Testing of Statistical Hypothesis

Adejumo Ridwan Suleiman

2022-06-30

What is a Statistical Hpothesis

- This is a formal statement about the nature of a population.
- Example;
  - Are all Africans rich?
  - Do all footballers make above $100k/week.
  - Do Ronaldo and Messi have the same goal ratio.

Hypothesis Testing

- This is the use of statistical tests to make an inference about a population.
- The method of inference used depends on the nature of the data and reason for analysis.

Steps in conducting a Statistical Hypothesis Test

- Stating the Null and Alternative Hypothesis
- Stating the Level of Significance
- Test Statistics to be used
- Critical Region
- Decision Rule

Null Hypothesis

- Know as the Hypothesis of no difference.
- This Hypothesis states that the relationship the researcher is trying to investigate does not exist.
- OUr hope is always to prove the **Null hypothesis** wrong.
- The **Null Hypothesis** is represented mathematically as H_0

Alternative Hypothesis

- Opposite of the **Null Hypothesis**
- This is usually the hypothesis of the researcher.
- Represented as H_A

Case Study

- Let's say the mean price of diamonds is less than \$10000
- In this case the **Null Hypothesis** can be stated as;
  "The mean price of diamonds is \$10000"
- The **Alternative hypothesis** can be stated as ;
  "The mean price of diamonds is less than \$1000"
- The parameter of interest in this example is the **Mean**.

Type 1 and Type 2 error

- Situation when you reject the **Null Hypothesis** when it is actually true is know as **Type 1** error.
- Situation when you fail to reject the **Alternative Hypothesis** when it is actually false is know as **Type 2** error.

Level of Significance (α)

- Situation where an experiment occurred beyond chance, then it is said to **Statistically Significant**.
- If we are to perform an experiment a number of times and the result of that experiment does not occurs by chance the experiment is declared **Statistically Significant**
- The level of significance denoted by α is the probability of rejecting the null hypothesis when it is actually true.
- The level of significance is mostly used as **0.05**.
- We can say that this is the probability of committing a **Type 1** error.

p-Value

- This is the probability of having the result of an experiment by chance when reproduced many times.
- For example a p-value of 0.32 means that when the experiment is repeated a number of times, there is a 32% chance that the result of your experiment actually occurred by chance.
- The lower the p-value the higher the statistical significance of the test.

Correlation Test

- This statistics test if two continuous variables are related to each other.
- The value ranges from -1 to +1 and the higher the value the stronger the relationship.
- Let's say from the diamonds data example we want to see if there is a relationship between diamonds carat and price.

```
## 
##  Pearson's product-moment correlation
## 
## data:  diamonds$price and diamonds$carat
## t = 551.41, df = 53938, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.9203098 0.9228530
## sample estimates:
##       cor 
## 0.9215913
```

One-Sample t-test

- Use to see if the mean of a sample from a population is equal to a specific value.
- Let's assume someone argues with me that the mean price of diamonds is \$6000 and I am not capable to get the mean price of all diamonds from the population.
- Instead I take a random sample of 100 diamonds from the population and perform a t-test and compare it with the mean which we don't know if it is true or not.
- We can state the **Null hypothesis** as:
  The mean price of diamonds is \$6000
- We can state the **Alternative hypothesis** as:
  The mean price of diamonds is not \$6000

One-Sample t-test(continuation)

```
##
##  One Sample t-test
##
## data:  diamonds[sample(nrow(diamonds), 100), ]$price
## t = -5.9537, df = 99, p-value = 3.999e-08
## alternative hypothesis: true mean is not equal to 6000
## 95 percent confidence interval:
##  2696.175 4347.865
## sample estimates:
## mean of x
##   3522.02
```

One-Sample t.test(continuation)

What if I am to use the compare the actually mean price of diamonds or a value closer to that. Let's try 4000(the actual mean is 3900)

```
##
##  One Sample t-test
##
## data:  diamonds[sample(nrow(diamonds), 100), ]$price
## t = -1.1484, df = 99, p-value = 0.2536
## alternative hypothesis: true mean is not equal to 4000
## 95 percent confidence interval:
##  2696.175 4347.865
## sample estimates:
## mean of x
##   3522.02
```

- Can you see the difference?

Two-Sample t-test

- This is like the one-sample t-test but is used to compare the means of two samples.
- This time around I have 100 samples of diamonds from two different store.
- I want to know if the mean price from these store are the same.
- We can state our **Null Hypothesis** as:
  The average price of diamonds from Store A is the same as that of Store B
- The **Alternative Hypothesis** takes the form:
  The average price of diamonds from Store A is not the same as that of Store B

Two-Sample t-test(continuation)

```
##
##  Welch Two Sample t-test
##
## data:  Store_A and Store_B
## t = -0.17614, df = 196.93, p-value = 0.8604
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -1220.246  1020.146
## sample estimates:
## mean of x mean of y
##   3522.02   3622.07
```

Analysis of Variance (ANOVA)

- Now I have four samples from four different Store, what do I do?
- The **ANOVA** is used when you want to compare the mean of more than two samples.
- We can state our **Null Hypothesis** as:
  The mean price of diamonds is the same across all Stores
- The **Alternative Hypothesis** takes the form:
  The average price of diamonds is not the same across all Stores

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## df$Store      1 6.630e+05   662953   0.047  0.829
## Residuals   398 5.631e+09 14149331
```