Statistics II
Distribution of the sample means of random variable *X*

Two examples of random variable *X*:
The IQ of a person in a population (e.g. the Perth population)
The length of a fish in a lake

Population:
Consider *X*, a random variable in the population, with *any type* of probability distribution.
Population mean (μ) of *X*, and population standard deviation (σ) of *X* are known constants for a particular population

Sampling: Take random samples *A*, *B*, *C*, … of the same size n from the population.

Consider random sample *A* $x_{A1}\ ,\ x_{A2}\ ,\ x_{A3}\ ,\ .....\ ,\ x_{An}$

In the random sample, the mean value of the content is called the sample mean of *A* and it is denoted by $\bar{x}_A$.

It is calculated by $\bar{x}_A = \dfrac{x_{A1} + x_{A2} + \cdots + x_{An}}{n}$.

In random sample *B*, $\bar{x}_B = \dfrac{x_{B1} + x_{B2} + \cdots + x_{Bn}}{n}$, etc.

The sample mean depends on the content of a sample.
.: it varies from sample to sample, .: we can consider sample mean as a random variable denoted by $\overline{X}$ having values $\bar{x}_A$, $\bar{x}_B$, … etc.

Notations: $\overline{X}$ represents the sample mean random variable and $\bar{x}$ represents the value of $\overline{X}$ for a random sample.

Irrespective of the distribution of *X* in the population, if the sample size n is large enough,

(1) the distribution of the sample mean $\overline{X}$ is approximately *normal*

(2) the mean of $\overline{X}$ is given by $\mathrm{E}(\overline{X})$ which is exactly equal to the population mean μ

(3) the standard deviation of $\overline{X}$ is given by $\mathrm{sd}(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of *X* in the population

In the above discussion, the population constants μ and σ of *X* are known. Use them to find $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{sd}(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$, and hence the probability of $\overline{X}$ in the random samples,
e.g. $\Pr(\overline{X} > x)$ using the normal distribution if n is large.

Inference about the population from a random sample

Very often we don't know what the population μ and σ are.
To learn about the population we take a random sample of size *n* from it. $\bar{x}$ of the random sample can then be used as an estimate of the population mean μ, i.e. we carry out sampling and make **inference** about the population from a random sample.

Consider random sample $x_1\ ,\ x_2\ ,\ x_3\ ,\ .....\ ,\ x_n$

Calculate the sample mean (if it is not known): $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

The value of $\bar{x}$ is a reasonable estimate of the population mean μ of random variable *X*. The larger the sample size *n*, the better is the estimation. $\bar{x}$ is called a **point estimator** of μ.

A better alternative to $\bar{x}$ as an estimator of μ is to give an interval of *X* values that we are 95% sure contains the population mean μ. This interval is called a 95% **confidence interval** for μ.

Its calculation is $\left(\bar{x} - 1.96\dfrac{\sigma}{\sqrt{n}},\ \ \bar{x} + 1.96\dfrac{\sigma}{\sqrt{n}}\right)$ approximately.

Interpretation of 95% confidence interval: About 95 out of 100 *confidence intervals of X values* calculated from the random samples contain the population μ.

The 99% confidence interval is $\left(\bar{x} - 2.85\dfrac{\sigma}{\sqrt{n}},\ \ \bar{x} + 2.85\dfrac{\sigma}{\sqrt{n}}\right)$.

Note 1: The larger the sample size *n*, the distribution is closer to normal and .: the confidence interval is more precise.

Note 2:
$\Pr(-1.96 < Z < 1.96) \approx \dfrac{95}{100}$, $\Pr(-2.85 < Z < 2.85) \approx \dfrac{99}{100}$
where random variable Z has a standard normal distribution.

Note 3: In general, an *A*% confidence interval is approximately $\left(\bar{x} - z\dfrac{\sigma}{\sqrt{n}},\ \ \bar{x} + z\dfrac{\sigma}{\sqrt{n}}\right)$ where $\Pr(-z < Z < z) \approx \dfrac{A}{100}$, $z > 0$

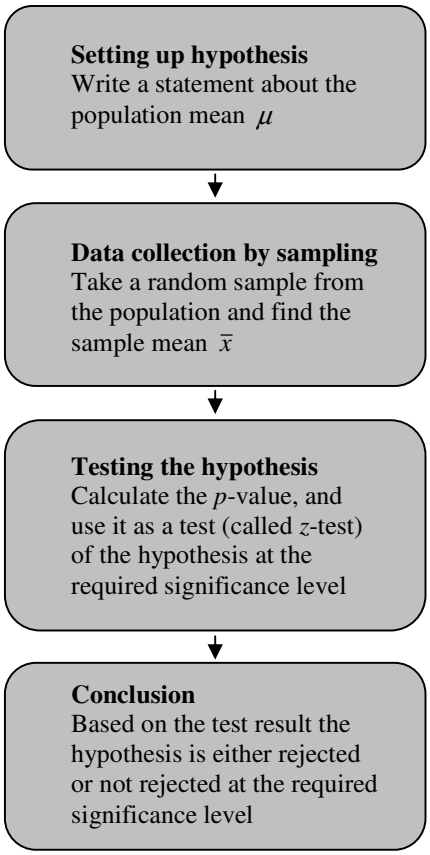
Note 4:
For constant sample size *n*, the higher the required confidence level, the wider is the required interval.

For a required confidence level, the larger the sample size *n*, the narrower is the required interval.

However, the population standard deviation σ is generally unknown.
In the absence of any other information, **the sample standard deviation** s can be used instead of σ in the calculation, e.g.
$\left(\bar{x} - 1.96\dfrac{s}{\sqrt{n}},\ \ \bar{x} + 1.96\dfrac{s}{\sqrt{n}}\right)$ for the 95% confidence interval.

Statistical Testing

Here we concentrate on testing of the population mean.
The sequence in statistical testing:



Setting up hypothesis

Over time a population may change.
A **hypothesis**: *The population mean remains the same.*
A 'no change' hypothesis is called a null hypothesis.
It is denoted by $\mathbf{H}_0$.
Alternative hypothesis: *The population mean is less than before.*
The alternative hypothesis is denoted by $\mathbf{H}_1$.

For example, suppose $\mu = c$ originally,
then $\mathbf{H}_0 : \mu = c$, $\mathbf{H}_1 : \mu < c$

Data collection by taking a random sample of size *n*

Calculate the sample mean $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ if it is not known.

The sample mean random variable $\overline{X}$ has a normal distribution (or a distribution approximately normal), then

$$\Pr\left(\overline{X} \le \bar{x} \mid \mu = c\right) = \Pr\left(Z \le \frac{\bar{x} - c}{\mathrm{sd}\left(\overline{X}\right)}\right) \text{ or } \mathrm{normalcdf}\left(-\infty, \bar{x}, c, \mathrm{sd}\left(\overline{X}\right)\right)$$

If the alternative hypothesis is: *The population mean is greater than before,* then $\mathbf{H}_0 : \mu = c$, $\mathbf{H}_1 : \mu > c$

$$\Pr\left(\overline{X} \ge \bar{x} \mid \mu = c\right) = \Pr\left(Z \ge \frac{\bar{x} - c}{\mathrm{sd}\left(\overline{X}\right)}\right) \text{ or } \mathrm{normalcdf}\left(\bar{x}, \infty, c, \mathrm{sd}\left(\overline{X}\right)\right)$$

The value of each probability above is called the ***p*-value** of the test.

Testing the hypothesis

p-value is used to *decide* whether the null hypothesis $\mathbf{H}_0$ is to be rejected or not rejected. The test itself is called a ***z*-test**.

By convention,
p-value > 0.05 means insufficient evidence against $\mathbf{H}_0$
p-value < 0.05 means good evidence against $\mathbf{H}_0$
p-value < 0.01 means strong evidence against $\mathbf{H}_0$
p-value < 0.001 means very strong evidence against $\mathbf{H}_0$

Conclusion

If the null hypothesis $\mathbf{H}_0$ is rejected in favour of the alternative hypothesis $\mathbf{H}_1$ when the *p*-value < 0.05 (the condition for rejection), we say $\mathbf{H}_0$ is rejected at the 0.05 significance level.

0.05 (5%) is the **most commonly used value** for significance level. Other levels like 0.01 (1%) and 0.001 (0.1%) are sometimes used.

One-tail test and two-tail test

For $\mathbf{H}_0 : \mu = c$, $\mathbf{H}_1 : \mu < c$ OR $\mathbf{H}_0 : \mu = c$, $\mathbf{H}_1 : \mu > c$
a one-tail test is applied to test the hypothesis as discussed above

For $\mathbf{H}_0 : \mu = c$, $\mathbf{H}_1 : \mu \ne c$, a two-tail test is applied to test the hypothesis.

The difference between the two tests is in the *p*-values.
p-value (two-tail test) $= 2 \times$ *p*-value (one-tail test)

When the *p*-value (two-tail test) < 0.05, we say $\mathbf{H}_0$ is rejected at the 0.05 significance level.

A 95% confidence interval for the population mean μ, given by $\left(\bar{x} - 1.96\dfrac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\dfrac{\sigma}{\sqrt{n}}\right)$, can also be used as a two-tail test for the hypothesis.

If $\bar{x} - 1.96\dfrac{\sigma}{\sqrt{n}} < c < \bar{x} + 1.96\dfrac{\sigma}{\sqrt{n}}$, $\mathbf{H}_0$ is *not* rejected, otherwise it would be rejected at the 5% significance level.

Note: Experience tells whether $\mu < c$, $\mu > c$ or $\mu \ne c$, hence dictates the decision to do a one-tail test or a two-tail test.

Types of error in statistical testing

Type I error: $\mathbf{H}_0$ is rejected when $\mathbf{H}_0$ is true
Type II error: $\mathbf{H}_0$ is not rejected when $\mathbf{H}_0$ is not true

	$\mathbf{H}_0$ is true	$\mathbf{H}_0$ is not true
$\mathbf{H}_0$ is rejected	*Type I error*	*Correct decision*
$\mathbf{H}_0$ is not rejected	*Correct decision*	*Type II error*

Note 1: A type I error is viewed to be more serious than a type II error.

Note 2: For a good statistical test, we would want the probabilities of making type I and type II errors to be small.

Note 3: The probability of making type I error can be reduced by increasing the significance level. However, it will increase the probability of making type II error.
Using 0.05 as the significance level is a compromise.

Example 1 (2016 VCAA SM Sample Exam 2 SECTION A Q 19)

The mean study score for a large VCE study is 30 with a standard deviation of 7. A class of 20 students may be considered as a random sample drawn from this cohort.
The probability that the class mean for the group of 20 exceeds 32 is

A. 0.1007
B. 0.3875
C. 0.3993
D. 0.6125
E. 0.8993

$\text{sd}(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{20}} \approx 1.56525$, $\mu = 30$

Normal: $\Pr(\overline{X} > 32) \approx 0.1007$ A

Example 2 (2016 VCAA SM Sample Exam 1 Q3b, modified)

A coffee machine dispenses volumes of coffee that are normally distributed. The owner has been told that the mean volume is 240 mL. The owner is concerned that the coffee machine is, on average, dispensing less coffee than the mean 240 mL.
A sample of 16 cups of coffee (with no milk) is dispensed and it is found that the mean volume of all coffees served in this sample is 235 mL.
Assume that the population standard deviation of 8 mL.

i. State appropriate null and alternative hypotheses for the volume V in this situation. 1 mark

ii. The p value for this test is given by the expression $\Pr(Z \le a)$, where Z has the standard normal distribution. Find the value of a and hence determine whether the null hypothesis should be rejected at the 0.05 level of significance. 2 marks

Solutions
i Null hypothesis: The second machine is, on average, dispensing **not** less coffee than the first.
OR $\mathbf{H}_0$: $\mu = 240$ mL

Alternative hypothesis: The second machine is, on average, dispensing less coffee than the first.
OR $\mathbf{H}_1$: $\mu < 240$ mL

ii $\text{sd}(\overline{X}) = \frac{8}{\sqrt{16}} = 2$, $a = \frac{235 - 240}{2} = -2.5$,

$p = \Pr(Z \le -2.5) \approx 0.0062$

Since $p < 0.05$, the null hypothesis should be rejected at the 0.05 significance level.

Example 3 (2016 VCAA SM Sample Exam 2 SECTION B Q6)

A certain type of computer, once fully charged, is claimed by the manufacturer to have $\mu = 10$ hours lifetime before a recharge is needed. When checked, a random sample of $n = 25$ such computers is found to have an average lifetime of $x = 9.7$ hours and a standard deviation of $s = 1$ hour. To decide whether the information gained from the sample is consistent with the claim $\mu = 10$, a statistical test is to be carried out. Assume that the distribution of lifetimes is normal and that s is a sufficiently accurate estimate of the population (of lifetimes) standard deviation σ.

a. Write down suitable hypotheses H_0 and H_1 to test whether the mean lifetime is less than that claimed by the manufacturer. 2 marks

b. Find the p value for this test, correct to three decimal places. 2 marks

c. State with a reason whether H_0 should be rejected or not rejected at the 5% significance level. 1 mark

Let the random variable X denote the mean lifetime of a random sample of 25 computers, assuming $\mu = 10$.

d. Find C^* such that $\Pr(\overline{X} < C^* \mid \mu = 10) = 0.05$. Give your answer correct to three decimal places. 2 marks

e. i. If the mean lifetime of all computers is in fact $\mu = 9.5$ hours, find $\Pr(\overline{X} > C^* \mid \mu = 9.5)$, giving your answer correct to three decimal places, where C* is your answer to part d. 2 marks

e ii. Does the result in part e.i. indicate a type I or type II error? Explain your answer. 1 mark

Solutions
a H_0: The mean lifetime is that claimed by the manufacturer. ($\mu = 10$)

H_1: The mean lifetime is less than that claimed by the manufacturer. ($\mu < 10$)

b $\sigma \approx s = 1$, $\text{sd}(\overline{X}) = \frac{\sigma}{\sqrt{n}} \approx \frac{1}{\sqrt{25}} = 0.2$, $\text{E}(\overline{X}) = \mu = 10$

$p\text{-value} = \Pr(\overline{X} \le 9.7 \mid \mu = 10) = \Pr\left(Z \le \frac{9.7 - 10}{0.2}\right) \approx 0.067$

c Since $p\text{-value} > 0.05$, $\therefore$ H_0 should not be rejected at the 5% level of significance.

d $\Pr(\overline{X} < C^* \mid \mu = 10) = 0.05$, $\Pr\left(Z < \frac{C^* - 10}{0.2}\right) = 0.05$

$\therefore \frac{C^* - 10}{0.2} \approx -1.64485$, $C^* \approx 9.671$

e i $\Pr(\overline{X} > 9.671 \mid \mu = 9.5) = \Pr\left(Z > \frac{9.671 - 9.5}{0.2}\right)$

$\approx \Pr(Z > 0.855) \approx 0.196$

e ii Type II error:
$p\text{-value} \approx 0.196 > 0.05$, it supports that the actual mean lifetime is $\mu = 9.5$, $\therefore$ H_0 is not true but it is not rejected.