Inferential Statistics

It is a branch of statistics that uses data from a sample to draw conclusions or make predictions about a larger population. It involves using mathematical analysis, such as hypothesis testing and confidence intervals, to make generalizations and assess the validity of hypotheses about the whole group from which the sample was drawn

Central Limit Theorem

The Central Limit Theorem states that when plotting a sample distribution of means the means of the sample will be equal to the population mean and the sample distribution will approach sample distribution with variance equal to standard error.

There are a few assumptions behind the CLT :

- The sample data must be sampled and selected randomly from the po
- There should not be any multicollinearity in the sampled data whi sample should not influence the other samples.
- The sample size should be no more than 10% of the population. Ge sample size greater than 30 (n>30) is considered good.

In [25]:
```
import pandas as pd
import numpy as np
import random
import seaborn as sns
import matplotlib.pyplot as plt
```

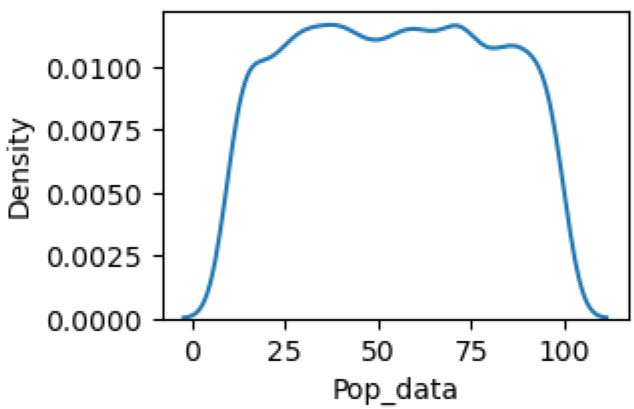
In [26]:
```
pop_data = [np.random.randint(10, 100) for i in range(10000)]
pop_table=pd.DataFrame({'Pop_data': pop_data})
pop_table
```

Out[26]:

	Pop_data
0	73
1	49
2	75
3	75
4	21
...	...
9995	24
9996	60
9997	31
9998	75
9999	42

10000 rows × 1 columns

In [27]:
```
plt.figure(figsize=(3,2))
sns.kdeplot(x="Pop_data",data=pop_table)
plt.show()
```



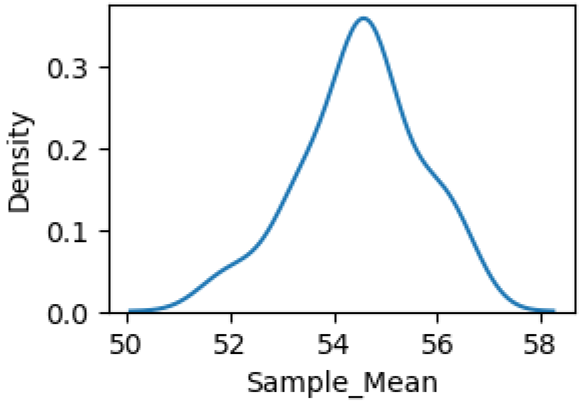
In [28]:
```
sam_mean=[]

for no_sample in range (50):
    sample_data=[]
    for data in range(500):
        sample_data.append(np.random.choice(pop_data))
    sam_mean.append(np.mean(sample_data))
```

In [29]:
```
sample_M=pd.DataFrame({'Sample_Mean':sam_mean})
```

In [30]:
```
plt.figure(figsize=(3,2))
sns.kdeplot(x="Sample_Mean",data=sample_M)
```

```
plt.show()
```



Hypothesis Testing

- It is a part of statistical analysis, where we tet the assumptions made regarding a population parameter.
- It is generally used when we were to compare a single group with an external standard and two or more groups with each other

Null Hypothesis Testing

Null Hypothesis is a stastical theory that suggests there is no statistical significance exists between the population. It is denoted by HO and reaad as H-naught.

Alternative Hypothesis:

An Alternative hypothesis suggests there is a statistical difference between the population parameters. It could be greater. Basically, it is the contrast of the Null Hypothesis, it is denoted by H1 or Ha

Types of Hypothesis Testing

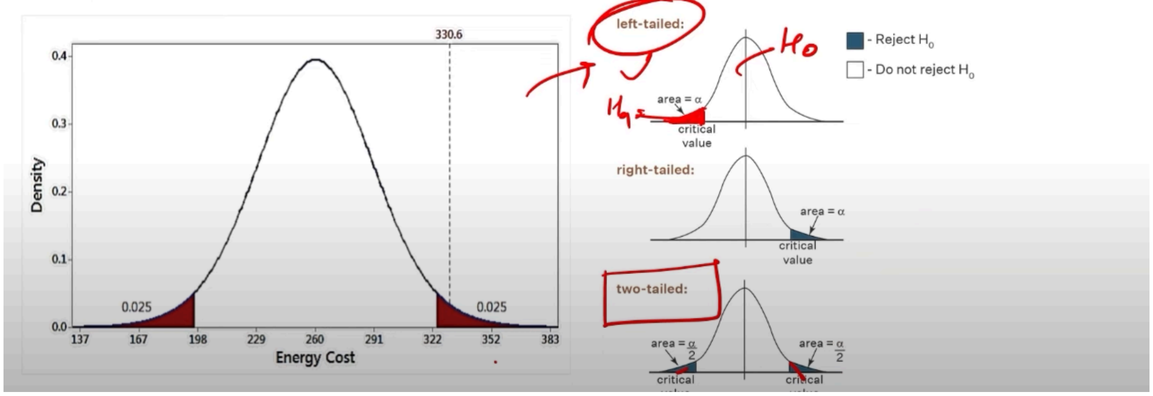
- Z – Test → Average
- T – Test
- Chi – Square Test → Relationship, Goodness

Steps of Hypothesis Testing:

- State null(H0) and alternative(H1) hypothesis
- Choose level of significance(a)
- Find critical values
- Find test statistic
- Draw your conclusion

Choose level of significance (α): 1% → 5%

Denoted by alpha or α. It is a fixed probability of wrongly rejecting a True Null Hypothesis.



Z-Test

A z-test checks if a sample mean differs from a known/target population mean when the population standard deviation (σ) is known (or n is large so σ≈s works via CLT).

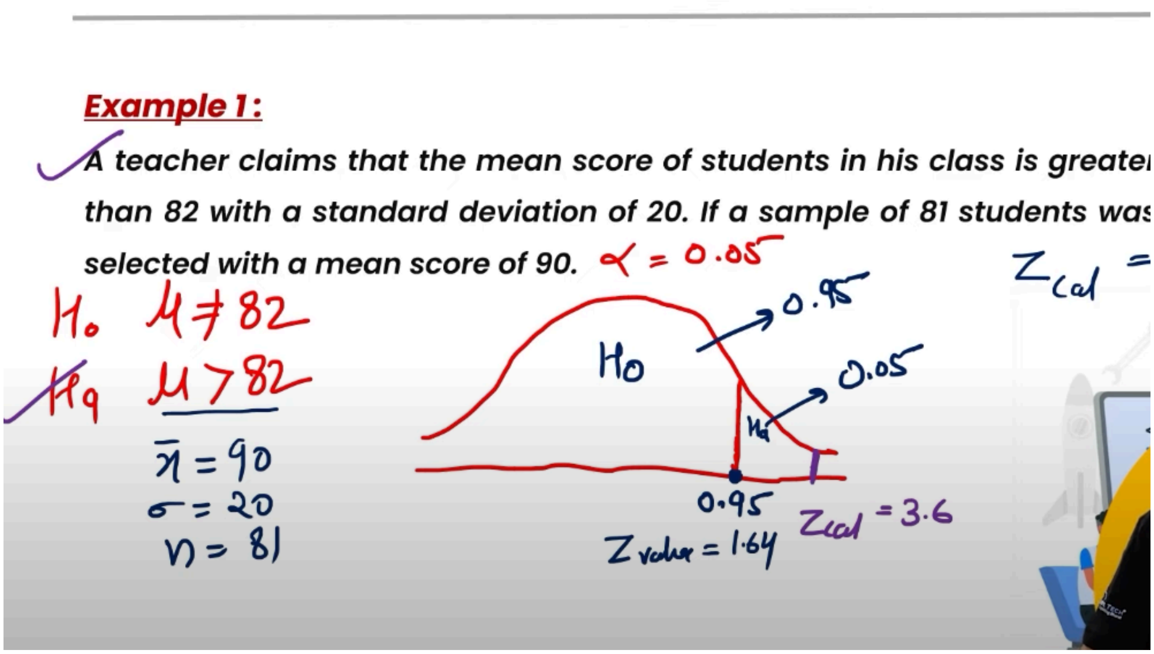
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$\bar{x}$ = sample mean μ_0= hypothesized mean, σ = population SD, n = sample size.

Compare |z| to the standard normal (N(0,1)) to get the p-value.

Assumptions (practical):

Random/independent sample, 2) σ known (or n large), 3) Data roughly normal or n large (CLT)



Example:

Suppose the average exam score of students in a university is known to be 70 with a standard deviation of 10. A professor believes his class performs better, so he takes a sample of 30 students with an average score of 74. We want to test (using z-score) if his class mean is significantly higher.

In [31]:
```python
import numpy
import math
import scipy.stats as st
from scipy.stats import norm

alpha=0.05
mu0=70 #sample mean
sigma=10 #population SD(assumed known)
n=30 #sample size
x_bar=74  #observed_mean

# Step 1: Calculate z-score
z_score = (x_bar - mu0) / (sigma / math.sqrt(n))

# Step 2: Get critical z for two-tailed test
z_critical = norm.ppf(1 - alpha/2)  # positive side

# Step 3: Apply if-else logic
if abs(z_score) > z_critical:
    print(f"Reject H0 | z_score={z_score:.3f}, critical={z_critical:.3f}")
else:
    print(f"Fail to Reject H0 | z_score={z_score:.3f}, critical={z_critical:.3f}")
```

```
Reject H0 | z_score=2.191, critical=1.960
```

Since Z-score is greater than Critical Z-score, we reject H0.

Example 2 (Z Test)

- Scenario: Imagine you work for an ecommerce comapny, and your team is responsible for analyzing customer purchase data. You want to determine whether new website design has led to significant increase in the average purchase amoumt as comapred to old design.

- Data: You have collected data from a random sample of 30 customers who made the purhcase on the old website design and 30 customers who made purchase on the new website design. You have the sample mean , sample S.D, and sample size for both groups.

- H0 (new website == old website)

- H1 (new website > old website)

In [32]:
```
old_data= np.array([51.96, 46.03, 60.04, 48.90, 42.19, 54.52, 40.66, 63.90, 47.40,
        62.03, 65.60, 37.14, 51.13, 45.24, 55.74, 44.91, 49.61, 61.75, 52.94,
        44.74, 42.48, 42.42, 52.93, 47.75, 55.20, 33.34, 42.56, 52.48, 46.54])

new_data=np.array([56.35, 52.13, 62.76, 52.92, 66.60, 58.44, 49.97, 67.54, 61.04, 5
    50.72, 48.77, 63.33, 58.41, 54.96, 64.83, 53.12, 57.89, 61.12, 55.66,
    60.51, 56.22, 45.88, 59.74, 57.98, 52.87, 63.14, 54.39, 58.30, 60.51])
```

In [33]:
```
pop_std=2.5
n_sp = len(new_data)
mean_new=np.mean(new_data)
mean_old=np.mean(old_data)
ap=0.05
```

In [34]:
```
z_table=st.norm.ppf(1-ap)
print(f"{z_table:.4f}")
```

```
1.6449
```

In [35]:
```
z_cal=(mean_new - mean_old) /(pop_std / np.sqrt(n_sp))
print(f"The calculated Z-score is: {z_cal:.3f}")
```

```
The calculated Z-score is: 16.323
```

In [36]:
```
if z_cal > z_table:
    print ("HA is correct. The new website design has led to significant increase i
else:
    print("H0 is correct. Old design is still leading in sales.")
```

```
HA is correct. The new website design has led to significant increase in purchase am
ount.
```

T-test

A t-test is a statistical hypothesis test used to determine whether there is a significant difference between the means of one or two groups, especially when the population standard deviation (σ) is unknown and the sample size is small (n < 30).

Formula (One-Sample t-test)

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where:

- $\bar{X}$ = sample mean
- μ_0 = population mean (hypothesized)
- s = sample standard deviation
- n = sample size

This compares the sample mean to a hypothesized population mean.

Example

- A manufacturer claims that the average weight of a bag of potato chips is 150 grams. A sample of 25 bags is taken, and the average weight is found to be 148 grams, with a standard deviation of 5 grams. Test the manufacturer's claim using a one-tailed t-test with a significance level of 0.05

In [37]:
```
import scipy.stats as st
```

In [38]:
```
t =st.t.ppf(0.05, 24)
```

In [39]:
```
x_bar=148
pop_mean=150
s=5
sample_size=25
```

In [40]:
```
t_cal=(x_bar - pop_mean) / (s/np.sqrt(sample_size))
t_cal
```

Out[40]:
```
np.float64(-2.0)
```

In [41]:
```
if t_cal > t:
    print("HA is right")
else:
    print("H0 is right")
```

```
H0 is right
```

Example 2

- A company wants to test whether there is a difference in productivity between two teams. They randomly select 20 employees from each team and record their productivity scores. The mean productivity score for Team A is 80 with a standard devition of 5, while the mean productivity score for Team B is 75 with a S.D of 6. Test at a 5% level of significance whether there is a differnce in productivity betweent two team A.
- H0 => PA-Pb = 0
- HA => PA - PB != 0

In [42]:
```
t_table_2= st.t.ppf(1-0.02, 38)
print(f"{t_table_2:.4f}")
```

```
2.1267
```

In [43]:
```
t_cal2= (80 -75) / (np.sqrt((25/20) + (36/20)))

print(f"{t_cal2:.4f}")
```

```
2.8630
```

In [44]:
```
if t_cal2 > t_table_2:
    print("HA is right. There is difference in productivity between Team A and Team
else:
    print("H0 is right.There is no difference.")
```

```
HA is right. There is difference in productivity between Team A and Team B.
```

Chi-Square Test

- The Chi-Square test (χ^2 test) is used to check if there's a significant relationship between categorical variables or if the observed frequencies differ from expected frequencies.

Example

A study was conducted to investigate whether there is a relationship between gender and preferred genre of music. A sample of 235 people were selectd, and the data collected is shown below. Test at a 5% level of significance whether there is a significant association between gender and music preference.

Formula for Chi-Square

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O = Observed frequency
- E = Expected frequency

The bigger the difference between O and E, the larger χ^2 will be.

In [45]:
```
st.chi2.ppf(1-0.025,3)
```

Out[45]:
```
np.float64(9.348403604496148)
```

In [46]:
```
row_1=np.array([40,45,25,10])
row_2 = np.array([35,30,20,30])
```

In [48]:
```
sum_r_1=np.sum(row_1)
sum_r2=np.sum(row_2)
sum_row=np.array([sum_r_1,sum_r2])
sum_row
```

Out[48]:
```
array([120, 115])
```

In [49]:
```
sum_cal= row_1+row_2
sum_cal
```

Out[49]:
```
array([75, 75, 45, 40])
```

In [51]:
```
exp=[]
for i in sum_row:
    for j in sum_cal:
        value = (i * j)/235
        exp.append(value)
```

In [52]:
```
obj=np.array([40,45,25,10,35,30,20,30])
```

In [53]:
```
result= (np.sum(np.square(obj-exp) / exp))
print(f"{result:.4f}")
```

```
13.7887
```

H0 is correct.

In [ ]: