Introduction to Inferential Statistics

semicolon

Learning Objectives

- At the end of the lesson, participants should know different types of inferential to use when faced with different analysis problem.
- Participants should understand what hypothesis test is as well as different tests in hypothesis.
- Participants should be able to state an hypothesis statement based on their analysis problem.
- Participants should be able to know when to apply Chi-square test and Anova test in carrying out their analysis.

What is Inferential Statistics

Inferential statistics, is an aspect of statistics that uses the result from a sample data to draw conclusion about the population since it is expensive, time consuming and capital intensive to carry out an experiment on the population most times. We use inferential statistics to assess the likelihood that an observed difference between groups is reliable or that it occurred by chance in this study.

Inferential statistics is classified into two namely:

a. Hypothesis Testing
b. Regression Analysis

Hypothesis Testing

Hypothesis testing is a type of statistical analysis that involves testing your assumptions about a population parameter. It is used to calculate the correlation between two statistical variables.

Example: a teacher assuming 70% of his students come from lower-middle-class families

Types of Hypothesis Testing

Null Hypothesis and Alternate Hypothesis

The assumption that the event will not occur is known as the Null Hypothesis. Unless and until it is rejected, a null hypothesis has no bearing on the study's outcome.
H0 is the symbol for it, and it is pronounced H-naught.

The null hypothesis logical opposite is the Alternate Hypothesis. Following the rejection of the null hypothesis, the alternative hypothesis is accepted. It is represented by the symbol H1.

Example A sanitizer manufacturer claims that its product kills 95 percent of germs on average.
To put this company's claim to the test, create a null and alternate hypothesis.
H0 (Null Hypothesis): Average = 95%.
Alternative Hypothesis (H1): The average is less than 95%.

One-Tailed and Two-Tailed Hypothesis Testing

The One-Tailed test, also known as a directional test, considers a critical region of data that, if the test sample falls into it, would result in the null hypothesis being rejected, implying acceptance of the alternate hypothesis.

In a one-tailed test, the critical distribution area is one-sided, meaning the test sample is either greater or lesser than a specific value.

Example:

Supposed H0: mean >=50, then H1: mean <50

Here the mean is less than 50. It is called a One-tailed test.

One-Tailed and Two-Tailed Hypothesis Testing

In a two-tailed test, the test sample is checked to see if it is greater or less than a range of values, implying that the critical distribution area is two-sided.
If the sample falls within this range, the alternate hypothesis will be accepted, and the null hypothesis will not be rejected.

Example:
Suppose H0: mean = 50 and H1: mean not equal to 50
According to the H1, the mean can be greater than or less than 50. This is an example of a Two-tailed test.

Type 1 and Type 2 Error

Type 1 and type two error are some of the basic and fundamental error sometimes being made by data scientist or data analyst most time due to their own perception of that particular scenarios.

- Type 1 Error: A Type-I error occurs when sample results reject the null hypothesis despite being true.
- Type 2 Error: A Type-II error occurs when the null hypothesis is not rejected when it is false, unlike a Type-I error.

Level of Significance

The alpha value is used to determine whether or not a test is statistically significant. In a statistical test, Alpha represents the probability of a Type I error that is acceptable. Because alpha is a probability, it can range from 0 to 1. The most common alpha values in practice are 0.01, 0.05, and 0.1, which represent a 1%, 5%, and 10% chance of a Type I error, respectively (i.e. rejecting the null hypothesis when it is in fact correct)

Confidence Interval

A confidence interval is useful for estimating the parameters of a population. A 95 percent confidence interval, for example, indicates that if a test is repeated 100 times with new samples under the same conditions, the estimate can be expected to fall within the given interval 95 times. A confidence interval can also be used to calculate the critical value in hypothesis testing.

P-Value

A p-value is a metric that expresses the probability that an observed difference happened by chance. The statistical significance of the observed difference increases as the p-value decreases. If the p-value is too low, the null hypothesis is rejected.

P-Value

Example:
Assume you want to see if the new advertising campaign has increased product sales. The p-value represents the probability that the null hypothesis, which states that there is no change in sales as a result of the new advertising campaign, is correct. If the p-value is.30, there is a 30% chance that the product's sales will not increase or decrease. If the p-value is 0.03, there is a 3% chance that there will be no increase or decrease in sales value as a result of the new advertising campaign. As you can see, the lower the p-value, the greater the likelihood that the alternate hypothesis is correct, implying that the new advertising campaign causes an increase or decrease in sales.

Type of Test Statistics

Z-test

The **z test** is applied to data that has a normal distribution and a sample size greater than or equal to 30. When the population variance is known, it is used to test if the sample and population means are equal. The right-tail hypothesis can be constructed as follows:

Null Hypothesis: $H0 : \mu = \mu0$

Alternate Hypothesis: $H1 : \mu > \mu0$

Type of Test Statistics

Z-test

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

X: mean of the sample

μ: population mean

σ: standard deviation of the population

n: length of values

Decision Criteria: If the z statistic > z critical value then reject the null hypothesis.

Type of Test Statistics

T-test

When the data follows a student t distribution and the sample size is less than 30, the t test is used. When the variance of the population is unknown, it is used to compare the sample and population means. The following is the hypothesis test for inferential statistics:

$$t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \quad , \quad df = n - 1$$

Type of Test Statistics

T-test

Null Hypothesis: H0 : $\mu = \mu 0$

Alternate Hypothesis: H1 : $\mu > \mu 0$

n is the size of the sample,
S_d standard deviation,
X_d is the mean of the sample,
μ_d is the specified mean.
df = n−1 : The degree of freedom formula.
Decision Criteria: If the t statistic > t critical value then reject the null hypothesis.

Assumptions of T-Tests

a. The samples are randomly sampled from their population.
b. Continuous dependent variable
c. Dependent variable has a normal distribution in each group

Type I and Type II Errors

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

Example 1

Assuming you have being Employed by Semicolon Africa as a data scientist to check if their new products will thrive better than the others already existing in the market. Given the current rate of usage by customers in the market of both the new and already existing products. They randomly sample weekly usage, is given as follows:

data: 45, 42, 64, 54, 58, 49, 48, 56

The average usage of the products is 60 users per hour Test the null hypothesis at the 0.05 level of significance.

Statement of Hypothesis

- H0: Usage of new products = 60
- H1: Usage of new products differs by 60

```
Product <- t.test(data, mu = 60, alternative = "less", conf.level = 0.95)
```

T-test Result

One Sample t-test
data: data
t = -3.0956, df = 7, p-value = 0.008714
alternative hypothesis: true mean is less than 60
95 percent confidence interval:
    -Inf 56.89615
sample estimates:
mean of x
      52

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

Where
X_1 is the mean of one sample,
X_2 is the mean of the second sample,
S_1^2 is the size of sample 1,
S_2^2 is the size of sample 2
N_1 is the size of sample 1
N_2 is the size of sample 2

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

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
72.6724
```

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 will be rejected. The alternate hypothesis,

μA – μB ≠ 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 T-Test

The Paired Samples t Test compares the means of two measurements taken from the same person, object, or unit of measurement. These "paired" measurements could represent: A measurement taken at two distinct points in time (e.g., pre-test and post-test score with an intervention administered between the two time points).

S is the estimator of the common variance of the two samples, and the formula is

$$s = \sqrt{\frac{(x_i - \bar{x})^2}{n-1}}$$

- The degrees of freedom formula is

Two Sample Paired Test

Used when two samples are not independent of each other Observations in one sample can be paired with observations in the other sample For example:

- Before and after observations on the same subjects
- A comparison of two different measurements or treatments on the same subjects

Example

Four individuals with high levels of cholesterol went on a special diet, avoiding high cholesterol foods and taking special supplements. Using the.05 level of significance, was there a significant decrease in cholesterol level?

Their total cholesterol levels before and after the diet were as follows:

Before	After
287	255
305	269
243	245
309	247

Two Sample Paired Test

- # Weight participants before treatment
- before <- c(287, 305, 243, 309)
- # Weight participants after treatment
- after <-c(255, 269, 245, 247)
- paired_1<-t.test(x, y, paired=TRUE)
- paired_1

Results

Paired t-test

data: before and after

t = 2.4353, df = 3, p-value = 0.09289

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-9.8182 73.8182

sample estimates:

mean of the differences

32

F-test

The f test is used to determine whether or not there is a difference in the variances of two samples or populations. The right tailed f hypothesis test is as follows:

Null Hypothesis: H0 : $\sigma^2_1 = \sigma^2_2$

Alternate Hypothesis: H1 : $\sigma^2_1 > \sigma^2_2$

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

$$where\ \sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

where σ^2_1 is the variance of the first population and σ^2_2 is the variance of the second population.

Decision Criteria: If the f test statistic > f test critical value then reject the null hypothesis.

- **Method 1:** var.test(x, y, alternative = "two.sided")

Chi square Test

Contingency analysis is a hypothesis test used to determine whether or not two categorical variables are independent. In other words, we're asking, "Can we predict the value of one variable if we know the value of the other?" If the answer is yes, we can conclude that the variables in question are not independent. If the answer is no, the variables under consideration are said to be independent. The test is known as 'Contingency Analysis' because it makes use of contingency tables. The test is also known as the 'Chi-square test of independence' because the test statistic has a chi-square distribution and it is used to determine whether two categorical variables are independent or not.

Chi square Test

The null hypothesis of the test is that the two variables are independent and the alternative hypothesis is that the two variables are not independent.

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying. Therefore, a chi square test is an excellent choice to help us better understand and interpret the relationship between our two categorical variables.

Chi square Test

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.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

Chi square Test

Given the mtcars dataset from R environment, find out if the 'cyl' and 'carb' variables are in 'mtcars' dataset and whether it is dependent or not.

```
data("mtcars")
table(mtcars$carb, mtcars$cyl)    #Author DataFlair
chisq.test(mtcars$carb, mtcars$cyl)
```

From the output from our chisq calculation, we have a high chi-squared value and a p-value of less than 0.05 significance level. So we reject the null hypothesis and conclude that carb and cyl have a significant relationship.

Chi square Test

Example:
A Human Resources department of an organization wants to check whether age and experience of the employees are dependent on each other. For this purpose, a random sample of 1470 employees is collected with their age and experience

Alpha_value = 0.05

Statement of Hypothesis
H0: Age and Experience are two independent variables
H1: Age and Experience are two dependent variables

Chi square Test

```
data<-read.csv(file.choose())
dim(data)
head(data)

ct <- prop.table(data$age.intervals,data$suicides.100k.pop)

#calculating the chi-squ test
chisq.test(ct)
#output
Pearson's Chi-squared test

data: ct
X-squared = 679.97, df = 9, p-value < 2.2e-16
```

Chi square Test

The p-value here is less than 0.05. Therefore, we will reject our null hypothesis. We can conclude that age and experience are two dependent variables, aka as the experience increases, the age also increases (and vice versa).

Analysis of Variance(ANOVA)

ANOVA is the process of testing the means of two or more groups. It also checks the impact of factors by comparing the means of different samples. In a t-test, you test the means of two samples; in a chi-square test, you test categorical attributes or variables; in ANOVA, you test means of two or more groups

To perform an anova, you must have a continuous response variable and at least one categorical factor with two or more levels. anova requires data from approximately normally distributed populations with equal variances between factor levels. However, anova.

Analysis of Variance(ANOVA)

ANOVA is the process of testing the means of two or more groups. It also checks the impact of factors by comparing the means of different samples. In a t-test, you test the means of two samples; in a chi-square test, you test categorical attributes or variables; in ANOVA, you test means of two or more groups

Analysis of Variance(ANOVA)

To perform an anova, you must have a continuous response variable and at least one categorical factor with two or more levels. anova requires data from approximately normally distributed populations with equal variances between factor levels. However, anova. procedures work quite well even if the normality assumption has been violated unless one or more of the distributions are highly skewed or if the variances are quite different.

Analysis of Variance(ANOVA)

Grand Mean

In ANOVA, you use two kinds of means, sample means and a grand mean. A grand mean is the mean of all of the samples" means.

Hypothesis

In ANOVA, a null hypothesis means that the sample means are equal or do not have significant differences. The alternate hypothesis is when the sample means are not equal..

Assumptions of Anova

You assume that the variables are randomly sampled, independent, and selected or sampled from a population that is normally distributed with unknown but equal variances

Terminology Related to Anova
Between group variability

Between Group Variation: The difference in mean between each group and the overall mean. The study of how the means of groups differ from one another is known as between group variation.

$$SS(B) = \sum n(\bar{x} - \bar{X}_{GM})^2$$

$$MS_{between} = \frac{n1(x_1 - x)^2 + n2(x_2 - x)^2 + n3(x_3 - x)^2}{df}$$

This measures the interaction between the groups or samples. If the group means don't differ greatly from each other and the grand mean, the SS(B) will be small.

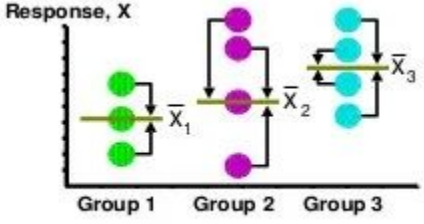
Terminology Related to Anova

Within-group variability

Within-Group Variation: It refers to variations caused by differences between individuals in different groups (or levels). In other words, not all values (e.g., means) within each group are the same. These are differences that are not due to the independent variable.

$$SSW = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2$$

$$SSW = (x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_1)^2 + \ldots + (x_{Kn_K} - \bar{x}_K)^2$$



Types of Anova Test

One-Way Anova

There are many situations where you need to compare the mean between multiple groups. For instance, the marketing department wants to know if three teams have the same sales performance.

- Team: 3 level factor: A, B, and C
- Sale: A measure of performance

The ANOVA test can tell if the three groups have similar performances.

Types of Anova Test

One-Way Anova

To clarify if the data comes from the same population, you can perform a **one-way analysis of variance** (one-way ANOVA hereafter). This test, like any other statistical tests, gives evidence whether the H0 hypothesis can be accepted or rejected.

- H0: The means between groups are identical
- H1: At least, the mean of one group is different

In other words, the H0 hypothesis implies that there is not enough evidence to prove the mean of the group (factor) are different from another.
This test is similar to the t-test, although ANOVA test is recommended in situation with more than 2 groups.

Case Study 1

Using the poison dataset, determine if there is difference in survival time average between groups. The dataset contains 48 rows and 4 variables:

- Time: Survival time of the animal
- poison: Type of poison used: factor level: 1,2 and 3
- treat: Type of treatment used: factor level: 1,2 and 3
- X: continuous values

Steps Taken in performing the anova case study

- Step 1: Check the format of the variable poison
- Step 2: Print the summary statistic: count, mean and standard deviation
- Step 3: Plot a box plot
- Step 4: Compute the one-way ANOVA test
- Step 5: Run a pairwise t-test

Case Study 1

```
library(dplyr)

PATH <- "https://raw.githubusercontent.com/guru99-edu/R-Programming/master/poisons.csv"
df <- read.csv(PATH) %>%
  select(-X) %>%
  mutate(poison = factor(poison, ordered = TRUE)) #
glimpse(df)
```

Our objective is to test the following assumption:

- H0: There is no difference in survival time average between group
- H3: The survival time average is different for at least one group.

```
levels(df$poison)
[1] "1" "2" "3"

df %>%
 group_by(poison) %>%
 summarise(
  count_poison = n(),
  mean_time = mean(time, na.rm = TRUE),
  sd_time = sd(time, na.rm = TRUE)
 )
```

A tibble: 3 x 4

	poison	count_poison	mean_time	sd_time
	<ord>	<int>	<dbl>	<dbl>
1	1	16	0.617500	0.20942779
2	2	16	0.544375	0.28936641
3	3	16	0.276250	0.06227627

Case Study 1

```
ggplot(df, aes(x = poison, y = time, fill = poison)) +
  geom_boxplot() +
  geom_jitter(shape = 15,
              color = "steelblue",
              position = position_jitter(0.21)) +
  theme_classic()

anova_one_way <- aov(time~poison, data = df)

summary(anova_one_way)
```

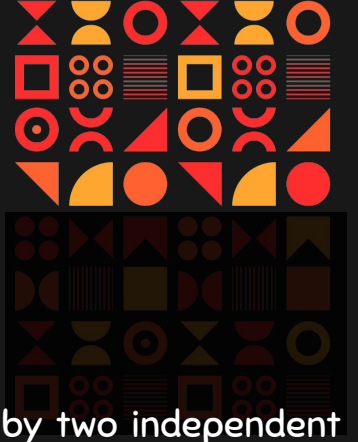
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
poison	2	1.033	0.5165	11.79	7.66e-05 ***
Residuals	45	1.972	0.0438		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Case Study 1

```
summary(anova_one_way)
            Df Sum Sq Mean Sq F value   Pr(>F)
poison       2  1.033  0.5165   11.79 7.66e-05 ***
Residuals   45  1.972  0.0438
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is lower than the usual threshold of 0.05. You are confident to say there is a statistical difference between the groups, indicated by the "*".



Two Way Anova

The two-way ANOVA compares mean differences between groups divided by two independent variables (called factors). The primary goal of a two-way ANOVA is to determine whether or not the two independent variables interact with the dependent variable. For example, you could use a two-way ANOVA to see if there is an interaction between gender and educational level on test anxiety among university students, where gender (males/females) and educational level are variables.

Example 1

The objective of the ANOVA test is to analyse if there is a (statistically) significant difference in breast cancer, between different continents.
Statement of Hypothesis

- Null Hypothesis: all seven continents means are equal –> there is no relationship between continents and new cases of breast cancer, which we can write as follows:

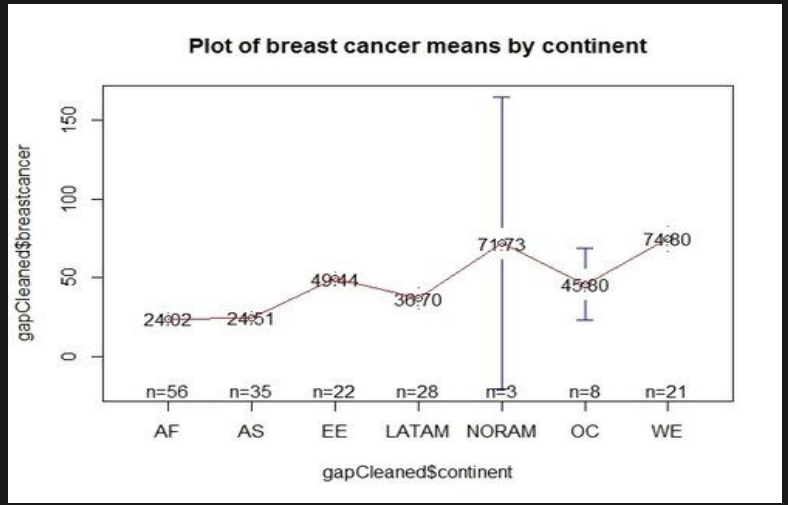
  H0: U1 = U2 = U3 = U4 = U5 = U6 = U7

- Alternative Hypothesis: not all seven continents means are equal –> there is a relationship between continents and new cases of breast cancer:

  H1: not all U are equal

```
means<- round(tapply(gapCleaned$breastcancer, gapCleaned$continent, mean), digits=2)  # note that I I round values to just 2 decimal places
```

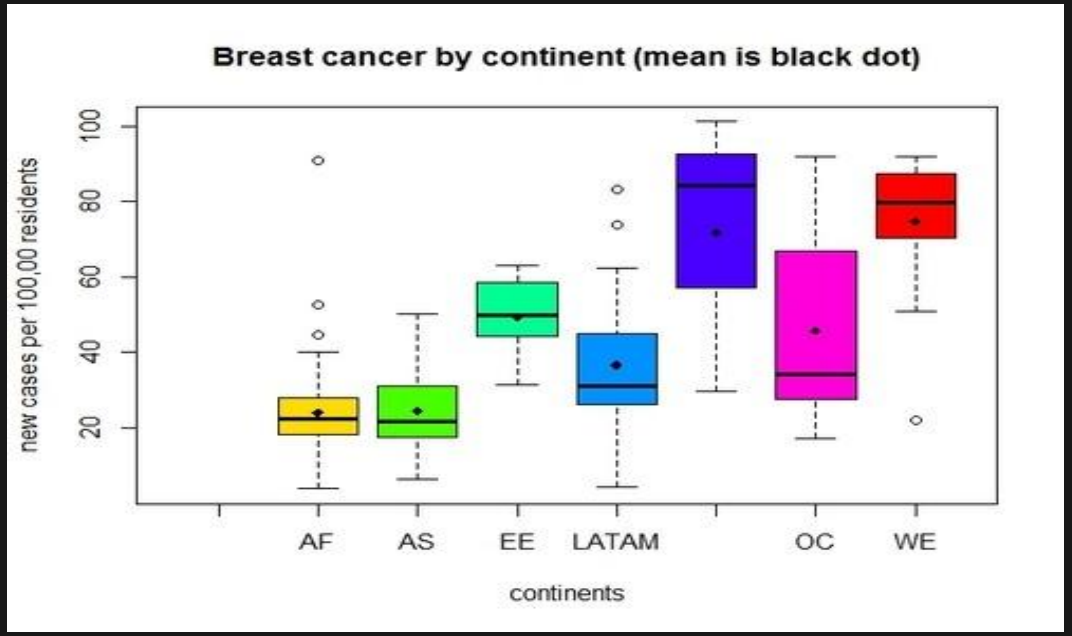
```
library(gplots)
plotmeans(gapCleaned$breastcancer~gapCleaned$continent, digits=2, ccol="red", mean.labels=T,
main="Plot of breast cancer means by continent")
```



The previous graph shows how breast cancer means change between continents, as well as the number of countries taken into account for calculating the mean of each continent. From the plot it is visible that the means differ among continents, with Africa presenting the lowest value and West Europe the highest.

Using boxplot to Confirm

```
boxplot(gapCleaned$breastcancer ~ gapCleaned$continent, main="Breast cancer by continent (mean is black dot)", xlab="continents", ylab="new cases per 100,00 residents", col=rainbow(7))
points(means, col="black", pch=18)
```


Breast cancer by continent (mean is black dot)
new cases per 100,00 residents
20
40
60
80
100
AF
AS
EE
LATAM
OC
WE
continents

(* the blue boxplot with missing label, refers to North America).
The boxplot shows that means are different (some less, others more). But it also shows that each continent present a different amount of variation/spread in breast cancer, so that there is much overlap of values between some continents (e.g. Africa&Asia or North America & West Europe). Hence, differences in means could have come about by chance (and we shouldn't reject the null hypothesis case)

The question we are answering with ANOVA is: are the variations between the continents means due to true differences about the populations means or just due to sampling variability? To answer this question, ANOVA calculates a parameter called F statistics, which compares the variation among sample means (among different continents in our case) to the variation within groups (within continents).

F statistics = Variation among sample means / Variation within groups
Through the F statistics we can see if the variation among sample means dominates over the variation within groups, or not. In the first case we will have strong evidence against the null hypothesis (means are all equals), while in the second case we would have little evidence against the null hypothesis.

```
aov_cont<- aov(gapCleaned$breastcancer ~ gapCleaned$continent)

summary(aov_cont) # here I see results for my ANOVA test
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gapCleaned$continent	6	52531	8755	**40.28**	**<2e-16 *****
Residuals	166	36083	217		

———

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Interpretation of the Result

F value is 40.28, and p-value is very low too. In other words, the variation of breast cancer means among different continents (numerator) is much larger than the variation of breast cancer within each continents, and our p-value is less than 0.05 (as suggested by normal scientific standard). Hence we can conclude that for our confidence interval **we accept the alternative hypothesis H1** that there is a significant relationship between continents and breast cancer.

Interpretation of the Result

From the anova test we can see that NOT ALL THE MEANS ARE EQUAL. However my categorical variable "continents" has more than two levels (actually it has 7), and it might be that it's just one continent that is not equal to the others. ANOVA doesn't tell me which groups (continents) are different from the others. In this sense we will have to see each pair of continents.

Interpretation of the Result

To determine which groups are different from the others **I need to conduct a POST HOC TEST** or a post hoc pair comparison (note we can't perform multiple anova tests one for each pair, as this would increase our error) which is designed to evaluate pair means. There are many post hoc tests available for analysis of variance and in this case I will use the Tukey post hoc test, calling with R the function "TukeyHSD" as follows:

```
tuk<- TukeyHSD(aov_cont)
tuk
        Tukey multiple comparisons of means 95% family-wise confidence level
Fit: aov(formula = gapCleaned$breastcancer ~ gapCleaned$continent)
$`gapCleaned$continent`
```

From the table above (looking at "diff" and "p adj" columns) I can see which continents have significant differences in breast cancer from others. For example I can conclude that:

- **there is no significant difference** in breast cancer new cases between Asia and Africa (ρ =0.99 > 0.05), as well as between West Europe and North America (ρ=0.99) or Oceania and Latin America (ρ=0.72), etc.
- **THERE IS A SIGNIFICANT DIFFERENCE** in breast cancer new cases between East Europe and Africa (ρ= 0.00) as well as between Latin America and Africa (ρ=0.005) or West Europe and Oceania (ρ=0.00)

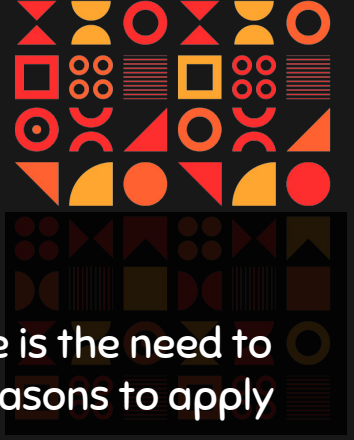
Finally, I can also visualize continent pairs and analyse significant differences by plotting the the "tuk" object in R (sorry the y axis is not displayed properly). Significant differences are the ones which not cross the zero value.

plot(tuk)

Non-Parametric test

Nonparametric tests are methods of statistical analysis that do not require a distribution to meet the required assumptions to be analyzed (especially if the data is not normally distributed). Due to this reason, they are sometimes referred to as distribution-free tests.

Nonparametric tests serve as an alternative to parametric tests such as T-test or ANOVA that can be employed only if the underlying data satisfies certain criteria and assumptions.



Reasons to Use Nonparametric Tests.

In order for us to make a good and correct statistical analysis, there is the need to know what type of non-parametric test is appropriate. The main reasons to apply the nonparametric test include the following:

- **The underlying data do not meet the assumptions about the population sample**

Generally, the application of parametric tests requires various assumptions to be satisfied. For example, the data follows a normal distribution and the population variance is homogeneous. However, some data samples may show skewed distributions.

The skewness makes the parametric tests less powerful because the mean is no longer the best measure of central tendency because it is strongly affected by the extreme values. At the same time, nonparametric tests work well with skewed distributions and distributions that are better represented by the median.

Reasons to Use Nonparametric Tests.

- **The population sample size is too small**

The sample size is an important assumption in selecting the appropriate statistical method. If a sample size is reasonably large, the applicable parametric test can be used. However, if a sample size is too small, it is possible that you may not be able to validate the distribution of the data. Thus, the application of nonparametric tests is the only suitable option.

- **The analyzed data is ordinal or nominal**

Unlike parametric tests that can work only with continuous data, nonparametric tests can be applied to other data types such as ordinal or nominal data. For such types of variables, the nonparametric tests are the only appropriate solution.

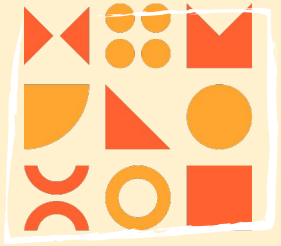
Types of Test

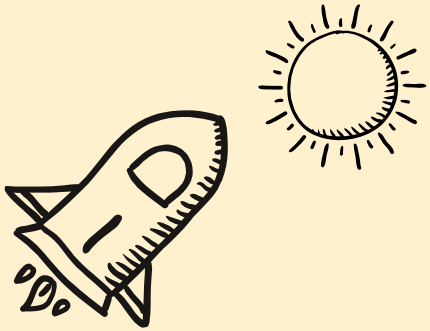
1. **Mann–Whitney U Test**

The Mann–Whitney U Test is a nonparametric version of the independent samples t–test. The test primarily deals with two independent samples that contain ordinal data.

2. **Wilcoxon Signed Rank Test**

The Wilcoxon Signed Rank Test is a nonparametric counterpart of the paired samples t–test. The test compares two dependent samples with ordinal data.





TO TALENT

++

INNOVATION

semicolon