Inferential Statistics

Agenda

01 What is Inferential Statistics?

02 Confidence Interval

03 Hypothesis Testing

04 Errors in Hypothesis Testing

05 t-Test

06 f-test

07 z-test

08 Chi-Square Test & ANOVA

What is Inferential Statistics?

What is Inferential Statistics?

While descriptive statistics describes the data, inferential statistics is used to draw conclusions about the population based on statistical findings on sample analysis.

Confidence Interval

Confidence Interval

Confidence interval assumes certainty of population parameter falling in the given intervals i.e. 95%, 99%, etc.

For example: If a point estimate 10.0 from the sample statistics for the confidence interval 95% falls into 9.5 to 10.5, we can infer that there is a 95% certainty that the true or population estimate will fall in the same interval.

Confidence Interval

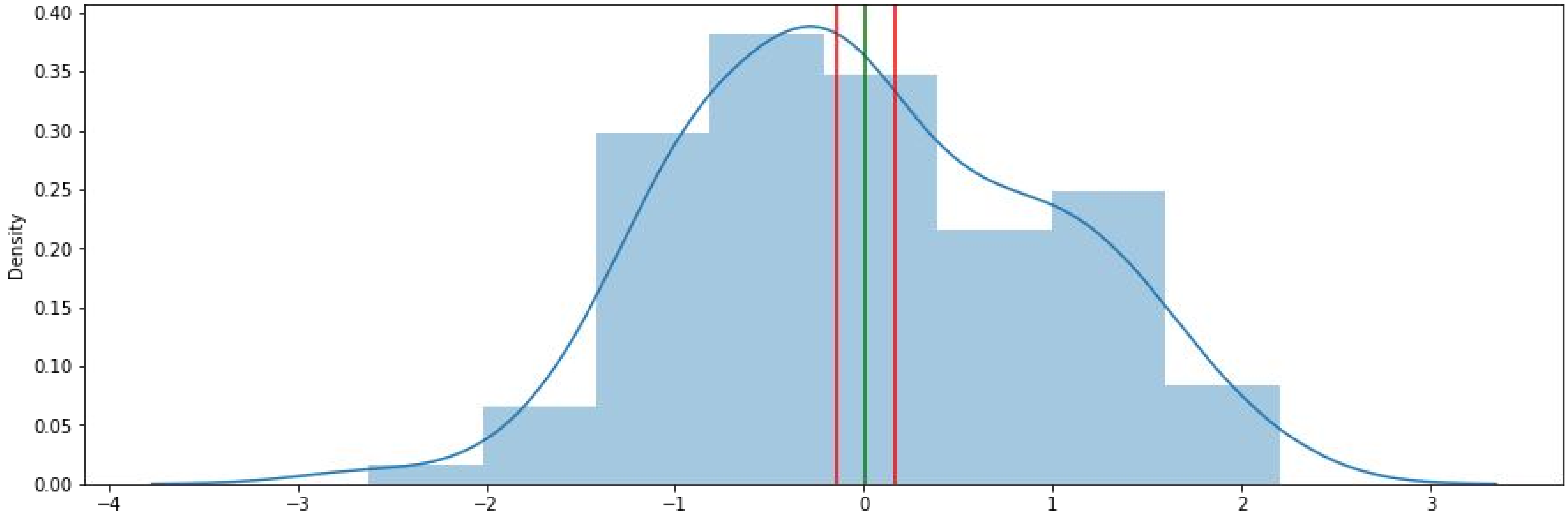
```
#confidence interval
import scipy.stats as st
import statistics as s
x = np.random.normal(size=100)
sample_mean = np.mean(x)
sample_std = s.stdev(x)
std_err = st.sem(x)
Z_value = st.norm.ppf(1 - 0.05)

lowerCi = sample_mean - (Z_value * std_err)
upperCi = sample_mean + (Z_value * std_err)

#plotting CI
plt.figure(figsize=(15,5))
sns.distplot(x)
plt.axvline(x=lowerCi, color='red')
plt.axvline(x=upperCi, color='red')
plt.axvline(x=sample_mean, color='green')
```

We have taken a random normal sample of size 100, and calculated the lower confidence interval and upper confidence interval with interval value 95%.

Confidence Interval



According to our analysis, there is 95% certainty that the population will have the mean in the given interval.

Hypothesis Testing

Hypothesis Testing

Hypothesis testing is the analysis where the plausibility of an assumption for a population parameter is tested on the sample, and statistical evidence is used to verify the hypothesis.

Steps involved in Hypothesis Testing

01 Formulate Two Hypothesis for analysis

02 Draw samples from population for analysis

03 Perform appropriate statistical test

04 Accept or reject hypothesis based on evidence

Hypothesis Testing

Null Hypothesis

Null Hypothesis states that there is no effect on the population mean.

Alternate Hypothesis

Alternate Hypothesis states that there is effect on the population mean

Errors in Hypothesis Testing

Errors in Hypothesis Testing

Type 1 Error

The Type 1 error is the false positive error where we have rejected the null hypothesis but it is actually true.

Type 2 Error

Type 2 error is a false negative conclusion where we have not rejected the null hypothesis but it is actually false.

T-Test

T-Test

T-test is a parametric test, that compares the means of the two samples. Ideally, a sample for t-test should have less than 30 values. There are a few other assumptions that are taken before we can conduct a t-test.

Assumptions

1. The samples are independent
2. Homogeneity in sample variances
3. The Data is assumed to be normally distributed.

Types of t-test

One-sample

If we are comparing the sample against a standard value.

Two-sample

If both the samples are taken from two different populations.

Paired

If the samples are taken from the same population.

One-Tailed vs Two-Tailed T-Test

One-tailed

If we want to check whether the population means are greater than or smaller than, we will use one-tailed test.

Two-tailed

If we want to check whether the population means differ significantly, we will use a two tailed test.

One Sample t-test

The average height of Indian adult males is 165cm.

Null hypothesis: The average height is 165cm.
Alternate Hypothesis: The average height is not 165cm.

One Sample t-test

We will use python programming to perform a one sample test on a random sample taken from adult Indian males, where each of the 30 samples have their heights in cm.

```
#one sample t-test
from scipy.stats import ttest_1samp
from random import sample

#generating a random sample to get the heights
sample = sample(range(145, 180), 30)
#calculating the sample mean
sample_mean = np.mean(sample)

#one-sample t-test parameters
ttest_1samp(a=sample, popmean=165)
```

```
Ttest_1sampResult(statistic=-2.060128462146794, pvalue=0.04845967670620546)
```

Since the p-value is less than 0.05, we can reject the null hypothesis.

Two Sample t-test

We have to check whether the mean height of adult males in both the schools is same or not.

Null hypothesis: The means are equal.
Alternate Hypothesis: The means are not equal.

Two-Sample t-test

We will check the variances of each groups and then perform a two-sample t-test for equal variances, otherwise a Welch's t-test will be conducted by not taking into consideration – the unequal population variances.

```
#two sample t-test
from random import sample
sample_1 = sample(range(140, 184), 30)
sample_2 = sample(range(140, 184), 30)

var_1 = np.var(sample_1)
var_2 = np.var(sample_2)
print(var_1, var_2)
```

```
169.0 164.56555555555553
```

```
from scipy.stats import ttest_ind

ttest_ind(sample_1, sample_2, equal_var = True)
```

```
Ttest_indResult(statistic=0.28502643634986835, pvalue=0.7766392074708405)
```

We have insufficient evidence to reject the null hypothesis.

Two Sample t-test

We have to check if the mean of heights of males and females are same in the school?

Null hypothesis: The means are equal.
Alternate Hypothesis: The means are not equal.

Paired t-test

We will use the paired sample t-test for the groups because the samples come from the same population.

```
#paired t-test
from random import sample
from scipy.stats import ttest_rel

sample_female = sample(range(135, 170), 30)
sample_male = sample(range(145, 180), 30)

ttest_rel(sample_female, sample_male)
```

```
Ttest_relResult(statistic=-4.284988931336786, pvalue=0.00018363298182473822)
```

We have sufficient evidence to reject the null hypothesis.

F-Test

F-Test

F-test is a statistical test that is used to compare the variances of two populations. There are several assumptions that are made about the data before we can begin the F-test.

Assumptions

1. Data is normally distributed

2. The data is independent

f-test

We have to check if the variances of the two populations where the groups are taken from equal or not.

Null hypothesis: The variances are equal.
Alternate Hypothesis: The variances are not equal.

F-test

We will calculate the variances of the two samples and compute the f-statistic and p-value to gather statistical evidence to reject the null hypothesis.

```
#f-test
from random import sample
import scipy

sample_1 = sample(range(0,100), 30)
sample_2 = sample(range(0,100), 30)
f = np.var(sample_1)/np.var(sample_2)
p = 1 - scipy.stats.f.cdf(f, (len(sample_1)-1), (len(sample_2)-1))
print(f, p)
```

```
1.0351902254518512 0.46322108632360104
```

Not enough evidence to reject the null hypothesis.

ANOVA

ANOVA

ANOVA or Analysis of Variance is a statistical test that compares the means or two or more groups to find significance or either groups on one another or how different they are from each other.

Assumptions

1. Independent Samples

2. All populations have common variance

3. Samples are drawn from normally distributed population

One-Way ANOVA

We have to check if the effect of 4 different performance enhancers on an electric vehicle is same or not?

Null hypothesis: The performance averages are equal.
Alternate Hypothesis: The performance averages are not equal.

One-Way ANOVA

We have taken 4 random samples that has performance values, we will calculate the test statistics and p-value to reject or fail to reject he null hypothesis.

```
#One-factor ANOVA
from random import sample
from scipy.stats import f_oneway

sample_1 = sample(range(0,100), 20)
sample_2 = sample(range(0,95), 20)
sample_3 = sample(range(0,120), 20)
sample_4 = sample(range(0,145), 20)

f_oneway(sample_1, sample_2, sample_3, sample_4)
```

```
F_onewayResult(statistic=3.1076995586786063, pvalue=0.03133772988980599)
```

P-value is less than 0.05, we can reject the null hypothesis.

Two-Way ANOVA

Two way ANOVA checks how two factors will affect the response variable.

Null hypothesis: There is no significance of the two factors on response variable.
Alternate Hypothesis: There is significance of the two factors on response variable.

One-Way ANOVA

```python
#Two-factor ANOVA
import statsmodels.api as sm
from statsmodels.formula.api import ols

x = {'Lectures': np.repeat(["Daily", "Weekly"], 20),
     'Tuition': np.repeat(["Daily","Weekly"], 20),
     'Marks': sample(range(33, 100), 40)}

data = pd.DataFrame(x)

# Performing two-way ANOVA
model = ols('Marks ~ C(Lectures) + C(Tuition) + C(Lectures):C(Tuition)', data=data).fit()
sm.stats.anova_lm(model, typ=2)
```

	sum_sq	df	F	PR(>F)
C(Lectures)	349.601151	1.0	1.125000	0.295540
C(Tuition)	349.601151	1.0	1.125000	0.295540
C(Lectures):C(Tuition)	570.025000	1.0	1.834314	0.183618
Residual	11808.750000	38.0	NaN	NaN

There is no evidence to reject the null hypothesis.

Z-Test

Z-Test

Z-test is a statistical test to compare the means of populations where the variances are known and sample sizes are considerably larger compared to t-test.

Assumptions

1. Standard Deviation and variances are known.

2. Population should be 10 times as much as the sample size.

3. Samples are drawn at random from the population.

One Sample z-test for Means

The average weight of the high-schoolers pre pandemic was 55Kg with a standard deviation of 8. Has it changed post pandemic?

Null hypothesis: The average weight is same.
Alternate Hypothesis: The average weight is not same.

One Sample z-test for Means

We will use a one sample z-test for this problem, where we will take weights of 50 high schoolers randomly and perform the z-test using python.

```
#one-sample z-test
from random import sample, choices
from statsmodels.stats.weightstats import ztest

sample = sample(range(30, 80), 50)
ztest(sample, value=55)
```

```
(-0.24253562503633297, 0.8083651559145103)
```

Not enough evidence to reject the null hyptohesis

Two Sample z-test for Means

Is the average height post pandemic for high schoolers going to school A and school B is same, given that the standard deviation of the populations is known.

Null hypothesis: The mean difference is zero.
Alternate Hypothesis: The mean difference is not zero.

Two Sample z-test for Means

We will take one sample from each of the populations with 50 individuals each. And then perform a two-sample z-test using python.

```
#two-sample z-test
from random import sample, choices
from statsmodels.stats.weightstats import ztest

sample_1 = sample(range(130, 185), 50)
sample_2 = sample(range(130, 185), 50)

ztest(sample_1, sample_2, value=0)
```

```
(0.5098286102416721, 0.6101715399231471)
```

Not enough evidence to reject the null hypothesis

One Sample z-test for Proportion

It was observed from a purchase case study, that 35% of women spend more than 10000. Is it true for our population in analysis?

Null hypothesis: The proportion is same.
Alternate Hypothesis: The proportion is not same.

One Sample z-test for Proportion

```
data_new = data.loc[(data['Purchase'] > 10000)]

#No of women in the sample
count = data_new['Gender'].value_counts()[0]

#number of observations
nobs = len(data_new['Gender'])

#hypothesised value
p0 = 0.35

#Z-test
from statsmodels.stats.proportion import proportions_ztest

z_stat, p_val = proportions_ztest(count=count,
                                  nobs=nobs,
                                  value=p0,
                                  alternative="two-sided",
                                  prop_var=False)
print(z_stat, p_val)
```

```
478.72085551496957 0.0
```

We will perform a one sample z-test for proportion to check the test statistics in order to reject or fail to reject the null hypothesis. Since the p-value is less than 0.05, we can reject the null hypothesis.

Two Sample z-test for Proportion

Is the percentage of men who have spend more than 10000 sam
e for the ages 18-25 and 26-35

Null hypothesis: The proportion is same.
Alternate Hypothesis: The proportion is not same.

z-test for Proportion

```
#two-sample test of proportion
data_age1 = data.loc[(data['Age'] == 1) & (data['Purchase'] > 10000)]
data_age2 = data.loc[(data['Age'] == 2) & (data['Purchase'] > 10000)]

#sampling
data_age1_sample = data_age1.sample(1000, random_state=0)
data_age2_sample = data_age2.sample(1000, random_state=0)

#count
count = [(data_age1_sample['Gender'] == 1).sum(), (data_age2_sample['Gender'] == 1).sum()]

#nobs
nobs = [(len(data_age1_sample)), len(data_age2_sample)]

#Z-test
from statsmodels.stats.proportion import proportions_ztest
stat_2sample, p_value_2sample = proportions_ztest(count=count,
                                                  nobs=nobs,
                                                  value=0,
                                                  alternative='two-sided',
                                                  prop_var=False)
print(stat_2sample, p_value_2sample)
```

```
0.5084344113930828 0.6111487252921447
```

We will perform a two sample z-test for proportion to check the test statistics in order to reject or fail to reject the null hypothesis. Not sufficient evidence to reject the null hypothesis.

Chi-Square Test

Chi-Square Test

Chi-Square test for categorical data that can be used to check the goodness of fit or test of independence.

Assumptions

1. The features are categorical in Nature

2. The samples are taken at random.

3. Minimum of five observations expected in each group.

Chi-Square Test of Independence

Is Purchase independent of Product_Category_1?

Null hypothesis: Purchase and product_category_1 are not related
Alternate Hypothesis: Purchase and product_category_1 are related

Chi-Square Test of Independence

```python
#chi-square test of independence
data['Purchase'].max()
data['Purchase'] = pd.cut(data['Purchase'], bins=[0, 10000, 23961], labels=[0,1])

#making a cross table
cross_table = pd.crosstab(data['Purchase'], data['Product_Category_1'])
```

```python
scipy.stats.chi2_contingency(cross_table)
```

```
(359770.82102148153,
 0.0,
 19,
 array([[92030.13737211, 15644.95290037, 13251.40097952,  7705.12619167,
         98949.86909618, 13417.26475272,  2439.4430834 , 74687.86704553,
           268.79109492,  3359.88868649, 15922.26663976,  2587.60597962,
          3637.85801392,   998.46057942,  4123.64874888,  6443.11922162,
           378.92988503,  2048.71261371,  1050.90762233,  1671.74949279],
        [48347.86262789,  8219.04709963,  6961.59902048,  4047.87380833,
         51983.13090382,  7048.73524728,  1281.5569166 , 39237.13295447,
           141.20890508,  1765.11131351,  8364.73336024,  1359.39402038,
          1911.14198608,   524.53942058,  2166.35125112,  3384.88077838,
           199.07011497,  1076.28738629,   552.09237767,   878.25050721]]))
```

We will perform chi-square test of independence and validate our assumptions based on statistical evidence. P-value is less than 0.05, we can reject the null hypothesis.

India: +91-7847955955

US: 1-800-216-8930 (TOLL FREE)

support@intellipaat.com

24/7 Chat with Our Course Advisor