SciPy Statistical Significance Tests

Read the file of hypotheis test file first

What is Statistical Significance Test?

In statistics, statistical significance means that the result that was produced has a reason behind it, it was not produced randomly, or by chance.(there is a relation b/w the result_value/assumtion_result and hypothesis value/)

SciPy provides us with a module called "scipy.stats", which has functions for performing statistical significance tests.

Statistically Significance:-

Statistical significance is the likelihood that the difference in conversion rates between a given variation and the baseline() is not due to random chance.(means the difference/variation b/w result and hypothesis are not due to random/by chance. but may be due to relation b/w them i,e,,greater than ,less than or not equal to,,ect. it should be contrast to null_hypothesis)

A result of an experiment is said to have statistical significance, or be statistically significant, if it is likely not caused by chance for a given statistical significance level.

Your statistical significance level reflects your risk tolerance and confidence level. For example, if you run an A/B testing experiment with a significance level of 95%, this means that if you determine a winner, you can be 95% confident that the observed results are real and not an error caused by randomness(but remaining 5% causes error caused by randomness). It also means that there is a 5% chance that you could be wrong.
,,
,,,

Population proportion & population mean:-See this vedio(https://youtu.be/fd-RcDYRENc)

1)population proportion:-(it is used when 2 population means are different.Therefore z-value is used)is a parameter that describes a percentage value associated with a population. For example, the 2010 United States Census showed that 83.7% of the American Population was identified as not being Hispanic or Latino; the value of . 837 is a population proportion.
*it was indicated as [p′ = x / n] where
x represents the number of successes and
n represents the sample size.
p′ is the sample proportion and serves as the point estimate for the true population proportion.

2)Population Mean:-The population mean is the mean or average of all values in the given population and is calculated by the sum of all values in population denoted by the summation of X divided by the number of values in population which is denoted by N.

z-value & t-value:-

1)z-test/Normal test:-
The standard normal distribution is also called the 'Z-distribution' and the values are called 'Z-values' (or Z-scores).
*Z-values express how many standard deviations from the mean a value is.If we don't have standard deviation then we will go for T-test.
*If a z-score is equal to 0, it is on the mean. A positive z-score indicates the raw score is higher than the mean average. For example, if a z-score is equal to +1, it is 1 standard deviation above the mean.

2)T-Test:-If we don't have standard deviation then we will go for T-test.
T-tests are used to determine if there is significant deference between means of two variables. and lets us know if they belong to the same distribution.The
*t-distribution is used for estimation and hypothesis testing of a population mean (average).
*The t-distribution is adjusted for the extra uncertainty of estimating the mean.
*If the sample is small, the t-distribution is wider. If the sample is big, the t-distribtution is narrower.
*The bigger the sample size is, the closer the t-distribution gets to the standard normal distribution.
*For the t-distribution this is expressed as 'degrees of freedom' (df), which is calculated by subtracting 1 from the sample size (n).
*For example a sample size of 30 will make 29 degrees of freedom for the t-distribution.
*Finding the critical t-values and p-values of the t-distribution is similar z-values and p-values of the standard normal distribution. But make sure to use the correct degrees of freedom.

How to choose z-value & t-value?
Commonly estimated parameters are:
1)Population Proportions:- (uesd for qualitative data)
Real Example:-15 peoples out of 20 in America knows English.(this is the claim for Null_hypothesis)
Population Proportion=15/20=0.75

2)Proportion Mean:- values (used for numerical data)
Real Example:-In America average 20 peoples knows English.
Population Mean=20

3)Standard Normal Distribution (Z): used for Testing Population Proportions(population proportion uses Z-values for hypothesis testing)

4)Student's T-Distribution (T): used for Testing Population Means(Population Mean uses T-values for hypothesis testing)

,,

,,,

T-Test in Scipy:-

T-tests are used to determine if there is significant deference between means of two variables. and lets us know if they belong to the same distribution. It is a two tailed test.(i,e,, both sides of hypothesis curve or A!=B)
The function "ttest_ind()" takes two samples of same size and produces a tuple of t-statistic and p-value.

Example:-

```
import numpy as np
from scipy.stats import ttest_ind

v1 = np.random.normal(size=100)       ##ramdomly selected values from normal
distribution curve.
v2 = np.random.normal(size=100)

result = ttest_ind(v1, v2)
print(result)
```

,,,,,,,,,,,,,,,,,,,,,,,,
o/p is Ttest_indResult(statistic=0.88243958372222664, pvalue=0.37860920117232288) , once we get p-value we can test hypothesis
,,

If you want to return only the p-value, use the "pvalue" property:

```
import numpy as np
from scipy.stats import ttest_ind

v1 = np.random.normal(size=100)
v2 = np.random.normal(size=100)

res = ttest_ind(v1, v2).pvalue
print(res)
```

,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,O/P is 0.68346891833752133 this is only p-value.

,,
,,,

Probability Density Function & Cummulative Distribution Function:-

1)Probability Density Function(pdf):-

In probability theory, a probability density function (PDF), or density of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would be close to that sample.The Probability Density Function(PDF) defines the probability function representing the density of a continuous random variable lying between a specific range of values. In other words, the probability density function produces the likelihood of values of the continuous random variable. Sometimes it is also called a probability distribution function or just a probability function. see image
(https://cdn1.byjus.com/wp-content/uploads/2021/10/Probability-Density-Function.png)

2)Cumulative Distribution Function(cdf):-

It gives the probability density of left side from the given value.See the image(https://support.minitab.com/en-us/minitab-express/1/cdf_def.png)
The cumulative distribution function (CDF) calculates the cumulative probability for a given x-value. Use the CDF to determine the probability that a random observation that is taken from the population will be less than or equal to a certain value.see image
(https://www.graduatetutor.com/wp-content/uploads/2021/03/Cumulative-Density-Function-of-a-dice-6.jpg)

Application of Cummulative Distribution Function:-
*It gives probability of graph from fixed value, from that we can find p-value in hypothesis test.In Scipy we have "KS test" is used to check if given values follow a distribution.

TS TEST in Scipy:-

KS test is used to check if given values follow a distribution.
The function takes the value to be tested, and the CDF as two parameters.
A CDF can be either a string or a callable function that returns the probability.
It can be used as a one tailed or two tailed test.
By default it is two tailed. We can pass parameter alternative as a string of one of two-sided, less, or greater.

Example:-
Find if the given value follows the normal distribution:

```
import numpy as np
from scipy.stats import kstest

v = np.random.normal(size=100)

result = kstest(v, 'norm')              ##norm means "NORMAL DISTRIBUTED FUNCTION" 
or we have other options like "UNIFORM" & "Exponential Function".

print(result)
,,,,,,,,,,,,,,,,,,,,,,,,,,O/P IS KstestResult(statistic=0.04779870122195684l, 
pvalue=0.97630967161777515)
```

,,
,,,

Statistical Description of Data:-

In order to see a summary of values in an array, we can use the "describe()" function.

It returns the following description:

1.number of observations (nobs)
2.minimum and maximum values = minmax
3.mean
4.variance
5.skewness
6.kurtosis

Example
Show statistical description of the values in an array:

```
import numpy as np
from scipy.stats import describe

v = np.random.normal(size=100)
res = describe(v)

print(res)
```

,,,,,,,,,,,,,,,,,,,,,,,,O/P IS

```
 DescribeResult(
    nobs=100,
    minmax=(-2.0991855456740121, 2.1304142707414964),
    mean=0.11503747689121079,
    variance=0.99418092655064605,
    skewness=0.013953400984243667,
    kurtosis=-0.671060517912661
  )
```

,,,
,,,

Normality Tests (Skewness and Kurtosis):-

Normality tests are based on the skewness and kurtosis.

The "normaltest()" function returns p value for the null hypothesis:

"x comes from a normal distribution".

SKEWNESS & KURTOSIS:-

Skewness:
A measure of symmetry in data.
For normal distributions it is 0.
If it is negative, it means the data is skewed left.
If it is positive it means the data is skewed right.
SEE THIS IMAGE (https://cdn.analyticsvidhya.com/wp-content/uploads/2020/06/sk1.png)

Kurtosis:
A measure of whether the data is heavy or lightly tailed to a normal distribution.
Positive kurtosis means heavy tailed.
Negative kurtosis means lightly tailed.
SEE THIS IMAGE (https://miro.medium.com/max/830/0*WzpbLu8KqMAryg-9)

Example:-

Find skewness and kurtosis of values in an array:

```
import numpy as np
from scipy.stats import skew, kurtosis

v = np.random.normal(size=100)

print(skew(v))
print(kurtosis(v))
,,,,,,,,,,,,,,,,,,,,,,,,,,,o/p is 0.11168446328610283
                                 -0.1879320563260931
```

Application of SKEWNESS & KURTOSIS:-

*Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.

*In finance, kurtosis is used as a measure of financial risk. Learn risk analysis.

*linear models work on the assumption that the distribution of the independent variable and the target variable are similar. Therefore, knowing about the skewness of data helps us in creating better linear models.

*A positive mean with a positive skew is good, while a negative mean with a positive skew is not good. If a data set has a positive skew, but the mean of the returns is negative, it means that overall performance is negative, but the outlier months are positive.