Markdown

JupyterLab

Python 3 (ipykernel)

Hypothesis Testing

- Hypothesis is a claim made by a person / organization.
- The claim is usually about the population parameters such as mean or proportion and we seek evidence from a sample for the support of the claim (Example: average salary of Data Scientist with 1 year experience is Rs 5 Lakhs per annum).
- Hypothesis testing is a process used for either rejecting or retaining null hypothesis.
- There are two types of hypotheis:
  - Null Hypothesis, H_0: Hypothesis of no difference
  - Alternative Hypothesis, H_1: Hypothesis of difference

** Examples of some claims:**

- If you drink Horlicks, you can grow taller, stronger and sharper.
- Two - minute for cooking noodles. (or eating !!)
- Married people are happier than singles (Anon - 2015).
- Smokers are better sales people.

Hypothesis testing is used for checking the validity of the claim using evidence found in sample data.

Example:

- pop mean=350
- pop std=11.3
- Sample size n=120
- Sample mean=320

- Sample size n=120
- Sample mean=320
- H_0: Sample mean is same as population mean (Sample is representative of the population)
- H_1: Sample mean is less than the population mean.(Sample is not representative of the population)

- Type I error: Occurs when we reject a true null hypothesis
- Type II error: Not rejecting a false null hypothesis

- Knowing the Probability Dist - Normal Dist (z dist), (<30 data points - t dist), chi sq, f dist
- (FYI - when the sample (# of data points) starts to go above 30, a t dist starts to follw z dist)
- Significance level: Measure how frequently the conclusion will be wrong. (Type I error - alpha)
- Rejection Region: A range of values such that if the test statistics (say z score) falls into that range, we decide to reject null hypothesis in favour of alternative hypothesis
  - If calculated test statistics < table value then accept the null hypothesis
  - If calculated test statistics > table value(falls in rejection region) then reject null hypothesis
- P value: "Strength of evidence in support of a null hypothesis"
  - P value is the probability of observing a test statistics as extreme as the one computed, assuming that the null hypothesis is true

Parametric and Non-parametric tests

- A parametric test makes assumptions about a population's parameters, and a non-parametric test does not assume anything about the underlying distribution.
- A parametric test makes assumptions about a population's parameters:
  - Normality: Data in each group should be normally distributed.

- A parametric test makes assumptions about a population's parameters:
  - Normality: Data in each group should be normally distributed.
  - Independence: Data in each group should be sampled randomly and independently.
  - No outliers: No extreme outliers in the data.
  - Equal Variance: Data in each group should have approximately equal variance.
- However, a non-parametric test (sometimes referred to as a distribution free test) does not assume anything about the underlying distribution (for ex. data follows normal distribution is an assumption)
- For ordinal and categorical data, we use non-parametric tests.
- Also, if we doubt about the normality of our data, then we use non-parametric tests
- Examples of parametric tests: z-test, t-test
- Examples of non-parametric tests: χ^2 test, Kruskal Walli's test etc.

One sample z test

Population mean and standard deviation is known, population is normal(or approx)

- $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- To apply normal distribution: population mean, population standard deviation, N>=30

```
import pandas as pd
import numpy as np

data=pd.read_excel('Hypothesis Testing.xlsx',sheet_name='One Sample z').iloc[:,0]
data.head()
```

```
data=pd.read_excel('Hypothesis Testing.xlsx',sheet_name='One Sample z').iloc[:,0]
data.head()
```

```
0    55.000000
1    54.000000
2    62.589736
3    53.000000
4    59.372035
Name: Minutes, dtype: float64
```

```
data.shape
```

```
(35,)
```

- A random sample of 35 young adult men was sampled.
- Each person was asked how many minutes of sports he watched on television daily. The responses are listed here. It is known that $\sigma = 10$.
- At 5% significance level can we conclude that mean amount of television watched daily by all young adult men is greater than 50 minutes.

```
#N=35, pop std=10, pop mean=50, alpha=0.05
```

```
data.describe()
```

```
count    35.000000
mean     51.058857
std       7.445037
min      37.347848
25%      46.160073
50%      50.400279
75%      54.869239
```

Let us define the hypothesis:

- H_0: Average amount of daily televious minutes watched by young adult men is 50. (Sample mean is same as population mean)
- H_1: Average amount of daily televious minutes watched by young adult men >50. (Sample mean is >50)
- This can be written as -
  - $H_0 : \mu = 50$
  - $H_1 : \mu > 50$

```
]: from statsmodels.stats import weightstats as wst
   wst.ztest(data,value=50,alternative='larger') #RIght tailed test
```

```
]: (0.8414036993777245, 0.20006090686584627)
```

- z-statistic is 0.8414
- p value is 0.2001
- p-value > 0.05 (alpha)
- So, accept the null hypothesis. That is, we don't have enough evidence to reject the null hypothesis.

Example with two-way hypothesis

- Consider a dataset containing BP of 34 patients.
- Null Hypothesis: The average BP of patients before the treatment is 146. That is, $H_0 = 146$
- Alternative Hypothesis: Average BP is not 146. That is, $H_1 \neq 146$

```
[7]: df= pd.read_excel('BP.xlsx')
     df.head()
```

[7]:

	patient_name	patient_ sex	patient_agegrp	patient_bp_before	patient_bp_after
0	1	Male	30-45	142	153
1	2	Male	30-45	163	170
2	3	Male	30-45	143	168
3	4	Male	30-45	153	142
4	5	Male	30-45	146	141

```
[8]: df.shape
```

```
[8]: (34, 5)
```

```
[15]: df['patient_ sex'].value_counts()
```

```
[15]: Male    34
      Name: patient_ sex, dtype: int64
```

```
[9]: df.describe()
```

[9]:

	patient_name	patient_bp_before	patient_bp_after
count	34.000000	34.000000	34.000000
mean	17.500000	155.764706	150.647059

max	34.000000	184.000000	184.000000

2]:
```
ztest ,propability_value = wst.ztest(df['patient_bp_before'], value=146)
print(float(propability_value))
if propability_value<0.05:
    print("Null hyphothesis rejected ")
else:
    print("Null hyphothesis accepted ")
```

```
7.039677630242457e-07
Null hyphothesis rejected
```

Inferences:

- There is no enough evidance to say that the the sample has the average BP (before the treatment) as 146
- That is, there is no guarantee that the given sample has been drawn from the population where population has average BP(before) as 146.
- Or, the sample drawn may be biased - more of a data analyst inference.
  - As a data analyst, we can see that all 34 records here are of Males.
  - Hence, the sample is not a true representaiton of the population, because, there is a sample bias.
  - However, stratified sampling should have been benefial.

[ ]:

Tasks:

1. Consider the UsedCarsPrice dataset. Check whether this dataset is drawn from the population where the average kilometer is 60000.
2. Consider the UsedCarsPrice dataset. Check whether this dataset is drawn from the population where the average price of the car is 12000

View Run Kernel Settings Help

Markdown

One sample t test

- Student's t-test is used to check whether the sample is a true representation of popultation.
- That is, whether the sample has features similar to that of population

```
data=pd.read_excel('Hypothesis Testing.xlsx',sheet_name='One sample T').iloc[:,0]
data
```

```
0     3.03
1     6.33
2     6.50
3     5.22
4     3.56
5     6.76
6     7.98
7     4.82
8     7.96
9     4.54
10    5.09
11    6.46
Name: D.Time, dtype: float64
```

Check whether this sample of size n=12 has been drawn from the population which has a mean =6 at alpha=0.05

localhost:8888/notebooks/Day5_Stats.ipynb

pyter Day5_Stats Last Checkpoint: last month

it View Run Kernel Settings Help

Markdown JupyterLab Python 3 (ipy

Population mean is known and population standard deviation is unknown, population is normal.

- $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{(n)}}}$

In statistics, μ and σ are used to represent population mean and population standard deviation respectively. And, $\bar{x}$ and s are used to represent sample mean and samp

standard deviation

]:
```
data.describe()
```

]:
```
count    12.000000
mean      5.687500
std       1.580369
min       3.030000
25%       4.750000
50%       5.775000
75%       6.565000
max       7.980000
Name: D.Time, dtype: float64
```

Define null and alternative hypothesis

- H0: Average delivery time is 6 hours.
- H1: Average delivery time is less than 6 hours

]:
```
#Sample mean < population mean : left tailed --> p(x)
```

```
2]: from scipy.stats import ttest_1samp
    ttest_1samp(data,6)    #pass data and population mean as parameters
```

```
2]: Ttest_1sampResult(statistic=-0.6849867420895185, pvalue=0.5075293854463145)
```

```
3]: # The p-value obtained above is for two-tailed test.
    # So, One tailed probability is
    0.5075/2
```

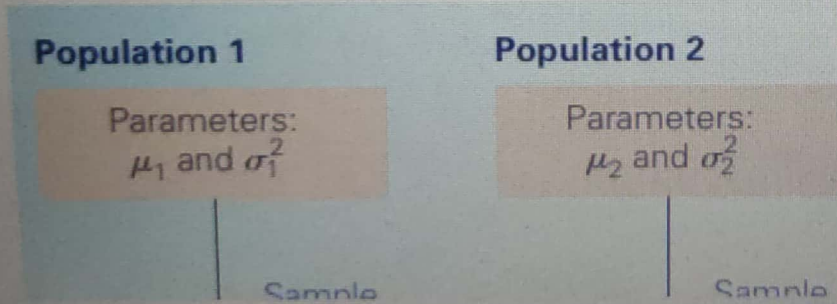
```
3]: 0.25375
```

- P value 0.25375 > 0.05 (alpha)
- Accept the H0.
- The given sample is drawn from the population with mean 6

Two sample Tests ¶

```
[14]: from IPython.display import Image
      Image(filename='IDS.png')
```

[14]:



Inference about difference between two means

- Independent samples
- Sampling distribution of $\bar{x}_1 - \bar{x}_2$ is normally distributed if populations are normally distributed or approx. normally distributed.

[15]:
```
#Independent samples of large size n>=30
Image(filename='z test.png')
```

[15]:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The interval estimator is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Question:

- Recent studies seem to indicate that using a cell phone while driving is dangerous.
- One reason for this is that a drivers reaction times may slow while he or she is talking on the phone.
- Researchers at Ohio University measured the reaction times of a sample of drivers who owned a car, phone.
- Half the sample was tested while on the phone and the other half was not on the phone.
- Can we conclude that reaction times are slower for drivers using cell phones?

Define the hypothesis

- Null Hypothesis: Reaction time is same for drivers who use cell phones compared to who don't.
- Alternative Hyp: Reaction time is more for drivers who use cell phones compared to who don't.

```
5]: two=pd.read_excel('Hypothesis Testing.xlsx',sheet_name='Two sample z').iloc[:,0:2]
```

```
6]: two.shape
```

```
6]: (125, 2)
```

```
7]: two.head()
```

```
7]:
```

	Phone	Not
0	0.596	0.620
1	0.708	0.523
2	0.646	0.652
3	0.725	0.652
4	0.649	0.506

```
[28]: from statsmodels.stats import weightstats as test
      test.ztest(x1=two['Phone'],x2=two['Not'], alternative= 'larger')
```

```
[28]: (7.0668652773518135, 7.923633314310826e-13)
```

- P value = 7.923633314310826e-13
- So, p- value < 0.05 (alpha)
- Reject H0
- We do not have enough evidance to say the response time of the drivers using mobile and not using mobile are same

[ ]:

Two Sample T Test

22]:
```
Image(filename='t test equal var.png')
```

22]:

Test Statistic for $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad v = n_1 + n_2 - 2$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$n_1 + n_2 - 2$$

If n1,n2<30, compare two means of population assuming that variance of the population are equal

Question:

- A number of restaurants feature a device that allows credit card users to swipe their cards at the table.
- It allows the user to specify a percentage or a dollar amount to leave as a tip.
- In an experiment to see how it works, a random sample of credit card users was drawn.
- Some paid the usual way, and some used the new device.
- The percent left as a tip was recorded and listed below.
- Can we infer that users of the device leave larger tips?

```
df=pd.read_excel('Hypothesis Testing.xlsx',sheet_name='Two sample t').iloc[:,0:2]
df
```

	Usual	Device
0	10.3	13.6
1	15.2	15.7
2	13.0	12.9
3	9.9	13.2
4	12.1	12.9
5	13.4	13.4

Define Hypothesis:

- Null hypothesis: Percentage of tip paid are same
- Alternative hyp: Percentage of tip paid by device are higher

28]:
```
n1=10    # number of people who are paying by usual method
n2=11    # number of people who are paying using device
```

35]:
```
#Two sample t test: Independent samples assuming equal variance
from scipy.stats import ttest_ind
ttest_ind(df['Usual'].dropna(),df['Device'], equal_var= True)
```

35]:
```
Ttest_indResult(statistic=-2.17300971633649, pvalue=0.04263401479924383)
```

- The p value 0.0426 obtained here is for two-sided test.
- So, for one-sided test (as per our H0 and H1), the actual p-value = 0.0426/2 = 0.0213
- As p-value < alpha=0.05, reject H0
- So, tips paid by people with cash/usual paymnet method and through device are not same.

Non parametric: Chi square test of association/independence

Understand the relationship between Gender and Campaign Response?

30]:
```
data=pd.read_excel("CreditCardData.xlsx")
data.head()
```



[30]:
```
data=pd.read_excel("CreditCardData.xlsx")
data.head()
```

[30]:

	Card_ID	Campaign_Responce	Registration_Date	Gender	Birth_Date
0	100005950	False	1998-11-18	M	1984-02-06
1	100022191	True	1999-09-15	F	1959-09-11
2	100025442	False	1998-05-12	M	1970-08-25
3	100026513	False	1999-02-12	M	1951-03-12
4	100039145	False	2000-08-12	M	1949-06-08

- Is there a relationship between Gender and Campaign_Responce?

[31]:
```
data.Campaign_Responce.value_counts(normalize= True)
```

[31]:
```
False    0.835017
True     0.164983
Name: Campaign_Responce, dtype: float64
```

- Null: There is no relationship between Gender and Campaign Responce. (independent)
- Alternate: There is a relatinoship between Gender and Campaign Responce. (dependent)

Edit View Run Kernel Settings Help

Markdown

JupyterLab Python 3 (ipyke

```
#Contingency table :
obs=pd.crosstab(data['Gender'],data['Campaign_Responce'])
obs
```

Campaign_Responce	False	True
Gender		
F	102	25
M	146	24

```
from scipy.stats import chi2_contingency,chisquare

chi_sq_Stat, p_value, deg_freedom, exp_freq=chi2_contingency(obs)
print('Chi-square statistic %3.5f P value %1.6f Degrees of freedom %d'
      %(chi_sq_Stat, p_value,deg_freedom))
```

```
Chi-square statistic 1.25639 P value 0.262335 Degrees of freedom 1
```

- p value is 0.2623
- p-value > 0.05 (alpha)
- Accept H0
- ***We do not have enough evidance to say that gender has an influence on campaign response. That means, gender and response are independent of each oth***

```
#Degrees of freedom: (r-1)*(c-1)
(2-1)*(2-1)
```

Tasks:

Search