R-HypoTest

Hypothesis Testing Toolkit in R

Group 7

27 January 2024

Group Members

Fahmida Akter Keya

Kazi Anika Hayder

Shafrin Mardia

Md Omar Faruk

Montasir Affan

Objective

> A function that will do hypothesis testing and show the test statistic , the p value , confidence interval of the estimate that will assist the user to draw conclusions according to the input data set.

Hypothesis

- A concept or idea that you test through research and experiments.

Null Hypothesis (H_0)

- The null hypothesis is the statement or claim being made about population. (which we are trying to disprove)

Alternative Hypothesis (H_1)

- The hypothesis that we are trying to prove and which is accepted if we have sufficient evidence to reject the null hypothesis.

Hypothesis

"The average monthly salary of a garments employee is more than 7000 taka"

"The grades of the students of AST 230 are associated to their performance in R Fest 2024"

Are these assumptions correct?

Hypothesis Testing

- A form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability.

Test can be of two types:

Parametric

Parametric tests are those that make assumptions about the parameters of the population distribution from which the sample is drawn.

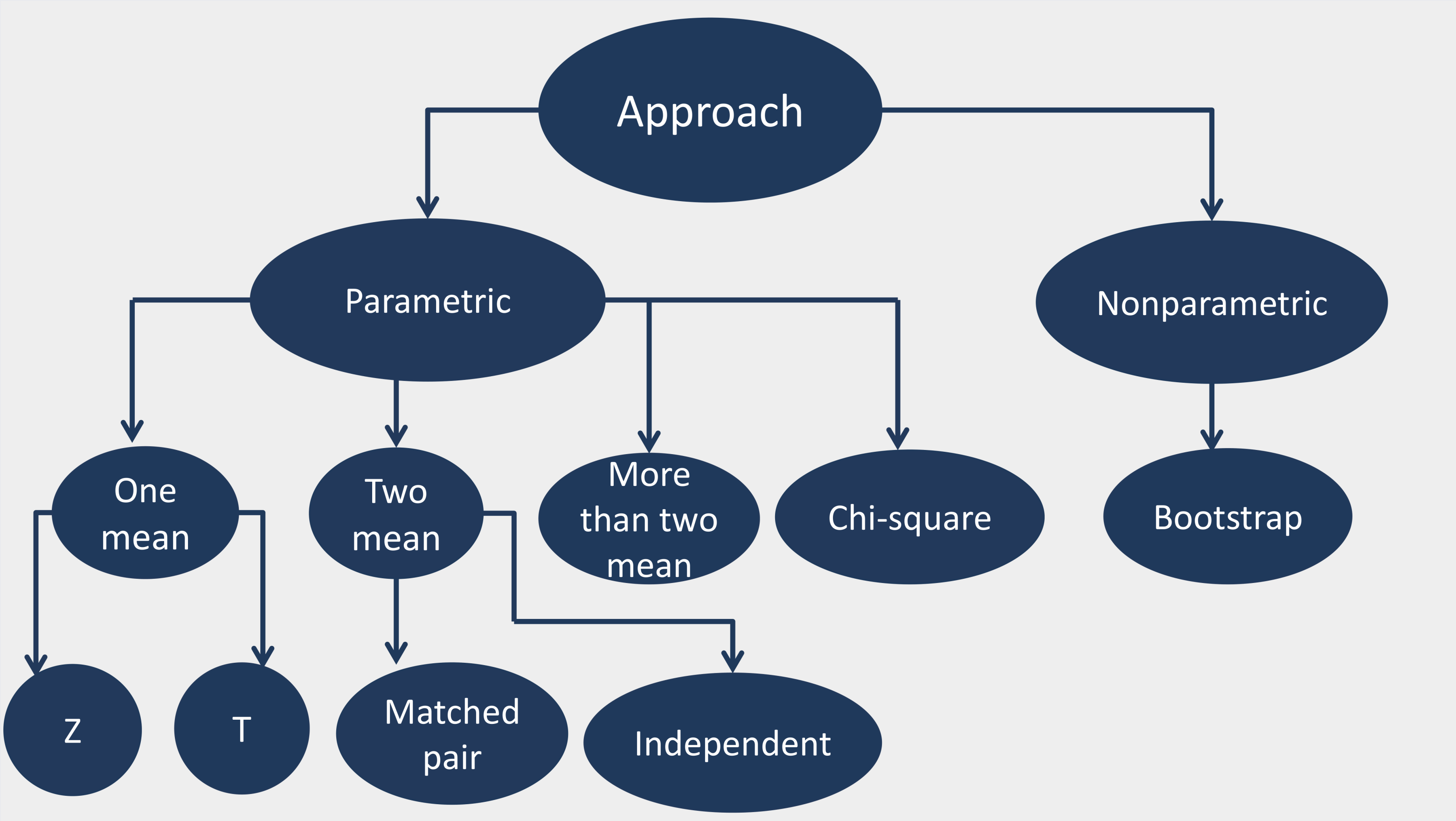
Example- one mean test , two mean test , analysis of variance test

Non-parametric

Non-parametric test does not assume anything about the underlying distribution.

Example - bootstrap test

Gist in a Flowchart



One-Sample Test of Means

- Used to determine whether the mean of a single sample is significantly different from a known or hypothesized population mean.

Hypothesis

- Null hypothesis: $H_0 : \mu = \mu_0$
- Alternative hypothesis: $H_1 : \mu \neq \mu_0 \quad or \quad \mu < \mu_0 \quad or \quad \mu > \mu_0$

Test Statistic

When standard deviation is known,

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad \sim \quad N(0,1)$$

When standard deviation is unknown,

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}} \quad \sim \quad t(df)$$

Two-Sample Test of Means

Independent

- Compares the means of two independent samples to assess whether they are significantly different from each other.

Hypothesis

- Null hypothesis: $H_0 : \mu_1 = \mu_2$
- Alternative hypothesis: $H_1 : \mu_1 \neq \mu_2 \quad or \quad \mu_1 < \mu_2 \quad or \quad \mu_1 > \mu_2$

Test Statistic

When standard deviation is known,

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

When standard deviation is unknown,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Two-Sample Test of Means

Matched Pair

- Assesses whether the mean of the differences between paired observations is significantly different from zero. Commonly used for pre- and post-treatment comparisons.

Hypothesis

- Null hypothesis: $H_0 : \mu_d = 0$
- Alternative hypothesis: $H_1 : \mu_d \neq 0 \quad or \quad \mu_d < 0 \quad or \quad \mu_d > 0$

Test Statistic

When standard deviation is known,

$$z = \frac{\bar{d}}{\frac{\sigma_d}{\sqrt{n}}}$$

When standard deviation is unknown,

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}}$$

More Than Two Means Test

- Used to compare means across more than two groups to determine if there are significant differences.

Hypothesis

- Null hypothesis: $H_0 : \mu_1 = \mu_2 = \quad \ldots \quad = \mu_n$
- Alternative hypothesis: $H_1 : \mu_i \neq \mu_j; for \quad at \quad least \quad one \quad pair$

Test Statistic

$$F_0 = \frac{MS_{treatment}}{MS_{error}} \quad \sim \quad F(d_1, d_2)$$

Assumptions

- The responses for each factor level have a normal population distribution.
- These distributions have the same variance.
- The data are independent.

Pearson's Chi-Square Test

- Examines the association between two categorical variables in a contingency table.

Hypothesis

- Null hypothesis:
  $H_0: \quad There \quad is \quad no \quad association \quad between \quad two \quad groups$
- Alternative hypothesis:
  $H_1: \quad There \quad is \quad association \quad between \quad two \quad groups$

Test Statistic

$$\chi^2 = \sum \frac{(Observed_i - Expected_i)^2}{Expected_i} \quad \sim \quad \chi^2(k)$$

Assumptions

- The data in the cells should be frequencies, or counts of cases.
- There are 2 variables, and both are measured as categories, usually at the nominal level. However, data may be ordinal, interval or ratio data that have been collapsed into ordinal categories.

Bootstrap Test

- Bootstrapping is any test or metric that uses random sampling with replacement, and falls under the broader class of resampling methods. Bootstrapping assigns measures of accuracy to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods.

Hypothesis

- Null hypothesis:

$$H_0: \quad There \quad is \quad no \quad significant \quad difference$$

- Alternative hypothesis:

$$H_1: \quad There \quad is \quad significant \quad difference$$

Test Statistic

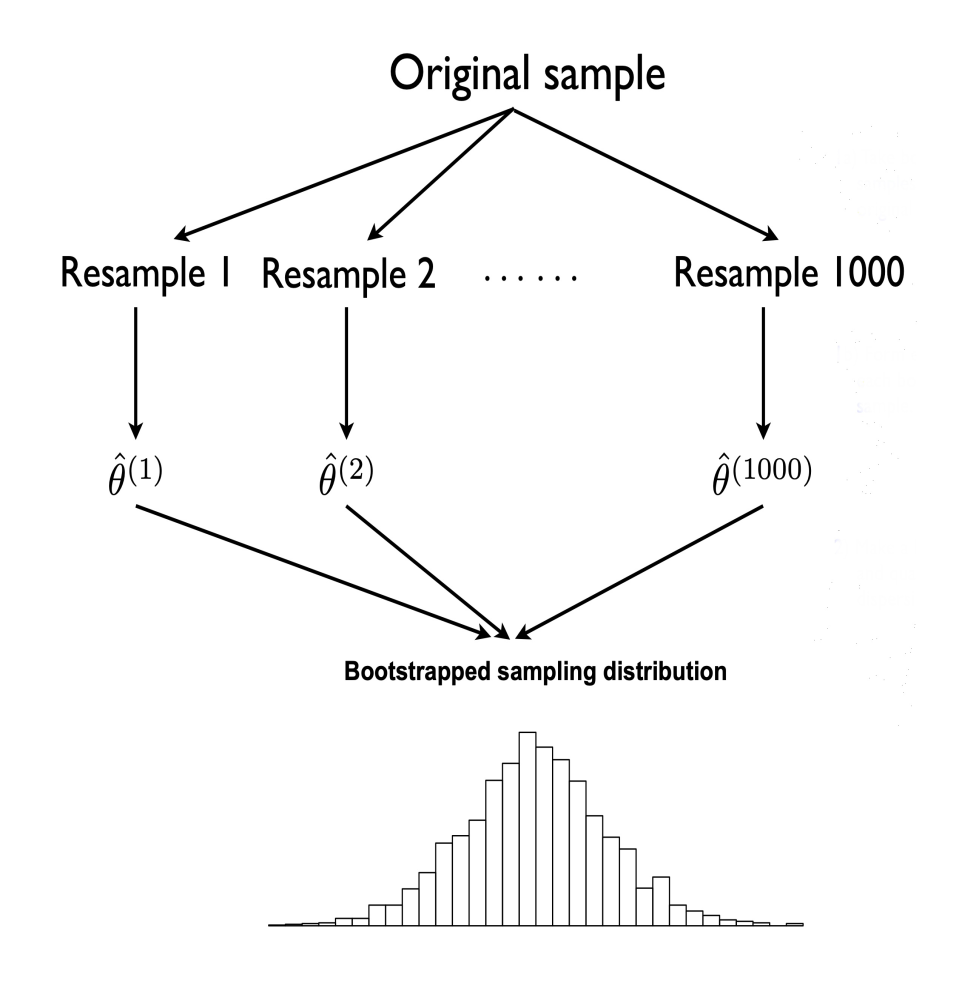
$$Mean \quad difference \quad of \quad two \quad groups$$

Bootstrap Test

Algorithm

- Choose a number of bootstrap samples.
- For each bootstrap sample, draw a sample with replacement with the given sample size.
- Calculate the test statistic of the given sample and every bootstrapped samples.
- Finding the proportion of test statistic of bootstrapped samples that were greater or equal to the given sample's test statistic.
- Inspect the p-value and draw conclusion thereby.

Bootstrap Test



P-value

> The p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.

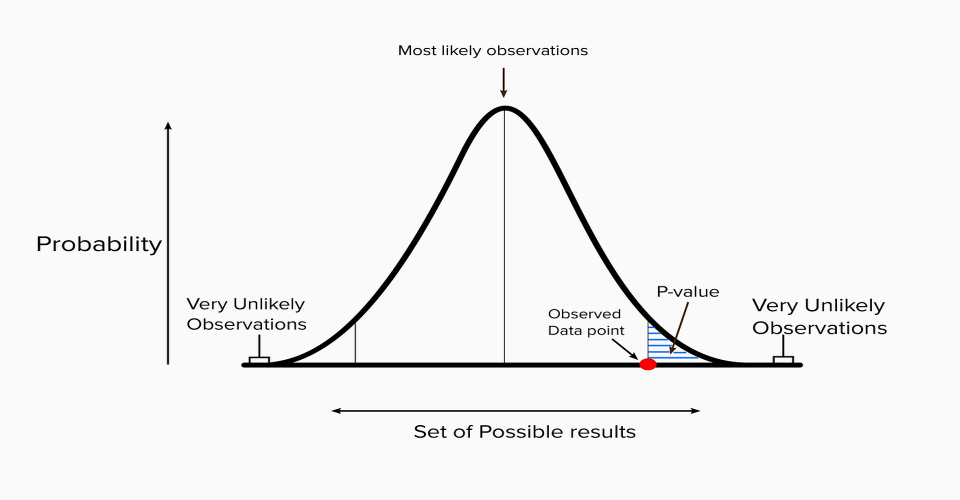
$$p - value = \quad P(E/H_0)$$

Confidence Interval

> A confidence interval refers to the probability (often 95% is used) that a population parameter will fall between a set of values for a certain proportion of times.

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

Interpreting P-value



Under the null the hypothesis if -

- P value > alpha : we fail to reject the null hypothesis.
- P value < alpha : we may reject the null hypothesis.

Function

The Function

```
test_of_hypothesis <-
  function(approach,independent=NULL,x,y=NULL,mu_0=NULL,sigma_x=NULL,
           sigma_y=NULL,sigma_d=NULL,n1=NULL,n2=NULL,B=NULL,alpha,H1)
```

Description

Arguments	Identities
approach	parametric or non-parametric
independent	True or False
x	sample data
y	sample data
mu_0	hypothesized mean
sigma_x	population standard deviation of x
sigma_y	population standard deviation of y
sigma_d	population standard deviation of paired differences
n1	number of rows/level of treatment
n2	number of columns/number of observations
B	number of replication
alpha	level of significance
H1	alternate hypothesis

Uses:

One-Sample Test of Means: Z test

```
x <- rnorm(50, mean = 3.5, sd = 1.7)
sigma <- 1
mu_0 <- 3
alpha <- 0.05
H1 <- "mu>mu_0"
```

```
test_of_hypothesis(approach = "parametric", x = x, sigma_x = sigma,
    mu_0 = mu_0, alpha = alpha, H1 = H1)
```

```
# A tibble: 1 × 4
  p_value Test_statistics upper_bound H1
    <dbl>           <dbl>       <dbl> <chr>
1   0.072            1.46        3.44 mu>mu_0
```

One-Sample Test of Means: T test

```r
x <- rnorm(20, mean = 3.5, sd = 1.7)
mu_0 <- 3
alpha <- 0.05
H1 <- "mu>mu_0"
```

```r
test_of_hypothesis(approach = "parametric", x = x, mu_0 = mu_0,
    alpha = alpha, H1 = H1)
```

```
# A tibble: 1 × 5
  p_value Test_statistics degrees_of_freedom upper_bound H1
    <dbl>           <dbl>              <dbl>       <dbl> <chr>
1   0.084            1.43                 19        4.18 mu>mu_0
```

Two-Sample Test of Means(Independent): T test

```
x <- rnorm(20, mean = 3.5, sd = 1.7)
y <- rnorm(20, mean = 5.4, sd = 2.6)
alpha <- 0.05
H1 <- "mu>mu_0"
```

```
test_of_hypothesis(approach = "parametric", independent = TRUE,
    x = x, y = y, alpha = alpha, H1 = H1)
```

```
# A tibble: 1 × 5
  p_value Test_statistics degrees_of_freedom upper_bound H1
    <dbl>           <dbl>              <dbl>       <dbl> <chr>
1   0.638          -0.357                 38       -1.50 mu>mu_0
```

Two-Sample Test of Means(Matched Paired): T Test

```r
x <- rnorm(20, mean = 3.5, sd = 1.7)
y <- rnorm(20, mean = 5.4, sd = 2.6)
alpha <- 0.05
H1 <- "mu>mu_0"
```

```r
test_of_hypothesis(approach = "parametric", independent = FALSE,
    x = x, y = y, alpha = alpha, H1 = H1)
```

```
# A tibble: 1 × 5
  p_value Test_statistics degrees_of_freedom upper_bound H1
    <dbl>           <dbl>              <dbl>       <dbl> <chr>
1   0.007            2.68                 19        2.81 mu>mu_0
```

More Than Two Means Test

```r
y <- c(22, 42, 44, 52, 45, 37, 52, 33, 8, 47, 43, 32, 16, 24,
    19, 18, 34, 39)
x <- rep(c("A", "B", "C"), each = 6)
n1 <- 3
n2 <- 6
alpha <- 0.05
H1 <- "means differ"
```

```r
test_of_hypothesis(approach = "parametric", x = x, y = y, n1 = n1,
    n2 = n2, alpha = alpha, H1 = H1)
```

```
  p_value test_statistic df1 df2 x CI_lower CI_upper           H1
1   0.112          2.541   2  15 A 29.79477 50.87190 means differ
2   0.112          2.541   2  15 B 25.29477 46.37190 means differ
3   0.112          2.541   2  15 C 14.46143 35.53857 means differ
```

Pearson's Chi-Square Test

```
x <- matrix(c(8, 4, 13, 9, 16, 14, 10, 16, 3, 7), 2, 5)
n1 <- 2
n2 <- 5
```

```
test_of_hypothesis(approach = "parametric", x = x, n1 = n1, n2 = n2,
    alpha = 0.05, H1 = "There is association")
```

```
  p_value Test_statistic degrees_of_freedom                    H1
1   0.269          5.179                  4 There is association
```

Bootstrap Test

```r
x <- rnorm(12, 10, 1)
y <- rnorm(1000, 20, 51)
B <- 1000
```

```r
test_of_hypothesis(approach = "non_parametric", x = x, y = y,
    B = B, alpha = 0.05, H1 = " Treatment effect exists ")
```

```
  p_val test_statistic                          H1
1 0.484       10.78548  Treatment effect exists
```

Thank you