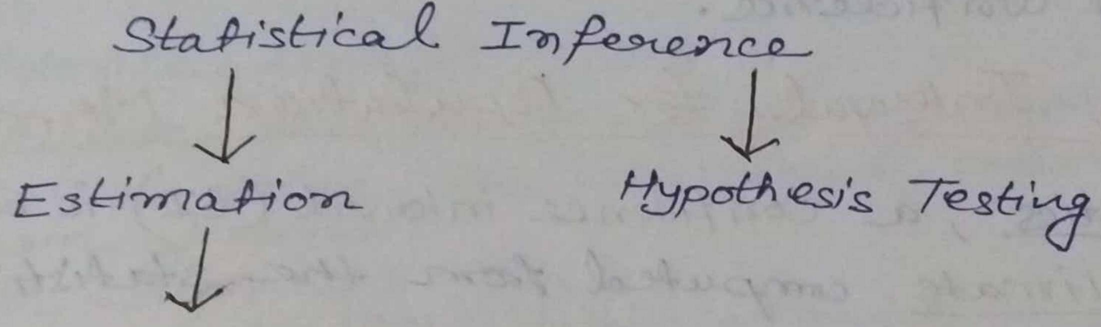
Statistical Inference


Statistical Inference

Estimation

Hypothesis Testing

1. Point estimates
2. Confidence Interval

Point Estimation

- Point estimation involves calculating a single descriptive statistic to estimate the population parameter.
- For example, if we calculate the mean entrance exam score for a sample of 25 students, then this would be the point estimate of the population mean.
- The proportion of students qualified the entrace examination, this would be a point estimate of the proportion.

Interval Estimation

- Point estimates does not convey any information about margin of error, so inference about the accuracy of the parameter estimate cannot objectively be made.
- Interval estimation indicates a range of values (upper and lower) within which the parameter has a specified probability of lying.
- Constructing a confidence inteaval (CI) around a statistic establishes a range of values as well as the probability of being right. This.
- means that the CI is made with a certain degree of confidence.

Confidence Interval for Population Mean

- In statistics, a confidence interval (CI) is a type of estimate computed from the statistics of the observed data.
- This proposes a range of plausible values for an unknown parameter. The interval has an associated confidence level that the true parameter is in the proposed range.
- This is more clearly stated as: the confidence level represents the probability that the unknown parameter lies in the stated interval. The level of confidence can be chosen by the investigator.
- In general terms, a confidence interval for an unknown parameter is based on sampling the distribution of a corresponding estimator

Mostly, the confidence level is selected before examining the data. The commonly used confidence level is 95% confidence level. However, other confidence levels are also used, such as 90% and 99% confidence levels.

Confidence Interval formula

$$\text{Confidence Interval} = \left(\bar{x} - z \times \frac{\sigma}{\sqrt{n}}\right) \text{ to } \left(\bar{x} + z \times \frac{\sigma}{\sqrt{n}}\right)$$

$$= \bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$$

```
import scipy.stats as stats
import math
x = 6
n = 30
confidence_level = 0.95
# Calculate the point estimate, alpha, the
critical z-value, the standread errors, and the
margin of error
point_estimate = x/n
alpha = (1 - confidence_level)
critical_z = stats.norm.ppf(1-alpha/2)
 standard_error = math.sqrt(point_estimate*
                  (1-point_estimate)/n))
 margin_of_error = critical_z * standard_error
# Calculate the lower and upper bound of the
                       confidence interval
lower_bound = point_estimate - margin_of_erro
upper_bound = point_estimate + margin_of_erro
# Point the results
 print("Point Estimate: {:.3f}".format(point_estimate)
print("Critical Z-value: {:.3f}".format(critical_z))
print("Margin of Error: {:.3f}".format(margin_of_erro
point("Confidence Interval: [{:.3f}, {:.3f}]".format
                       (lower_bound, upper_bound))
point("The {:.1%} confidence interval for the
            population proportion is:".format(confidence
                                   level))
```

```
print ("between {:.3f} and {:.3f}".format
                    (lower_bound, upper_bound))
```

Errors:

Absolute Error:- the amount of error in your measurement. For example, if you step on a scale and it says 150 pounds but you know your true weight is 145 pounds, then the scale has an absolute error of 150lbs - 145lbs = 5lbs

Relative Error:- the ratio of the absolute error to the accepted measurement. As a formula, that's

$$E_{relative} = \frac{E_{absolute}}{E_{measured}}$$

Imp

Hypothesis in Statistics

Hypothesis is an assumption about a parameter in population.

Null Hypothesis

It assumes that the observation is not ~~static~~ stastically significant.

Alternate Hypothesis

It assumes that the observations are due to some reason.

Its alternate to Null Hypothesis.

Example

~~For~~ an assessment of a student we would take:

"Student is worse than average"- as a null hypothesis, and:

"Student is better than average- as an alternate hypothesis

One tailed test

When our hypothesis is testing for one side of the value only, it is called "one tailed test".

Example

For the null hypothesis:
"the mean is equal to K", we can have alternate hypothesis
"the mean is less than K", or;
"the mean is greater than K"

Alpha value

Alpha value is the level of significance.

Example

How close to extremes the data must be for null hypothesis to be rejected.
He is usually taken as 0.01, 0.05, or 0.1. It means only rejected value for eg ⇒ 95% is condience interval and 5% is alpha value. $\frac{5}{100} = 0.05$.

P value

P value tells how close to extreme the data actually is:

P value and alpha values are compared to establish the statistical significance.

If pvalue <= alpha we unable ~~[illegible]~~ to accept ~~reject~~ the null hypothesis.

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.

It is a two tailed test.

The function ttest_ind() takes two samples of same size and produces a tuple of t-statistic and

P-value.

Find if the given values v1 and v2 are from same distribution;

```
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)
print (res)
```

Output

```
TtestResult (statistic = -1.4931638 , pvalue = 0.1369860,
df = 198.0 )
```

Note

It is only ttest

Null Hypothesis - There ~~are~~ is no difference between in mean. It means ~~[illegible]~~ they are ~~follow~~ from same population

Alternative Hyothesis - There ~~are~~ is difference between in mean. It means ~~that is in differe~~ they are from difference population.

~~Since~~

1) If you want to return ~~it~~ only the p-value, use the pvalue property

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

Output

```
0.13698606
```

KS - Test

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.

eg → Find if the given value follows the normal distribution.

```
import numpy as np
from scipy.stats import kstest
v= np.random.normal(size=100)
res = kstest(v,'norm')
print (res)
```

Output

```
KstestRsult(statistic=0.016 ..., pvalue=0.8239,
statistic_location=0.8463. ..., statistic_sign=1)
```

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

⊙ 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)
```

Output

```
DescribeResult(nobs=100, minmax=(-2.0991855456740
2.13041427), mean=0.1150374, variance=0.99418,
skewness= 0.013953, Kurtosis=-0.6710605)
```

Normality Test (Skewness and Kurtosis)

Normality tests are based on the skewness and Kurtosis. The normaltest( ) function returns p value for the null ~~p~~ hypothesis:
"x comes from a normal distribution".

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 positi
it means the data is skewed right.

e.g. 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))
```

Output

```
 0.111 68 44
-0.187932
```

e.g Find if the data comes from a normal
distribution:

```
import numpy as np
from scipy.stats import normaltest
v = np.random.normal (size = 100)
print (normaltest(v))
```

Output

```
NormaltestResult (statistic = 4.478 . .  pvalue=0.10
```

Outlier

Outlier is those data which is not behaving like a normal data of those data set.