Problem Statement 1:

In each of the following situations, state whether it is a correctly stated hypothesis
testing problem and why?

1. H0: $\mu = 25$, H1: $\mu \neq 25$
2. H0: $\sigma > 10$, H1: $\sigma = 10$
3. H0: $x = 50$, H1: $x \neq 50$
4. H0: $p = 0.1$, H1: $p = 0.5$
5. H0: $s = 30$, H1: $s > 30$

Solution:
1. Correct. The alternate of case Mean equal to 50 is always Mean not equal to 50
2. Correct: This is one tail hypothesis testing(Right tail)
3. Not correct. Hypothesis statement should be taken with statistical data, not with random variable
4. Not correct.Hypothesis statement should be taken with statistical data, not with probality of distribution
5.Correct. This is one tail hypothesis testing(Left tail)

Problem Statement 2:
The college bookstore tells prospective students that the average cost of its textbooks is Rs. 52 with a standard deviation of Rs. 4.50. A group of smart statistics students thinks that the average cost is higher. To test the bookstore's claim against their alternative, the students will select a random sample of size 100. Assume that the mean from their random sample is Rs. 52.80. Perform a hypothesis test at the 5% level of significance and state your decision.

$$\mu = 52$$
$$\sigma = 4.5$$
$$\alpha = 0.05$$

n = 100
Sample mean = 52.80

Null hypothesis , H0)$\mu = 52$
Alternate hypothesis , H1) $\mu > 52$

Decision rule:
This is right tail test
If Z value is greater than 1.68 then accept H1
Z = (52.80 - 52)/(4.5/sqrt(100)) = 1.78

Since 1.78 falls outside critical region. So we will accept H1

Problem Statement 3:

A certain chemical pollutant in the Genesee River has been constant for several years with mean μ = 34 ppm (parts per million) and standard deviation σ = 8 ppm. A group of factory representatives whose companies discharge liquids into the river is now claiming that they have lowered the average with improved filtration devices. A group of environmentalists will test to see if this is true at the 1%

level of significance. Assume \ that their sample of size 50 gives a mean of 32.5 ppm. Perform a hypothesis test at the 1% level of significance and state your decision

Solution:

Null hypothesis , H0)µ = 34
Alternate hypothesis , H1) µ < 34

Alpha = 0.01

Decision rule:
If Z value is less than -2.3   reject null hypothesis

```
[3]: p_mean = 34
p_sd = 8
s_mean = 32.5
n = 50
z = (s_mean - p_mean)/(8/50**0.5)
print("Z value is :",z)

Z value is : -1.3258252147247767
```

Since -1.32 Falls outside critical region, so we will accept null hypothesis

Problem Statement 4:

Based on population figures and other general information on the U.S. population, suppose it has been estimated that, on average, a family of four in the U.S. spends about $1135 annually on dental expenditures. Suppose further that a regional dental association wants to test to determine if this figure is accurate for their area of country. To test this, 22 families of 4 are randomly selected from the population in that area of the country and a log is kept of the family's dental expenditure for one year. The resulting data are given below. Assuming, that dental expenditure is normally distributed in the population, use the data and an alpha of 0.5 to test the dental association's hypothesis. 1008, 812, 1117, 1323, 1308, 1415, 831, 1021, 1287, 851, 930, 730, 699, 872, 913, 944, 954, 987, 1695, 995, 1003, 994

Solution:
Null hypothesis , H0)µ = 1135
Alternate hypothesis , H1) µ ≠ 1135

Decision rule:

```
In [17]: import numpy as np
from scipy import stats as st
import statistics
p_mean = 1135
n =22
alpha = 0.5
data = [1008,812,1117,1323,1308,1415,831,1021,1287,851,930,730,699,872,913,944,954,987,1695,995,1003,994]
s_mean = statistics.mean(data)
s_std = statistics.stdev(data)
Degree_of_freedom = 22-1
print("Crtical region:",st.t.ppf(alpha/2,Degree_of_freedom),"<t<",-(st.t.ppf(alpha/2,Degree_of_freedom)))

Crtical region: -0.6863519891164291 <t< 0.6863519891164291
```

If t value is less than -0.69 and greater than 0.69 then reject null hypothesis

```
[19]: import numpy as np
from scipy import stats as st
import statistics
p_mean = 1135
n =22
alpha = 0.5
data = [1008,812,1117,1323,1308,1415,831,1021,1287,851,930,730,699,872,913,944,954,987,1695,995,1003,994]
s_mean = statistics.mean(data)
s_std = statistics.stdev(data)
Degree_of_freedom = 22-1
print("Crtical region:",st.t.ppf(alpha/2,Degree_of_freedom),"<t<",-(st.t.ppf(alpha/2,Degree_of_freedom)))
t = (s_mean - p_mean)/(s_std/22**0.5)
print("t value is :",t)

Crtical region: -0.6863519891164291 <t< 0.6863519891164291
t value is : -2.023137479931484
```

The t value is inside critical region and hence we will reject null hypothesis

Problem Statement 5:

In a report prepared by the Economic Research Department of a major bank the Department manager maintains that the average annual family income on Metropolis is $48,432. What do you conclude about the validity of the report if a random sample of 400 families shows and average income of $48,574 with a standard deviation of 2000?

Solution:
Null hypothesis , H0)$\mu = 48,432$
Alternate hypothesis , H1) $\mu \neq 48,432$

Decision rule:

Let alpha = 0.05
Here we use Z stats. Because population size is very large

Decision rule:
If z value is less than -1.96 or greater than 1.96, them we will reject null hypothesis

```
In [2]: p_mean = 48432
s_sd =2000
s_mean = 48574
n = 400
z = (s_mean - p_mean)/(s_sd/n**0.5)
print("Z value is :",z)

Z value is : 1.42
```

Since 1.42 is outside critical region, we will accept null hypothesis

Problem Statement 6:

Suppose that in past years the average price per square foot for warehouses in the United States has been $32.28. A national real estate investor wants to determine whether that figure has changed now. The investor hires a researcher who randomly samples 19 warehouses that are for sale across the United States and finds that the mean price per square foot is $31.67, with a standard deviation of $1.29. assume that the prices of warehouse footage are normally distributed in population. If the researcher uses a 5% level of significance, what statistical conclusion can be reached? What are the hypotheses?

Solution:
Null hypothesis , H0)µ = 32.28
Alternate hypothesis , H1) µ ≠ 32.28

```
In [1]: import numpy as np
        from scipy import stats as st
        import statistics
        p_mean = 32.28
        n =19
        alpha = 0.05
        s_mean = 31.67
        s_std = 1.29
        Degree_of_freedom = n-1
        print("Crtical region:",st.t.ppf((1-alpha)/2,Degree_of_freedom),"<t<",-(st.t.ppf((1-alpha)/2,Degree_of_freedom)))
        t = (s_mean - p_mean)/(s_std/n**0.5)
        print("t value is :",t)

        Crtical region: -0.06358688254735774 <t< 0.06358688254735774
        t value is : -2.06118477175179
```

Since t value is inside the critical region, we will reject null hypothesis

Problem Statement 8:

Find the t-score for a sample size of 16 taken from a population with mean 10 when the sample mean is 12 and the sample standard deviation is 1.5.

Solution:
n =16
P_mean = 10
Standard deviation = 1.5
S_mean = 12

```
: p_mean = 10
  n =16
  s_mean = 12
  s_std = 1.5
  t = (s_mean - p_mean)/(s_std/n**0.5)
  print("t score is: ", t)

  t score is:  5.333333333333333
```

Problem Statement 9:

Find the t-score below which we can expect 99% of sample means will fall if samples of size 16 are taken from a normally distributed population.

Solution:
Alpha = 1-0.99 = 0.01
Degree of freedom = 15

From t table , t score is -2.602

Problem Statement 10:

If a random sample of size 25 drawn from a normal population gives a mean of 60 and a standard deviation of 4, find the range of t-scores where we can expect to find the middle 95% of all sample means. Compute the probability that (−t0.05 <t<t0.10).

Solution:
Alpha = (1- 0.95)/2 = 0.025
So t score is 2.064

```
5]: sd = 4
s_mean = 60
n = 25
t_score = 2.064
print("range is :",s_mean - t_score*(sd/n**0.5),",",s_mean + t_score*(sd/n**0.5))

range is : 58.3488 , 61.6512
```

```
8]: from scipy import stats
p=stats.t.cdf(0.1,df=24)-stats.t.cdf(-0.05,df=24)
print(f"probability that (-t0.05 <t<t0.10) is {p}")

C:\Users\91984\anaconda3\lib\site-packages\numpy\_distributor_init.py:30: UserWarning: loaded more than 1 DLI
C:\Users\91984\anaconda3\lib\site-packages\numpy\.libs\libopenblas.el2c6ple4zyw3eceviv3oxxgrn2nrfm2.gfortran
C:\Users\91984\anaconda3\lib\site-packages\numpy\.libs\libopenblas.WCDJNK7YVMPZQ2ME2ZZHJJRJ3JIKNDB7.gfortran
  warnings.warn("loaded more than 1 DLL from .libs:"

probability that (-t0.05 <t<t0.10) is 0.05914441613731247
```

Problem Statement 11:

Two-tailed test for difference between two population means
Is there evidence to conclude that the number of people travelling from Bangalore to Chennai is different from the number of people travelling from Bangalore to Hosur in a week, given the following:
Population 1: Bangalore to Chennai n1 = 1200
x1 = 452
s1 = 212
Population 2: Bangalore to Hosur n2 = 800
x2 = 523
s2 = 185

Solution:
H0: Mean is same
H1: Mean is not same

```
In [13]: from scipy import stats
alpha = 0.05
print("Crtical region:",stats.norm.ppf((alpha)/2),"<t<",-(stats.norm.ppf((alpha)/2)))
n1 = 1200
x1 = 452
s1 = 212
n2 = 800
x2 = 523
s2 = 185
SE = ((((s1)**2)/n1) + (((s2)**2)/n2))**0.5
z = (x1-x2)/SE
print("Z score is ", z)

Crtical region: -1.959963984540054 <t< 1.959963984540054
Z score is  -7.926428526759299
```

Since z score fall inside critical region, we will reject null hypothesis. That means people travelling from Bangalore to Chennai is different from the number of people travelling from Bangalore to Hosur in a week

Problem Statement 12:

Is there evidence to conclude that the number of people preferring Duracell battery is different from the number of people preferring Energizer battery, given the following:

Population 1: Duracell
n1 = 100
x1 = 308
s1 = 84
Population 2: Energizer
n2 = 100
x2 = 254
s2 = 67

Solution:
H0: Peoples preferring Duracell and Energizer battery is same
H1: Peoples preferring Duracell and Energizer battery is different.

[14]:
```python
from scipy import stats
alpha = 0.05
print("Crtical region:",stats.norm.ppf((alpha)/2),"<t<",-(stats.norm.ppf((alpha)/2)))
n1 = 100
x1 = 308
s1 = 84
n2 = 100
x2 = 254
s2 = 67
SE = ((((s1)**2)/n1) + (((s2)**2)/n2))**0.5
z = (x1-x2)/SE
print("Z score is ", z)
```

```
Crtical region: -1.9599639845400545 <t< 1.9599639845400545
Z score is  5.025702668336442
```

Since z score fall in the critical region, we will reject null hypothesis

Problem Statement 13:

Pooled estimate of the population variance
Does the data provide sufficient evidence to conclude that average percentage increase in the price of sugar differs when it is sold at two different prices?
Population 1: Price of sugar = Rs. 27.50 n1 = 14
x1 = 0.317%
s1 = 0.12%
Population 2: Price of sugar = Rs. 20.00 n2 = 9
x2 = 0.21%
s2 = 0.11%
H0 : Average Sugar price has no change(Mean1 = Mean2)
H1 : Sugar price has decrease(Mean1 > Mean2)

In [16]:
```python
from scipy import stats
alpha = 0.05
n1 = 14
x1 = 0.317
s1 = 0.12
n2 = 9
x2 = 0.21
s2 = 0.11
s_1 = s1**2
s_2 = s2**2
df = n1 + n2 -2
print("Crtical region:",stats.t.ppf((1-alpha),df))
s = (n1-1) * s_1 + (n2-1) * s_2
n = n1 + n2 -2
sp = (s/n)**0.5
print("Pooled variance is :",sp)
n_1=((1/n1)+(1/n2))**0.5
t = (x1-x2)/(sp*((1/n1) + (1/n2))**0.5)
print("t score is ", t)
```

```
Crtical region: 1.7207429028118775
Pooled variance is : 0.1162919151265879
t score is  2.15355322387416
```

Since t score is in critical region, we will reject null hypothesis

Problem Statement 14:

The manufacturers of compact disk players want to test whether a small price reduction is enough to increase sales of their product. Is there evidence that the small price reduction is enough to increase sales of compact disk players?

Population 1: Before reduction

n1 = 15

x1 = Rs. 6598 s1 = Rs. 844

Population 2: After reduction n2 = 12

x2 = RS. 6870

s2 = Rs. 669

H0: There is no change in sales(Mean1 = Mean2)

H1: Sales increase after reduction(Mean1 < Mean2)

```
In [4]: from scipy import stats
alpha = 0.05
n1 = 15
x1 = 6598
s1 = 844
n2 = 12
x2 = 6870
s2 = 669
s_1 = s1**2
s_2 = s2**2
df = n1 + n2 -2
print("Crtical region:",stats.t.ppf((alpha),df))
s = (n1-1) * s_1 + (n2-1) * s_2
n = n1 + n2 -2
sp = (s/n)**0.5
print("Pooled variance is :",sp)
n_1=((1/n1)+(1/n2))**0.5
t = (x1-x2)/(sp*((1/n1) + (1/n2))**0.5)
print("t score is ", t)

Crtical region: -1.708140761251899
Pooled variance is : 771.9034913769985
t score is  -0.9098300343990459

In [ ]:
```

T score falls outside critical region, we will accept null hypothesis

Problem Statement 15:

Comparisons of two population proportions when the hypothesized difference is zero

Carry out a two-tailed test of the equality of banks' share of the car loan market in 1980 and 1995.

Population 1: 1980

n1 = 1000

x1 = 53

p1 = 0.53

Population 2: 1985

n2 = 100

x2 = 43

p2= 0.53

H0: Proportions are same

H1: Proportions are not same

```
In [1]: from scipy import stats
n1 = 1000
x1 = 53
p1 = 0.53
n2 = 100
x2 = 43
p2= 0.53
p = (x1 + x2)/(n1+n2)
z = (p1 - p2)/(p*(1-p)*(1/n1 + 1/n2))**0.5
print("Z score is ", z)
alpha = 0.05
print("Crtical region:",stats.norm.ppf((alpha)/2),",",-(stats.norm.ppf((alpha)/2)))

C:\Users\91984\anaconda3\lib\site-packages\numpy\_distributor_init.py:30: UserWarning: loaded more than 1 DLL from .libs:
C:\Users\91984\anaconda3\lib\site-packages\numpy\.libs\libopenblas.el2c6ple4zyw3eceviv3oxxgrn2nrfm2.gfortran-win_amd64.dll
C:\Users\91984\anaconda3\lib\site-packages\numpy\.libs\libopenblas.WCDJNK7YVMPZQ2ME2ZZHJJRJ3JIKNDB7.gfortran-win_amd64.dll
  warnings.warn("loaded more than 1 DLL from .libs:"

Z score is  0.0
Crtical region: -1.959963984540054 5 , 1.959963984540054 5
```

Since z score is outside critical region, we will accept null hypothesis

Problem Statement 16:

Carry out a one-tailed test to determine whether the population proportion of traveler's check buyers who buy at least $2500 in checks when sweepstakes prizes are offered as at least 10% higher than the proportion of such buyers when no sweepstakes are on.
Population 1: With sweepstakes
n1 = 300
x1 = 120
p1= 0.40
Population 2: No sweepstakes n2 = 700
x2 = 140
p2= 0.20

H0: P1-P2 < 0.10
H1: P1-P2> 0.10

```
In [4]: from scipy import stats
n1 = 300
x1 = 120
p1= 0.40
n2 = 700
x2 = 140
p2= 0.20
p = (x1 + x2)/(n1+n2)
z = (p1-p2-0.10)/(p*(1-p)*(1/n1 + 1/n2))**0.5
print("Z score is ", z)
alpha = 0.05
print("Crtical region:",(stats.norm.ppf(1-alpha)))

Z score is  3.303749523611152
Crtical region: 1.6448536269514722
```

Since z score is in the critical region we will reject null hypothesis

Problem Statement 17:

A die is thrown 132 times with the following results: Number turned up: 1, 2, 3, 4, 5, 6
Frequency: 16, 20, 25, 14, 29, 28
Is the die unbiased? Consider the degrees of freedom as p^−1 .

```
In [10]: f_obs= [16, 20, 25, 14, 29, 28]
         f_exp= [22,22,22,22,22,22]
         result=stats.chisquare(f_obs,f_exp)
         print("P value is :", result[1])
         print("Chisquare value is: ", result[0])

P value is : 0.1090641579497725
Chisquare value is:  9.0
```

Since chi square value is in the critical region, we can say that die is unbiased

Problem Statement 18:

In a certain town, there are about one million eligible voters. A simple random sample of 10,000 eligible voters was chosen to study the relationship between gender and participation in the last election. The results are summarized in the following 2X2 (read two by two) contingency table:

	Men	Women
Voted	2792	3591
Not voted	1486	2131

We would want to check whether being a man or a woman (columns) is independent of having voted in the last election (rows). In other words, is “gender and voting independent”?

Solution:
H0: Gender is dependant on voting
H1: Gender is not dependant on voting

```
[18]: o_m_v = 2792
      o_w_v = 3591
      o_m_nv = 1486
      o_w_nv = 2131
      t_m = 2792 + 1486
      t_w = 3591 + 2131
      t_v = 2792 + 3591
      t_nv = 1486 + 2131
      e_m_v = (t_v * t_m)/10000
      e_w_v = t_v * (t_w/10000)
      e_m_nv = t_nv * (t_m/10000)
      e_w_nv = t_nv * (t_w/10000)
      f_obs = [o_m_v,o_w_v,o_m_nv,o_w_nv]
      f_exp = [e_m_v,e_w_v,e_m_nv,e_w_nv]
      result=stats.chisquare(f_obs,f_exp)
      print("P value is :", result[1])
      print("Chisquare value is: ", result[0])

P value is : 0.08354483102552426
Chisquare value is:  6.660455899328039
```

We will reject null hypothesis

Problem Statement 19:

A sample of 100 voters are asked which of four candidates they would vote for in an

Higgins	Reardon	White	Charlton
41	19	24	16

election. The number supporting each candidate is given below:
Do the data suggest that all candidates are equally popular? [Chi-Square = 14.96, with 3 df, 0.05 .

Solution:

H0: All candidates are equally popular
H1: All candidates are not equally popular

In [9]:
```
from scipy import stats
obs = [41,19,24,16]
exp = [25,25,25,25]
result=stats.chisquare(obs,exp)
print("Chisquare value is: ", result[0])
p = stats.chi2.ppf(1-0.05,df=3)
print("p value is :",p)
```

```
Chisquare value is:  14.959999999999999
p value is : 7.814727903251179
```

So all candidates are note equally popular

Problem Statement 20:

Children of three ages are asked to indicate their preference for three photographs of adults. Do the data suggest that there is a significant relationship between age and photograph preference? What is wrong with this study? [Chi-Square = 29.6, with 4 df: < 0.05].

###		Photograph		
		A	B	C
Age of child	5 – 6 years	18	22	20
	7 – 8 years	2	28	40
	9 – 10 years	20	10	40

Solution:

H0: There is a no significant relation
H1: There is a significant relation

In [6]:
```
def find_exp_obs(obs):
    array = np.asarray(obs)
    print(array)
    shape = array.shape
    m = shape[0]
    n = shape[1]
    column_total = []
    row_total = []
    f_exp = []
    f_obs = []
    #Find the column totaland row total list
    for i in range(m):
        sum1 = 0
        sum2 = 0
        for j in range(n):
            sum1 = array[i][j] + sum1
            sum2 = array[j][i] + sum2
        column_total.append(sum1)
        row_total.append(sum2)
    total_value = sum(row_total)
    #Finding the expected value
    for i in range(m):
        for j in range(n):
            expected = (row_total[j] * column_total[i])/total_value
            f_exp.append(expected)
            f_obs.append(array[i][j])
    print(f_obs)
    print(f_exp)
    return f_obs,f_exp
```

In [7]:
```
obs=[[18,22,20],[2,28,40],[20,10,40]]
f_obs,f_exp = find_exp_obs(obs)
result=stats.chisquare(f_obs,f_exp)
print("Chisquare value is: ", result[0])
p = stats.chi2.ppf(1-0.01,df=4)
print("p value is :",p)
```

```
[[18 22 20]
 [ 2 28 40]
 [20 10 40]]
[18, 22, 20, 2, 28, 40, 20, 10, 40]
[12.0, 18.0, 30.0, 14.0, 21.0, 35.0, 14.0, 21.0, 35.0]
Chisquare value is:  29.603174603174608
p value is : 13.276704135987622
```

So we will reject null hypothesis

Problem Statement 21:
A study of conformity using the Asch paradigm involved two conditions: one where one confederate supported the true judgement and another where no confederate gave the correct response.

	Support	No support
Conform	18	40
Not conform	32	10

Is there a significant difference between the "support" and "no support" conditions in the frequency with which individuals are likely to conform? [Chi-Square = 19.87, with 1 df: < 0.05].

Solution:

H0: There is no significant difference
H1: There is a significant deference

```
In [9]: obs=[[18,40],[32,10]]
        f_obs,f_exp = find_exp_obs(obs)
        result=stats.chisquare(f_obs,f_exp)
        print("Chisquare value is: ", result[0])
        p = stats.chi2.ppf(1-0.01,df=1)
        print("p value is :",p)

        [[18 40]
         [32 10]]
        [18, 40, 32, 10]
        [29.0, 29.0, 21.0, 21.0]
        Chisquare value is:  19.868637110016422
        p value is : 6.6348966010212145
```

We will reject null hypothesis

Problem Statement 22:
We want to test whether short people differ with respect to their leadership qualities (Genghis Khan, Adolf Hitler and Napoleon were all stature-deprived, and how many midget MP's are there?) The following table shows the frequencies with which 43 short people and 52 tall people were categorized as "leaders", "followers" or as "unclassifiable". Is there a relationship between height and leadership qualities?
[Chi-Square = 10.71, with 2 df: < 0.01]

##	**Height**	
	Short	Tall
Leader	12	32
Follower	22	14
Unclassifiable	9	6

H0: There is no relationship
H1: There is a relationship

```
In [20]: obs=[[12,32],[22,14],[9,6]]
f_obs,f_exp = find_exp_obs(obs)
result=stats.chisquare(f_obs,f_exp)
print("Chisquare value is: ", result[0])
p = stats.chi2.ppf(1-0.001,df=2)
print("p value is :",p)
```

```
[[12 32]
 [22 14]
 [ 9  6]]
[12, 32, 22, 14, 9, 6]
[19.91578947368421, 24.08421052631579, 16.294736842105262, 19.705263157894738, 6.7894736842105265, 8.210526315789474]
Chisquare value is:  10.712198008709638
p value is : 13.815510557964274
```

We will accept null hypothesis

Problem Statement 23:

Each respondent in the Current Population Survey of March 1993 was classified as employed, unemployed, or outside the labor force. The results for men in California age 35-44 can be cross-tabulated by marital status, as follows:Men of different marital status seem to have different distributions of labor force status. Or is this just chance variation? (you may assume the table results from a simple random sample.)

	Married	Widowed, divorced or separated	Never married
Employed	679	103	114
Unemployed	63	10	20
Not in labor force	42	18	25

Solution:
H0: Men of different marital status has same distribution
H1: Men of different marital status has different distribution

```
1]: obs=[[679,103,114],[63,10,20],[42,18,25]]
f_obs,f_exp = find_exp_obs(obs)
result=stats.chisquare(f_obs,f_exp)
print("Chisquare value is: ", result[0])
p = stats.chi2.ppf(1-0.001,df=2)
print("p value is :",p)
```

```
[[679 103 114]
 [ 63  10  20]
 [ 42  18  25]]
[679, 103, 114, 63, 10, 20, 42, 18, 25]
[654.0633147113595, 109.28864059590316, 132.64804469273744, 67.88826815642459, 11.343575418994414, 13.768156424581006, 62.04841
713221602, 10.36778398510242, 12.583798882681565]
Chisquare value is:  31.61310319407798
p value is : 13.815510557964274
```

We will reject null hypothesis