Data Science with R

Lesson 6— Statistics for Data Science – II

Learning Objectives

- Discuss Hypothesis Test
- Explain Parametric test and its types
- Explain Non-Parametric test and its types
- Perform Hypothesis Tests on Population Means
- Perform Hypothesis Tests on Population Variance
- Perform Hypothesis Tests on Population Proportions

Statistics for Data Science – II

Topic 1—Hypothesis Test

What Is Hypothesis Test?

A hypothesis test is a formal procedure in statistics used to test whether a hypothesis can be accepted or not.

It is used to infer the results of a hypothesis performed on sample data to a large population.

The testing methodology depends on the data used and the reason for the analysis.

Types of Hypothesis Test

Simple Hypothesis Test

Complex Hypothesis Test

Null Hypothesis Test

Alternative Hypothesis Test

Statistical Hypothesis Test

Parametric Test

Non-Parametric Test

You have already learned about simple, complex, null, alternative, and statistical hypotheses in the previous lesson. This lesson will focus on discussing parametric and non-parametric tests.

Statistics for Data Science – II

Topic 2—Parametric Test

What Is a Parametric Test?

A parametric statistical test is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which one's data is drawn.

In these tests, inferences are based on the assumptions made about the nature of the population distribution. The tests are used for normal data.

Types of Parametric Tests

Z-Test and T-Test

Two population means or proportions are compared and tested.

Analysis of Variance (ANOVA) Test

Equality of several population means is tested.

There are many tests that are parametric. We will limit our attention to the tests mentioned above.

Types of Parametric Tests

Z-TEST

Z-Test

T-Test

ANOVA

Z-Test is performed in cases where the test statistic is t, σ is known, the population is normal, and the sample size is at least 30.

The formula to calculate z (standard statistic) is:

$$z = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Where,
n: Sample number
$\bar{X}$: Sample mean from a sample X_1, X_2, ..., X_n
μ: Population mean
σ: Standard Deviation

Types of Parametric Tests

EXAMPLE IN R

Z-Test

T-Test

ANOVA

Problem statement

Calculation on R

Solution

The test scores of an entrance exam fit a normal distribution with the mean test score of 72 and a standard deviation of 15.2. Compute the percentage of students scoring 84 or more.

Types of Parametric Tests

EXAMPLE IN R

Z-Test

T-Test

ANOVA

Problem statement

Calculation on R

Solution

Let’s use the pnorm (probability normal distribution) function to find the required percentage of students and the upper tail of the normal distribution (since the given score criteria is 84 or more).

```
pnorm(84, mean = 72, sd = 15.2, lower.tail = FALSE)

[1] 0.21492
```

lower.tail = TRUE is used to find the probability of values no larger than z, whereas lower.tail = FALSE is used to find the probability of values z or larger.

Types of Parametric Tests

EXAMPLE IN R

Z-Test

T-Test

ANOVA

Problem statement

Calculation on R

Solution

The required percentage is 21.5%.

Types of Parametric Tests

Z-Test

T-Test

ANOVA

T-Test is performed in cases where the test statistic is t, σ is unknown, sample standard deviation is known, and the population is normal.

The formula to calculate t is:

$$t = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

Where,
n: Sample number
$\bar{X}$: Sample mean from a sample X_1, X_2, …, X_n
μ: Population mean
σ: Standard Deviation

Types of Parametric Tests

EXAMPLE IN R

Z-Test

T-Test

ANOVA

Problem statement

Calculation on R

Solution

Find out the 2.5th and 97.5th percentiles of the Student's t-distribution, assuming 5 degrees of freedom.

Types of Parametric Tests

EXAMPLE IN R

Z-Test

T-Test

ANOVA

Problem statement

Calculation on R

Solution

Let's use the quantile function (applied to compute percentiles) "qt" against the decimal values 0.025 and 0.975.

```
qt(c(.025, .975), df = 5)  # 5 degrees of freedom
```

```
[1] -2.5706 2.5706
```

Degree of freedom refers to the number of values in the final calculation of a test statistic that varies freely. It is calculated using the formula df = N-1 (where N is the number of values in a dataset).

Types of Parametric Tests

EXAMPLE IN R

Z-Test

T-Test

ANOVA

Problem statement

Calculation on R

Solution

The required 2.5th and 97.5th percentiles are -2.5706 and 2.5706, respectively.

Types of Parametric Tests

Z-Test

T-Test

ANOVA

The ANOVA test is used for hypothesis tests that compare the averages of two or more groups.

For example, consider the following statements:

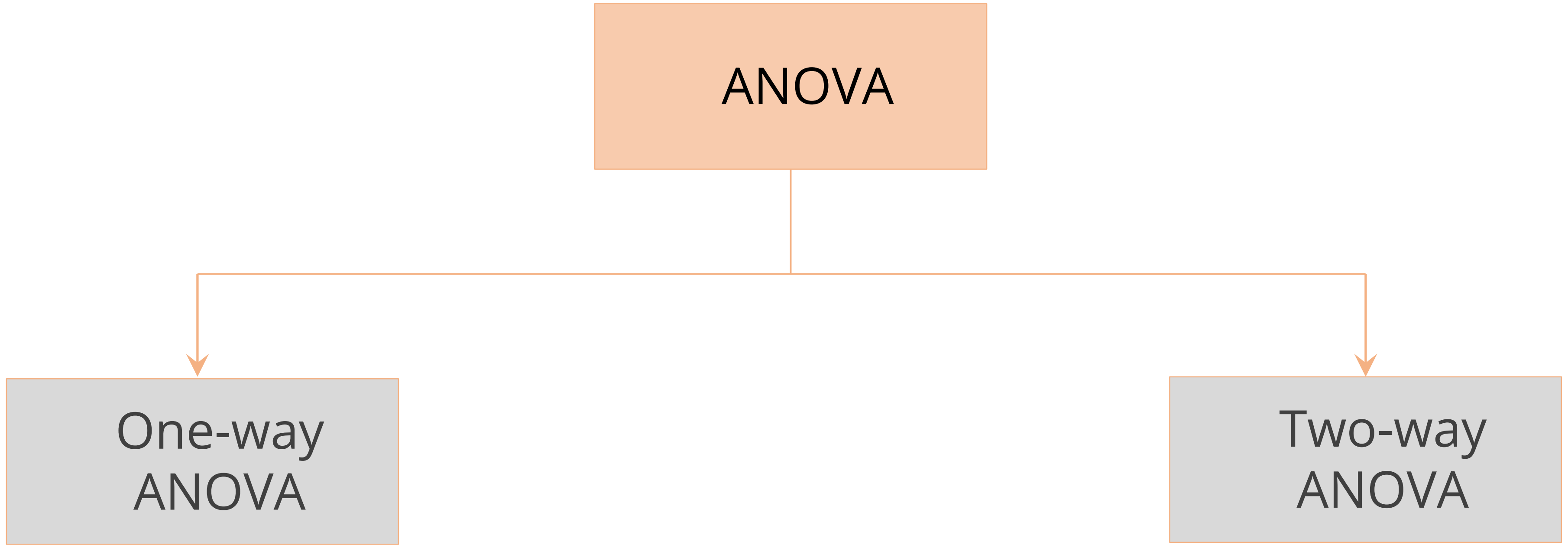
- An environmentalist wants to know if the average amount of pollution varies in several bodies of water.

- A sociologist wants to find out if a person’s income varies according to his/her upbringing.

Types of Parametric Tests

TYPES

- Z-Test
- T-Test
- ANOVA

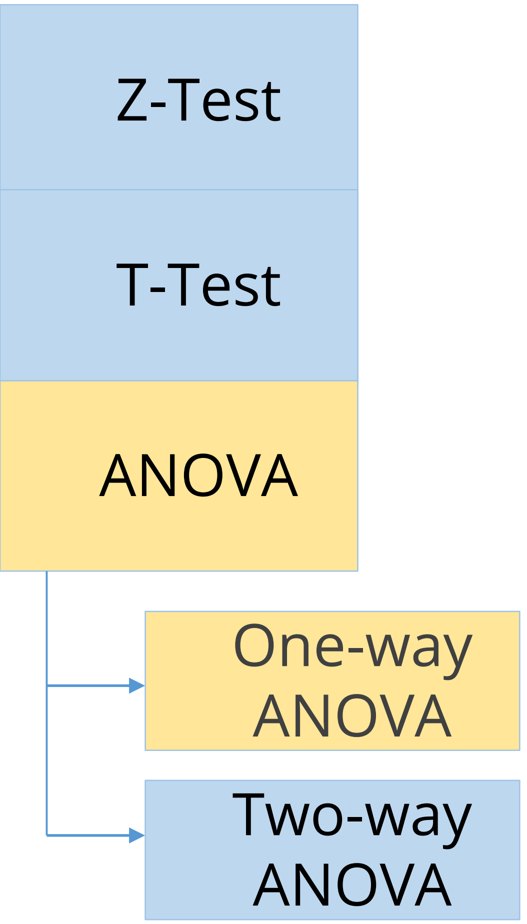


ANOVA

One-way ANOVA

Two-way ANOVA

Types of Parametric Tests



Z-Test

T-Test

ANOVA

One-way ANOVA

Two-way ANOVA

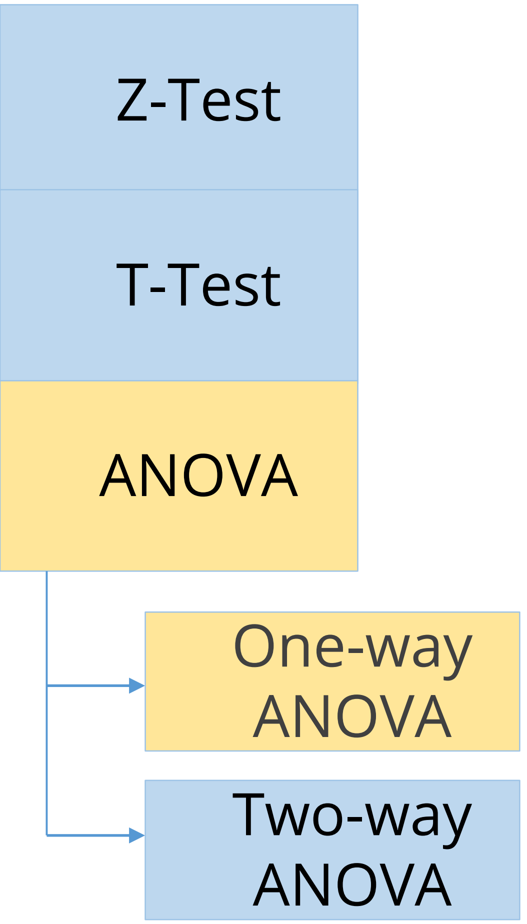
One-way Anova:

- Uses variances to determine if a statistically significant difference exists among several group means or not
- Tests H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ (where, μ = group mean and k = number of groups)

For one-way ANOVA, the ratio of the between-group variability to the within-group variability follows an **F-distribution** when the null hypothesis is true.

Types of Parametric Tests

ASSUMPTIONS

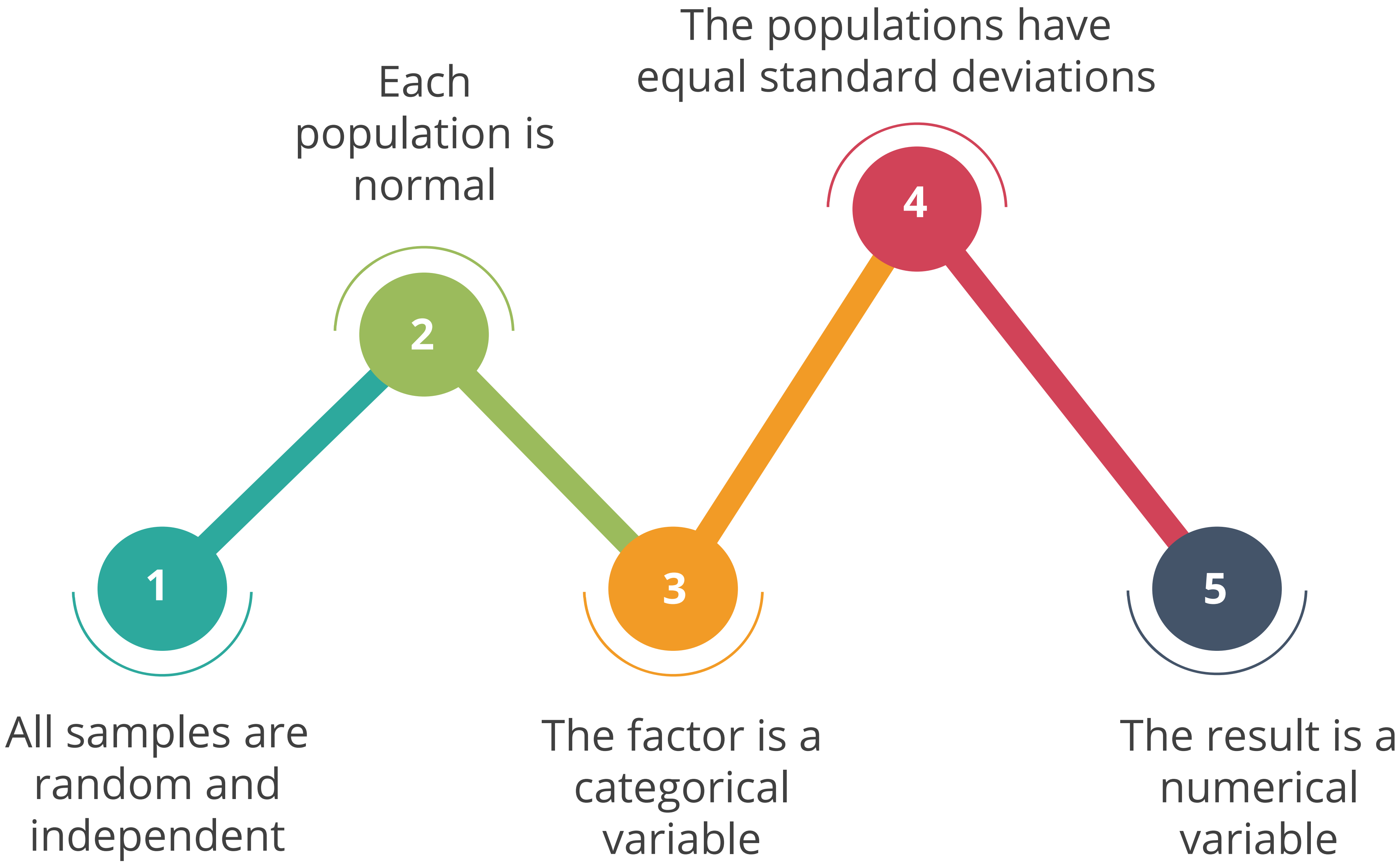


Z-Test

T-Test

ANOVA

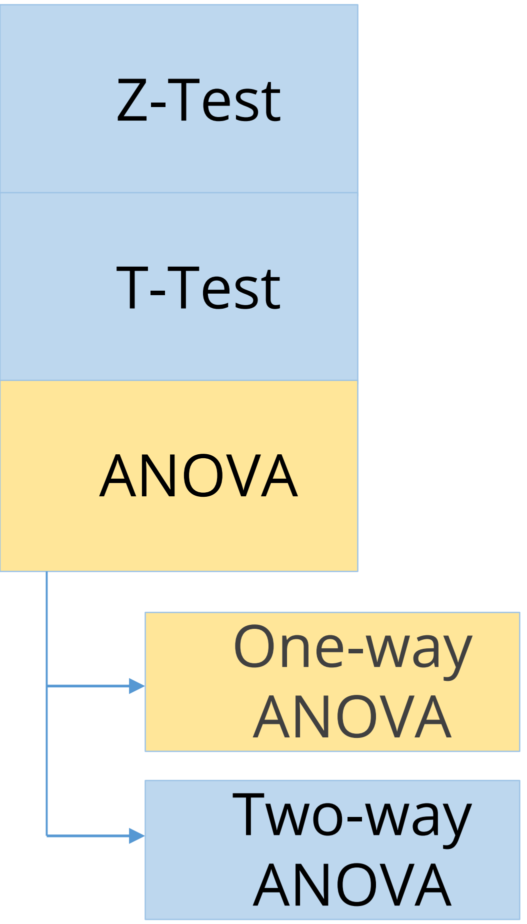
- One-way ANOVA
- Two-way ANOVA



1. All samples are random and independent
2. Each population is normal
3. The factor is a categorical variable
4. The populations have equal standard deviations
5. The result is a numerical variable

Types of Parametric Tests

EXAMPLE 1



Z-Test

T-Test

ANOVA

One-way ANOVA

Two-way ANOVA

Problem statement

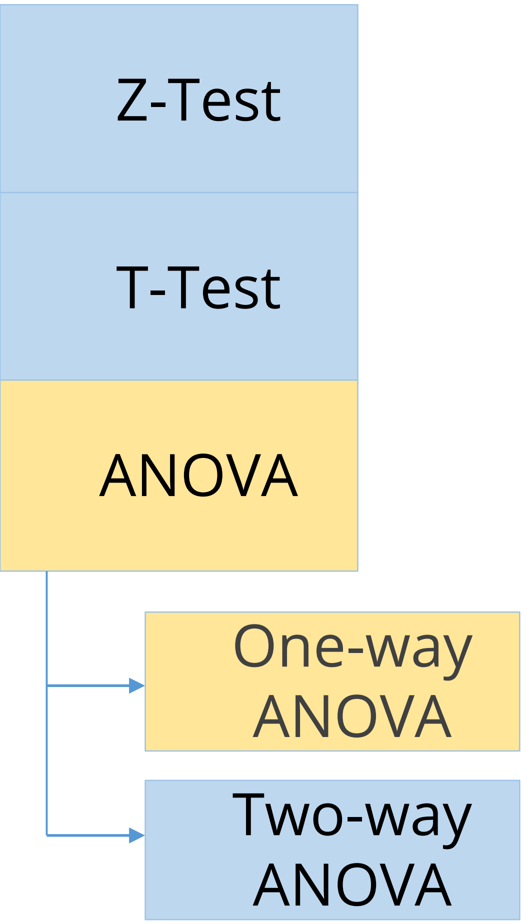
Calculation on R

Solution

Find out if there is a difference in the mean grades among the sororities, assuming μ_1, μ_2, μ_3, and μ_4 are the population means of the sororities.

Types of Parametric Tests

EXAMPLE 1



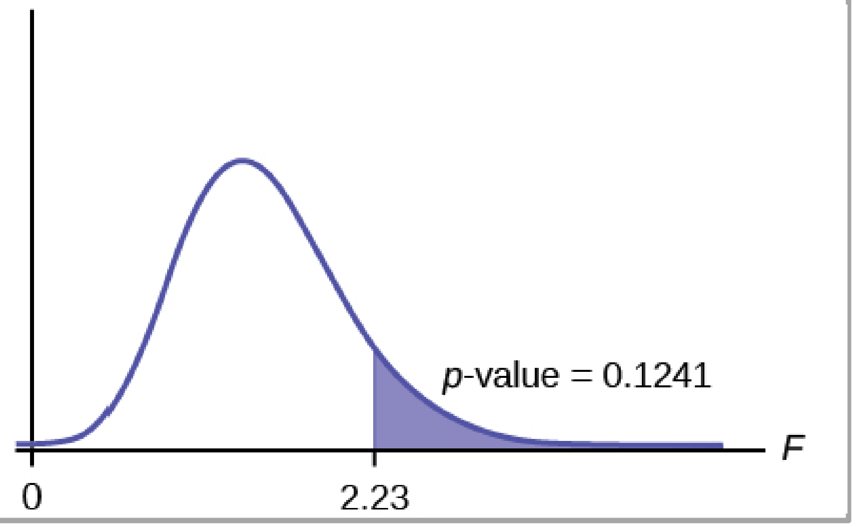
Problem statement

Calculation on R

Solution

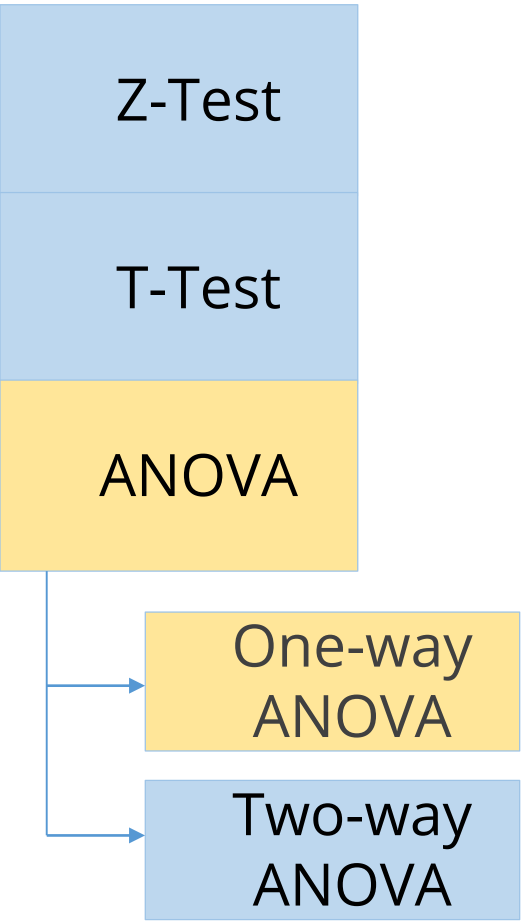
Test:

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_1: Not all of the means μ_1, μ_2, μ_3, and μ_4 are equal
- **Distribution for the test**: F3,16
  - df(num)= k – 1 = 4 – 1 = 3
  - df(denom) = n – k = 20 – 4 = 16
- **Calculate the test statistic**: F = 2.23
- **Define probability statement**: p-value = P(F > 2.23) = 0.1241
- **Compare α and the p-value**: α = 0.01
  - p-value = 0.1241
  - α < p-value
- **Decide**: Since α < p-value, you cannot reject H_0.



Types of Parametric Tests

EXAMPLE 1



Problem statement

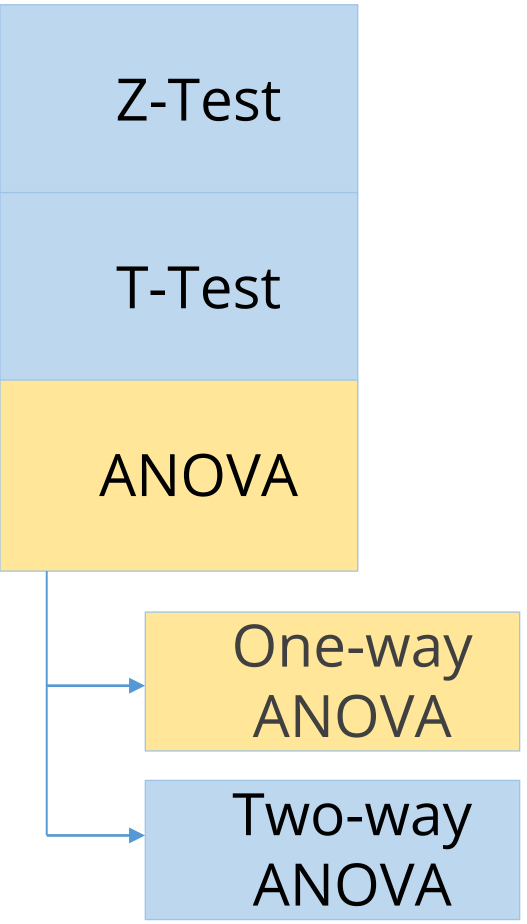
Calculation on R

Solution

Without sufficient evidence, you cannot conclude that there is a difference among the mean grades for the sororities.

Types of Parametric Tests

EXAMPLE 2



Z-Test

T-Test

ANOVA

One-way ANOVA

Two-way ANOVA

Problem statement

Calculation on R

Solution

A fast food chain wants to test and market three of its new menu items. To analyze if they are equally popular, consider:

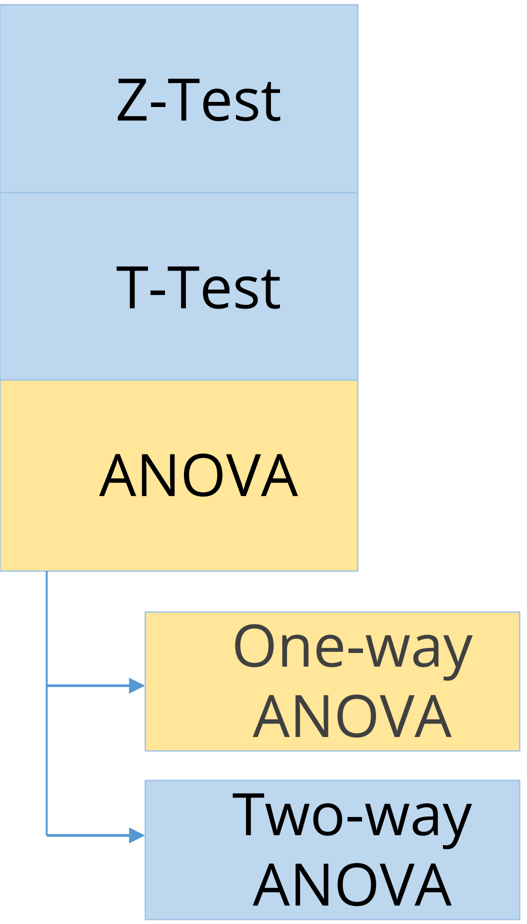
- 18 random restaurants for the study
- 6 of the restaurants to test market the first menu item, another 6 for the second one, and the remaining 6 for the last one

The table below shows the sales figures of the menu items in the 18 restaurants. At .05 level of significance, test whether the mean sales volumes for these menu items are equal.

Item 1	Item 2	Item 3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39

Types of Parametric Tests

EXAMPLE 2



Z-Test

T-Test

ANOVA

One-way ANOVA

Two-way ANOVA

Problem statement

Calculation on R

Solution

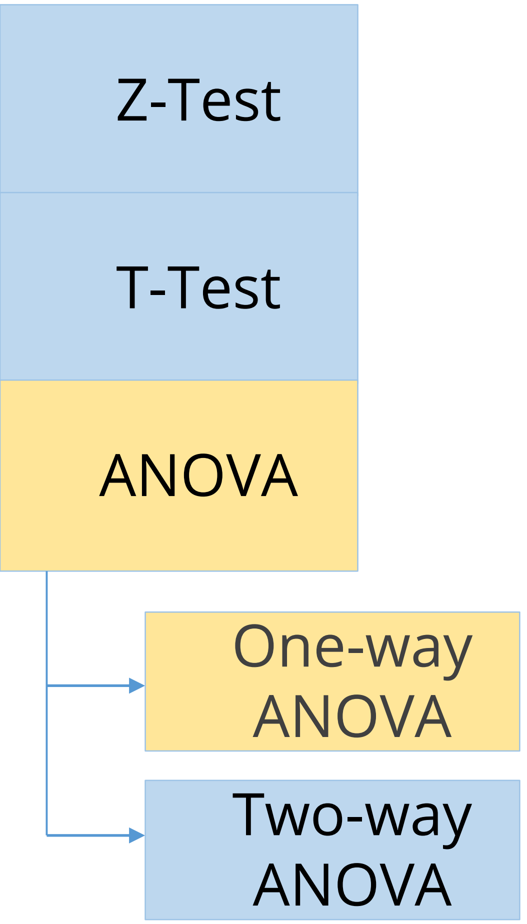
1. Copy and paste the sales figures in a table file "fastfood-1.txt" using a text editor.

2. Load the file into a data frame df1 using the read.table function.

```
  df1 = read.table("fastfood-1.txt", header = TRUE); df1
   Item1 Item2 Item3
  1  22  52  16
  2  42  33  24
  3  44   8  19
  4  52  47  18
  5  45  43  34
  6  37  32  39
```

Types of Parametric Tests

EXAMPLE 2



Problem statement

Calculation on R

Solution

3. Concatenate the data rows of df1 into a single vector r.

```
r = c(t(as.matrix(df1))) # response data
r
[1] 22 52 16 42 33 …
```

4. Assign new variables for the treatment levels and number of observations.

```
f = c("Item1", "Item2", "Item3")  # treatment levels
k = 3          # number of treatment levels
 n = 6          # observations per treatment
```

Types of Parametric Tests

EXAMPLE 2

- Z-Test
- T-Test
- ANOVA
  - One-way ANOVA
  - Two-way ANOVA

Problem statement

Calculation on R

Solution

5. Create a vector of treatment factors, corresponding to each element of R in step 3, using the gl function.

```
tm = gl(k, 1, n*k, factor(f)) # matching treatments
Tm
[1] Item1 Item2 Item3 Item1 Item2 ...
tm = gl(k, 1, n*k, factor(f)) # matching treatments
tm
 [1] Item1 Item2 Item3 Item1 Item2 ...
```

Apply the function aov to a formula that describes the response r by the treatment factor tm.

```
av = aov(r ~ tm)
```

Print out the ANOVA table with the summary function.

```
summary(av)

            Df Sum Sq Mean Sq F value Pr(>F)
tm        2  745   373  2.54  0.11
Residuals  15  2200   147
```

Types of Parametric Tests

EXAMPLE 2

Z-Test

T-Test

ANOVA

One-way ANOVA

Two-way ANOVA

Problem statement

Calculation on R

Solution

p-value of 0.11 > .05 significance level. Do not reject $H_{0.}$ This means that the mean sales volumes of the new menu items are all equal.

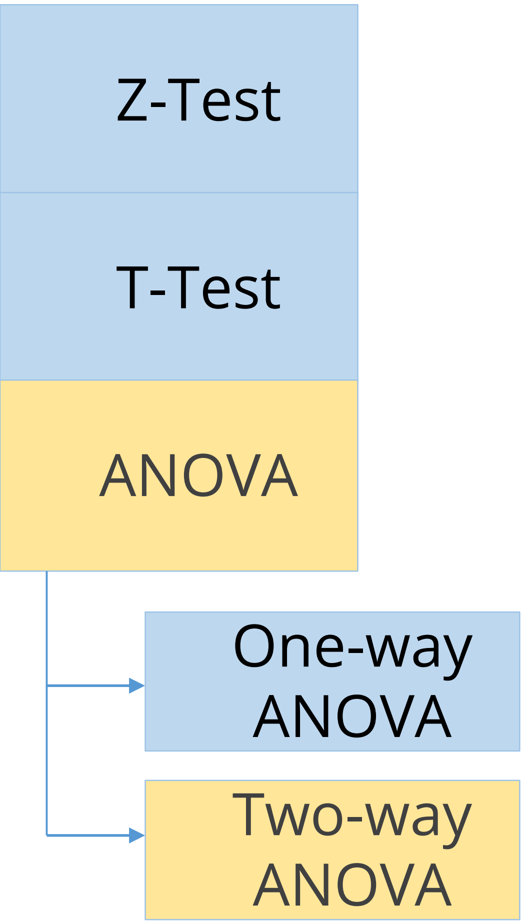
F- Distribution

F distribution or the Fisher–Snedecor distribution is a continuous probability distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA).

F-Ratio refers to the value derived from two estimates of the variance, as described below:

- **Variance between samples (SSbetween)**: It is an estimate of σ^2: variance of the sample means * n, when the sample sizes are the same. When sizes are different, the variance is weighted to account for different sample sizes.
- **Variance within samples (SSwithin)**: It is an estimate of σ^2: average of sample variances. When sizes are different, the variance within samples is weighted.

Types of Parametric Tests



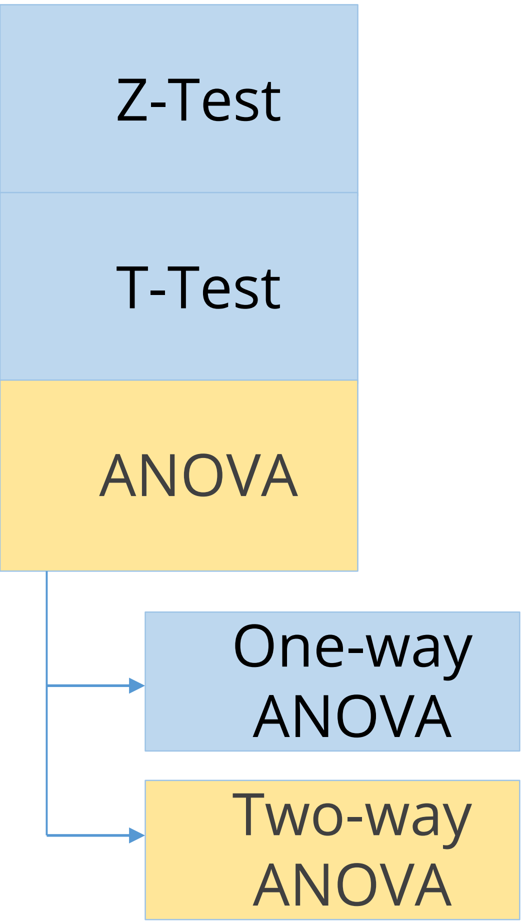
Two-way ANOVA refers to a hypothesis test where the classification of data is based on two independent variables

For example:

A company bases its sales classification by identifying the sales by a salesman and sales by region.

Types of Parametric Tests

ASSUMPTIONS



Z-Test

T-Test

ANOVA

One-way ANOVA

Two-way ANOVA

1 Normal distribution of the population sample

2 Measurement of dependent variable at continuous level

3 Categorical independent groups that have the same size

4 Independence of observations

5 Homogeneity of the variance of the population

https://keydifferences.com/difference-between-one-way-and-two-way-anova.html

Statistics for Data Science – II

Topic 3—Non-Parametric Test

What Is a Non-Parametric Test?

A non-parametric test (sometimes called a distribution free test) does not assume anything about the underlying distribution. It is used when the data is not distributed normally.

It refers to a null category, since virtually all statistical tests assume one thing or another about the properties of the source population(s).

http://www.statisticshowto.com/parametric-and-non-parametric-data/

Types of Non-Parametric Tests

- Kruskal Willis test (alternative to the One way ANOVA)
- Mann Whitney test (alternative to the two sample t test)
- **Chi-square test**

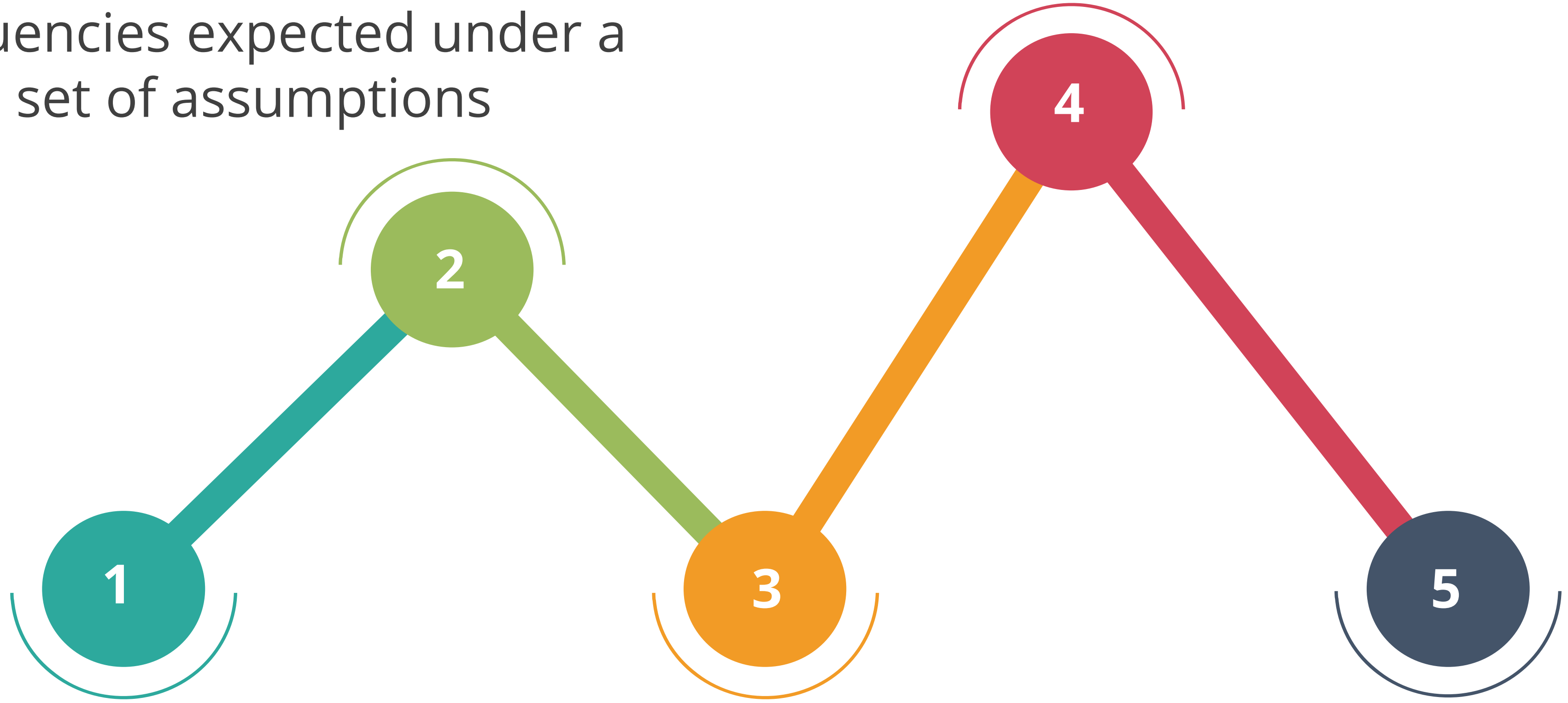
Chi-square test is the most commonly used non-parametric test. We will limit our scope to learning chi-square test in this course.

What Is Chi-square Test?

Chi-square test is a nonparametric test used to compare two or more variables for randomly selected data.

Chi-Square Test

FEATURES



1. Considers the square of a standard normal variate

2. Evaluates if frequencies observed in different categories vary significantly from the frequencies expected under a specified set of assumptions

3. Determines how well an assumed distribution fits the data

4. Uses contingency tables (in market researches, these tables are called cross-tabs)

5. Supports nominal-level measurements

Types of Chi-square Test

1. Chi-square test for goodness of fit
2. Chi-square test for independence of two variables

Types of Chi-square Test

Chi-square test for goodness of fit

Chi-square test for independence of two variables

It is used to observe the closeness of a sample that matches a population. The Chi-square test statistic (χ^2) is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

with **k-1** degrees of freedom.

Where O_i is the observed count, k is categories, and E_i is the expected counts

Goodness of fit of a statistical model refers to the understanding of how well sample data fits a set of observations.

https://www.chegg.com/homework-help/definitions/chi-square-test-14

Types of Chi-square Test

USE CASES

Chi-square test for goodness of fit

Chi-square test for independence of two variables

Goodness of fit test is used to identify the relation between two attributes, as in the cases below:

- Credit worthiness of borrowers based on their age groups and personal loans
- Relation between the performance of salesmen and training received
- Return on a single stock and on stocks of a sector like pharmaceutical or banking
- Category of viewers and impact of a TV campaign

Types of Chi-square Test

Chi-square test for goodness of fit

Chi-square test for independence of two variables

It is used to check whether the variables are independent of each other or not. The Chi-square test statistic (χ^2) is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

With **(r-1) (c-1)** degrees of freedom

Where O_i is the observed count, r is number of rows, c is the number of columns, and E_i is the expected counts

Two random variables are called independent if the probability distribution of one variable is not affected by the other.

https://www.chegg.com/homework-help/definitions/chi-square-test-14

Types of Chi-square Test

USE CASES

Chi-square test for goodness of fit

Chi-square test for independence of two variables

Test of independence is suitable for the following situations:

- There is one categorical variable.
- There are two categorical variables, and you will need to determine the relation between them.
- There are cross-tabulations, and relation between two categorical variables needs to be found.
- There are non-quantifiable variables (For example, answers to questions like, do employees in different age groups choose different types of health plans?)

Types of Chi-square Test

EXAMPLE

Chi-square test for goodness of fit

Chi-square test for independence of two variables

Problem statement

Calculation on R

Solution

The manager of a restaurant wants to find the relation between customer satisfaction and the salaries of the people waiting tables.

- She takes a random sample of 100 customers asking if the service was excellent, good, or poor.
- She then categorizes the salaries of the people waiting as low, medium, and high.

Her findings are shown in the table below:

	Salary			
Service	Low	Medium	High	Total
Excellent	9	10	7	26
Good	11	9	31	51
Poor	12	8	3	23
Total	32	27	41	100

Types of Chi-square Test

EXAMPLE

Chi-square test for goodness of fit

Chi-square test for independence of two variables

Problem statement

Calculation on R

Solution

Assume the level of significance is 0.05. Here, H_0 and H_1 denote the independence and dependence of the service quality on the salaries of people waiting tables.

Test: DF = (3-1) (3-1) = 4

- Under H_0, expected frequencies are:
  - E_{11} = (26X32)/100 = 8.32, E_{12} = 7.02, E_{13} = 10.66
  - E_{21} = 16.32, E_{22} = 13.77, E_{23} = 20.91
  - E_{31} = 7.36, E32 = 6.21, E_{33} = 9.41

Therefore, $\aleph^2$(calculated) = (9-8.32)2/8.32+(10-7.02)2/7.02+(7-10.66)2/10.66 +(11-16.32)2/16.32+(9-13.77)2/13.77+(31-20.91)2/20.91+(12-7.36)2/7.36+(8-6.21)2/6.21+(3-9.43)2/9.43 = 18.658

- $\aleph^2$ 0.05,4 = 9.48773

$\aleph^2$ (Calculated) > $\aleph^2$(Tabulated)

- Reject H_0, accept H_1.

Types of Chi-square Test

EXAMPLE

Chi-square test for goodness of fit

Chi-square test for independence of two variables

Problem statement

Calculation on R

Solution

Service quality is dependent on the salaries of the people waiting.

Types of Chi-square Test

EXAMPLE IN R

Chi-square test for goodness of fit

Chi-square test for independence of two variables

Problem statement

Calculation on R

Solution

To perform this test in R, let's consider a table that is a result of a survey conducted among students about their smoking habits.

This tables has:

"Smoke" variables, which record the smoking habits of students (Allowed values: "Heavy," "Regul," "Occas," and "Never")
"Exer" variables, which record the exercise levels of smoking (Allowed values: "Freq," "Some, " and "None")

Assuming .05 as the significance level, test the hypothesis whether the smoking habits of students are independent of their exercise levels or not.

Types of Chi-square Test

EXAMPLE IN R

Chi-square test for goodness of fit

Chi-square test for independence of two variables

Problem statement

Calculation on R

Solution

Let’s build the contingency table in R:

```
library(MASS)     # load the MASS package
head(survey)
tbl = table(survey$Smoke, survey$Exer)
tbl
 Freq   None Some
 Heavy  7   1   3
 Never  87  18  84
 Occas  12  3   4
 Regul  9   1   7
```

Let’s use the chisq.test function for the contingency table and find the value of p (calculated probability).
chisq.test(tbl)

Output: data: table(survey$Smoke, survey$Exer)
X-squared = 5.4885, df = 6, p-value = 0.4828

Types of Chi-square Test

EXAMPLE IN R

Chi-square test for goodness of fit

Chi-square test for independence of two variables

Problem statement

Calculation on R

Solution

As p > significance level, H_0 is not rejected. This means that the smoking habits of students are independent of their exercise levels.

Hypothesis Test around Mean, Variance, and Proportion

Both parametric and non-parametric hypothesis tests are used to check whether the **mean, variance, and proportion** of the population have pre-determined values or if the values need to be defined.

Let’s discuss them in detail.

Statistics for Data Science – II

Topic 4—Hypothesis Tests about Population Means

Hypothesis Tests about Population Means

Hypothesis tests about population means involve testing the hypothesis that compares the population mean of interest with a specified value.

Hypothesis Tests about Population Means

ASSUMPTION

X_1, X_2,……., X_n is a sample of size n from a normal population with mean μ and variance $σ^2$. The mean X is distributed normally with the mean μ and variance **$σ^2/n$ ($X \sim N(μ, σ^2/n)$)**.

If n is large, X will be calculated similarly, even if the sample is from a non-normal population.

Therefore, for large samples, the standard normal variable corresponding to X bar is Z (as calculated in the Z-test).

Hypothesis Tests about Population Means

WHEN POPULATION VARIANCE IS KNOWN

Consider a random large sample of size n, with a sample mean $\bar{X}$

Test the hypothesis that the sample mean X has been drawn from a population with the mean μ and a specified value μ_0, that is:

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$
- $H_1 : \mu > \mu_0$
- $H_1 : \mu < \mu_0$

Under null hypothesis, $Z = (\bar{X} - \mu_0)/S.E.(X)$ follows Standard Normal Distribution approximately.

When population variance is unknown, Z test is used.

Hypothesis Tests about Population Means

WHEN POPULATION VARIANCE IS UNKNOWN

Consider the following hypothesis formation:

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$

If μ_0 falls in the confidence interval, the test result is “failing to reject the null hypothesis”; if not, the result is “reject the null hypothesis.”

When population variance is unknown, T test is used.

Statistics for Data Science – II

Topic 5—Hypothesis Tests about Population Variance

Hypothesis Tests about Population Variance

Hypothesis test about population variance involves finding the squared deviation of a random variable from its mean. It measures how far a set of (random) numbers are spread out from their average value.

Hypothesis Tests about Population Variance

FORMULA

Consider the case where data consists of a simple random sample drawn from a normally distributed population. The test statistic for testing hypotheses about a single population variance is calculated as:

$$\chi^2 = (n-1)\, s^2 / \sigma^2$$

Chi-square test is used in hypothesis tests of population variance.

Statistics for Data Science – II

Topic 6—Hypothesis Tests about Population Proportions

Hypothesis Tests about Population Proportions

Hypothesis Tests about population proportions are defined as the ratio of the values in a subset S to the values in a set R.

Hypothesis Tests about Population Proportions

FORMULA

Consider a random sample of the size n and the proportion of members with a certain attribute p.

You need to test the hypothesis that the proportion P in the population has a specified value P_0, that is:

- $H_0 : P = P_0$
- $H_1 : P \neq P_0$
- $H_1 : P > P_0$
- $H_1 : P < P_0$

For a large sample, $Z = (p - P_0)/S.E.(p) \sim N(0,1)$ (under H_0)

Where,
p = X/n = Number of successes in sample/Sample size
P_0 = Hypothesized proportion of successes in the population

Key Takeaways

- ✓ Hypothesis test is a formal procedure in statistics used to test whether a hypothesis can be accepted or not.
- ✓ The Z-test is performed in cases where the test statistic is t and σ is known.
- ✓ The T-test is performed in cases where the test statistic is t and σ is unknown.
- ✓ The degree of freedom is the number of independent variates that make up the statistic.
- ✓ The Chi-Square Test considers the square of a standard normal variate.
- ✓ The ANOVA test is used for such hypothesis tests that compare the averages of two or more groups.
- ✓ Both parametric and non-parametric tests of the population have a pre-determined value, or the values need to be defined.