Statistics Software Lab Report - 8

Name of the Student: Shatansh Patnaik
Roll No: 20MA20067

IIT Kharagpur
Statistics Software Lab

Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about a population parameter based on sample data. The process typically involves setting up a null hypothesis (H_0) and an alternative hypothesis (H_1), and then using sample data to assess the evidence against the null hypothesis.

The general steps for hypothesis testing are as follows:

1. Formulate the null hypothesis (H_0) and the alternative hypothesis (H_1).
2. Choose an appropriate test statistic, which depends on the type of data and the hypotheses being tested.
3. Specify the level of significance, denoted by α, which represents the probability of rejecting the null hypothesis when it is actually true.
4. Collect sample data and compute the test statistic.
5. Determine the critical region or critical value based on the chosen level of significance.
6. Compare the test statistic to the critical region or critical value.
7. Make a decision to either reject or fail to reject the null hypothesis based on the comparison.

The choice of test statistic and critical region depends on the specific hypothesis test being conducted. For example, in a hypothesis test concerning the population mean μ, the test statistic might be the sample mean $\bar{x}$, and the critical region might be determined using the standard normal distribution or the t-distribution, depending on whether the population standard deviation is known or unknown.

The decision to reject or fail to reject the null hypothesis is based on comparing the test statistic to the critical region or critical value. If the test statistic falls within the critical region, the null hypothesis is rejected in favor of the alternative hypothesis; otherwise, the null hypothesis is not rejected.

Hypothesis Testing for the Mean of a Normal Distribution (Unknown Variance)

Suppose we have a random sample $X_1, X_2, ..., X_n$ from a normal distribution with mean μ and known variance σ^2. We want to test hypotheses about the population mean μ.

Null and Alternative Hypotheses

The null hypothesis H_0 and alternative hypothesis H_1 can be stated as follows:

Test Statistic

The test statistic depends on the sample mean $\bar{X}$, the population mean under the null hypothesis μ_0, the sample standard deviation s, and the sample size n. For a sample from a normal distribution, the test statistic follows a Student's t-distribution:

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

Algorithm 1 Hypotheses

1: Null hypothesis: $H_0 : \mu = \mu_0$

2: Alternative hypothesis:

   1. $H_1 : \mu < \mu_0$ (left-tailed test)
   2. $H_1 : \mu \neq \mu_0$ (two-tailed test)
   3. $H_1 : \mu > \mu_0$ (right-tailed test)

Decision Rule

Algorithm 2 Decision Rule

1: For a left-tailed test ($H_1 : \mu < \mu_0$), reject H_0 if $t < -t_{\alpha,n-1}$, where $t_{\alpha,n-1}$ is the critical value from the t-distribution with $n-1$ degrees of freedom at significance level α.

2: For a two-tailed test ($H_1 : \mu \neq \mu_0$), reject H_0 if $|t| > t_{\alpha/2,n-1}$.

3: For a right-tailed test ($H_1 : \mu > \mu_0$), reject H_0 if $t > t_{\alpha,n-1}$.

Conclusion

Algorithm 3 Conclusion

1: If the null hypothesis is rejected, we have evidence to support the alternative hypothesis.

2: If the null hypothesis is not rejected, we do not have enough evidence to support the alternative hypothesis.

Hypothesis Testing for the Mean of a Normal Distribution (Known Variance)

Suppose we have a random sample $X_1, X_2, ..., X_n$ from a normal distribution with mean μ and known variance σ^2. We want to test hypotheses about the population mean μ.

Null and Alternative Hypotheses

The null hypothesis H_0 and alternative hypothesis H_1 can be stated as follows:

Test Statistic

Under the assumption of known variance σ^2, the test statistic is based on the sample mean $\bar{X}$ and follows a standard normal distribution:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Algorithm 4 Hypotheses

1: Null hypothesis: $H_0 : \mu = \mu_0$

2: Alternative hypothesis:

1. $H_1 : \mu < \mu_0$ (left-tailed test)
2. $H_1 : \mu \neq \mu_0$ (two-tailed test)
3. $H_1 : \mu > \mu_0$ (right-tailed test)

Decision Rule

Algorithm 5 Decision Rule

1: For a left-tailed test ($H_1 : \mu < \mu_0$), reject H_0 if $Z < -z_\alpha$, where z_α is the critical value from the standard normal distribution at significance level α.

2: For a two-tailed test ($H_1 : \mu \neq \mu_0$), reject H_0 if $|Z| > z_{\alpha/2}$.

3: For a right-tailed test ($H_1 : \mu > \mu_0$), reject H_0 if $Z > z_\alpha$.

Conclusion

Algorithm 6 Conclusion

1: If the null hypothesis is rejected, we have evidence to support the alternative hypothesis.

2: If the null hypothesis is not rejected, we do not have enough evidence to support the alternative hypothesis.

We can write the code for the same in R as follows:

```r
# Exercise - 1
# Assuming the value of variance for the data in unknown

get_value_of_t_given_mean_unknown_variance <- function(x, given_mean){
  t <- (mean(x) - given_mean) / (sd(x)/sqrt(length(x)))
  return(t)
}

alpha <- 0.05
deviations <- c(11.28, -9.48, -10.42, 6.25, -8.51, 10.11, 1.95, -8.65, 6.47,
     -0.68)
critical_value <- qt(1 - alpha/2, length(deviations)-1)
t <- get_value_of_t_given_mean_unknown_variance(deviations, 0)

if(abs(t) > critical_value){
  cat("Reject the Null Hypothesis that mean is equal to 0 with alpha = 0.05\
     n")
} else {
  cat("Accept the Null Hypothesis that mean is equal to 0 with alpha = 0.05\
     n")
}
```

```
# Exercise - 3
alpha <- 0.1
critical_value <- qt(1 - alpha/2, length(deviations)-1)

if(abs(t) > critical_value){
  cat("Reject the Null Hypothesis that mean is equal to 0 with alpha = 0.1\n
      ")
} else {
  cat("Accept the Null Hypothesis that mean is equal to 0 with alpha = 0.1\n
      ")
}

get_value_of_z_given_mean_known_variance <- function(x, given_mean, stdev){
  z <- (mean(x) - given_mean) / (stdev/sqrt(length(x)))
  return(z)
}

strengths <- c(578, 572, 570, 568, 572, 570, 570, 572, 596, 584)
z <- get_value_of_z_given_mean_known_variance(strengths, 570, 8.75)
alpha <- 0.01
critical_value <- qnorm(alpha)

if(z > critical_value){
  cat("Reject the Null Hypothesis that mean = 570 with alpha = 0.1\n")
} else {
  cat("Accept the Null Hypothesis that mean = 570 with alpha = 0.1\n")
}

# Exercise - 4
alpha <- 0.05
times <- c(9.85, 9.93, 9.75, 9.77, 9.67, 9.87, 9.67, 9.94, 9.85, 9.75)
t <- get_value_of_t_given_mean_unknown_variance(times, 10)
critical_value <- qt(1-alpha, length(times)-1)

if(t >= critical_value){
  print("Reject the Null Hypothesis that that the average time for the
      athelete to complete the race is less than 10 seconds")
} else {
  print("Accept the Null Hypothesis that that the average time for the
      athelete to complete the race is less than 10 seconds")
}
```

Hypothesis Testing for the Variance of a Normal Distribution

Suppose we have a random sample $X_1, X_2, ..., X_n$ from a normal distribution with mean μ and variance σ^2. We want to test hypotheses about the population variance σ^2.

Null and Alternative Hypotheses

The null hypothesis H_0 and alternative hypothesis H_1 can be stated as follows:

Algorithm 7 Hypotheses

1: Null hypothesis: $H_0 : \sigma^2 = \sigma_0^2$
2: Alternative hypothesis:

1. $H_1 : \sigma^2 < \sigma_0^2$ (left-tailed test)
2. $H_1 : \sigma^2 \neq \sigma_0^2$ (two-tailed test)
3. $H_1 : \sigma^2 > \sigma_0^2$ (right-tailed test)

Test Statistic

The test statistic for testing the variance follows a chi-square distribution with $n - 1$ degrees of freedom:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

where s^2 is the sample variance.

Decision Rule

Algorithm 8 Decision Rule

1: For a left-tailed test ($H_1 : \sigma^2 < \sigma_0^2$), reject H_0 if $\chi^2 < \chi^2_{\alpha,n-1}$, where $\chi^2_{\alpha,n-1}$ is the critical value from the chi-square distribution with $n - 1$ degrees of freedom at significance level α.
2: For a two-tailed test ($H_1 : \sigma^2 \neq \sigma_0^2$), reject H_0 if $\chi^2 < \chi^2_{\alpha/2,n-1}$ or $\chi^2 > \chi^2_{1-\alpha/2,n-1}$.
3: For a right-tailed test ($H_1 : \sigma^2 > \sigma_0^2$), reject H_0 if $\chi^2 > \chi^2_{1-\alpha,n-1}$.

Conclusion

Algorithm 9 Conclusion

1: If the null hypothesis is rejected, we have evidence to support the alternative hypothesis.
2: If the null hypothesis is not rejected, we do not have enough evidence to support the alternative hypothesis.

The above testing can successfully be demonstrated using the following code in R:

```
# Exercise - 2
get_value_of_x_given_variance <- function(x, given_variance){
  chi_square <- (length(x) - 1) * var(x) / given_variance
  return(chi_square)
}
```

```r

lengths <- c(7.12, 7.13, 7.01, 6.95, 6.89, 6.97, 6.99, 6.93, 7.05, 7.02)
alpha <- 0.05
chi_square <- get_value_of_x_given_variance(lengths, 0.005)
critical_value <- qchisq(alpha, length(lengths)-1)
print(critical_value)
print(chi_square)

if(chi_square > critical_value){
  cat("Reject the Null Hypothesis that variance is less than equal to 0.005
      with alpha = 0.05\n")
} else {
  cat("Accept the Null Hypothesis that variance is less than equal to 0.005
      with alpha = 0.05\n")
}
```

Hypothesis Testing for the Difference Between Means of Two Normal Distributions (Same Variance Assumption)

Suppose we have two independent random samples $X_1, X_2, ..., X_{n_1}$ from a normal distribution with mean μ_1 and variance σ^2, and $Y_1, Y_2, ..., Y_{n_2}$ from another normal distribution with mean μ_2 and the same variance σ^2. We want to test hypotheses about the difference between population means μ_1 and μ_2.

Null and Alternative Hypotheses

The null hypothesis H_0 and alternative hypothesis H_1 can be stated as follows:

Algorithm 10 Hypotheses

1: Null hypothesis: $H_0 : \mu_1 - \mu_2 = \delta_0$

2: Alternative hypothesis:

1. $H_1 : \mu_1 - \mu_2 < \delta_0$ (left-tailed test)
2. $H_1 : \mu_1 - \mu_2 \neq \delta_0$ (two-tailed test)
3. $H_1 : \mu_1 - \mu_2 > \delta_0$ (right-tailed test)

Test Statistic

Under the assumption of the same variance σ^2, the test statistic for the difference between means follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom:

$$t = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means, s_1^2 and s_2^2 are the sample variances, and n_1 and n_2 are the sample sizes.

Decision Rule

Algorithm 11 Decision Rule

1: For a left-tailed test ($H_1 : \mu_1 - \mu_2 < \delta_0$), reject H_0 if $t < -t_{\alpha,n_1+n_2-2}$.
2: For a two-tailed test ($H_1 : \mu_1 - \mu_2 \neq \delta_0$), reject H_0 if $|t| > t_{\alpha/2,n_1+n_2-2}$.
3: For a right-tailed test ($H_1 : \mu_1 - \mu_2 > \delta_0$), reject H_0 if $t > t_{\alpha,n_1+n_2-2}$.

Conclusion

Algorithm 12 Conclusion

1: If the null hypothesis is rejected, we have evidence to support the alternative hypothesis.
2: If the null hypothesis is not rejected, we do not have enough evidence to support the alternative hypothesis.

The above algorithm can be demonstrated in the code given below:

```
# Exercise - 5
test_of_hypothesis_of_two_means_less_than <- function(mean1, variance1,
    mean2, variance2, n1, n2){
  pooled_variance <- ((n1-1)*variance1 + (n2-1)*variance2) / (n1 + n2 - 2)
  t <- (mean1 - mean2) / sqrt(pooled_variance * (1/n1 + 1/n2))
  return(t)
}

t <- test_of_hypothesis_of_two_means_less_than(16, 1.4, 20, 2, 15, 19)
alpha <- 0.05
n1 <- 15
n2 <- 19
df <- n1 + n2 - 2
critical_value <- qt(1-alpha, df)

if (t >= critical_value) {
  print("Reject the null hypothesis")
} else {
  print("Accept the null hypothesis")
}
```

Hypothesis Testing for the Difference Between Means of Two Normal Distributions (Different Variances)

Suppose we have two independent random samples $X_1, X_2, ..., X_{n_1}$ from a normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, Y_2, ..., Y_{n_2}$ from another normal distribution with mean μ_2 and variance σ_2^2. We want to test hypotheses about the difference between population means μ_1 and μ_2.

Null and Alternative Hypotheses

The null hypothesis H_0 and alternative hypothesis H_1 can be stated as follows:

Algorithm 13 Hypotheses

1: Null hypothesis: $H_0 : \mu_1 - \mu_2 = \delta_0$
2: Alternative hypothesis:

1. $H_1 : \mu_1 - \mu_2 < \delta_0$ (left-tailed test)
2. $H_1 : \mu_1 - \mu_2 \neq \delta_0$ (two-tailed test)
3. $H_1 : \mu_1 - \mu_2 > \delta_0$ (right-tailed test)

Test Statistic

Under the assumption of different variances σ_1^2 and σ_2^2, the test statistic for the difference between means follows a t-distribution with degrees of freedom calculated using the Welch-Satterthwaite equation:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

$$t = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means, s_1^2 and s_2^2 are the sample variances, and n_1 and n_2 are the sample sizes.

Decision Rule

Algorithm 14 Decision Rule

1: For a left-tailed test ($H_1 : \mu_1 - \mu_2 < \delta_0$), reject H_0 if $t < -t_{\alpha,\nu}$.
2: For a two-tailed test ($H_1 : \mu_1 - \mu_2 \neq \delta_0$), reject H_0 if $|t| > t_{\alpha/2,\nu}$.
3: For a right-tailed test ($H_1 : \mu_1 - \mu_2 > \delta_0$), reject H_0 if $t > t_{\alpha,\nu}$.

Conclusion

Algorithm 15 Conclusion

1: If the null hypothesis is rejected, we have evidence to support the alternative hypothesis.
2: If the null hypothesis is not rejected, we do not have enough evidence to support the alternative hypothesis.

Hypothesis Testing for the Difference Between Variances of Two Normal Distributions

Suppose we have two independent random samples $X_1, X_2, ..., X_{n_1}$ from a normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, Y_2, ..., Y_{n_2}$ from another normal distribution with mean μ_2 and variance σ_2^2. We want to test hypotheses about the difference between population variances σ_1^2 and σ_2^2.

Null and Alternative Hypotheses

The null hypothesis H_0 and alternative hypothesis H_1 can be stated as follows:

Algorithm 16 Hypotheses

1: Null hypothesis: $H_0 : \sigma_1^2 - \sigma_2^2 = \delta_0$
2: Alternative hypothesis:

1. $H_1 : \sigma_1^2 - \sigma_2^2 < \delta_0$ (left-tailed test)
2. $H_1 : \sigma_1^2 - \sigma_2^2 \neq \delta_0$ (two-tailed test)
3. $H_1 : \sigma_1^2 - \sigma_2^2 > \delta_0$ (right-tailed test)

Test Statistic

The test statistic for the difference between variances follows an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom:

$$F = \frac{s_1^2}{s_2^2}$$

where s_1^2 and s_2^2 are the sample variances.

Decision Rule

Algorithm 17 Decision Rule

1: For a left-tailed test ($H_1 : \sigma_1^2 - \sigma_2^2 < \delta_0$), reject H_0 if $F < F_{\alpha, n_1-1, n_2-1}$.
2: For a two-tailed test ($H_1 : \sigma_1^2 - \sigma_2^2 \neq \delta_0$), reject H_0 if $F < F_{\alpha/2, n_1-1, n_2-1}$ or $F > F_{1-\alpha/2, n_1-1, n_2-1}$.
3: For a right-tailed test ($H_1 : \sigma_1^2 - \sigma_2^2 > \delta_0$), reject H_0 if $F > F_{\alpha, n_1-1, n_2-1}$.

Conclusion

Algorithm 18 Conclusion

1: If the null hypothesis is rejected, we have evidence to support the alternative hypothesis.
2: If the null hypothesis is not rejected, we do not have enough evidence to support the alternative hypothesis.

The above can be demonstrated in the code given below:

```
# Exercise - 6
test_of_hypothesis_of_two_variances <- function(variance1, variance2, n1, n2
    , alpha){
  f <- variance1 / variance2
  critical_value1 <- qf(1-alpha/2, n1-1, n2-1)
  critical_value2 <- qf(alpha/2, n1-1, n2-1)

  if(f>critical_value1 ||  f<critical_value2){
    cat("Reject the Null Hypothesis\n")
  } else {
    cat("Accept the Null Hypothesis\n")
  }
}


test_of_hypothesis_of_two_means <- function(mean1, mean2, variance1,
    variance2, n1, n2, alpha){
  pooled_variance <- ((n1-1)*variance1 + (n2-1)*variance2) / (n1 + n2 - 2)
  t <- (mean1 - mean2) / sqrt(pooled_variance * (1/n1 + 1/n2))
  df <- n1+n2-2
  critical_value <- qt(alpha/2, df)

  if(abs(t) > abs(critical_value)){
    cat("Reject the Null Hypothesis")
  } else {
    cat("Accept the Null Hypothesis")
  }
}

test_of_hypothesis_of_two_variances(3.89, 4.02, 8, 8, 0.05)
test_of_hypothesis_of_two_means(91.73, 93.75, 3.89, 4.02, 8, 8, 0.05)

# Exercise - 7
test_of_hypothesis_of_two_variances_given_data <- function(x, y, alpha){
  variance1 <- var(x)
  variance2 <- var(y)
  f <-  variance1 / variance2
  n1 <- length(x)
  n2 <- length(y)
  critical_value1 <- qf(1-alpha/2, n1-1, n2-1)
  critical_value2 <- qf(alpha/2, n1-1, n2-1)

  if(f>critical_value1 ||  f<critical_value2){
    cat("Reject the Null Hypothesis\n")
  } else {
```

```
    cat("Accept the Null Hypothesis\n")
  }
}

test_of_hypothesis_of_two_means_given_data_less_than <- function(x, y, alpha
    ){
  mean1 <- mean(x)
  mean2 <- mean(y)
  variance1 <- var(x)
  variance2 <- var(y)
  n1 <- length(x)
  n2 <- length(y)
  pooled_variance <- ((n1-1)*variance1 + (n2-1)*variance2) / (n1 + n2 - 2)
  t <- (mean1 - mean2) / sqrt(pooled_variance * (1/n1 + 1/n2))
  df <- n1+n2-2
  critical_value <- qt(1-alpha, df)

  if (t >= critical_value) {
    print("Reject the null hypothesis")
  } else {
    print("Accept the null hypothesis")
  }
}

x<- c(150, 250, 240, 280, 290, 210, 220, 180)
y <- c(140, 230, 270, 190, 270, 200, 150, 200, 190, 170)

test_of_hypothesis_of_two_variances_given_data(x, y, 0.1)
test_of_hypothesis_of_two_means_given_data_less_than(x, y, 0.05)

# Exercise - 8

x <- c(12, 29, 16, 37, 28, 15)
y <- c(10, 28, 17, 35, 25, 16)

test_of_hypothesis_of_two_variances_given_data(x, y, 0.1)
test_of_hypothesis_of_two_means_given_data_less_than(x, y, 0.05)
```

Contingency Table Test (Chi-square Test for Independence)

The contingency table test, also known as the Chi-square test for independence, is used to determine whether there is a significant association between two categorical variables.

Null and Alternative Hypotheses

The null hypothesis H_0 assumes that the two categorical variables are independent, while the alternative hypothesis H_1 suggests that there is a dependency between them.

Algorithm 19 Hypotheses

Null hypothesis: H_0: The two categorical variables are independent.
Alternative hypothesis: H_1: There is a dependency between the two categorical variables.

Test Statistic

The test statistic for the Chi-square test for independence is calculated based on the observed and expected frequencies in a contingency table. Let O_{ij} represent the observed frequency in the i-th row and j-th column of the table, and E_{ij} represent the expected frequency under the assumption of independence. Then, the Chi-square statistic is given by:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Decision Rule

Algorithm 20 Decision Rule

Reject H_0 if χ^2 is greater than the critical value $\chi^2_{\alpha,(r-1)(c-1)}$, where r is the number of rows and c is the number of columns in the contingency table, and α is the chosen significance level.

Conclusion

Algorithm 21 Conclusion

If the null hypothesis is rejected, we conclude that there is a significant association between the two categorical variables.
If the null hypothesis is not rejected, we do not have enough evidence to conclude that there is a dependency between the two categorical variables.

Demonstration using R is as follows:

```r
# Exercise - 9
is_age_opinion_related <- function(data, alpha){
  O <- data
  E <- data
  chi_sq_value <- 0
  for(i in 1 : nrow(data)){
    for(j in 1 : ncol(data)){
      E[i,j] <- (sum(data[i,]) * sum(data[,j])) / sum(data)
      chi_sq_value  <- chi_sq_value  + ((O[i,j] - E[i,j])^2) / E[i,j]
    }
  }
  critical_value <- qchisq(alpha,(nrow(data)-1)*(ncol(data)-1))
  if(chi_sq_value  > critical_value){
    print("Reject the Null Hypothesis")
  } else {
```

```
    print("Accept the Null Hypothesis")
  }
}
is_age_opinion_related(matrix(c(400,600,100,400,500,500),nrow=2,ncol=3)
    ,0.05)
```