Inferential Statistics

Learning Objectives

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

Contingency Analysis

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

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 chisquare test is an excellent choice to help us better understand and interpret the relationship between our two categorical variables.

Chi-square Goodness of Fit Test

This is a non-parametric test. We typically use it to find how the observed value of a given event is significantly different from the expected value. In this case, we have categorical data for one independent variable, and we want to check whether the distribution of the data is similar or different from that of the expected distribution.

Example

A data scientist is interested in the relationship between the rating by users of an app in google or ios appstore and their frequency of rating. The

Formula for calculating the goodness of fit test.

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

Where:

- χ 2 = Chi-Square value
- Oi = Observed frequency
- Ei = Expected frequency

To calculate the expected value, use

Expected Frequency = (Row Total x Column Total)/Grand Total

Example 1

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 co

llected 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

Example1 continuation

```
#Importing the csv data

data<-read.csv(file.choose())

#Count of Rows and columns

dim(data)

#View top 10 rows of the dataset

head(data,10)
```

Example1 continuation

#Make a table and Calculate the proportion of experience of employees

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

Conclusion of analysis

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).

Exercise1 15mins

Image you are employed as a data scientist in an organization who claim that the experience of the employees of different department is distributed in the following categories

11 – 20 Years = 20%

21 – 40 Years = 17%

6 – 10 Years = 41% and

Up to 5 Years = 22%

A random sample of 1470 employees is collected. You are to

provide evidence against the organization's claim?

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

Anova Continuation

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.

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

It is an apriori statement.

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

H0: μ 1 = μ 2 = μ 3 =....= μl *Null hypothesis*

H1: μ 1 != μ 2 *Alternative hypothesis*

Assumptions of Anova

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

Types of Anova Test

A. One-way anova is use when you want to test two groups to see if there's a difference between them. Example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.
B. Two way anova is use when you have one group and you're double-testing that same group. For example, you're testing one set of individuals before and after they take a medication to see if it works or not.

Between group variability

The distribution of two samples, when they overlap, their means are not significantly different. Hence, the difference between their individual mean and the grand mean is not significantly different. The group and level are different groups in the same independent variable

Sum of Squares Between

To calculate the sum of the square of between the group variability, use

$$BSS = \eta_1 (\overline{X_1} - \bar{X})^2 + \eta_2 (\overline{X_2} - \bar{X})^2 + \eta_3 (\overline{X_3} - \bar{X})^2$$

This sum of squares has a number of degrees of freedom equal to the number of groups minus.

Degree of freedom refers to **the maximum number of logically independent values**, which are values that have the freedom to vary, in the data sample.

To calculate degree of freedom use the formula $df = n - 1$

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

We divide the BSS figure by the number of degrees of freedom to get our estimate of the variation between groups, referred to as "Between Mean Squares"

Sum of Squares Within

. You use the SS to divide by the degree of freedom, where the degree of freedom is the number of sample means(k) minus one

. Within-group variation refers to the variations caused by differences within individual groups or levels. To calculate the sum of squares of within-group variation, use

$$SS_{within} = \Sigma\Sigma(y_{ij} - y_{.j})^2$$

Example1

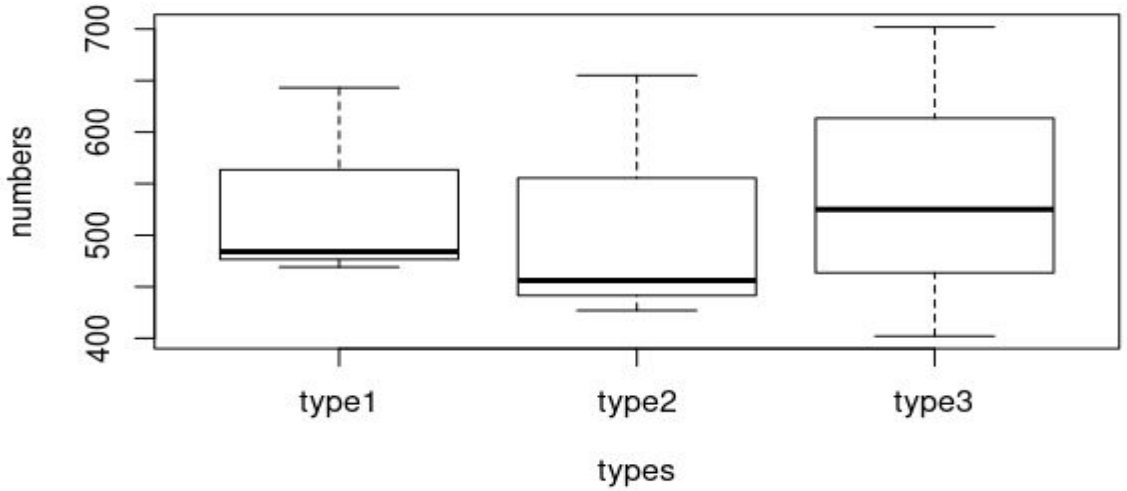
Suppose there are 3 chocolates in town and their sweetness is quantified by some metric (S). Data is collected on the three chocolates. You are given the task to identify whether the mean sweetness of the 3 chocolates are different. The data is given as below:

type1	type2	type3
643	469	484
655	427	456
702	525	402

Statement of Hypothesis

H0: Mean sweetness of the three chocolates are same.

H1: Mean sweetness of at least one of the chocolates is different.



	643	469	484
	655	427	456
	702	525	402
$\overline{X}$	666.67	473.67	447.33
S	31.18	49.17	41.68

```
res.aov <- aov(type1 ~ type2, data = my_data)
```

Example 2

The objective of the ANOVA test is to analyse if there is a (statistically) significant difference in breast cancer, between different continents. In other words, I am interested to see whether new episodes of breast cancer are more likely to take place in some regions rather than others

ANOVA is going to compare means of breast cancer among the seven continents, and **check if differences are statistically significant**. Here are my null and alternative 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:

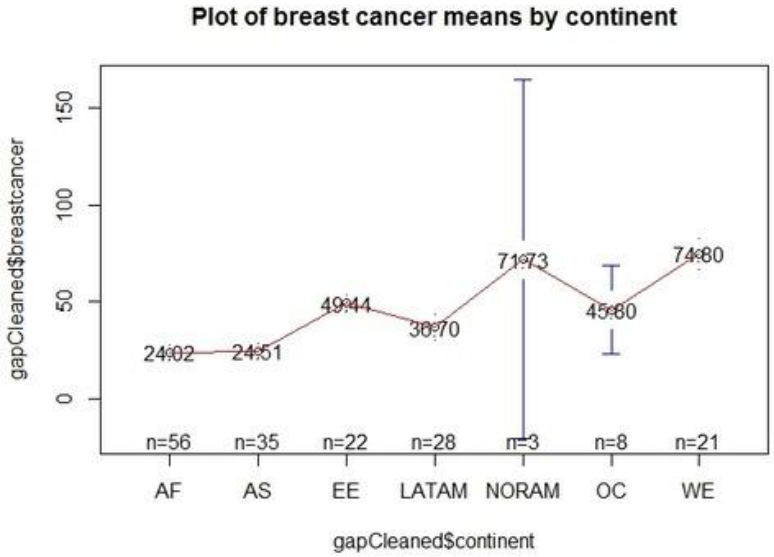
  H1: not all U are equal

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

Anova Cont'n

library(gplots) #I load the "gplots" package to plot means

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

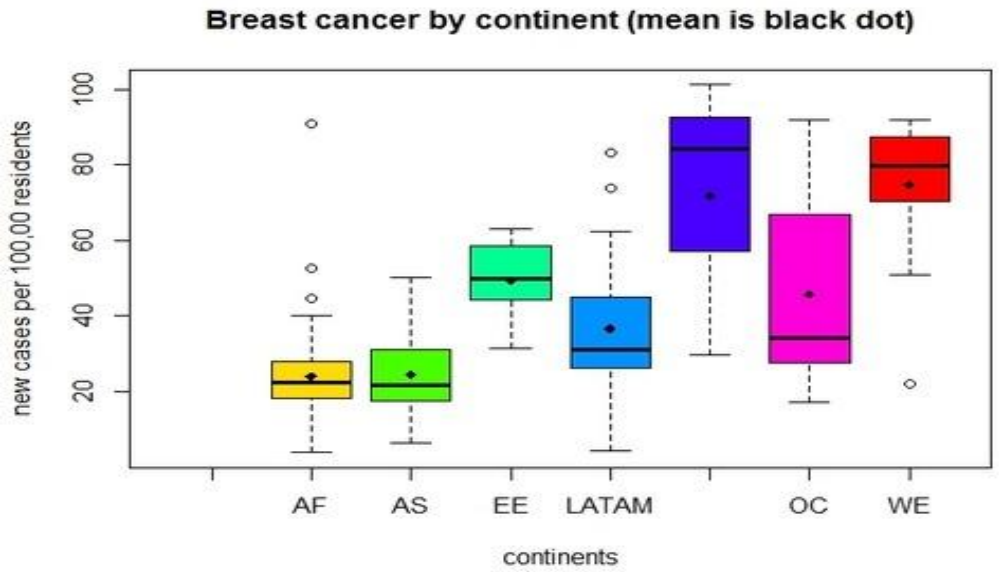


The above graph shows how breast cancer means change between continents, as well as the number of countries taken into account for calculating the mean of each continent. Cool, it looks like means differ among continents, with Africa presenting the lowest value and West Europe the highest.

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)



(* 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

RESULTS & INTERPRETATIONS

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.

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.

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`
```

RESULTS & INTERPRETATIONS

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 (p =0.99 > 0.05), as well as between West Europe and North America (p=0.99) or Oceania and Latin America (p=0.72), etc.
- **THERE IS A SIGNIFICANT DIFFERENCE** in breast cancer new cases between East Europe and Africa (p= 0.00) as well as between Latin America and Africa (p=0.005) or West Europe and Oceania (p=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)

Example2

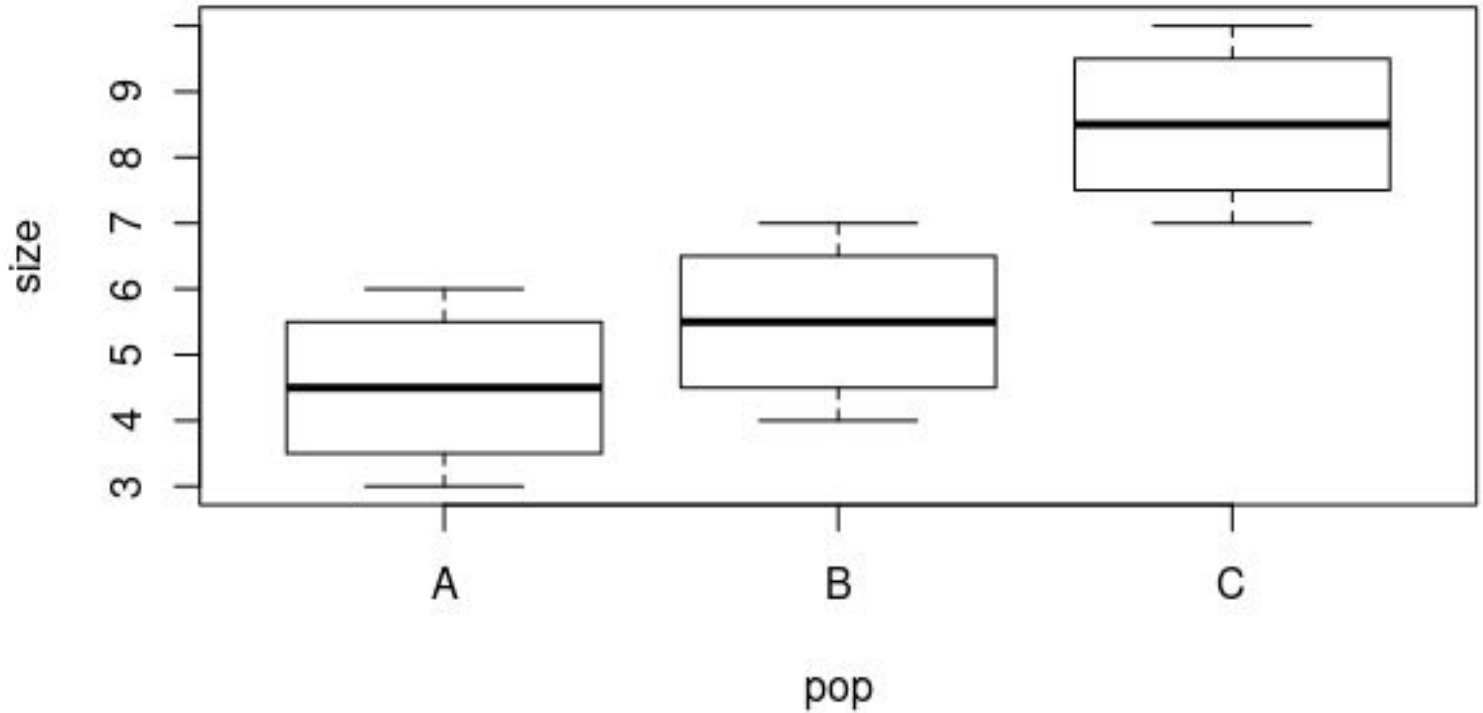
Consider the maximum size of 4 fish each from 3 populations (*n*=12). We want to use a model that will help us examine the question of whether the mean maximum fish size differs among populations.

```
size <-  c(3,4,5,6,4,5,6,7,7,8,9,10)

pop <-  c("A","A","A","A","B","B","B","B","C","C","C","C")

library(ggplot2)

boxplot(size~pop)
```


size
3
4
5
6
7
8
9
A
B
C
pop

Steps for Performing Anova

1. Find the mean for each of the groups.
2. Find the overall mean (the mean of the groups combined).
3. Find the Within Group Variation; the total deviation of each member's score from the Group Mean.
4. Find the Between Group Variation: the deviation of each Group Mean from the Overall Mean.
5. Find the F statistic: the ratio of Between Group Variation to Within Group Variation.