The Chai Point stall at Bengaluru airport estimates that each person visiting the store drinks an average of 1.7 small cups of tea.

Assume a population standard deviation of 0.5 small cups. A sample of 30 customers collected over a few days averaged 1.85 small cups of tea per person.

Test the claim using an appropriate test at an alpha = 0.05 significance value, with a critical z-score value of .

Note: Round off the z-score to two decimal places.

A) The computed z-score is 1.64, and since 1.64 is less than 1.96, the null hypothesis cannot be rejected.

B) The computed z-score is 1.64, and since 1.64 is less than 1.96, the null hypothesis is rejected.

C) The computed z-score is 2.33, and since 1.96 is less than 2.33, the null hypothesis cannot be rejected.

D) The computed z-score is 2.33, and since 1.96 is less than 2.33, the null hypothesis is rejected.

```
## H0:the average small number of cups that are cosnumed equals 1.7
## Ha: not equals 1.7
import pandas as pd
import numpy as np

# z=x-u/std

# z= 1.85-1.7/0.5/

sample_mean=1.85
population_mean=1.7
population_std=0.5
sample_size=30

# Calculate z score

z_score=(sample_mean-population_mean)/(population_std/np.sqrt(sample_size))
z_score=np.round(z_score,2)
print (z_score)
```

```
1.64
```

p_value approach

```
#P(Z< 1.64)
from scipy.stats import norm
1-norm.cdf(z_score)
```

```
np.float64(0.050502583474103746)
```

```
p_value=2*0.05
```

```
p_value
```

```
0.1
```

```
#is 0.1 <0.05?
```

The Zumba trainer claims to the customers, that their new dance routine helps to reduce more weight.

Weight of 8 people were recorded before and after following the new Zumba training for a month:

wt_before = [85, 74, 63.5, 69.4, 71.6, 65,90,78]

wt_after = [82, 71, 64, 65.2, 67.8, 64.7,95,77]

Test the trainer's claim with 90% confidence. Further, what would be the pvalue?

A) P value: 0.854, Customers did not reduce their weight

B) P value: 0.145, Customers did not reduce their weight

C) P value: 0.854, Customers have reduced their weight

D) P value: 0.145, Customers have reduced their weight

```
from os import terminal_size
# Ho: Customer did not reduce their weight average weight before and remains same
# Ha: Customer have reduce their weight ubefore > uafter
from scipy.stats import ttest_rel
wt_before = [85, 74, 63.5, 69.4, 71.6, 65,90,78]
```

```
wt_after = [82, 71, 64, 65.2, 67.8, 64.7,95,77]
alpha=0.10
t_stat,p_value=ttest_rel(wt_before,wt_after)
print(t_stat,p_value)
```

```
1.1421853793555032 0.2909361700265277
```

```
p_value =0.29/2
```

```
p_value
```

```
0.145
```

```
if p_value < alpha:
  print("Reject null hypothesis")
else:
  print("Fail to reject null hypothesis")
```

```
Fail to reject null hypothesis
```

```
from os import terminal_size
# Ho: Customer did not reduce their weight average weight before and remains same
# Ha: Customer have reduce their weight ubefore > uafter
from scipy.stats import ttest_rel
wt_before = [85, 74, 63.5, 69.4, 71.6, 65,90,78]

wt_after = [82, 71, 64, 65.2, 67.8, 64.7,95,77]
alpha=0.10
t_stat,p_value=ttest_rel(wt_before,wt_after,alternative="greater")
print(t_stat,p_value)
```

```
1.1421853793555032 0.14546808501326386
```

The quality assurance department claims that on average the non-fat milk contains more than 190 mg of Calcium per 500 ml packet.

To check this claim 45 packets of milk are collected and the content of calcium is recorded.

Perform an appropriate test to check the claim with a 90% confidence level.

data = [193, 321, 222, 158, 176, 149, 154, 223, 233, 177, 280, 244, 138, 210, 167, 129, 254, 167, 194, 191, 128, 191, 144, 184, 330, 216, 212, 142, 216, 197, 231, 133, 205, 192, 195, 243, 224, 137, 234, 171, 176, 249, 222, 234, 191]

Note: Round off the answer to four decimal places.

A)Test statistic: 1.3689 , Reject null hypothesis

B) Test statistic: 1.3689 , Fail to reject null hypothesis

C) Test statistic: 1.2851, Reject null hypothesis

D) Test statistic: 1.2851 , Fail to reject null hypothesis

```
#Ha: u > 190
#H0 : u <=190

from scipy.stats import ttest_1samp
data=pd.Series([193, 321, 222, 158, 176, 149, 154, 223, 233, 177, 280, 244, 138, 210, 167, 129, 254, 167, 194, 1
print("Obs sample mean",round(data.mean(),2))

t_stat,p_value=ttest_1samp(data,190,alternative="greater")
print(t_stat,p_value)
if p_value < 0.10:
  print("Reject null hypothesis")
else:
  print("Fail to reject null hypothesis")
```

```
Obs sample mean 199.49
1.3689029903414232 0.08898891556150607
Reject null hypothesis
```

Child development researchers studying growth patterns of children collect data on the height of fathers and sons.

Analyse the correlation between the father's height and their son's height using the given data

Father Height = [169.39, 161.91, 159.23, 161.72, 167.52, 152.13, 169.64, 162.56, 154.92, 158.57, 153.17, 159.56, 153.77, 168.02, 157.75, 157.42, 160.65, 160.09, 151.4, 151.05, 136.94, 163.56, 160.39, 146.92, 171.66, 150.48, 158.12, 157.83, 163.99, 164.95]

Son Height = [187.35, 177.8, 181.85, 190.69, 188.07, 168.16, 181.65, 173.94, 174.28, 177.87, 176.01, 185.18, 180.33, 175.85, 178.11, 177.34, 185.46, 173.56, 177.19, 169.02, 157.13, 179.58, 181.05, 169.8, 190.89, 164.82, 175.32, 173.69, 185.73, 185.29]

A) Negative Correlation

B) Positive Correlation

C) No Correlation

D) Correlation coefficient > 1

```
Father_Height = [169.39, 161.91, 159.23, 161.72, 167.52, 152.13, 169.64, 162.56, 154.92, 158.57, 153.17, 159.56,
```

```
Son_Height = [187.35, 177.8, 181.85, 190.69, 188.07, 168.16, 181.65, 173.94, 174.28, 177.87, 176.01, 185.18, 180
```

```
df=pd.DataFrame({"Father_Height":Father_Height,"Son_Height":Son_Height})
```

```
df.head()
```

	Father_Height	Son_Height
0	169.39	187.35
1	161.91	177.80
2	159.23	181.85
3	161.72	190.69
4	167.52	188.07

```
df.corr()
```

	Father_Height	Son_Height
Father_Height	1.000000	0.800205
Son_Height	0.800205	1.000000

A researcher is investigating the distribution of response times (in seconds) for two different versions of a mobile app, i.e. the time taken for a mobile app to respond to a user action, measured in seconds.

The goal is to determine if the response time distributions significantly differ between the two versions.

Data for 20 users for each app version is collected.

response_times_version_A = [1.2, 1.3, 1.1, 1.4, 1.2, 1.3, 1.0, 1.5, 1.2, 1.3, 1.2, 1.4, 1.1, 1.3, 1.2, 1.5, 1.3, 1.4, 1.2, 1.3]

response_times_version_B = [1.6, 1.2, 1.3, 1.4, 1.1, 1.3, 1.2, 1.5, 1.3, 1.4, 1.2, 1.3, 1.2, 1.4, 1.1, 1.3, 1.5, 1.2, 1.3, 1.4] Choose the appropriate test for the given scenario

A) One -Way ANOVA

B) Two Sample Z Proportion

C) Paired T-Test

D) Two-Sample Z-Test

E) KS Test

PS2

```
import pandas as pd
url ="https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/133/892/original/Ride_Sharing_Service.csv?1
df=pd.read_csv(url)
df.head()
```

	ride_id	user_id	ride_type	distance_km	duration_min	fare_amount	tip_amount	pickup_area	dropoff_area	use
0	RIDE1000	USER2000	Shared	6.85	58.72	257.55	10.0	Suburb	Suburb	
1	RIDE1001	USER2001	NaN	12.99	15.80	113.77	0.0	Suburb	Industrial	
2	RIDE1002	USER2002	Standard	22.80	21.01	141.85	0.0	Suburb	Suburb	
3	RIDE1003	USER2003	Shared	11.97	24.39	257.99	15.0	Airport	Uptown	
4	RIDE1004	USER2004	Shared	10.61	10.81	139.86	10.0	Airport	Airport	

Next steps: Generate code with df | New interactive sheet

You are a data analyst at a ride-sharing company that operates across multiple city zones (Airport, Suburb, Uptown, Industrial).

You've been handed a messy dataset from recent rides and asked to extract business insights by cleaning, transforming, and analysing this data. The company is trying to optimise pricing, understand customer satisfaction, predict tips/fares, and personalise service

Objective Help the marketing team identify generous customers by engineering a feature that flags riders who tip more than 5% of the fare amount. This enables targeted promotions and deeper insight into tipping behavior across ride types.

Question You are asked to create a new feature to predict generosity in tipping behavior.

A rider is considered "generous" if their tip percentage exceeds 5% of the fare amount.

Based on this analysis, which of the following conclusions is most accurate?

A) Premium rides show the highest generosity, suggesting customers value luxury service and tip more.

B) Shared rides have the lowest generous tipping behavior because tips are split among co-passengers.

C) Standard ride customers are the most generous tippers, while Premium customers rarely tip—likely due to already high fare amounts.

D) Tipping generosity is uniform across ride types, meaning tipping has no correlation with ride category.

```
df['tip_percentage']=(df['tip_amount']/df['fare_amount'])*100
```

```
### is generous
df['is_generous']=(df['tip_percentage']>5).astype(int)
```

```
## how generosity varies by ride_type
gen_analysis=df.groupby('ride_type')['is_generous'].mean().sort_values(ascending=False)
print(gen_analysis)
```

```
ride_type
Shared      0.461538
Standard    0.200000
Premium     0.000000
Name: is_generous, dtype: float64
```

There are 8 females and 12 males in a coaching class.

After a practice test, the coach wants to know whether the average score of females is greater than the average score of males.

Given data describes the scores of females and males in his class.

female_scores=[25,30,45,49,47,35,32,42]

male_scores=[45,47,25,22,29,32,27,28,40,49,50,33]

Use an appropriate test to check whether the assumption of the coach is significant or not, at a 2% significance level?

A) P_value = 0.580, There is a significant evidence that the average score of females is greater than the average score of males.

B) P_value = 0.285, There is no significant evidence that the average score of females is greater than the average score of males.

C) P_value = 0.285, There is a significant evidence that the average score of females is greater than the average score of males.

D) P_value = 0.580, There is no significant evidence that the average score of females is greater than the average score of males.

```
# H0 : u1<=u2
# Ha: u1>u2

from scipy.stats import ttest_ind
female_scores=pd.Series([25,30,45,49,47,35,32,42])
male_scores=pd.Series([45,47,25,22,29,32,27,28,40,49,50,33])
t_stat,p_value=ttest_ind(female_scores,male_scores,alternative="greater")
print(t_stat,p_value)
```

```
0.5795450171026676 0.2847023809445894
```

```
alpha=0.02
if p_value < alpha:
  print("Reject null hypothesis")
else:
  print("Fail to reject null hypothesis")
```

```
Fail to reject null hypothesis
```

A Mobile Retail store owner is interested in the distribution of popular smartphone brands among a group of 200 people.

They expect that 30% of people would prefer Brand A, 40% would prefer Brand B and 30% would prefer Brand C.

However, upon surveying the group, the results are as follows: 70 prefer Brand A, 80 prefer Brand B, and 50 prefer Brand C.

Conduct an appropriate test to see if the distribution of preferences matches the store owner's expectations at a 5% significance level.

Choose the correct option below:

A) P-value: 0.2048

We fail to reject the null hypothesis

Thus concluding that the observed distribution matches the expectations.

B) P-value: 0.2048

The null hypothesis is rejected

Thus concluding that the observed distribution does NOT match the expectations.

C) P-value: 0.1888

We fail to reject the null hypothesis

Thus concluding that the observed distribution matches the expectations.

D) P-value: 0.1888

The null hypothesis is rejected

Thus concluding that the observed distribution does not match the expectations.

```
#H0: obs=expec
#Ha obs!=expec
import numpy as np
from scipy.stats import chisquare
observed_counts=np.array([70,80,50])
expected_counts=np.array([0.3*200,0.4*200,0.3*200])
```

```
chi_sq,p_value=chisquare(observed_counts,expected_counts)
print(chi_sq,p_value)
```

```
3.3333333333333335 0.1888756028375618
```

```
if p_value < 0.05:
  print("Reject null hypothesis")
else:
  print("Fail to reject null hypothesis")
```

```
Fail to reject null hypothesis
```

Imagine a dataset with two features: 'Age' and 'Salary'.

'Age' has a mean of 30 and a standard deviation of 5, while 'Salary' has a mean of 80,000 and a standard deviation of 20,000. You decide to apply Standard Scaling to both features.

After applying Standard Scaling, what can be said about the transformed 'Age' and 'Salary' features?

A) Both features will have a mean of 0 and a standard deviation of 1.

B) 'Age' will have a mean of 0 and a standard deviation of 1, while 'Salary' will have a mean of 0 and a standard deviation of 20,000.

C) There will be no change in the mean and standard deviation for both features.

D) Both features will be standardized but will not necessarily have values between 0 and 1.

The Quidditch teams at Hogwarts conducted tryouts for two positions: Chasers and Seekers.

In Group Chasers, out of 90 students who tried out, 57 were selected. In Group Seekers, out of 120 students who tried out, 98 were selected.

Is there a significant difference in the proportion of students selected for Chasers and Seekers positions?

Conduct a test at 90% confidence level.

A) P-value: 0.00278, There is a significant difference in the proportion of students selected for Chasers and Seekers positions.

B)P-value: 0.00278, There is no significant difference in the proportion of students selected for Chasers and Seekers positions.

C)P-value: 0.00461, There is a significant difference in the proportion of students selected for Chasers and Seekers positions.

D) P-value: 0.00461, There is no significant difference in the proportion of students selected for Chasers and Seekers positions.

```
import statsmodels.api as sm

# Data for Chasers
selected_chasers = 57
total_chasers = 90

# Data for Seekers
selected_seekers = 98
total_seekers = 120

# Perform two-sample Z-proportion test
z_stat, p_value = sm.stats.proportions_ztest([selected_chasers, selected_seekers], [total_chasers, total_seekers

# Confidence level
confidence_level = 0.90
# Calculate the critical value for a two-tailed test
alpha = 1 - confidence_level

# Print the results
print(f"Z-statistic: {z_stat}")
print(f"P-value: {p_value}")

# Decision Rule
if p_value < alpha:
   print("Reject the null hypothesis. There is a significant difference in the proportion of students selected f
else:
   print("Fail to reject the null hypothesis. There is no significant difference in the proportion of students s
```

```
Z-statistic: -2.990306921349541
P-value: 0.002786972588958094
Reject the null hypothesis. There is a significant difference in the proportion of students selected for Chasers
```

A soft drink manufacturing company claims that the volume of drink of their bottles is 15 oz. A consumer group suspects the bottles are under-filled and plans to conduct a test.

What is the Type I error in this situation?

A) The consumer group has evidence that the volume of the bottles is not 15 oz.

B) The consumer group does not conclude that the soft drink bottles have less than 15 oz. when the mean actually is less than 15 oz.

C) The consumer group concludes that the soft drink bottles have less than 15 oz. when the mean actually is 15 oz.

D) The consumer group has evidence that the claim is correct.

```
np.bincount,np.eye,reshape
vecotrization and broadcasting
Handling nulls,merge, concat,
pd.cut,pd.qcut

DAV-2
Baye's Theorem and Conditional Probability (Mathematical Question)
CLT, Properties of Normal Distribution,Poisson, Binomial and Bernoulli
DAV-3
Power of test, Type1 and Type 2
Assumptions of T test , Mann Whitney.
Outliers,Encoding, Feature Scaling
Annova, Ttest, Ztest, Chisq, Kruskal
Diff between Correlation and Covariance(Pearson and Spearmann)
```