Statistical Inference

Ms Swapnil Shrivastava
C-DAC Bangalore

Content

- Statistical Inference
- Hypothesis Testing
  - Procedure
  - One sample testing
  - Two sample testing
- Analysis of Variance (ANOVA)
- Chi-square testing
- Summary
- References

Statistical Inference

- The managers may want to validate the effectiveness of decisions they have made or planning to make.
- Statistical Inference focuses on drawing conclusions about populations from samples.
- It includes estimation of population parameters and hypothesis testing.
  - Drawing conclusions about the value of the parameters of one or more populations based on sample data.
- The fundamental statistical approach for doing this is called hypothesis testing.

Examples

- Did an advertising campaign increase sales?
- Will product placement in a grocery store make a difference?
- Did a new assembly improve productivity or quality in a factory?

Hypothesis Testing

- To validate statistical conclusions about the value of population parameters or difference among them.
- It involves drawing inferences about two contrasting propositions each called a hypothesis.
  - Null hypothesis (H_0): describe existing theory or a belief.
  - Alternative hypothesis (H_1) : complement of the null hypothesis.
- They validate the values of one or more population parameters such as mean, proportion, standard deviation or variance.

Hypothesis Testing

- Alternative hypothesis must be true if the null hypothesis is false.
- Using sample data, we either
  - *Reject* the null hypothesis and conclude that the sample data provide sufficient statistical evidence to support the alternative hypothesis.
  - *Fail to reject* the null hypothesis and conclude that the sample data does not support the alternative hypothesis.
- If we fail to reject the null hypothesis, then we can only assume that the existing theory or belief is true, although we haven’t proven it.

A Legal Analogy

- In system of justice, a defendant is innocent until proven guilty.
  - Null hypothesis – is not guilty
  - Alternative hypothesis – is guilty
- Sample data is evidence.
  - If the evidence strongly indicates that defendant is guilty then we reject the null hypothesis.
  - If the evidence is not sufficient to indicate guilt, then we cannot reject the not guilty hypothesis.

Hypothesis Testing Procedure

1. Identifying the population parameter of interest and formulating the hypotheses to test.
2. Selecting a level of significance, which defines the risk of drawing an incorrect conclusion when the assumed hypothesis is actually true.
3. Determining a decision rule on which to base a conclusion.
4. Collecting data and calculating a test statistics.
5. Applying the decision rule to the test statistic and drawing a conclusion.

This procedure is applicable to both single population (one sample test) as well as more than one population (multiple sample tests).

One-sample Hypothesis Tests

- It involves a single population parameter such as the mean, proportion, standard deviation and so on.
- Use a single sample of data from the population.
- Three types of one-sample hypothesis tests
  - H_0: population parameter >= constant vs H_1: population parameter < constant
  - H_0: population parameter <= constant vs H_1: population parameter > constant
  - H_0: population parameter = constant vs H_1: population parameter <> constant

Formulating a One-Sample Test of Hypothesis

Cadsoft a producer of computer aided design software for the aerospace industry receives numerous calls for technical support. In the past, the average response time has been at least 25 minutes. The company has upgraded its information systems and believes that this will help reduce response time. As a result, it believes that the average response time can be reduced to less than 25 minutes.

The company collected the sample of 44 response times in the CadSoft Technical Support Response Times excel file.

- Null hypothesis: population mean response time >= 25 minutes
- Alternate Hypothesis: population mean response time < 25 minutes

H_0: $\mu >= 25$

H_1: $\mu < 25$

Potential Errors In Hypothesis Testing

- The null hypothesis is actually true, and the test correctly fails to reject it.
- The null hypothesis is actually false, and the hypothesis test correctly reaches this conclusion.
- The null hypothesis is actually true, but the hypothesis test incorrectly rejects it (called Type I error)
- The null hypothesis is actually false, but the hypothesis test incorrectly fails to reject it (called Type II error)

Understanding Type I error

- Level of significance(α) : probability of making a Type I error i.e. P(rejecting H_0| H_0 is true)
  - Likelihood that is acceptable in making the incorrect conclusion
  - Selected by decision maker. Common values are: 0.10, 0.05 and 0.01
- Confidence Coefficient: probability of correctly failing to reject the null hypothesis i.e. P(not rejecting H_0| H_0 is true)
  - Calculated as 1- α.

Understanding Type ll error

- β represents probability of Type II error i.e. P(not rejecting H0| H0 is false)
  - Cannot be controlled.
- 1- β represents the probability of correctly rejecting the null hypothesis when it is indeed false i.e. P(rejecting H0| H0 is false)
  - It is also called the power of the test
- The power of test depends upon the sample size.
  - Large sample size, large 1- β

β and True Population Mean

- The farther away the true mean response time is from the hypothesized value, the smaller is β.
- As α decreases β increases.
  - If α is 0.01 instead of 0.05 and sample size is constant, probability of Type I error would reduce but probability of Type II error would increase.
- If α is small, choose large sample size while conducting test.

Selecting the Test Statistics

- Collect the sample data and use the data to draw a conclusion.
- The decision to reject or fail to reject a null hypothesis is based on computing a test statistics from sample data.
- The test statistics used depends on the type of hypothesis test.
- For example, test statistics formulas
  - One-sample test for mean, σ known $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
  - One-sample test for mean, σ unknown $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Computing the Test Statistic

- In CadSoft example, the population standard deviation is not known.
- Hence test statistic is calculated using

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where

n = 44 customers

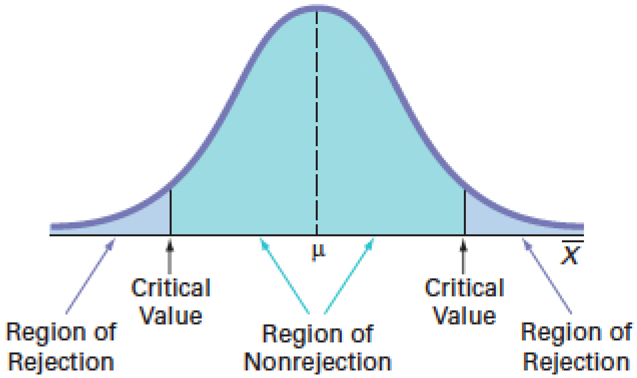
$\bar{x}$ = 21.91 minutes

s = 19.49

μ_0 = 25

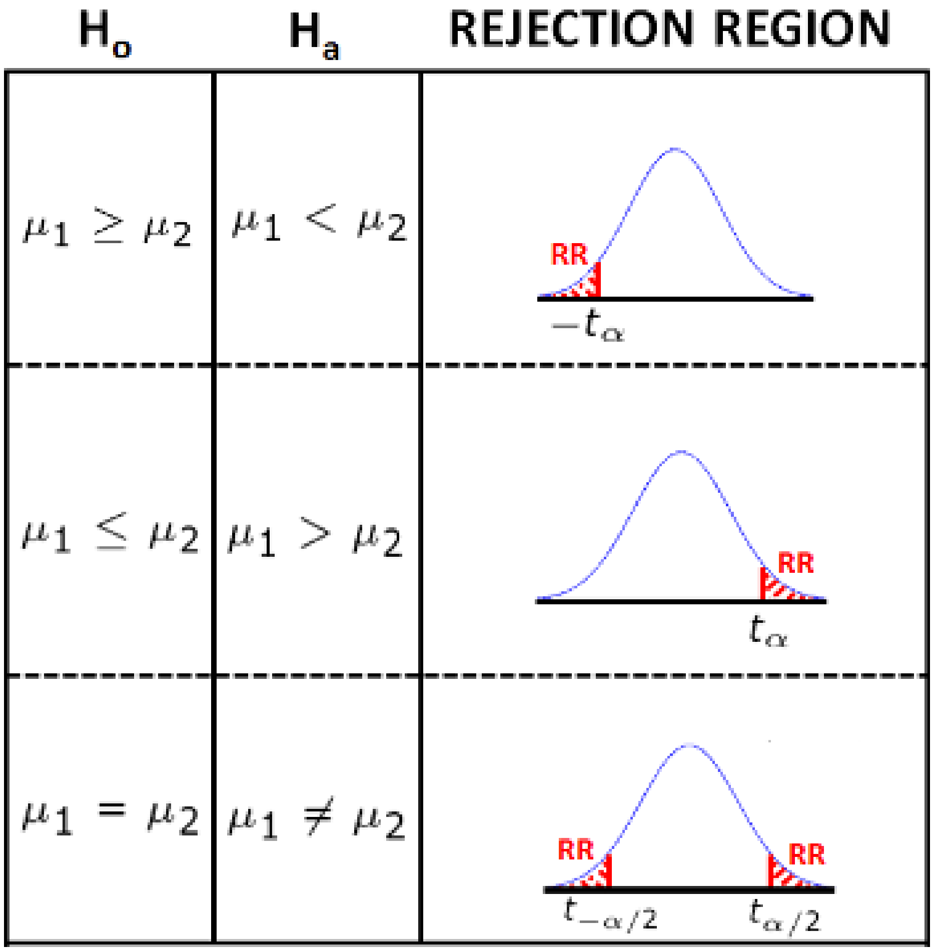
- The value of test statistic t is -1.05

Drawing a Conclusion



- The decision to reject or fail to reject H0 is based on comparing the value of the test statistic to a “critical value”
- The sampling distribution of test statistics could be normal distribution, t-distribution and more.
- The critical value divides the sampling distribution into two parts, a rejection region and a non rejection region.
- If null hypothesis is false, it is more likely that the test statistic will fall into the rejection region.

Rejection Region in Hypothesis Testing



H_o	H_a	REJECTION REGION
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	RR $-t_\alpha$
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	RR t_α
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	RR RR $t_{-\alpha/2}$ $t_{\alpha/2}$

P values

- To find the probability of obtaining a test statistic value equal to or more extreme than that obtained from the sample data when the null hypothesis is true.
- If p-values is less than α then
  - reject the null hypothesis
  - else fail to reject.

Finding the Critical Value

- If the level of significance is 0.05, then the critical value for a one-tail test is the value of the t-distribution with n-1 degrees of freedom that provides a tail area of 0.05 i.e. $t_{\alpha, n-1}$
- In CadSoft example, the critical value is $t_{0.05,43}$=1.68.
- The t-distribution is symmetric with a mean of 0 and is a low tail test. So -1.68 could be chosen as a critical value.
  - Test statistic is -1.05 and it is greater than – 1.68.
  - Is not in rejection region and therefore we cannot reject $H_{0.}$

One Sample T-Test

```
import pandas as pd
from scipy import stats

df=pd.read_excel("F:\Business Analytics\Data_Files\CadSoft Technical Support Response Times.xlsx", skiprows=2)
print(df)

print(stats.ttest_1samp(df[" Time (min)"],25))
```

Output

```
Ttest_1sampResult(statistic=-1.0521681183492575, pvalue=0.2985994510452377)
```

Two tailed test of Hypothesis for the Mean

- Vacation Survey excel file has data collected from 34 respondents. The travel agency wants to target individuals who were approximately 35 years old. Thus we wish to test whether the average age of respondents is equal to 35.

  H_0: mean age = 35

  H_1: mean age <> 35

- t-test statistics for unknown standard deviation is 2.73
- The critical value by calculating $T_{0.025,33}$ is 2.0345
- The test statistics fall in rejection region and hence we must reject the null hypothesis.

Two-sample Hypothesis Tests

- Many practical applications of hypothesis testing involve comparing two populations for differences in means, proportions or other population parameters.
- They include
  - Confirm differences between suppliers
  - Performance of two different factory locations
  - New and old work methods
  - New and old reward and recognition programs

Two Sample Test Hypothesis

- Lower-tailed test H0: population parameter (1) - population parameter(2) >= 0 vs H1 : population parameter (2) < 0. This test seeks evidence that population parameter (2) is larger than parameter(1)
- Upper-tailed test H0: population parameter (1) - population parameter(2) <= 0 vs H1 : population parameter (2) > 0. This test seeks evidence that population parameter (2) is smaller than parameter(1)
- Two-tailed test H0: population parameter (1) - population parameter(2) = 0 vs H1 : population parameter (2) <> 0. This test seeks evidence that population parameter (2) is the same as parameter(1)

Two sample tests for differences in Means

H_0: $\mu_1 - \mu_2$ {>=, <= or =} 0
H_1: $\mu_1 - \mu_2$ {<, > or <>} 0

- Type of test
  - Two-sample test for means, σ^2 known
  - Two-sample test for means, σ^2 unknown, assumed unequal
  - Two-sample test for means, σ^2 unknown, assumed equal
  - Paired two sample test for means
  - Two sample test for equality of variances

Comparing Supplier Performance

- Purchasing managers have noted that they order many of the same types of items from Alum Steering and Durrable Products.
- They are considering dropping Alum Sheeting from its supplier base if its lead time is significantly longer than that of Durrable Products.

$$H_0: \mu_1 - \mu_2 <= 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

- Where
  - μ_1 = mean lead time for Alum Sheeting
  - μ_2 = mean lead time for Durrable Products
- T test: two sample assuming unequal variance.

Two sample test for Means with Paired Samples

- In many situations, data from two samples are naturally paired or matched.
- In such cases, a paired t-test is more accurate than assuming that the data come from independent populations.
- The null hypothesis we test revolves around the (μ_D) between the paired samples.

  H_0: μ_D{>=, <= or =} 0

  H_1: μ_D{<, > or <>} 0
- The test uses the average difference between the paired data and the standard deviation of the differences.

Example

- The Excel file Pile Foundation contains the estimates used in a bid and actual auger-cast pile lengths that engineers ultimately had to use for a foundation engineering project.
- The contractor’s past experience suggested that the bid information was generally accurate.
- So the average of the paired differences between the actual pile lengths and estimated lengths should be close to zero.
- After the project was completed, the contractor found that the average difference between the actual lengths and the estimated lengths was 6.38.
- Could the contractor conclude that the bid information was poor?

Paired Two Sample Test for Means

```
import pandas as pd
from scipy import stats

df=pd.read_excel("F:\Business Analytics\Data_Files\Pile Foundation.xlsx", skiprows=3)
print(df)

print(stats.ttest_rel(df["Estimated"],df["Actual"]))
```

<u>Output</u>

```
Ttest_relResult(statistic=-10.912250253185432, pvalue=1.1188697222438401e-23)
```

F-test

- It is used to test equality of variance between two samples.
- To use this test, we assume that both samples are draw from normal populations.

$$H_0: \sigma^2{}_1 - \sigma^2{}_2 = 0$$
$$H_1: \sigma^2{}_1 - \sigma^2{}_2 > 0$$

- The F-test statistics F= $s^2{}_1$ /$s^2{}_2$ where s_1 and s_2 are variance calculated from sample 1 of population 1 and sample 2 of population 2 respectively.
- These samples follows F-distribution. It has two degree of freedoms viz n_1-1 and n_2-1
- Variance of lead time for Alum Sheeting and Durable Products in the Purchase Orders data.

Analysis of Variance (ANOVA)

- Compare the means of several different groups to determine if all are equal or if any are significantly different from the rest.
- This variable of interest is called factor.
- In the Excel data file Insurance Survey, the factor is educational level. There are three categorical levels and we may be interested in finding whether any significant difference exist in satisfaction among individuals with different levels of education.
  - Can mean be used as measure?
  - Pairwise test

ANOVA Test

- ANOVA: analyze variance in the data
  - The null hypothesis of ANOVA is that population mean of all groups are equal.
  - The alternative hypothesis is that at least one mean differs from the rest.

    $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_m$

    H_1: at least one mean is different from others
- It computes a measure of the variance between the means of each group and a measure of variance within the groups and examines a test statistic that is the ratio of these measures.
  - If F-statistic > F-critical reject null hypothesis
  - If p-value < level of significance reject null hypothesis

ANOVA Test

```
import pandas as pd
from scipy.stats import f_oneway

df=pd.read_excel("F:\Business Analytics\Data_Files\Insurance Survey.xlsx", skiprows=2, nrows=24)
print(df.columns)

df1=df.loc[df["Education"]=="Some college"]["Satisfaction* "]
df2=df.loc[df["Education"]=="Graduate degree"]["Satisfaction* "]
df3=df.loc[df["Education"]=="College graduate"]["Satisfaction* "]

print(f_oneway(df1,df2,df3))
```

Output

```
F_onewayResult(statistic=3.924651731927711, pvalue=0.03563539756488997)
```

Chi-Square Test for Independence

- The chi-square test for independence test the following hypotheses:
  - H0: the two categorical variables are independent
  - H1: the two categorical variables are dependent.
- An example of nonparametric test i.e. one that does not depend on restrictive statistical assumptions.
- It is used for understanding relationships among categorical data.

Chi-Square Test

- For the Energy drink survey data
- Step 1: Compute the expected frequency in each cell of the cross tabulation if the two variables are independent. This is done using the following:

  Expected frequency in row i and column j =

  $$\frac{\text{(grant total row i) (grand total column j)}}{\text{total number of observations}}$$

- Step 2: Chi-square statistics

  $$X^2 = \frac{\Sigma\ (f_0 - f_e)^2}{f_e}$$

  where f_o – original frequency

  f_e – expected frequency

- The sampling distribution for X^2 is chi-square ($X^{2)}$ distribution. Calculated in terms of degree of freedom (r-1)(c-1) and level of significance α.

Conducting the Chi-Square Test

```
import pandas as pd
from scipy.stats import chi2_contingency
import numpy as np

df=pd.read_excel("F:\Business Analytics\Data_Files\Energy Drink Survey.xlsx", skiprows=2)
print(df)

table=pd.crosstab(df["Gender"],df["Brand Preference"])

‘f_obs=table.loc["Female"].tolist()+table.loc["Male"].tolist()
f_obs_arr=np.array(f_obs)
f_obs_arr.shape=( 2, 3 )

chi2_contingency(f_obs_arr)
```

Output

```
(6.4924250792329055, 0.038921342064441915, 2, array([[12.58,  8.51, 15.91],
       [21.42, 14.49, 27.09]]))
```

Summary

- Hypothesis testing allows to draw valid statistical conclusions about the value of population parameters or differences among them.
- Hypothesis is a claim and hypothesis testing is a process used to either reject or retain the claim.
- It forms the basis for many predictive analytics algorithms.

References

- James Evans, Business Analytics: Methods, Models and Decisions, Second Edition, Pearson Publication, 2017.
- Manaranjan Pradhan and U Dinesh Kumar, Machine Learning using Python, Wiley Publication, 2019.
- U Dinesh Kumar, Business Analytics- The Science of Data-Driven Decision Making, Wiley Publication, 2017.
- Python pandas: https://pandas.pydata.org/
- Matplotlib: Visualization with Python https://matplotlib.org/