# CA200 – Quantitative Analysis for Business Decisions

**File name:** CA200_Section_05A_Statistical Inference

## Table of Contents

# 5. Statistical Inference

## 5.1 Overview

**Statistical Inference** is the process by which conclusions are drawn about some measure of the data e.g. the mean, variance (or Standard deviation) or proportion of the **population** …based upon analysis of the **sample** data.

Taking a sample is a '*real-world'* requirement because it is rare that one can ever work with the population as a whole; this is often too costly, time-consuming or difficult , e.g. if testing a component for faults means testing destructively, for example.

**Probabilistic Sampling** involves selection and examination of a set of items from the population, which are **representative** of the population as a whole.

- **Simple Random Sampling** means 'equal probability of selection'. This is not haphazard, but is based on systematic preparation of the sampling frame.
  <u>Note</u>: Variants of **SRS** - based on idea of **known** probabilities of selection
- **Objective**: Sample characteristics used to **estimate** population characteristics.

**Accuracy** depends on:

(i) Size of sample

(ii) How sample selected

(iii) Extent of variability in the population

**Sampling Types**

A *probability sample* may be drawn in a number of ways of which the simplest conceptually (and closest to true randomness) is a simple random sample. In

- **Simple Random Sampling,** establishing the **sampling frame** is the hardest part as this must be comprehensive. Hence:

**Quasi-random Sampling** is more usual. Again, there are a number of possibilities: *e.g.*

- **Systematic Sampling:** – used in production, Quality control etc. In this case the selection *starting point* is randomly chosen, then *every kth* item from that point is selected( sampling interval can be varied, but care is needed to ensure that the interval chosen does not follow a pattern in the data, e.g. highs and lows)

- **Stratified Sampling** : – here, a population has natural groups or *strata*, where members of a stratum are similar, but diversity exists between strata; random samples are taken *within each stratum* in the *same proportions* that apply in the *population*.
- **Multi-stage Sampling**: this is similar to stratified sampling but *groups* and *sub-groups*, *sub-sub-groups etc.* are selected e.g. on geographical /regional/ area/town/ street etc. basis, so strata have *natural hierarchy* – again similarity within strata, diversity between.
- **Cluster Sampling** : is a cost-effective way of dealing with lack of a comprehensive *sampling frame*. The method uses selection of a few *e.g. geographical areas* at *random* and then drills down comprehensively (i.e. examines every single unit).

**Types of Inference**

**1. Estimation**: involves the use of a sample statistic to estimate a population parameter. The statistic can be a mean, variance, proportion, etc. The quality of the estimator is of obvious interest and a number of properties contribute to good estimators, i.e.
Good estimators are

- *'unbiased'* : this implies that they are 'on target' : hence if the mean is of interest then the mean of all possible sample means (i.e. the expected value or on average value) is the population mean
- *'consistent'* : this means that the precision improves as the sample size increases
- *'efficient'* : this means that the variability improves with repeated sampling
- *'sufficient'* : this means that *all information available* in the *sample* is used in estimating the population parameter, so e.g. the mean is a better estimator than the median of the population 'average' because it uses all the numerical information, (i.e. *actual values*, not just the rank order).

**Point and Interval estimate**:

A *single* calculation of a mean is a **point estimate**.

In practice, if we know how the mean (or other sample statistic) is distributed, we can divide up its distribution in the usual way, such that 90% (or 95%) of the time, the mean should lie within an interval or range, based on the single sample value.
This interval estimate is then considered established with 90% (95%) confidence

**2. Hypothesis Testing**:

In Hypothesis Testing, a **statement is made** about a *population parameter*, such as the population mean, variance, proportion.

The statement validity is then **tested**, based on the **sample data** and a **decision** is made on the **basis of the result** obtained.

If we can assume *large samples* drawn, then the **Normal distribution** can be used to test for means and proportions.

If samples are *not large*, **or** we want to talk about variances, then other distributions will apply.

(Collectively, the various distributions used in statistical inference are sometimes grouped under the heading of ***Sampling Distributions*** and include the Normal as the common basis).

## 5.2 Additional distributions

On the web page for course notes : http://www.computing.dcu.ie/~hruskin/newteach.htm see CA200_Statisticaltables.pdf and also CA200_Sectio04_ExtraReStatsTablesEtc.pdf

**Students t-distribution**

- If population variance *unknown*, but population *large*, can still use the **Normal**.
- For *small samples*, and population variance $\sigma^2$ *unknown* (i.e. must be estimated from the sample), a slightly more **conservative distribution** than Normal applies = the *Student's T* or just *'t'-distribution*. Introduces the **degrees of freedom** concept.
- A random variable X is described as having a *t-distribution* with *$\nu$ degrees of freedom* (d.o.f.) and denoted ( $t_\nu$ ) . The degrees of freedom depend on sample size n and correspond to the degree of independence in the data.
- The t-distribution is *symmetric* about the origin, just like the Normal and also has mean zero, i.e. E[X] = 0. However, spread and shape change, depending on d.o.f. For *small values* of $\nu$, the $t_\nu$ (or $t_n$ ) distribution is very flat. As $\nu$ is increased, the curve becomes bell- shaped.

For values of sample size n > 25, the $t_v$ distribution is practically *indistinguishable* from the **Standardised Normal** distribution, so the latter can be used in the usual way.

Generally,

- If $x_1, x_2, \ldots, x_n$ is a random sample from a Normal distribution, with mean m and variance $s^2$ and sample size n *large*, can use the **Normal** for sampling distributions of means, with the

  standardised Normal $$U = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}$$

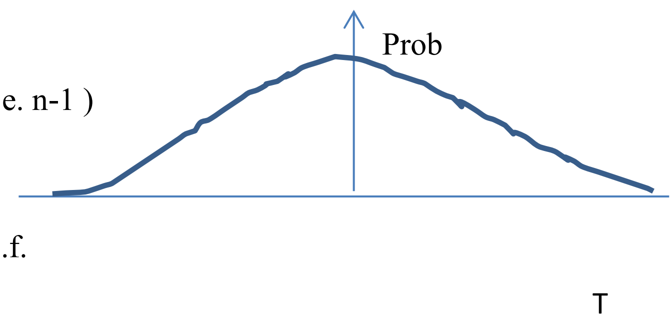
- If $x_1, x_2, \ldots, x_n$ is a random sample from a Normal distribution, with mean m and variance $s^2$ and if we define $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

  i.e. **Estimated Sample variance -** see Statistical tables

  then $$T = \frac{(\bar{x}-\mu)}{s/\sqrt{n}}$$ has a $t_{n-1}$ distribution

  where the denominator is the standard error and the d.o.f. for $t_{n-1}$ come from the estimated sample variance (i.e. n-1 )

  Note: Family of t-distributions, with members labelled by d.o.f.




**Chi-Square**

A random variable $X$ with a Chi-square distribution with $\nu$ degrees of freedom ($\nu$ a positive integer, related to sample size n) [Expectation and variance in tables] has the following important features:

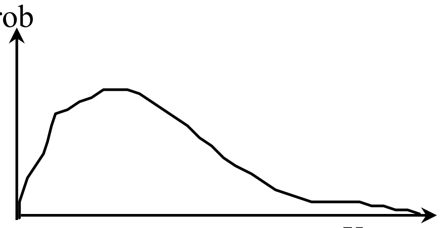
- If $X_1, X_2, \ldots, X_n$ are **Standardised Normal** Random Variables,

  $X_1^2 + X_2^2 + \ldots + X_n^2 \sim \chi^2$ **distribution** (on $n$ degrees of freedom)

- So, if $x_1, x_2, \ldots, x_n$ is a *random sample* of values for random variable X ~ N($\mu$ ,$s^2$), then

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = U \sim N(0,1) \qquad\qquad s^2 \sim \chi^2_{n-1}$$

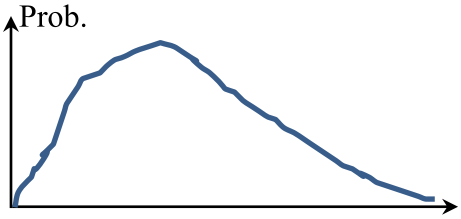
for sample mean $\bar{x}$, population mean μ ; sample variance as before, population variance $\sigma^2$ and where the denominator for the U transform is again the standard error (*standard deviation of the distribution for the* ***mean***).




Note: squared distribution, so all positive values.
d.o.f. again depend on sample size as $s^2$ (based on n-1)
: usually have to estimate population variance by sample variance.
Again, family of $\chi^2$ distributions, labelled by d.o.f.

**F-distribution**

A random variable V has an *F* distribution with *m* and *n* degrees of freedom if it has a distribution of form shown
[Expectation and variance – see tables]
Illustrated is $F_{m,n}$




For *X* and *Y* *independent* random variables, s.t. $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ then F = ratio of chi-squareds

$$F_{m,n} = \frac{X/m}{Y/n}$$

One consequence: *if* $x_1, x_2, \ldots, x_m$ $(m > 2)$ is a random sample from $N(\mu_1, \sigma_1^2)$, and $y_1, y_2, \ldots, y_n$ $(n > 2)$ a random sample from $N(\mu_2, \sigma_2^2)$, then

$$\frac{\sum (x_i-\bar{x})^2/(m-1)}{\sum (y_i-\bar{y})^2/(n-1)} \sim F_{m-1,\,n-1}$$

in other words can **compare two variances** using the F-distribution.

# 5.3 Sampling Theory

Simple random sample assumes *equal probability* of item selection. If the same element can not be selected more than once, then say that the sample is drawn **without replacement**; otherwise, the sample is said to be drawn **with replacement**.

The usual convention is to use lower case letters s, x, n are used for the sample characteristics with capital letters S, X, N or greek letters for the parent population. Thus, if sample size is n, its elements are designated, $x_1$, $x_2$, …, $x_n$, its mean is $\bar{x}$ and its variance is

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

If repeatedly draw random samples, size n (with replacement) from a population distributed with mean μ and variance $s^2$, then $\bar{x}_1$, $\bar{x}_2$ … is the set of sample means and

$$U = \frac{\bar{x}_j - \mu}{\sigma / \sqrt{n}} \qquad j = 1,2,3,....$$

is the *sampling distribution of the means*, i.e the standard Normal, with denominator equal to the standard error

Note: This important theorem underpins Statistical Inference and the importance of the Normal distribution for inferences made.

**Central Limit Theorem.**

In the limit, as sample size n tends to infinity, the *sampling distribution of means* has a **Standard Normal** distribution.

**Attribute and Proportionate Sampling**

- If sample elements are *measurements* = **attribute sampling**. If all sample elements are *1 or 0 (success/failure, agree/disagree)* = **proportionate sampling**.
- Sample average $\bar{x}$ and sample proportion p are handled in the same way, replacing the mean and its standard error in the U transform by the proportion and its standard error in the latter case.

Note 1: We can generalise the concept of the sampling distribution of means for the sampling distribution of *any statistic*, but may need to change distribution.

- We say that the sample characteristic is an *unbiased* estimator of the parent population characteristic if *the expectation* (average) of the corresponding sampling distribution equals the parent characteristic, (see previously).

So

$E\{\bar{x}\} = \mu$ ; $E[p] = P$ ; $E\{s^2\} \approx \sigma^2$ [ actually = $(n-1)\sigma^2$ ]

where these refer to sample and population mean, sample and population proportion and sample and population variance respectively.

Note 2:

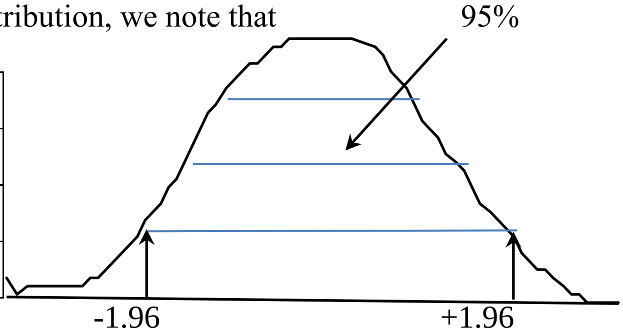
Although the *population* may be very large, (effectively infinite), in which case, we can use the expressions as discussed, if it is *small relative to sample size*, we make a correction.
Denoting size of finite population 'N'
The quantity $\sqrt{[(N - n) / (N - 1)]}$ is the **finite population correction (fpc)**

- If sample size n **large** relative to population size, **n> 0.05N**, we **should use fpc**. so e.g. $E\{s\} = \sigma \times$ fpc for estimated sample S.D. (with fpc needed).
- If sample size n **small** relative to population size N , i.e. n **< 0.05N** (or we have sampling *with replacement* ), then effectively **fpc = 1**, i.e. no correction made.

## 5.4 Estimation : Confidence Intervals

From the statistical tables for a Standard Normal distribution, we note that

| | From U= | To U = |
|---|---|---|
| 0.90 | -1.64 | +1.64 |
| 0.95 | -1.96 | +1.96 |
| 0.99 | -2.58 | +2.58 |




From *central limit theorem*, if $\bar{x}$ and $s^2$ are mean and variance of a random sample of size n, (**n > 25**), drawn from a **large population,** then use Normal distribution and the following statement applies w.r.t unknown population mean μ

$$PROB\{-1.64 \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq +1.64\} = 0.90$$

where U = term between inequalities

and if population σ *unknown*, can replace by s without needing another distribution, so, re-arrange to get

$$PROB\{\bar{x} - 1.64\ s/\sqrt{n} \leq \mu \leq \bar{x} + 1.64\ s/\sqrt{n}\} = 0.90$$

So, the range $\bar{x} \pm 1.64\ s/\sqrt{n}$ is called a **90% confidence interval** for population mean μ.

Example 1:

A random sample of **size 25** weights of bales from a production process has mean $\bar{x}$ = 15 kg. and standard deviation (s) = 2 kg. Find a **95% confidence interval** for the population mean weight μ of bales in the production process.

Solution:

**Note**: - **Large sample** – use **Normal** (i.e. can use s for σ without needing to change distribution)

- **Large parent population** relative to sample – do *not* have to worry about **fpc**.

Then a **95% confidence interval** for the mean weight (μ) of bales in the production process as a whole, is $\bar{x} \pm 1.96\ s/\sqrt{n}$ i.e. 95% of Normal distribution falls between

-1.96 and +1.96, so substituting for sample mean and S.D. here gives

$15 \pm 1.96\left(\frac{2}{5}\right)$ (i.e.) 95% confidence interval is
14.22 kg. to 15.78 kg.

This **can be interpreted** as : ‘ 95% of the time the mean weight of bales in the production process will fall between 14.22 kg. and 15.78 kg.’

Example 2:

An opinion poll is taken to determine the proportion of voters who are likely to vote for a given political party in a pending election.
A random sample of size n = 1000 indicates that this proportion $p = 0.40$.
Obtain (i) **90%** and (ii) **95% confidence intervals** of reliability of the poll.

Solution:

**Note**: - Proportion – dealt with in the same way as the mean.

- **Large sample** – use **Normal** (i.e. can use sample standard deviation for the population standard deviation again)
- **Large parent population** – do *not* have to worry about **fpc**.
- **Standard error** of the mean has general form $s/\sqrt{n}$ as we have seen.

For a proportion, we also had S.E. of a proportion, (parallelling the case for a mean), is the S.D. of the sampling distribution of proportions
so just replace ‘**s**’ by $\sqrt{pq}$ , so S.E. (proportion) = $\sqrt{pq/n}$, where $q = 1 - p$ as usual

(i) So for **90% Confidence Interval**

Substitute sample values in
$$p \pm 1.64\sqrt{pq/n}$$
$$= 0.40 \pm 1.64\sqrt{(0.40)(0.60)/1000}$$

(ii) For **95% Confidence Interval**

Substitute sample values in
$$p \pm 1.96\sqrt{pq/n}$$
$$= 0.40 \pm 1.96\sqrt{(0.40)(0.60)/1000}$$

So, a **90% C.I.** for population proportion, P is $0.40 \pm 0.025 = 0.375$ to $0.425$ and a **95%** confidence interval for P is $0.40 \pm 0.030 = 0.37$ to $0.43$.

Note: for n = 1000, $1.96 \sqrt{[p(1-p)]/n} \gg$ **0.03** for values of p between 0.3 and 0.7. This is the basis for statement that public opinion polls have an **"inherent error of 3%".** This simplifies calculations in the case of e.g. opinion polls for large political parties

**Small Samples**

Summary

For reference purposes, it is useful to regard the expression $\bar{x} \pm 1.96 \, s/\sqrt{n}$
as the "default formula" for a confidence interval and modify it to suit circumstances.

- If dealing with proportionate sampling, use the Normal: the sample proportion is the sample 'mean' $\bar{x} \rightarrow p$ and the **standard error** (s.e.) term $s/\sqrt{n}$

  simplifies as: $$s/\sqrt{n} \rightarrow \sqrt{p(1-p)/n} = \sqrt{pq/n}$$

- Changing from 95% to 90% confidence interval (say) in Normal $\rightarrow$swap 1.96 for 1.64.
- If n < 25, the **Normal** distribution is replaced by **Student's $t_{n-1}$** distribution.
- In sampling *without replacement* from a **finite** population, (i.e. if sample size more than 5% of population size, then **fpc** term used).

Note: width of the confidence interval **increases** with confidence level.

Example 3.

A random sample, size n = 10, drawn from a large parent population, has mean $\bar{x} = 12$, standard deviation, s = 2. Obtain 99% and 95% C.I. for population mean μ

Solution:

- **Small sample** – can not use **Normal** (must use **t distribution**, with n-1 = **9 d.o.f.** here. Values between which 95% of population falls (see t-distribution tables) are $\pm 2.262$.

  Values between which 99% of population falls = ±3.25)
- **Large parent population** – do *not* have to worry about **fpc**.

Then **99% C.I.** for parent population mean is $\bar{x} \pm 3.25 \, {}^{s}\!/\!\sqrt{n} = 12 \pm 3.25\left({}^{2}\!/\!{}_{3}\right)$

so from 9.83 to 14.17

and a **95% confidence interval** for the parent mean is $\bar{x} \pm 2.262 \, {}^{s}\!/\!\sqrt{n} = 12 \pm 2.262\left({}^{2}\!/\!{}_{3}\right)$

so from 10.492 to 13.508.

Example 4.

A department store chain has 10,000 credit card holders, who are billed monthly for purchases. The company want to take a sample of these credit card customers to determine average amount spent each month by all those holding credit cards. A random sample of 25 credit card holders was selected and the sample average was €75, with a sample standard deviation of €20.

(i) Obtain a 95% Confidence interval using the *Normal distribution*.

(ii) Obtain a 95% confidence interval using the *t-distribution*.

Solution

(i) Referring to the distribution diagram at the top of this section (or to standard Normal distribution tables direcly), 95% of the distribution lies between ±1.96 of the mean

The 95% confidence interval is given by $\bar{x} \pm 1.96 \, {}^{s}\!/\!\sqrt{n}$

$= 75 \pm 1.96\left({}^{20}\!/\!\sqrt{25}\right)$

= **74.43** to **76.57**

The situation now is that we must use the t-distribution on n-1 d.o.f. (= 25-1 = 24). The picture is very similar to (i) , but the distribution is a bit flatter and we must replace values of ±1.96 (for the Normal) by those that apply for $t_{24}$ . 95% of the $t_{24}$ distribution, (see tables), falls between ± 2.064.

The 95% confidence interval now is: $\bar{x} \pm 2.064 \, {}^{s}\!/\!\sqrt{n}$

$= 75 \pm 2.064\left({}^{20}\!/\!\sqrt{25}\right)$

= **73.35** to **76.65**

Note**:**

Since sample size (n) is 25, these are *very close*, as expected from the **summary points** previously.

The t-interval remains slightly wider (i.e. slightly more **conservative**).

Example 5 **:** on finite population correction, (fpc).

For a product, package weight is to be checked as part of the quality control. A sample of 80 is drawn at random from a batch of 100. The sample mean was found to be 25g. and the standard deviation was found to be 6g.

(i) What is the finite population correction factor here?

(ii) What effect does it have on a **99% confidence interval** for the mean?

Solution

Note: Sample size (n) **large** relative to **finite size** of population, (N), i.e. **sample size more than 5% of population size** (n > 0.05N), so the fpc applies

(i) The correction factor is $\sqrt{(N-n)/(N-1)} = \sqrt{(100-80)/(100-1)} = 0.449$

(ii)

Usual Standard Error is $s/\sqrt{n} = \left(6/\sqrt{80}\right) = 0.671$

S.E. with fpc applied is: $s/\sqrt{n}\left(\sqrt{(N-n)/(N-1)}\right) = (0.671)(0.449) = 0.301$

**99% Confidence Interval** for usual standard error is then

$$\bar{x} \pm 2.58\, s/\sqrt{n} = 25 \pm 2.58\left(6/\sqrt{80}\right) = 25 \pm 2.58\,(0.671) \qquad \text{i.e.} = 23.27 \text{ to } 26.73$$

**99% Confidence Interval** for S.E, *with fpc applied* is

$$\bar{x} \pm 2.58\, s/\sqrt{n}\left(\sqrt{(N-n)/(N-1)}\right) = 25 \pm 2.58\,(0.671)(0.449) \qquad \text{i.e.} = 24.22 \text{ to } 25.78$$

Note: So, the precision of the *sample estimate*, measured by the **standard error**, is determined not only by the absolute ***size*** of the sample, but also to an extent by the *proportion of the population sampled*.

## 5.5 Hypothesis Testing Summarised (Complement to Confidence Intervals)

Example 6.

The average grade of all 19 year old students for a particular aptitude test is thought to be 60%. A random sample of n= 49 students gives mean $\bar{x}$ = 55% with S.D. s = 2%. Is the sample result consistent with the claim?

Original claim is the **null hypothesis ($H_0$)**

$$H_0 : \mu = 60. \quad \text{Alternative then is } H_1 : \mu \neq 60.$$

If true, **test statistic** $U = \dfrac{(\bar{x}-\mu)}{s/\sqrt{n}}$ follows the standard Normal as usual, (see distribution as before, (beginning previous section), or from Normal tables directly. *Rejection regions* for the null hypothesis are regions outside ±1.96, i.e. < -1.96 or > +1.96, (*complementary* to Confidence Intervals, which define the *acceptance region* for the null hypothesis).

Substituting $$U = \frac{(55-60)}{2/\sqrt{49}} = -17.5$$

This lies *well outside* the 95% confidence interval (i.e. falls in the **rejection region)**, so either

(i) The null hypothesis is *incorrect*

or (ii) An event with probability of *at most* 0.05 has occurred (size of rejection region)

Hence, **Reject** $H_0$, knowing probability of 0.05 exists that we are in error. Technically, we say we reject the null hypothesis at the 0.05 ***level of significance ($\alpha$).***

Note: The level of significance reflects the amount of *risk* that a decision-maker is willing to take of being wrong, i.e. of rejecting the null hypothesis (that things are working fine), in favour of an alternative – that they are not. Obviously, if the wrong decision is made, this has cost implications.

**Steps in Hypothesis Testing**

1. State the **null hypothesis, $H_0$** (operation as usual)
2. State the **alternative hypothesis, $H_1$** (operation not as usual)
3. Specify the **level of significance**, $\alpha$, (i.e. risk prepared to take of being wrong in decision on accepting )
4. <u>Note:</u> the *sample size n* , *what is known* about sample and population, and what interested in measuring; hence **decide on distribution**, setting up the ***critical values*** that determine rejection and acceptance regions, based on $\alpha$ chosen
5. Determine the Test Transformation statistic (**Test Statistic**, i.e. U, T, $\chi^2$,F)
6. **Compare** value from sample calculation against critical values dividing acceptance/rejection regions.
7. <u>**Decision Rule**</u>: If the value of the Test Statistic, based on sample data, falls into *non-rejection region*, we have no evidence against $H_0$ , so **Accept $H_0$**

   If value of Test Statistic *does* fall into *rejection region*, the sample provides evidence against $H_0$ being true, so **Reject $H_0$**
8. Express result in terms of problem, giving *the risk* of getting decision on $H_0$ wrong, (i.e. the level of significance, $\alpha$ ).

   (*e.g.* if 95% confident, then 5% not confident, so level of significance $\alpha$ is 5%

   ($\alpha$ has probability = 0.05).

One –sided Tests: One and Two sample H.T. and further Examples_05B