean.

Challenge 1 - Calculating point estimates

A group of 20 adults that work in the same organisation have the following shoe sizes:

4	7	8	4	5	6	7	3	6	7
9	8	6	4	5	7	3	6	5	8

Input the data into R into a variable called x

Calculate the mean of the shoe size of these 20 people who work in the organisation.

```
> mean(x)
```

Suppose we take samples of size 5 from these 20 individuals. Calculate the sample mean of these samples.

```
> z<-sample(x,5,replace=F)
> mean(z)
```

What do you notice about the values of the sample means you calculate?

Task 2 – Interval Estimation for the Population Mean

Instead of assigning a single value to a population parameter, we can construct an interval around the point estimate, and then a probabilistic statement that this interval contains the corresponding population parameter is made.

An interval is constructed with regard to a given **confidence level** and is called a **confidence interval.** The confidence interval is given as

Point estimate $\pm$ Margin of Error

The confidence interval associated with a confidence interval states how much confidence we have that this interval contains the true population parameter. The confidence interval is denoted by (1-α)100%.

Although any value of the confidence level can be chosen to a construct a confidence interval the more common values are 90%, 95% and 99%. α is called the **significance level** - we will return to this in Section 10.

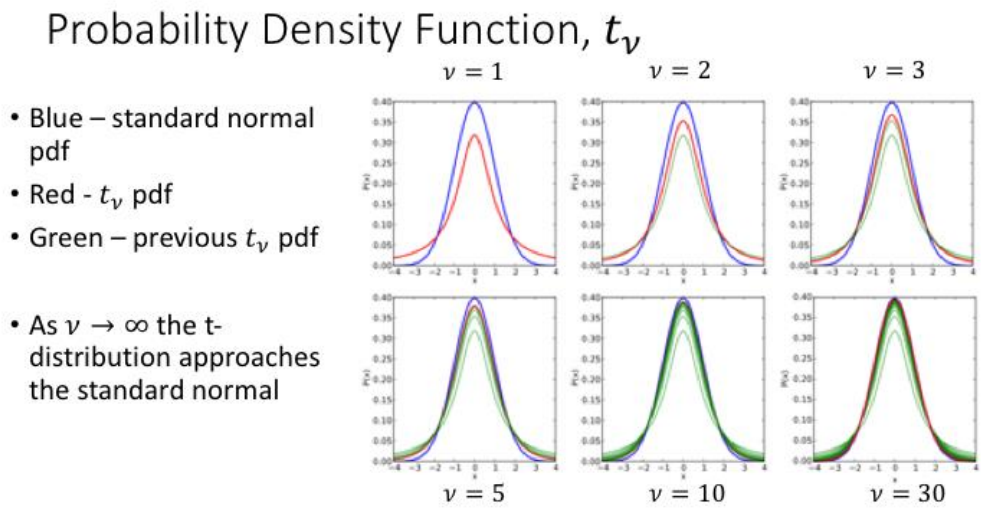
If σ, the population standard deviation, is known and if n $\geq$ 30 (or if n< 30 if we can assume the population from which the sample is selected is approximately normally distributed) then the (1-α)100% confidence interval for μ is

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

The quantity $z\frac{\sigma}{\sqrt{n}}$ is called the **margin of error** and is denoted by E. The value of z is obtained from the Normal distribution.

If the population variance is unknown, then instead of using the normal distribution to construct a confidence interval for μ, we use a **t-distribution**.

The t-distribution is similar to the normal distribution in some respects. Like the normal distribution curve, the t distribution is symmetric about the mean and never meats the horizontal axis. The t-distribution is flatter and wider than the standard normal distribution curve. The shape of a particular t-distribution depends on the number of degrees of freedom. For the purpose of this section, the degrees of freedom of the distribution is equal to the sample size minus 1, that is $df = n - 1$. As the sample size becomes larger, the t distribution approaches the standard normal distribution.



If σ, the population standard deviation, is unknown and if n ≥ 30 (or if n< 30 if we can assume the population from which the sample is selected is approximately normally distributed) then the (1-α)100% confidence interval for μ is

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

where s is the sample standard deviation and t is obtained from the t-distribution with n-1 degrees of freedom and the given confidence level.

How do we interpret a confidence interval?

Let us consider the example above. Suppose we took all possible samples of 25 such college textbooks and construct a 90% confidence interval for μ around each sample mean, we can expect 90% of these intervals will include μ and 10% won't.

In general, we can state that for a 90% confidence interval if we take **samples of the same size** from a population and construct 90% confidence intervals, then we expect 90% of these confidence intervals will include μ.

Note that it is **NOT** correct to say that there is a 90% chance that the confidence interval includes μ.

The **width** of a confidence interval depends on the size of the margin of error, which depends on the values z, σ and n. However, σ is not under the control of the investigator, therefore in order to decrease the width of the confidence interval we can either:

(a) Lower the confidence level
(b) Increase the sample size

Lowering the confidence interval is not a good choice because a lower confidence level may give less reliable results. Therefore, the preference is always to increase the sample size if we want to decrease the width of a confidence interval.

R has a function `t-test` that calculates the confidence interval for you.

Example 2

Let us return to the parkrun data set. Load the data into R. Suppose we want to construct a 95% confidence interval for the mean first finish times of female runners. What do each of the following commands do?

```
> attach(parkrun)
```

```
> t.test(Female,conf.level=0.95)
```

The output is:

```
     One Sample t-test

data:  Female
t = 235.86, df = 229, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1368.049 1391.099
sample estimates:
mean of x
 1379.574
```

At the moment you do not need to worry about the output which I have put in grey (we will cover that in the next worksheet). We want to focus on the output in red. This tells us that the lower confidence limit of our 95% confidence interval of mean finish times of female runners at parkrun is 1368.0 and the upper confidence limit is 1391.1.

Note: In order to construct the confidence interval we have had to assume that either the sample size is sufficiently large (so that the central limit theorem holds) or that the mean finish times of female runners at parkrun are normally distributed. In this case we have n>30. If n were small you could use QQ plots to visually check for normality.

Challenge 2

The `ewr` data set contains monthly information about the taxi in and taxi out times for 8 airlines at EWR (Newark) airport during 1999-2001. Let us assume that the taxi in and taxi out times of airline AA follows a normal distribution. We want to find a 95% confidence interval for the mean μ.

First we need to load the data:

```
>install.packages("UsingR")
```

```
>library(UsingR)
```

Let us have a look at the data:

```
> ewr
```

```
> data1<-ewr$AA
```

Now let us construct the 95% confidence interval for the mean taxi in and taxi out times for AA:

```
> t.test(data1,conf.level=0.95)
```

What is the 95% confidence interval?

Now calculate a 90% confidence interval and a 99% confidence interval. Compare the widths of the confidence intervals.

Task 3 – Interval Estimation for the Population Proportion

Often we want to estimate the population proportion or percentage (recall that a percentage is obtained by multiplying the proportion by 100). For example, the production manager of a company may want to estimate the proportion of defective items produced on a machine. A bank manager may want to find the percentage of customers who are satisfied with the service provided by the bank.

The population proportion is denoted by π and the sample proportion is denoted by p. Here we explain how to estimate the population proportion π using the sample proportion, p. The sample proportion, p is a sample statistic and it possesses a sampling distribution.

For a large sample, the (1-α)100% **confidence interval for the population proportion, π**, is

$$p \pm zs_p$$

The value of z is obtained from the standard normal distribution tables for the given confidence interval. In the case of a proportion, a sample is considered to be large if np and nq are both greater than 5.

Example 3

Suppose we want to know the proportion of the population who believe the government's policies will boost the economy. For this reason we will survey 1000 people; the selection process will ensure we have a random sample and that 1000 people is enough to approximate the normal distribution. The true population proportion is denoted by p. Suppose that 567 people are positive that the economy will grow following the proposed economic policies. Find a 95% confidence interval for the proportion p.

There is a built in R function, `prop.test` which will calculate confidence intervals for a population proportion.

For this question we would use the command:

```
> prop.test(567,1000,conf.level=0.95)
```

The 95% confidence interval for the proportion of people who are positive that the economy will grow following the proposed economic policies is 0.535-0.598.

Challenge 3

Suppose that a survey taken of a random sample of 350 students gives that the sample proportion is$p = 0.42$. The same survey is repeated and a random sample of 1000 students gives the same values for p. Find 95% confidence intervals for the population proportion π based on each of the two random samples. Compare the two confidence intervals.

Note: the first 2 arguments of the prop.test function need to be the “number of successes” and the “number of trials”. Therefore, we have to calculate these from the information in the question. In this case we are told that a proportion of 0.42 out of 350 students, therefore the number of successes will be 0.42x350.