Inferential Statistics

semicolon

Learning Outcomes

- Participants should understand the theory and application of inferential statistics, Parameter estimation, Hypothesis testing, Analysis of Variance

Inferential Statistics

- Inferential statistics and descriptive statistics are the two main branches of statistics. Descriptive statistics derives a summary from the data by using central tendencies like mean and median, dispersions like variance and standard deviation, and skew-ness
- Inferential statistics describes and makes inferences about the population and the sampled data. In inferential statistics, you use hypothesis testing and estimating of parameters. By estimating parameters, you try to answer the population parameters.

Hypothesis Testing

- In hypothesis testing, you try to answer a research question. In hypothesis testing, a research question is a **hypothesis** asked in question format. A research question can be, Is there a significant difference between the grades of class 1 and class 2 for their engineering math exams? A hypothesis can be, There is a significant difference between the grades of class 1 and class 2 for their engineering math exams.

Hypothesis Testing

- The research question begins with Is there and the hypothesis begins with There is. Based on the research question, the hypothesis can be a null hypothesis, H0, and an alternate hypothesis, Ha. A null hypothesis, H0, can be **μ1** = **u2** and an alternate hypothesis, Ha, can be μ1 ≠ u2. So μ1 is the mean of the grades of class 1 and μ2 is the mean of the grades of class 2. You can then use inference tests to get the p-value. If the p-value is less than or equal to alpha, which is usually 0.05, you reject the null hypothesis and say that the alternate hypothesis is true at the 95% confidence interval. If the p-value is more than 0.05, you fail to reject the null hypothesis.

Hypothesis Testing

For estimating parameters, the parameters can be the mean, variance, Standard deviation, and others. If you want to estimate the mean of heights of the whole population (and by the way, it is impossible to measure everyone in the population), you can use a sampling method to select some people from the population. Subsequently, you calculate the mean of the height of the samples and then make an inference on the mean of the height of the population. You can then construct the confidence intervals, which is the range in which the mean of the height of the population will fall. You construct a range because the sample cannot derive the exact mean of the height of the population.

Hypothesis Testing

A hypothesis can also be a null hypothesis, H0, and an alternate hypothesis, Ha. You can write the null hypothesis and alternate hypothesis as follows:

H0: $\mu 1 = u2$

Ha: $\mu 1 \neq u2$

where $\mu 1$ is the mean of one data, and $\mu 2$ is the mean of another data. You can use statistical tests to get your p-value. You use a t-test for continuous variables or data, and you use a chi-square test for categorical variables or data. For more complex testing, you use ANOVA. If data is not normally distributed, you use non-parametric tests. A P-value helps you

determine the significance of your statistical test results.

Hypothesis Testing

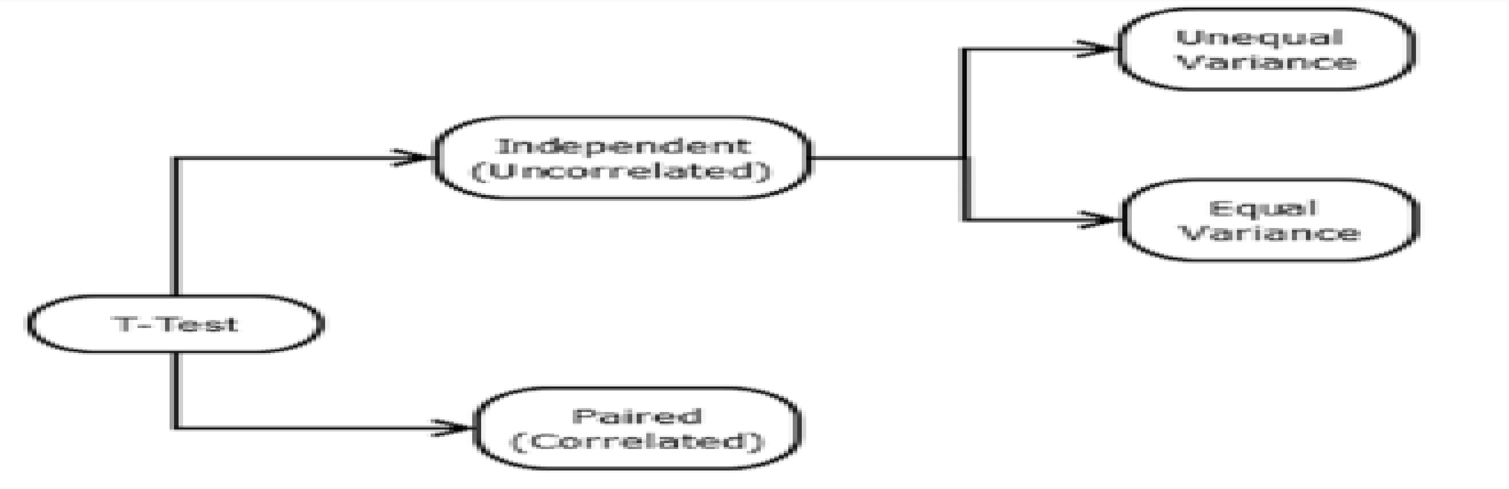
Your claim in the test is known as a null hypothesis and the alternate hypothesis means that you believe the null hypothesis is untrue.

A small p-value <= alpha, which is usually 0.05, indicates that the observed data is sufficiently inconsistent with the null hypothesis, so the null hypothesis may be rejected. The alternate hypothesis is true at the 95% confidence interval.

A larger p-value means that you failed to reject null hypothesis.

T-Test

- T-test is one of the more important tests in data science. A t-test is used to determine whether the mean between two samples/groups are equal to each other. The null hypothesis means that the two means are equal, and the alternative means that the two means are different.



Assumptions of T-Tests

The samples are randomly sampled from their population..
Continuous dependent variable

Dependent variable has a *normal distribution* in each group

Type I and Type II Errors

A type I error is a rejection of the null hypothesis when it is really true.
A type II error is a failure to reject a null hypothesis that is false.

One-Sample T-Test

- A one-sample t-test is used to test whether the mean of a population is equal to a specified mean.
  The formula of a one-sample t-test is

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

- where s is the standard deviation of the sample, n is the size of the sample, m is the mean of the sample, and u is the specified mean.
  The degree of freedom formula is *df* = n-1 .We can use the t statistics and the degree of freedom to estimate the p-value using a t-table

One Sample T-test

```
> dt<-read.csv(file.choose(),header = T)
> t.test(dt$price,mu=3000)
               One Sample t-test
data:  dt$price
t = 54.304, df = 53939, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 3000
95 percent confidence interval:
 3899.132 3966.467
sample estimates:
mean of x
•          3932.8
```

One Sample T-test

- In the example above , we tested a one-sample t-test
  $H0: \mu = 3000$
  $Ha\ \mu \neq 3000$
-

  m is 3000 in the above R code. The p-value is 2.2e-16, so the p-value is less than 0.05, which is the alpha value(level of Significance). Therefore, the null hypothesis may be rejected. The alternate hypothesis, $\mu \neq 3000$, is true at the 95% confidence interval.

Two-Sample Independent T-Test

The two-sample unpaired t-test is when you compare two means of two independent samples. The formula of the two-sample independent t-test is

$$t = \frac{\mu_A - \mu_B}{\sqrt{\frac{s^2}{n_A} + \frac{s^2}{n_B}}}$$

where
μA is the mean of one sample,
μB is the mean of the second sample,
nA is the size of sample A, and
nB is the size of sample B

Two-Sample Independent T-Test

- S2 is the estimator of the common variance of the two samples, and the formula is

$$s^2 = \frac{\Sigma\left(x-\mu_A\right)^2 + \Sigma\left(x-\mu_B\right)^2}{n_A - n_B - 2}$$

- The degrees of freedom formula is *df* = *nA*-*nB*- 2

Two-Sample Independent T-Test

```
>ct<-read.csv(file.choose(),header = T)
>t.test(ct$area_se~ct$diagnosis, alternative="two.sided")

            Welch Two Sample t-test

data:  ct$area_se by ct$diagnosis
t = -12.156, df = 216.22, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -59.89391 -43.18061
sample estimates:
mean in group B mean in group M
       21.13515        72.67241
```

Two-Sample Independent T-Test

- The p-value < 2.2e-16, so it is less than 0.05, which is the alpha value. Therefore, the null hypothesis may be rejected. The alternate hypothesis, $\mu A - \mu B \neq 0$, is true at the 95% confidence interval.

- Two Sample Independent T-test is of two types,

Equal Variance

Unequal Variance

Most times we assume unequal variance but to be sure, we can always carry out variance-test to determine if the variance in both groups are equal. We use the function “var.test” for variance test in R

Two Sample Paired Test

- A two-sample paired t-test is used to test the mean of two samples that depend on each other. The t-test formula is

$$t = \frac{\bar{d}}{\sqrt{s^2 / n}}$$

- where
  d is the mean difference,
  s is the sample variance, and
  n is the sample size.
  The degree of freedom formula is

$$df = n\text{-}1$$

Two Sample Paired Test

- > # Weight participants before treatment
- > x<-c(239, 170, 182, 143, 141, 170, 160, 154, 185, 130)
- > # Weight participants after treatment
- > y<-c(362, 343, 330, 340, 314, 567, 292, 439, 312, 292)
- > paired_1<-t.test(x, y, paired=TRUE)
- > paired_1

Two Sample Paired Test

- Paired t-test

- data: x and y
- t = -7.0343, df = 9, p-value = 6.09e-05
- alternative hypothesis: true difference in means is not equal to 0
- 95 percent confidence interval:
- -253.3491 -130.0509
- sample estimates:
- mean of the differences
- -191.7

Two Sample Paired Test

- The p-value is 6.09e-05, so it is less than 0.05, which is the alpha value. Therefore, the null hypothesis may be rejected. The alternate hypothesis, *μA* - *μB* ≠ 0, is true at the 95% confidence interval.

Reference

- All IT eBooks. “Learn R for Applied Statistics - PDF EBook Free Download.” *Allitebooks.In*, 13 Feb. 2019, www.allitebooks.in/learn-r-for-applied-statistics/.

Questions ??

semicolon