MODULE 5

Hypothesis testing is a statistical procedure that uses sample data to evaluate an assumption about a population parameter.

Statistical significance is the claim that the results of a test or experiment are not explainable by chance alone.

Steps for performing a hypothesis test:

1. **State the null hypothesis (H_0) and the alternative hypothesis (H_a)**
    - **Null hypothesis (H_0)**
        - Statement that is assumed to be true unless there's convincing evidence to the contrary.
        - Often include phrases such as "no effect", "no difference", "no relationship", or "no change"
        - Equality symbols (=, ≤, ≥)
    - **Alternative hypothesis (H_a)**
        - Statement that contradicts the null hypothesis.
        - Often include phrases such as "an effect", "a difference", "a relationship", or "a change"
        - Inequality symbols (≠, <, >)
2. **Choose a significance level**
    - **Significance level** is the threshold at which you will consider the result statistically significant
    - Typically data professional uses significance level at 5%
3. **Find the p-value**
    - **p-value** refers to the probability of observing results as or more extreme than those observed when the null hypothesis is true.
4. **Reject or fail to reject the null hypothesis**
    - **Reject** when: $p - value < significance\ level$
    - **Fail to reject** when: $p - value > significance\ level$
    - The decision also depends on chosen significance level.

Statistically significant results cannot prove with 100% certainty that our hypothesis is correct because hypothesis testing is based on probability. There is always a chance of drawing the wrong conclusion about the null hypothesis.

Two types of error when drawing a conclusion:

1. **Type I error** (False positive)
    - Reject a null hypothesis that is actually true (the result is statistically significant when in fact it occurred by chance)
    - Choose a lower significance level to reduce type I error
2. **Type II error** (False negative)
    - Fail to reject a null hypothesis, which is actually false (conclude the result occurred by chance when it's in fact statistically significant)
    - Choosing a lower significance level increase the likelihood of type II error

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True Negative)	Type II Error (False Negative)

Two types of hypothesis test:

1. **One sample**
   Determines whether or not a population parameter such as mean or proportion is equal to a specific value
2. **Two sample**
   Determines whether or not two population parameters such as two means or two proportion are equal to each other

Common types of hypothesis test:

- z-tests
- t-tests

One sample z-test:

- Assumes the data is a random sample of a normally-distributed population, the population standard deviation is known.
- The test statistic is z-score
- $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

Two-sample mean test

- Frequently used in A/B testing
- Two-sample t-test is the standard approach for comparing two means. It assumes that the two samples are independent of each other and the population standard deviation is unknown.
- The test statistic is t-score
- $$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

Typically data professionals uses:

- z-test when the population standard deviation is known
- t-test when the population standard deviation is unknown.

The population standard deviation is usually unknown because it's difficult to get complete data on large populations.

Two-sample proportion test

Data professional can do z-test, since t-test do not apply to proportions

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 (1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Where: p̂ = sample proportion, n = number of sample

There are at least 3 steps in designing experiment:

1. Define the variables
   Independent variable or dependent variable
2. Formulate hypothesis
   Null and alternate hypothesis
3. Assign test subjects to treatment and control groups
   Treatment group – exposed to the treatment
   Control group – not exposed to the treatment