Inferential statistics

Using pandas,matplotlib,numpy,seaborn,scipy.stats in python.

HYPOTHESIS TESTING:

A statement about the population parameter. It may be true or may not be true.

To verify the truthfulness of a statement hypothesis testing is done.

Hypothesis testing concerns using a sample as evidence that supports a hypothesis or not.

Null hypothesis: status quo.

Based on old known information.

Alternative hypothesis:

It is an assertion.

TYPES OF HYPOTHESIS TESTING:

Some of the types of hypothesis testing are

- Test of Means
- Test of Independence
- Analysis of variance

TEST OF MEANS:

- **One Sample:**
    1. Left tailed
    2. Right tailed
    3. Two tailed
- **Two sample:**
    1. Paired samples

2. Unpaired samples

RIGHT TAILED TEST:

Population mean> hypothesized value.

For example,

A new weight loss program advertises an average weight loss of 8 pounds after one month. A random sample of 30 participants shows an average weight loss of 9.5 pounds with a standard deviation of 2.2 pounds. Can we conclude that the program leads to a significant increase in weight loss compared to the advertised average?

Null hypothesis=average weight loss 8 pounds

n=30

Sample mean=9.5

Standard deviation=2.2

h0=population mean<=8

h1=population mean>8

```
1 tcritic=ss.t.ppf(q=0.95,df=29)
```

```
1 round(tcritic,2)
```

```
1.7
```

```
1 tstat=(9.5-8)/(2.2/np.sqrt(30))
2 round(tstat,2)
```

```
3.73
```

```
1 tstat>tcritic
2 reject null hypothesis
```

Thus we conclude that the program leads to a significant increase in weight loss compared to the advertised average.

LEFT TAILED TEST:

Population mean<hypothesized value

Question:

A restaurant claims that their new cooking technique reduces calorie content in dishes to an average of 300 calories. A sample of 40 dishes shows an average of 280 calories with a standard deviation of 20 calories. Is there evidence to suggest that the dishes contain fewer calories than claimed?

Sample (n)=40

Sample mean =280

s=20

h0>=300

h1<300

By performing Left tailed test as we are gathering evidence to prove hypothesized population mean is lesser than the reference value.

```
1 tcritic=ss.t.ppf(q=0.05,df=39)
```

```
1 round(tcritic,2)
```

```
-1.68
```

```
1 df=pd.read_csv("M2_T2_V1Movies.csv")
```

```
1 tstat=(280-300)/(20/np.sqrt(40))
```

```
1 round(tstat,2)
```

```
-6.32
```

t stat<t critical

we can reject null hypothesis.There is enough evidence to suggest that the dishes contain fewer calories than claimed

TWO TAILED TEST:

Population mean not equal to hypothesized value.

A bakery advertises that their cakes have an average sugar content of 20 grams. A sample of 30 cakes is tested, revealing an average sugar content of 22 grams with a standard deviation of 3 grams. Is there evidence to suggest that the cakes' actual sugar content is not equal to the advertised 20 grams?

ho=mean=20

h1=mean not equal to 20

Sample (n)=30

Sample mean=22

s=3

```
1 tcrit1=ss.t.ppf(q=0.05/2,df=29)
2 tcrit2=ss.t.ppf(q=1-0.05/2,df=29)
3 (tcrit1,tcrit2)
```

```
(-2.0452296421327034, 2.045229642132703)
```

```
1 tstat=(22-20)/(3/np.sqrt(30))
```

```
1 round(tstat,2)
```

```
3.65
```

tstat>tcrit2

we reject null hypothesis

there is evidence to suggest that the cakes' actual sugar content is not equal to the advertised 20 grams.

UNPAIRED SAMPLE TEST:

An unpaired sample test, also known as an independent sample test, is a statistical test used to compare the means of two separate groups or samples. These groups are independent of each other, meaning that the data points in one group are not related or matched to the data points in the other group.

Is there a significant difference in the average marks of a particular subject between two different classes, considering that their variances are not equal?

Null hypothesis: there is no significant difference between two sample means.

Alternate hypothesis: There is significant difference between two sample means.

```
rollno=[1,2,3,4,5,6,7,8,9,10]
class_a=[89,29,44,32,23,44,33,88,66,55]
class_b=[89,78,67,80,90,77,65,55,78,58]
data={"rollno":rollno,"class a":class_a,"class b":class_b}
df=pd.DataFrame(data)
df.set_index('rollno', inplace=True)
df.head(5)

```

rollno	class a	class b
1	89	89
2	29	78
3	44	67
4	32	80
5	23	90

Data is generated with the help of pandas.

From scipy.stats ttest_ind helps in giving p value and t statistic for two different sample groups.

```
from scipy.stats import ttest_ind as ti
ti(a=df["class a"],b=df['class b'])
```

```
Ttest_indResult(statistic=-2.7739885810777563, pvalue=0.012514445839302153)
```

p value< 0.05

reject null hypothesis

there is a significant difference between the marks of the two class groups.

PAIRED SAMPLE TEST:

A paired sample test, also known as a dependent sample test or matched-pairs test, is a statistical test used to compare the means of two related groups. The key characteristic of paired samples is that each data point in one group is matched or related to a specific data point in the other group. Paired sample tests are used when the samples are not independent of each other.

Is there a significant improvement in students' test scores after they undergo a tutoring program?

Null hypothesis: there is no significant difference between two sample means.

Alternate hypothesis: There is significant difference between two sample means

```
df.head(5)
```

rollno	before	after	difference
1	89	89	0
2	29	78	49
3	44	67	23
4	32	80	48
5	23	90	67

```
: 1 ttest_1samp(df["difference"],10)
```

```
: Ttest_1sampResult(statistic=1.457977073986065, pvalue=0.17884048344237602)
```

```
1 pvalue lesser than alpha (significance value)
2 reject null hyothesis
3 there is significant difference between the marks
```

TEST OF INDEPENDENCE:

data=TITANIC DATASET

To analyze whether there is association between the passenger class and survived column.

Dataset sample:

```
1 df=pd.read_csv("M5_T3_V2_TitanicSurvival.csv")
```

```
1 df.sample(5)
```

	Name	survived	gender	passengerClass
1118	Peltomaki, Mr. Nikolai Johannes	no	male	3rd
150	Harrison, Mr. William	no	male	1st
921	Keefe, Mr. Arthur	no	male	3rd
830	Goodwin, Mr. Charles Edward	no	male	3rd
956	Lefebre, Miss. Jeannie	no	female	3rd

Using pivot table function to calculate the total count of values under each category it takes four parameters

Index,columns,values,aggfunc

```
dframe=pd.pivot_table/
(data=df,index="passengerClass",columns="survived",aggfunc="count",values="Name")
```

```
1 dframe
```

survived	no	yes
passengerClass		
1st	123	200
2nd	158	119
3rd	528	181

To analyze association importing chi square contingency from scipy stats module.

```
1 from scipy.stats import chi2_contingency as chi
```

```
1 chi(dframe)
```

```
(127.85915643930326,
 1.7208259588256175e-28,
 2,
 array([[199.62337662, 123.37662338],
        [171.19404125, 105.80595875],
        [438.18258212, 270.81741788]]))
```

pvalue < alpha

reject null hypothesis

there is association between the survived and passenger class column.

ANALYSIS OF VARIANCE:

statistical test used to analyze the difference between the means of more than two groups. A one-way ANOVA uses one independent variable, while a two-way ANOVA uses two independent variables.

One way anova:one factor

to analyze if the factor gender effects weightloss

Dataset:

It is the data about the weight loss who had undertaken different diet plan

gender	Age	Height	preweight	Diet	weight6weeks
Male	47	179	73	3	72.1
Female	56	171	73	3	68.9
Male	39	166	87	1	81.9
Male	45	160	78	2	72.7
Male	50	160	78	1	73.9

```
1 grp=df["wtloss"].groupby(df["gender"])
```

```
1 ss.f_oneway(*[i[1] for i in grp])
```

```
F_onewayResult(statistic=0.04378592096006073, pvalue=0.8348273886348012)
```

p value> alpha

cannot reject null hypothesis

there is no significant difference on weight loss based on gender.

Two way anova: two factors affecting dependent variable

We can perform two way anova using statsmodels

```
from statsmodels.formula.api import ols
```

```
from statsmodels.stats.anova import anova_lm
```

```
formula="wtloss~C(gender)+C(Diet)"
```

```
model=ols(formula,df).fit()
```

```
anova_lm(model)
```

	df	sum_sq	mean_sq	F	PR(>F)
C(gender)	1.0	0.278485	0.278485	0.048877	0.825655
C(Diet)	2.0	60.417220	30.208610	5.301912	0.007112
Residual	72.0	410.233112	5.697682	NaN	NaN

Conclusion: gender does not affect weight loss whereas diet plan contributes weight loss.

To see the significant difference within groups can make use of bioinfokit.analys module.

Tukey_hsfd method:

Takes four parameters: dataframe, res_var (dependent variable), x_fac_var(independent variables),anova_model(formula)

```
a=stat()
```

```
a.tukey_hsd(df,res_var="wtloss",xfac_var=["gender","Diet"],anova_model=formula)
s=a.tukey_summary
```

```
1 s[s["p-value"]<0.05]
```

	group1	group2	Diff	Lower	Upper	q-value	p-value
1	(Female, 1)	(Female, 3)	2.830000	0.232983	5.427017	4.511936	0.024717
5	(Female, 2)	(Female, 3)	3.272857	0.675840	5.869874	5.217994	0.005583

Can conclude there is significant difference in weight loss between these groups.