Statistical Inference

Parametric Tests - II

Contents

1. Independent samples t-test
2. Paired sample t-test
3. t test for correlation

Independent samples t-test

- The independent-samples t-test compares the means of two independent groups on the same continuous variable.

- The following hypotheses are tested in an independent samples t test
  - H_0: Two population means are equal
  - H_1: Two population means are not equal

Data Snapshot

INDEPENDENT SAMPLES t
TEST

Variables

time_g1	time_g2
85	83
95	85
105	96
85	94
90	102

Observations

Columns	Description	Type	Measurement	Possible values
time_g1	Time to complete MIS report by group1	Numeric	Hours	Positive Values
time_g2	Time to complete MIS report by group2	Numeric	Hours	Positive Values

Case Study - 1

Background

The company is assessing the difference in time to complete an MIS report between two groups of employees :
Group I: Experience(0-1 years)
Group II: Experience(1-2 years)

Objective

To test whether the average time taken to complete the MIS by both the groups is same.

Sample Size

Sample size: 14
Variables: time_g1, time_g2

Assumptions for independent samples t-test

- The assumptions for the independent samples t-test are listed below :

    - The samples drawn are random samples.
        (Employees are selected at random from the company)

    - The populations from which samples are drawn have equal & unknown variances.
        (F-test is used to validate this assumption which will be covered in next presentation)

    - The populations follow normal distribution.
        (Time taken to complete MIS report should be normally distributed for both groups.

    - A normality assumption can be validated using a method explained earlier.

Independent sample t-test

Testing whether the means of the two groups are equal.

Null Hypothesis (H_0): $\mu 1 = \mu 2$

Alternate Hypothesis (H_1): $\mu 1 \neq \mu 2$

$\mu 1 =$ average time taken by group1 to complete MIS
$\mu 2 =$ average time taken by group2 to complete MIS .

Objective	To test the average time taken to complete the MIS by both the groups is same.
Test Statistic	$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$ $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Group I	Group II
Sample Size	n1=12	n2=14
Mean		
Variance	$S_1^2=41.9015$	$S_2^2=27.1483$
Pooled Variance	$Sp^2=33.9102$	
Difference		
t	$t = \dfrac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_P\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ =0.22345	

Independent samples t-test in R

```r
# Import data

data<-read.csv("INDEPENDENT SAMPLES t TEST.csv", header=TRUE)


# t-test for independent samples

t.test(data$time_g1,data$time_g2, alternative="two.sided",
       var.equal=TRUE)
```

- ❑ *data$time_g1 and data$time_g2 are the variables to be compared.*
- ❑ *alternative="two.sided" since H1 is* μ1 ≠ μ2
- ❑ *var.equal=TRUE assumes for equality of variance of two groups.*
  (there is a test which validates this assumption which is explained in
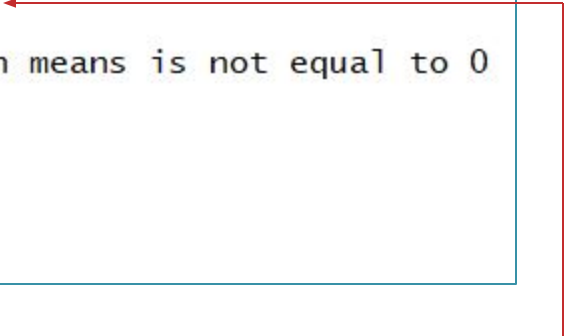 *subsequent slides.)*

* Before performing the t test, a normality test is done to ensure time variable is normally distributed in both the groups.

Independent samples t-test in R

`# Output:`

```
        Two Sample t-test

data:   data$time_g1 and data$time_g2
t = 0.22346, df = 24, p-value = 0.8251
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.216185  5.239994
sample estimates:
mean of x mean of y
  93.58333  93.07143
```

Interpretation :

- *Since p-value is >0.05, do not reject H0. There is no significant difference in average time taken to complete the MIS between both groups of employees.*
- *95% CI contains value 0 which is value under H0.($\mu1 = \mu2 \rightarrow \mu1\text{-}\mu2=0$). Hence, do not reject H0.*

Independent samples t-test when variances are not equal

- Welch's t test is used to test the equality of two means if the variances of two groups can not be assumed to be equal.

- Welch's t-test defines the statistic t by the following formula:

$$t \;=\; \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{N_1} + \dfrac{s_2^2}{N_2}}}$$

- The denominator is not based on a pooled variance estimate.

- If two variances are not equal, the t test syntax in R below is used:

```
data<-read.csv("INDEPENDENT SAMPLES t TEST.csv", header=TRUE)

t.test(data$time_g1,data$time_g2,alternative="two.sided",
var.equal=FALSE)
```

Paired samples t-test

- The paired sample t-test is used to determine whether the mean difference between two sets of observations is zero ,where each subject or entity is measured twice resulting in a pair of observations.

- Commonly used when observations are recorded 'before' and 'after' training and the objective is to test whether the training is effective.

Case Study - 2

Background

The company organized a training program to improve efficiency. The time taken to complete an MIS report before and after training is recorded for 15 employees.

Objective

To test whether the average time taken to complete the MIS before and after training is not different.
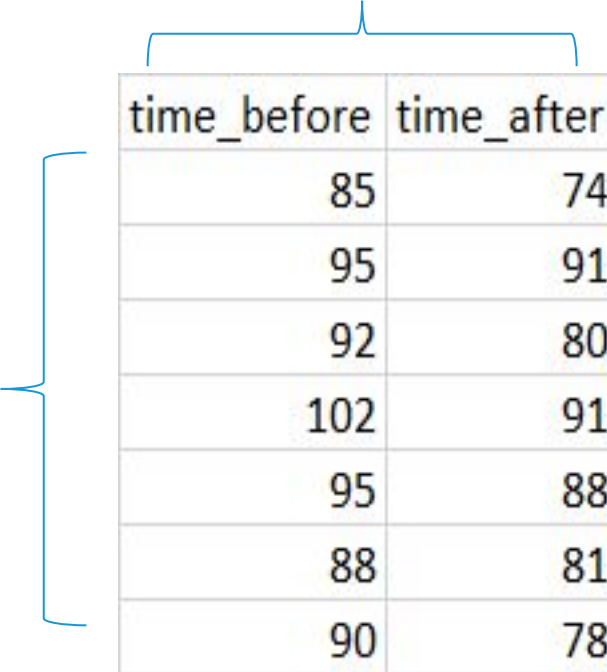
Sample Size

Sample size: 15
Variables: time_before, time_after

Data Snapshot

Variables

Observations

time_before	time_after
85	74
95	91
92	80
102	91
95	88
88	81
90	78

Columns	Description	Type	Measurement	Possible values
time_before	Time to complete MIS report before training	Numeric	Hours	Positive values
time_after	Time to complete MIS report after training	Numeric	Hours	Positive values

Assumptions for paired sample t-test

- The assumptions of the paired-sample t-test are listed below:

  - Random sampling from a defined population
    (employees are selected at random from the company)

  - Population of the testing variable is normally distributed
    (Difference time taken to complete MIS report should be normally distributed).

- A Normality test can be performed by any of the methods explained earlier.

- The validity of the test is not seriously affected by moderate deviations from the 'Normality' assumption.

Paired sample t-test

Testing whether means of two dependent groups are equal.

Objective	To test the average time taken to complete MIS before and after training is not different.

Null Hypothesis (H_0): There is no difference in average time before and after the training. i.e. D=0
Alternate Hypothesis (H_1):Average time is less after the training. (Training is effective.) D>0
D= µBefore – µAfter

Test Statistic	$t = \dfrac{\bar{d}}{s_d/\sqrt{n}}$ of	Where $\bar{d}$ is the sample mean of the difference before-after , s_d is the sample standard deviation of the difference, n is the sample size of difference. The quantity t follows a distribution called as 't distribution' with n-1 degrees of freedom.
Decision Criteria	Reject the null hypothesis if p-value < 0.05	

Computation

	Notation	Value
Sample Size	n	12
Mean difference (before-after)	$\bar{d}$	8.3333
Standard Deviation	s_d	3.9219
t	$t = \dfrac{\bar{d}}{s_d/\sqrt{n}}$	8.2295

Paired sample t-test in R

```
# Import data
```

```
data<-read.csv("PAIRED t TEST.csv",header=TRUE)
```

```
# t-test for paired samples
```

```
t.test(data$time_before,data$time_after,
       alternative="greater", paired=TRUE)
```

- ❑ *data$time_before and data$time_after are the variables under study.*
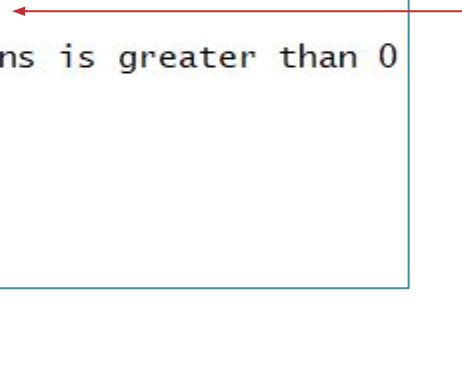- ❑ *alternative="greater" ,Since under H1, difference of before –after will be large.*

* Before performing a t test, a normality test is carried out to ensure difference variable is normally distributed.

Paired sample t-test in R

\# Output:

```
        Paired t-test

data:   data$time_before and data$time_after
t = 8.2295, df = 14, p-value = 4.919e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 6.549798       Inf
sample estimates:
mean of the differences
        8.333333
```

Interpretation :
- *Since p-value is <0.05, reject H0. Average time taken to complete the MIS report after the training is less. Hence, training is effective.*
- *95% C.I does not contain value of D=0 (under H0),reject H0.*

t-test for Correlation

- The Correlation coefficient summarizes the strength of a linear relationship between two variables.

- A t-test is used to check if there is significant correlation between two variables.

- Sample correlation coefficient (r) is calculated using bivariate data.

- Null hypothesis of this test is
  H0: there is no correlation between 2 variables under study ($\rho=0$)

Case Study - 3

Background

A company with 25 employees has calculated a job proficiency score & an aptitude test score for its employees

Objective

To test if there is significant correlation between the job proficiency and aptitude test scores.

Sample Size

Sample size: 25
Variables: Empcode, Aptitude, Job_prof

Data Snapshot

Correlation test

Variables

Observations

Empcode	aptitude	job_prof
E101	86	88
E102	62	80
E103	110	96
E104	101	76
E105	100	80
E106	78	73
E107	120	58
E108	105	116
E109	112	104

Columns	Description	Type	Measurement	Possible values
Empcode	Employee code	Numeric	-	
Aptitude	Score of aptitude test	Numeric	-	Positive values
Job_prof	Job proficiency score	Numeric	-	Positive values

Correlation t-test

Testing for correlation coefficient value.

Objective	To test whether there exists significant correlation between job proficiency and aptitude score.

Null Hypothesis (H$_0$): **There is no significant correlation between Job proficiency and Aptitude test (ρ=0).**
Alternate Hypothesis (H$_1$):**There is correlation between Job proficiency and Aptitude test (ρ≠0)**

Test Statistic	$$t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$$	where r is the sample correlation coefficient, the sample size. The quantity t follows a distribution called as 't distribution' with n-2 degrees of freedom.
Decision Criteria	Reject the null hypothesis **if p-value < 0.05**	

Computation

	Notation	Value
Sample Size	n	25
Sample correlation coefficient	r	0.514411
t	$t = \dfrac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$	2.8769

Correlation t-test in R

```
# Import data

data<-read.csv("Correlation test.csv", header=TRUE)


# t-test for correlation

cor.test(data$aptitude, data$job_prof, alternative="two.sided",
         method="pearson")
```
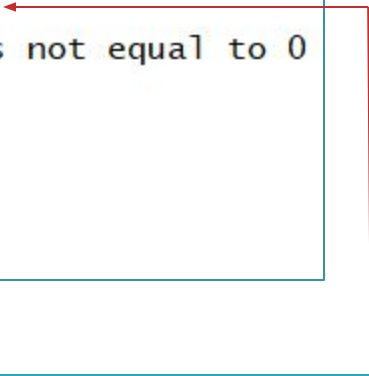
❑ *data$aptitude and data$job_prof are the variables under study.*
❑ *alternative="two.sided". Since under H1, ($\rho \neq 0$) .*

Correlation t-test in R

Output:

```
          Pearson's product-moment correlation

data:   data$aptitude and data$job_prof
t = 2.8769, df = 23, p-value = 0.008517
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1497097 0.7558981
sample estimates:
      cor
0.5144107
```

Interpretation :
- *Since p-value is <0.05, reject H0. There is correlation between aptitude test and job proficiency.*
- *95% C.I does not contain value $\rho=0$ (under H0), reject H0.*

Quick Recap

Independent sample t test	• It compares the means of two independent groups on the same continuous variable. • H0: $\mu_1 = \mu_2$
Paired sample t test	• Used to determine whether the mean difference between two sets of observations is zero ,where each subject or entity is measured twice resulting in pair of observations. • H0: $\mu_1-\mu_2=d=0$
t test for correlation	• Used to check if there is significant correlation between two variables. • H0: $\rho=0$