

## **ECO 274 LAB: Inferential Statistics, Chi-square-test, and ANOVA**

### **Learning Objectives**

By the end of this lecture students will be able to learn about the:

- What is the Chi-square test and how the test is related with categorical variables?
- Applications of Chi-Square test in applied economics research
- How to interpret Chi-Square-test results
- What is a Chi-square test used for?
- Chi-square test of independence
- Power analysis for the proportion
- One-way ANOVA in economics

### **Categorical Variables**

Categorical variables belong to a subset of variables that can be divided into discrete categories. Names or labels are the most common categories. These variables are also known as qualitative variables because they depict the variable's quality or characteristics.

Categorical variables can be divided into two categories:

**Nominal Variable:** A nominal variable's categories have no natural ordering. Example: Gender, Blood groups.

**Ordinal Variable:** A variable that allows the categories to be sorted is ordinal variables. Types of countries (developed, developing, emerging, low-income), credit borrower's (high risk, medium risk, low risk), customer satisfaction (Excellent, Very Good, Good, Average, Bad, and so on) are the example.

### **Example**

In her Senior Capstone project, a SUST economics student recorded the high-risk microcredit borrowers in 100 randomly selected credit receivers in two districts, Sunamgang and the Habigang. There was a total of six "credit groups" (one group was called "other") recorded (labeled as A, B, C, D, E, and F). In her project, researcher wants to determine, at the 5% level, if the frequency of high-risk borrowers differs between the two locations.

### **Example**

Let's say we want to know if the weather condition has anything to do with regional housing affordability. To see if the weather condition is linked to regional housing affordability, we perform a Chi-Square test of independence.

## Chi-square test

A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Karl Pearson introduced this test in 1900 for categorical data analysis and distribution. This test is also known as 'Pearson's Chi-Squared Test'. The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. This test can also be used to determine whether it correlates to the categorical variables in our data. It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them.

### Formula For Chi-Square Test

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

Where

c = Degrees of freedom

O = Observed Value

E = Expected Value

The Observed values are those we gather from samples and those are given. The expected values are the frequencies expected or predicted, based on the null hypothesis.

Here are some of the uses of the Chi-Squared test:

- The Chi-squared test can be used to see if your data follows a well-known theoretical probability distribution like the Normal or Poisson distribution.
- The Chi-squared test allows you to assess your trained regression model's goodness of fit on the training, validation, and test data sets.

There are two main types of Chi-Square tests namely -

- Independence
- Goodness-of-Fit

### Chi-square test of independence

The Chi-square test of independence tests whether there is a relationship between two categorical variables. The null and alternative hypotheses are:

H0 : the variables are independent, there is no relationship between the two categorical variables.  
Knowing the value of one variable does not help to predict the value of the other variable  
H1 : the variables are dependent, there is a relationship between the two categorical variables.  
Knowing the value of one variable helps to predict the value of the other variable

The Chi-square test of independence works by comparing the observed frequencies (so the frequencies observed in your sample) to the expected frequencies if there was no relationship between the two categorical variables (so the expected frequencies if the null hypothesis was true).

Example:

We will use iris dataset. Since there is only one categorical variable and the Chi-square test of independence requires two categorical variables, we add the variable size which corresponds to small if the length of the petal is smaller than the median of all flowers, big otherwise:

```
##Example of chi-square test

dat <- iris

dat$size <- ifelse(dat$Sepal.Length < median(dat$Sepal.Length),
                  "small", "big")

##We now create a contingency table of the two variables Species and size
with the table() function:

table(dat$Species, dat$size)

#draw a barplot to visually represent the data

library(ggplot2)

ggplot(dat) +
  aes(x = Species, fill = size) +
  geom_bar()

#visualize in terms of proportions

ggplot(dat) +
  aes(x = Species, fill = size) +
  geom_bar(position = "fill")

#If we prefer to have the bars next to each other:

ggplot(dat) +
  aes(x = Species, fill = size) +
  geom_bar(position = "dodge")
```

```
##we are going to test if there is a relationship between the variables
Species and size

test <- chisq.test(table(dat$Species, dat$size))
test
test$statistic
test$p.value
test$expected

#if the smallest expected frequencies is lower than 5, then use the Fisher's
exact test

# load packages
library(ggstatsplot)
library(ggplot2)

# plot
ggbarstats(
  data = dat,
  x = size,
  y = Species
) +
  labs(caption = NULL) # remove caption
```

we see that the p-value is less than the significance level of 5%. Like any other statistical test, if the p-value is less than the significance level, we can reject the null hypothesis. In this context, rejecting the null hypothesis for the Chi-square test of independence means that there is a significant relationship between the species and the size. Therefore, knowing the value of one variable helps to predict the value of the other variable.

## ANOVA (Analysis of Variance)

ANOVA is a statistical test to determine whether two or more population means are different. In other words, it is used to compare two or more groups to see if they are significantly different. In practice, however, the:

- Student t-test is used to compare 2 groups.
- ANOVA generalizes the t-test beyond 2 groups, so it is used to compare 3 or more groups.

Note that there are several versions of the ANOVA (e.g., one-way ANOVA, two-way ANOVA, mixed ANOVA, repeated measures ANOVA, etc.).

### One-way ANOVA

Data for this tutorial is the penguins dataset. The dataset contains data for 344 penguins of 3 different species (Adelie, Chinstrap and Gentoo). The dataset contains 8 variables, but we focus only on the flipper length and the species.

```
# load packages
#install.packages("palmerpenguins")
library(palmerpenguins)

library(tidyverse)

dat <- penguins %>%
  select(species, flipper_length_mm)
summary(dat)

library(ggplot2)

ggplot(dat) +
  aes(x = species, y = flipper_length_mm, color = species) +
  geom_jitter() +
  theme(legend.position = "none")
```

Flipper length varies from 172 to 231 mm, with a mean of 200.9 mm. There are respectively 152, 68 and 124 penguins of the species Adelie, Chinstrap and Gentoo. Here, the factor is the species variable which contains 3 modalities or groups (Adelie, Chinstrap and Gentoo). In this context and as an example, we are going to use an ANOVA to help us answer the question: “Is the length of the flippers different between the 3 species of penguins?”

The null and alternative hypothesis of an ANOVA are:

- $H_0: \mu_{\text{Adelie}} = \mu_{\text{Chinstrap}} = \mu_{\text{Gentoo}}$  ( $\mu_{\text{Adelie}} = \mu_{\text{Chinstrap}} = \mu_{\text{Gentoo}}$  (the 3 species are equal in terms of flipper length))
- $H_1$ : *at least* one mean is different (at least one species is different from the other 2 species in terms of flipper length)

## Underlying assumptions of one-way ANOVA

Below are the assumptions of the ANOVA, how to test them and which other tests exist if an assumption is not met:

- ANOVA requires a mix of one continuous quantitative dependent variable (which corresponds to the measurements to which the question relates) and one qualitative independent variable (with at least 2 levels which will determine the groups to compare).
- **Independence:** the data, collected from a representative and randomly selected portion of the total population, should be independent between groups and within each group.
- In case of small samples, residuals should follow approximately a normal distribution.
- Equality of variances: the variances of the different groups should be equal in the populations (an assumption called homogeneity of the variances, or even sometimes referred as homoscedasticity, as opposed to heteroscedasticity if variances are different across groups).
- There should be no significant outliers in the different groups, or the conclusions of your ANOVA may be flawed.
- If normality is not assumed, use the Kruskal-Wallis test.
- If variances are not equal, use the Welch ANOVA.

Before checking the normality assumption, we first need to compute the ANOVA

```
res_aov <- aov(flipper_length_mm ~ species,
              data = dat)

par(mfrow = c(1, 2)) # combine plots

# histogram
hist(res_aov$residuals)

# QQ-plot
library(car)
qqPlot(res_aov$residuals,
       id = FALSE) # id = FALSE to remove point identification

H0: data come from a normal distribution
H1: data do not come from a normal distribution

shapiro.test(res_aov$residuals)

##Equality of variance test

boxplot(flipper_length_mm ~ species,
       data = dat)

# Dotplot
library("lattice")

dotplot(flipper_length_mm ~ species,
       data = dat)
```

- $H_0$ : variances are equal
- $H_1$ : at least one variance is different

```
#the Levene's test can be performed thanks to the leveneTest() function from
the {car} package:

library(car)

leveneTest(flipper_length_mm ~ species,
            data = dat)

##Check outliers

library(ggplot2)

ggplot(dat) +
  aes(x = species, y = flipper_length_mm) +
  geom_boxplot()
```

From the histogram and QQ-plot above, we can already see that the normality assumption seems to be met. P-value of the Shapiro-Wilk test on the residuals is larger than the usual significance level of  $\alpha=5\%$ , so we do not reject the hypothesis that residuals follow a normal distribution (p-value = 0.261). Both the boxplot and the dotplot show a similar variance for the different species.

```
#ANOVA in R

# 1st method:
oneway.test(flipper_length_mm ~ species,
            data = dat,
            var.equal = TRUE) # assuming equal variances

# 2nd method:
res_aov <- aov(flipper_length_mm ~ species,
               data = dat)

summary(res_aov)
```

As you can see from the two outputs above, the test statistic ( $F$  = in the first method and  $F$  value in the second one) and the p-value (p-value in the first method and  $\Pr(>F)$  in the second one) are exactly the same for both methods, which means that in case of equal variances, results and conclusions will be unchanged.

The advantage of the first method is that it is easy to switch from the ANOVA (used when variances are equal) to the Welch ANOVA (used when variances are unequal). This can be done by replacing `var.equal = TRUE` by `var.equal = FALSE`, as presented below:

```
oneway.test(flipper_length_mm ~ species,
  data = dat,
  var.equal = FALSE # assuming unequal variances
)
```

### Interpretations of ANOVA results

Given that the p-value is smaller than 0.05, we reject the null hypothesis, so we reject the hypothesis that all means are equal. Therefore, we can conclude that at least one species is different than the others in terms of flippers length (p-value < 2.2e-16). If the null hypothesis is not rejected (p-value  $\geq$  0.05), it means that we do not reject the hypothesis that all groups are equal. The ANOVA more or less stops here. Other types of analyses can be performed of course, but—given the data at hand—we could not prove that at least one group was different so we usually do not go further with the ANOVA. On the contrary, if the null hypothesis is rejected (as it is our case since the p-value < 0.05), we proved that at least one group is different. We can decide to stop here if we are only interested to test whether all species are equal in terms of flippers length.

### Sources acknowledged:

- <https://statstutorial.com>
- <https://statsandr.com/blog/>
- <https://machinelearningmastery.com>
- Statistical Analysis with R For Dummies by Joseph Schmuller.
- The Book of R: A First Course in Programming and Statistics by Tilman M. Davies.
- <https://www.simplilearn.com>
- <http://derekogle.com/Book107>