## 16. ANOVA

Welcome to Trinity Term! We’ll have three weeks of classes this term, covering Analysis of Variance, Logistic Regression, and Research Design. In the third and last tutorial, you will do a practice exam question to help prepare you for the exam.

In week 1 this term, we are learning about Analysis of Variance, known as ANOVA. This method of statistical analysis has some overlap with Regression, and we will look at the similarities and differences of the two approaches.

### 16.1 Tasks for this week

Conceptual material is covered in the lecture. In addition to the live lecture, you can find the lecture recording and additional materials on Canvas.

Please work through the guided exercises in this section (everything except the page labelled “Tutorial Exercises”) in advance of the computer-based tutorial session.

### 16.2. Learning Objectives

#### CONCEPTUAL

After this week, you should be able to:

* Understand how and when to apply ANOVA.
* Understand the key concepts including mean squares and the F-statistic.
* Specify hypotheses and interpret the outcome of the test.
* Understand how ANOVA relates to regression.
* Understand when to use, and how to interpret, the Kruskal-Wallis test.

These points will be covered in the lecture.

#### PYTHON

Again, working with the “statsmodels” package in Python. This week, you will learn to:

* Run one-way and two-way ANOVAs and interpret the output.
* Compare the ANOVA table to the regression table.
* Run a Kruskal-Wallis test and interpret the output.

### 16.3 ANOVA Concepts

Analysis of Variance (ANOVA) is a test of independence where the outcome variable is continuous, and the explanatory variable is categorical. It is a way of comparing means across groups and is preferred where there are more than two groups. If we ran an experiment with three groups – say treatment1, treatment2, and control group - ANOVA enables the researcher to test the means of these three groups simultaneously.

Q: If there were only two groups, which test would you use to compare means?

A: The T-test. We use ANOVA when there are *more* *than two* groups.

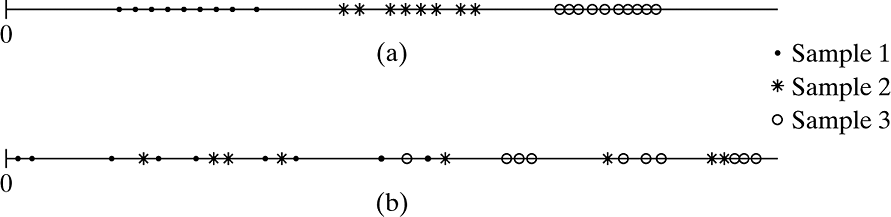
-----------------------------------

Q: Why is a comparison of means called Analysis of *Variance*?

A: ANOVA uses the variance to compare means. It compares the variability between the overall mean and the group means (where subscript means ‘in each group’) with the variability within each group 1 2 3 etc. (where subscript means ‘in group 1’ and so on).

--------------------------------------

In the below example, the means of the three groups in a) are clearly different to each other. In b) there is more overlap between the data points, and it is harder to tell just by eyeballing the data whether there is a difference in means in the three groups. In this example, the means of the three groups, and the overall mean, are the same in (a) and (b) which means that the *variability between groups* is the same (a) and (b). However, the *variability within* groups is lower in (a) than in (b). Low variability within groups suggests that the groups have means that are statistically different to each other.



(Source: Agresti Fig 12.1)

ANOVA uses the F-statistic to test whether there are significant differences in means across groups. The F-statistic is a ratio of the between-groups variance divided by the within-group variance as follows:

Q: What does it mean when the F-statistic is large?

A: The F test statistic is large if variability between groups is large relative to variability within groups. Looking back to the example above, where the within-group variability is lower in a) than in b), we would expect a larger F value in a) than in b). Larger F values have smaller p-values and indicate that there is a difference between groups. The cut-off value for F (i.e., what counts as ‘large’) depends on the degrees of freedom (sample size and number of groups). When you run ANOVA in a stats package such as Python, it will produce the p-value for you.

-----------------------------

Q: What are the assumptions of ANOVA?

A: The three assumptions are:

* For each group, the population distribution of the response variable is normal.
* The standard deviation of the population distribution is the same for each group. Denote the common value by .
* The samples from the populations are independent random samples.

-----------------------------

Q: What is the non-parametric equivalent of ANOVA?

A: The Kruskal-Wallis test. The Kruskal-Wallis test ranks the observations and compares mean ranks of the groups, thus uses only ordinal information in the data.

------------------------------------------

### 16.4 Longhand calculation example

In this worked example based on a small sample size, we can easily see how the key statistics for ANOVA are derived.

**Example**: A consumer protection group compares three types of front bumpers for a brand of automobile. A test is conducted by driving an automobile into a concrete wall at 15 miles per hour. The response variable is the amount of damage to the car, as measured by the repair costs, in hundreds of pounds. Due to the potentially large costs, the study conducts only three tests with each bumper type. The table below shows the raw data collected:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Bumper A** | **Bumper B** | **Bumper C** |
|  | 1 | 2 | 11 |
|  | 3 | 4 | 15 |
|  | 2 | 6 | 10 |
| **Mean** | **2** | **4** | **12** |

The grand mean across the three groups = **6**

Q: Specify the null and alternative hypotheses for an ANOVA test for differences in means.

A: (the group means are equal)

at least two of the population means are unequal.

Note: If our test of the null hypothesis is rejected, we can conclude that **not all the means are equal**: that is, at least one mean is different from the other means. The ANOVA test itself provides only statistical evidence of ***a*** difference, but not any statistical evidence as to ***which*** mean or means are statistically different.

--------------------------------

Q: How do we find the **Within-group** sum of squares?

A: The steps to compute Within-group sum of squares are as follows:

1. Compute each of the group means (these are provided for us in the table already!)
2. For each observation, compute the difference between the score and its group mean.
3. Square all these differences.
4. Sum the squared differences.

Within sum of squares =

*The colour coding shows where the figures are coming from!*

--------------------------------------

Q: How do we find the between sum of squares?

A: The steps to compute the Between-group sum of squares:

1. *For each car*, compute the difference between its group mean and the grand mean. The grand mean is the mean of all cars (here provided for us, grand mean = 6)
2. Square all these differences.
3. Sum the squared differences.

Between sum of squares =

-----------------------------------------

Q: Find the mean square within and the mean square between by dividing by degrees of freedom. How do we find the degrees of freedom (df)?

A: The within df = n – g (where g = the number of groups) = 9 – 3 = 6

The between df = g – 1 = 3 - 1 = 2

The mean square within = 24/6 = 4

The mean square between = 168/2 = 84

----------------------------------------

Q: Calculate the F-statistic and interpret.

A: = 84/4 = 42

The critical cut-off value for F here (df 6,2), for a significance level of 95% = 5.14. As 42 > 5.14 we can reject the null hypothesis. We conclude that there is a significant difference between the three bumper types. (You can take the cut-off value for F as given in this example. But one place you could look up cut-off of F for significance level 95% is [here](https://www.socscistatistics.com/tests/criticalvalues/default.aspx).)

--------------------------------------------

### 16.5 ANOVA equations

It would also be possible to calculate ANOVA without the raw data, if you knew the sample size, means, and standard deviations for each group and overall. Let’s take this example, where we have a set of summary statistics.

**Example:** In a study from 2014, researchers tested ‘contagious yawning’ among chimpanzees, i.e., whether they would copy yawns observed in others, as a measure of empathy. Chimps were shown videos of yawning among humans (who were familiar or unfamiliar to the chimps), baboons, and a control video of a human who was not yawning. The number of yawns per 10-minute session was recorded as shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Familiar humans | Unfamiliar humans | Baboons | Control | Total |
| Mean yawns | 3.40 | 3.75 | 1.20 | 0.80 | 2.21 |
| Std dev | 1.14 | 0.96 | 0.84 | 0.84 | 1.58 |
| N | 5 | 4 | 5 | 5 | 19 |

The equation for calculating Within sum of squares is:

And for calculating Total Sum of Squares:

Where Total Sum of Squares = Within Sum of Squares + Between Sum of Squares.

You should get the SS values in the ANOVA table below. Use the df to generate Mean Squares, and then find F: and .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **SS** | **df** | **MS** | **F** |
| Between groups | 31.61 | 3 | 10.54 | 11.66 |
| Within groups | 13.55 | 15 | 0.90 |  |
| Total | 45.16 | 18 | 2.51 |  |

The critical cut-off for F (3,15) at 95% significance = 3.29. (Again, you can take the cut-off as a given here, and remember that python will generate p-values for you).

As 11.66 is higher than 3.29, we conclude that at least two of the groups are different.

### 17.6 Kruskal-Wallis test longhand example

When the sample size is small, or when the data are not normally distributed, it is preferable to use the Kruskal-Wallis test which is the non-parametric equivalent of ANOVA. The equation for the K-W test is:

Where the test statistic, H, has a chi-squared distribution. In this equation, references the number of groups, indicates the number of observations per group, and is the sum of the ranks of observations in the group.

**Example:** A research paper published in the *Journal of Insect Conservation* in 2019 examined the attractiveness of different wasteland sites to bees. The table below shows a record of the richness of bee species in 10 wasteland sites, classified by the former land use into extractive industry, suburban/ residential and chemical industry. Test whether species richness differs by type of site using the Kruskal-Wallis test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Bee species richness** | | | |
|  | **Ob1** | **Ob2** | **Ob3** | **Ob4** |
| **Extractive industry** | 95 | 117 | 105 | 71 |
| **Suburban/ residential** | 55 | 73 | 67 |  |
| **Chemical industry** | 33 | 27 | 40 |  |

*Where ob = observation*

Q: State your hypotheses.

A: H0: the group rank means are equal, i.e., there is no association between site type and bee species richness.

H1: at least two of the rank means are unequal, i.e., there is an association between site type and bee species richness.

-----------------------------

Q: What are the rank sums for each group?

A:

|  |  |  |
| --- | --- | --- |
|  | n | Rank Sum |
| Chemical industry | 3 | 6 |
| Extractive industry | 4 | 33 |
| Suburban/ residential | 3 | 16 |

------------------------------

Q: Practice your ability to work with equations. Plug the values in and calculate H.

A:

H = \* ( + + )-3\*(11) = **7.318**

------------------------------------

With 2 df, the critical value of H = 5.99 (Look up [here again](https://www.socscistatistics.com/tests/criticalvalues/default.aspx), or refer to Jill’s lecture). As 7.318 > 5.99, we reject the null hypothesis and conclude that there is a difference between groups.

### 16.7 ANOVA and Kruskal-Wallis in Python

We’ll use the data from the chimps examples to demonstrate the python code. First, as usual, we’ll import the relevant packages and import the data as a csv file.

# Set-up Python libraries - you need to run this but you don't need to change it

import numpy as np

import matplotlib.pyplot as plt

import scipy.stats as stats

import pandas

import seaborn as sns

import statsmodels.api as sm

import statsmodels.formula.api as smf

Code for importing chimp.csv data.

We ask python for ANOVA with the following code:

Python code for ANOVA: y = yawns, x = group

The python output confirms our longhand example above. The p-value is 0.0003. There is a statistically significant difference in yawn rates between the groups. Add output table.

We can add a control variable in a “two-way ANOVA”. Here we want to control for age which is a categorical variable of the chimp’s age: young, middle-aged, and old.

Python code for ANOVA: y = yawns, x = group and age

The results show that the experimental treatment group is statistically significant (p=0.0002) but age is (just) not statistically significant (p=0.0533). Add output table.

ANOVA can also handle interaction terms (as we explored with Regression Analysis last term).

Python code for ANOVA: y = yawns, x = group and age and group\*age interaction.

The interaction is not statistically significant (p=0.4961), which we can interpret to mean that the effect of the treatment group was the same for chimps of different ages. Add output table.

Python will also run a Kruskal-Wallis test for us, with the following code:

Python code for Kruskal-Wallis: y = yawns, x = group

The Kruskal-Wallis test produces an H-statistic of 12.894 and a p-value of 0.0049. It therefore gives the same result as the one-way ANOVA, suggesting a statistically significant difference between treatment groups in the chimp experiment. Add python output.

### 16.8 ANOVA versus Regression

The similarities between ANOVA and regression analysis are:

* They both partition the variance using sum of squares i.e., they both depend on variance in their estimates, although in different ways:

In ANOVA the SS ratio is

In regression the SS ratio is =

* They can both include more than one -variable, and test for interactions.
* And, they could both be used to answer the same research question about differences between groups, but with differences in their approach…

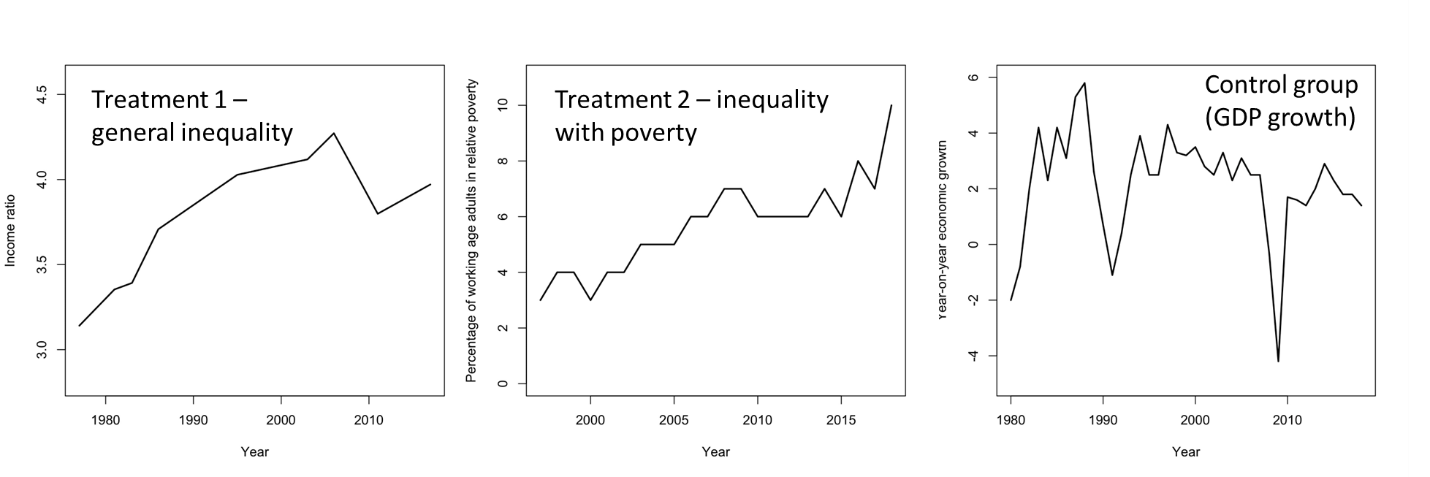
The differences between ANOVA and regression analysis are:

* ANOVA and Regression differ in their aims. ANOVA is a way of determining whether groups are different to each other, with a YES or NO answer. E.g., “Is personality trait the same or different among different age groups?” “Were the outcomes the same across treatment groups and control?” Regression, on the other hand, is a statistical model that approximates the relationship between and , telling us about the strength of the relationship.
* Traditionally, ANOVA was often used to analyse designed experiments while regression was used for inferential statistics.
* ANOVA is used for categorical variables and continuous variables, while linear regression is applied to continuous variables and variables of all types (binary, categorical, continuous). NB. Remember that regression would use ‘dummy’ variables to analyse categorical variables (taking values 0 and 1).

### 16.9 Tutorial exercises

The data this week were collected in 2019 for a project about the consequences of economic inequality (these data are real, from the same project as the data you used in Week 7 last term, but edited for this exercise). The data were collected online and consisted of two parts, a short survey, and an online experiment. The online experiment divided the participants randomly into one of three groups with the following labels: 1) general inequality, and 2) inequality with poverty, and 3) the control group. The participants were shown a graph showing the general trend in inequality (group 1) and a graph about increasing numbers in poverty (group 2), or a graph about GDP (the control group). These data were collected online by the polling company YouGov. The sample is intended to be representative of the UK adult population.

These were the ‘primes’ that were shown to the participants.



The data file is inequality.csv, and contains the following variables:

* ladder (a continuous measure of 1-11 where participants rate themselves in their standing in society, where the lowest rung on the ladder was labelled “bottom of society” and the top rung as “top of society”)
* treatment (a categorical measure of the treatment/control group)
* income (a categorical variable with four categories)
* age (a continuous measure in years)

Students’ own code for installing packages.

Students’ own Code for importing inequality data.

Before running an ANOVA, get to know your data. How many data points are there? Check how many values are in each experimental condition. Is this as expected?

The main research question for this exercise is this: **does the treatment (the experimental condition on inequality) influence perceptions of social standing?** Answer this question with ANOVA. Before you start, write down the null and alternative hypotheses.

Students own code for ANOVA.

What is the p-value and F-statistic? What does this tell us about the answer to the research question? Which hypothesis do you accept?

Next, run a two-way ANOVA including income as the control variable.

Students own code.

Interpret the results. Which of the -variables are statistically significant?

Now, compare the results of the ANOVA tests with linear regressions. Run two regression models. One with just treatment as the explanatory variable, and a second model that includes income as a control. What information do we get from regression that we do not get from ANOVA?

Students own code for linear regression models.

A member of the research team raised a concern that the outcome variable, ladder, does not seem to be normally distributed. Can you run a histogram, by experimental condition, to check this?

Students own code for running a histogram, by treatment

Do you think the normality assumption has been met?

Either way, run a Krusal-Wallis test with the same variables (ladder, treatment) to check if you get the same result as with ANOVA.

Students own code.

What is the result, and does it agree with the ANOVA?

Age is measured here as a continuous variable. How could we use ANOVA to test whether there are differences in ‘ladder’ by age? Try generating a new variable with 4 age groups. Refer to the python code in section xx for generating a new variable.

Students own code for generating a categorical variable for age.

Students own code for ANOVA with = age group.

**Extra Exercise**: Run two-way ANOVAs with interaction terms for a) treatment and age b) income and age. What are the questions you are trying to answer with the interaction terms? (Hint: “Is the effect of treatment the same for….”?)

Students own code.