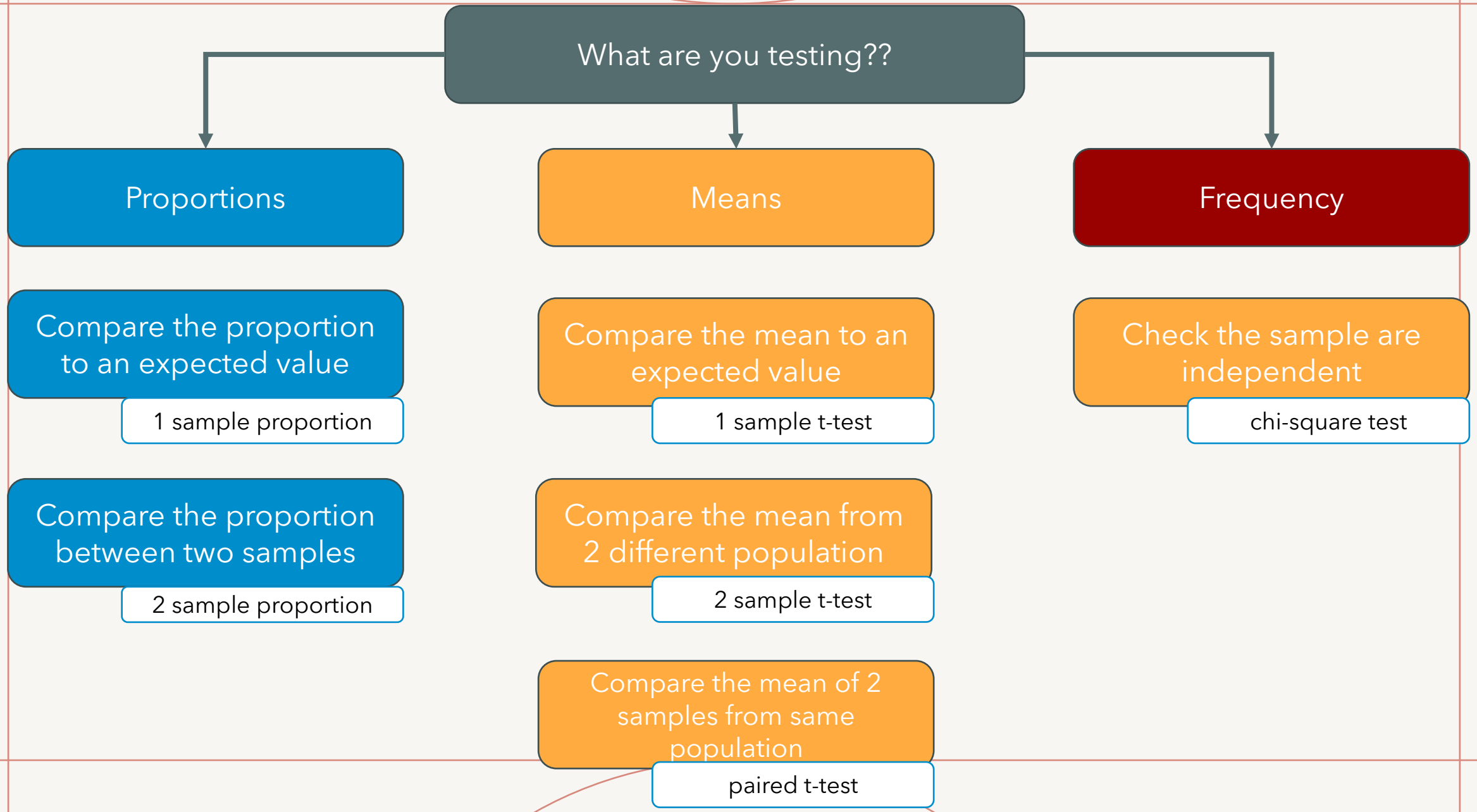
Week 9

Statistical Inference for Two Samples

Recap on week 8

- Statistical inference
- Hypothesis testing
- Significance test for one sample


What are you testing??
Proportions
Means
Frequency
Compare the proportion to an expected value
1 sample proportion
Compare the proportion between two samples
2 sample proportion
Compare the mean to an expected value
1 sample t-test
Compare the mean from 2 different population
2 sample t-test
Compare the mean of 2 samples from same population
paired t-test
Check the sample are independent
chi-square test

Statistical inference for two samples

- Independence Sample T-Test
- Paired T-Test
- 2 samples proportion
- Chi-Square Test

Paired T-Test

- It is commonly use to compare the same data but two different periods (before / after)
- The hypothesis will be stated by the difference between two periods
- Null hypothesis can be stated as: $H_0: \mu_d = 0$
- Alternative hypothesis can be stated as:

$H_a: \mu_d < 0$ $H_a: \mu_d \neq 0$ $H_a: \mu_d > 0$

Paired T-Test example

Swimming case: The professional coach try to train athletic with new technology by testing pre-and-post result. They try to collect the time speed of 50M freestyle from 10 athletics and the results is provided as the following table:

Pre	30	31	34	40	36	35	34	30	38	39
Post	30	31	32	38	32	31	32	29	28	30

Test at 5% significance level and interpret the pre-and-post result.

Paired T-Test in python

```
import scipy.stats as stats
alpha = 0.05
pre = [30, 31, 34, 40, 36, 35, 34, 30, 38, 39]
post = [30, 31, 32, 38, 32, 31, 32, 29, 28, 30]

t_value, p_value = stats.ttest_rel(pre, post)
print("T-value: ", t_value)
print("P-value: ", p_value)
if p_value < alpha:
        print("Reject Null Hypothesis")
else:
        print("Fail to Reject Null Hypothesis")
```

Paired T-Test practice

Training program: The organization found that their employees lack the functional skill on business intelligence. Then they decide to send their employees to learn on it. However, they would like to test the training course is good or not by testing their employee before and after training. The result is provided on the following table:

Before	67	74	58	45	78	79	61	83	70	69
After	89	75	64	71	80	82	92	81	73	75

Test at 10% significance level and interpret the results

Two samples proportion

- It is commonly to use to test two sample proportions variables
- Normally, the null hypothesis can state as equality of two proportion

$$H_0: p_1 = p_2$$

- An alternative hypothesis has three possibilities

$$H_a: p_1 < p_2 \qquad H_a: p_1 \neq p_2 \qquad H_a: p_1 > p_2$$

Two Sample Proportion Example

Google case: Google is the top search engine company. However, Microsoft Corporation try to develop web browser for search engine. Hence, researcher would like to collect the data from people randomly and it found that 410 of 500 use google search engine, while 370 from 500 use Microsoft search engine. Use five percent significance level to test the difference between both search engine.

Two samples proportion in Python

```
import numpy as np
from statmodels.stats.proportion import proportions_ztest
alpha = 0.05
success_a, size_a = (410, 500)
success_b, size_b = (370, 500)
success = np.array([success_a, success_b])
size = np.array([size_a, size_b])
stat, p_value = proportions_ztest(success, size, alternative = 'two-sided')
print("Z-value: ", stat)
print("P-value: ", p_value)
if p_value < alpha:
        print("Reject Null Hypothesis")
else:
        print("Fail to Reject Null Hypothesis")
```

Two sample proportion practice

Social Media: In the social media platform, TikTok user upload at least 1 millions short video in a day. However, the researcher would like to compare the humor and education video in TikTok platform, and they found that 795 of 1000 are humor and 178 of 500 are education. Test at 10% significance level.

Telecommunication: AIS and TRUE are the top telecommunication company in Thailand. Marketing research company try to collect the data randomly from 1000 people about the experience to use both service provider. However, it found 560 people have experienced to use AIS, whereas 486 people have experienced to use TRUE. Test at 2% significance level.

Chi-Square Test (x^2)

- It is commonly used for testing relationship between categorical variables
- It is also called a "goodness-of-fit" statistics
- The null and alternative hypothesis of Chi-Square test is stated as the following statement

H_0: There is no relationship exist in categorical data

H_a: There is related in the categorical data

Chi-Square test example

IT gadgets case: The IT company wishes to study the relationship between gender and IT gadget consumption. The survey randomly selected people are asked about the usage of IT gadget. The following results are obtained:

	Smart watch	Earbuds	Tablet
Male	63	37	45
Female	26	74	55

Test the hypothesis at 5% significance level

Chi-Square test in Python

```
import numpy as np
from scipy.stats import chi2_contingency

data = [[63, 37, 45], [26, 74, 55]]
stat, p, dof, expected = chi2_contingency(data)
alpha = 0.05
print("P value is " + str(p))
if (p <= alpha):
        print("Reject Null Hypothesis")
else:
        print("Fail to Reject Null Hypothesis")
```

Chi-Square Test practice

Netflix case: Netflix would like to study the relationship between viewership program and viewer age (18 years or less, 19 – 35 years, 36 – 59 years and 60 years or older). A sample of 250 Netflix viewers in each age group is randomly selected and the results was shown in the following table:

	18 years or less	19 – 35 years	36 – 59 years	60 years or older
Series	213	202	154	73
Movie	37	48	96	177

Test the independence between viewer program and age at 5% significance level.