

Unit 2: Analysis of Variance (ANOVA)

2.1 Basics of ANOVA

In this lesson we'll learn about the basic concepts about the Analysis of Variance model, in short, ANOVA, covering its simplest version, the one-way ANOVA. I will also illustrate this model with two examples.

Slide 3:

The data in an Analysis of Variance or ANOVA consist of **multiple samples** of data for a **response variable** of interest, **differentiated in groups or sub-populations** described by a **category or label**.

The simple model is so called one-way ANOVA. In one-way ANOVA we have **k different populations or groups**, and for each population we observe a sample of data for the response variable Y . We assume that the true mean and variance for the response variable are μ_1 and σ_1^2 squared for the first population, μ_2 and σ_2^2 squared for the second population, μ_k and σ_k^2 square for k th population. Based on the data sample of the response variable for each population, we can **obtain estimates for the mean and variance parameters for each population**. The mean estimate is the **sample mean denoted as \bar{Y}** , and the variance estimate is the **sample variance denoted as S^2** .

Population 1: $(\mu_1, \sigma_1^2) \rightarrow$ Sample 1: $(Y_{1,1}, \dots, Y_{1,n_1}) \rightarrow (\bar{Y}_1, s_1^2)$

Population 2: $(\mu_2, \sigma_2^2) \rightarrow$ Sample 2: $(Y_{2,1}, \dots, Y_{2,n_2}) \rightarrow (\bar{Y}_2, s_2^2)$

.....

Population k: $(\mu_k, \sigma_k^2) \rightarrow$ Sample k: $(Y_{k,1}, \dots, Y_{k,n_k}) \rightarrow (\bar{Y}_k, s_k^2)$

The overarching objective in the ANOVA is to compare the means across the k populations. Are the means equal? Which pairs of the means are different?

Slide 4:

Suicide is a worldwide phenomena and occurs across all population groups. Gaining an understanding of the underlying causes allows policy makers to develop strategies that might reduce or minimize suicide in populations that are negatively impacted. In the first illustrative example of one-way ANOVA, the population of interest consists of populations across multiple countries, for which we observe the rate of suicide for the population within each country. The countries included in this analysis are grouped by region. One can examine how suicide rates vary by region across the world.

Slide 5:

One approach to evaluate or study the response variable with respect to a categorical variable, in this case, the country level suicide rate, is to use the side by side boxplot. That is, for each category we plot a single boxplot of the response variable and we compare the boxplots across all categories, in this example, across all regions. In this case, because we have four categories, we have four boxplots.

In this example, we may be interested to address the following question, is there a difference, in the suicide rate by region of the country? Which regions have lower suicide rates? To answer these questions, we compare whether the mean suicide rates across all regions are different. Particularly in ANOVA, we compare the within-variability to the between variability of the response data. The within variability is the variability within regions, visually can be assessed that by the variability within each box. The between variability is the variability between the means across the groups, proxied by the middle lines in the boxplots. *We will find significant differences across the means if the between-variability is larger than the within-variability.*

Slide 6:

In the second example we're interested in the typing speed of three computer keyboard layouts by determining whether there are differences in their respective mean speeds. Is there a difference in the time taken to perform a test across the three layouts? Which layout is more effective?

Slide 7:

If we compare the three layouts using the side-by-side boxplots, we can see that the second layout has a much higher mean speed as compared to layouts one and three.

But this comparison based on the boxplots is only a visual assessment and it doesn't provide as statistical statements whether those differences are statistically significant?

Slide 8:

In ANOVA, the primary objectives are to:

1. **Analyze the variability in the data using the ANOVA table.** That means we compare the variability within each group to the variability between the means. We represent all the information in a table, laying out all the components needed to make the comparison.
2. **Use this analysis of variance** in order to test whether the means are equal. Specifically, we will test the null hypothesis that all means are equal versus the alternative that at least two of the means are not equal.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

3. **Estimate confidence intervals** for all the pairs of means, in order to identify which of the means are not equal, or which of the means are statistically significantly different. Specifically, we will consider a hypothesis test for each pair of means with the null hypothesis that the means in the pair are equal versus the alternative that are not equal. We will perform all the hypothesis tests across all pair jointly.

$$\mu_i - \mu_j \text{ for } i \text{ and } j = 1, \dots, k$$

2.2. Estimation Method

The topic of this lesson is parameter estimation in ANOVA. We'll learn about the estimation of both the means and the variance, which is assumed constant across the populations.

Slide 3:

As provided in the previous lesson of ANOVA, the data in the ANOVA model consist of a response variable of interest observed for multiple populations differentiated by a categorical variable or a label. For example, for the analysis of suicide rates, the categorization could be instead age group, year, weather type etc.

Let's begin with the model. In our notation, Y_{ij} are the response data differentiated across the k categories. j is the index within group and i is the index across groups. The model is: $Y_{ij} = \mu_i$ (mean of the group i) + ϵ_{ij} (the error term ϵ_{ij}).

Data: Y_{ij} for $j = 1, \dots, n_i; i = 1, \dots, k$

Model: $Y_{ij} = \mu_i + \epsilon_{ij}$ where ϵ_{ij} = error term

The assumptions on ANOVA are with respect to the error term:

- **Constant variance assumption.** Recall that we assume the response data has constant variance across all groups, and thus the variance of the error terms is constant, equal to sigma square: $\text{Var}(\epsilon_{ij}) = \sigma^2$
- **Independence assumption.** Response data and error terms (" ϵ ") are independent.
- **Normality assumption.** Error terms are normal, and thus the y_{ij} response data are normal as well.

Slide 4:

In ANOVA, we assume that the variance of the response variable is the same across all populations and equal to sigma square. Thus, we compare the means, assuming the variances are the same, and estimate the variance across all samples using the so-called **pooled variance estimator**.

Pooled Variance Estimator:

$$s_{\text{pool}}^2 = \frac{\sum_{i=1}^k (n_i - 1) s_i^2}{\sum_{i=1}^k (n_i - 1)} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2}{N - k}$$

In the formula for the pooled variance, s_i^2 are the sample variances of the data samples of the response variable. I will note again that a data sample corresponds to the response data for one group or population in the ANOVA. Thus if we have k groups, we will have k data samples. For each of the k data samples, we thus can compute the sample variance as shown in this formula. In addition, the n 's -- n_1, n_2, n_3, n_k , and so on -- are the sample sizes of the data samples. By adding up the weighted sample variances, we get the sum of the pool variance estimator. The big N is the total number of samples or the sum $n_1 + n_2 + n_3 + \dots + n_k$.

The degree of freedom in the estimation in the pooled variance estimator is $N - k$. When estimating the sample variance of one data sample for one group, we replace the true mean of that group, which is unknown, with its sample (estimated) mean. Because we replace k different means with their sample means, we lose k degrees of freedom hence we subtract k from the total sample size N . I'll come back to this aspect when we're going to discuss the sampling distribution for the pooled variance estimator.

Slide 5:

The formula for the pooled variance estimator is also called the **mean square error in ANOVA or abbreviated MSE**. The MSE is the mean of the sum of squared errors or SSE, where SSE is the sum of squares of the responses minus the sample means across the k groups or data samples. If we divide the sum of square of error by $N - k$ we end up with a mean square error.

Just like in simple regression analysis, we do not have the error terms in the ANOVA model and hence we replace the errors with the residuals. Hence, we will use errors and residuals interchangeably; in some contexts, sum of squared errors may be referred to as sum of squared residuals.

Slide 6:

The individual sample variance for each of the k samples has a chi-square distribution because we assume that the data are normally distributed. An important property of

the chi-square distribution is that, **if we have independent chi-squared random variables, their sum is also a chi-square distribution**. Let's take the first term in the sum. If we multiply the sample variance of the first sample by $n_1 - 1$ and divide by sigma squared, the result is a chi-square distribution where the number of degrees of freedom is $n_1 - 1$. Now if we add up all k such components, each with a chi-square distribution, the resulting distribution is a chi-square distr with the number of degrees of freedom being the sum across all degrees of freedom, in this case, $N-k$.

$$\frac{SSE}{\sigma^2} = \frac{(n_1-1)s_1^2}{\sigma^2} + \dots + \frac{(n_k-1)s_k^2}{\sigma^2} \sim \chi_v^2 \text{ where } v=N-k$$

In a nutshell, the sampling distribution of the pooled variance is a chi-square distribution with $N - k$ degrees of freedom.

Slide 7:

The means of the k sub-populations are the parameters of interest in ANOVA. To obtain estimates of these parameters, the sample mean of the individual samples are used to estimate the mean parameters. But what is the sampling distribution of those estimated means?

$$\hat{\mu}_i = \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

Remember that we **assume data are normally distributed for each sample** -- that is, they have a normal distribution with mean μ_i and sigma square. Since we estimate μ_i with μ_i hat which is the **average across the responses assumed normally distributed**, then the **sampling distribution of the sample mean is also normal within mean μ_i and variance sigma squared divided by the sample size, n_i for the i -th group**.

However, **we do not know sigma squared**, so we **replace sigma squared with the pooled variance estimator**, the **mean squared error**. With this, the **sampling distribution changes to be a t-distribution**.

$$\frac{\hat{\mu}_i - \mu_i}{\sqrt{\frac{MSE}{n_i}}} \sim t_{N-k}$$

We have a t distribution with N-k degrees of freedom because the sampling distribution of the estimated variance is a chi-square distribution with N – k degrees of freedom.

Slide 8:

Given the estimators for the mean and variance parameters, we now can get the confidence intervals for each individual mean. Similarly to the confidence intervals for sample means, we center the confidence intervals at the sample means, plus or minus the t-critical point with N-k degrees of freedom, multiplied by the standard error of the estimator μ_i hat which is square root of the mean square error divided by n_i . Again, the mean square error replaces sigma square. The mean square error is the estimator for the variance.

2.3. Estimation Data Examples

In this lesson, I'll illustrate parameter estimation using an R example. Particularly, I will show you with R, how to obtain the estimated means and how to interpret them.

Slide 3:

Let's go back to the example where we are interested in differences in suicide rates among regions of the countries for which we have data for. The question that we wanted to address is:

What are the estimates for the mean suicide rates for the different regions of countries?

Slide 4:

To fit an ANOVA model using the R statistical software, we can use the **R command `aov()`**. In this command, we need to input the response variable, the suicide rates for all the countries included in the study, all the samples of responses stacked in one vector. We also need to input the vector of labels, in this case, region, telling R which of the labels correspond to what responses. To get the estimated means, we can use the `model.table()` command in R. But you need to specify what kind of summary you want to obtain from the parameters, means or medians, for example. In this case we're interested in the estimated means. This table shows the results only from the output, specifically, the overall mean across all response data as well as the estimated means by region along with the sample sizes. It is important to note that the sample sizes are different across the regions; this means that we have an unbalanced ANOVA. Specifically, one region has only one observation. While we can still run the ANOVA for this example, I would point out that the sample sizes shall be larger than in this example to have reliable pairwise comparison. We will get back to this in a different lesson. Overall, there are differences in the suicide rates across the regions. The question will address in the next lesson is, are those differences statistically significantly different?

Slide 5:

In the second example, we're interested in the typing speed across three different keyboard layouts.

We want first to estimate the mean typing speeds across the three keyboard layouts and compare.

Slide 6:

Similarly as in the previous example, we can use the `aov` command, along with a `model.tables()` command provides us the estimating means. Below is the output.

For layout 1, the sample mean is 25.12 seconds. For the second layout it's 29.11 seconds. For the third layout it's 24.76 seconds. Thus there are differences in the mean speed times. Particularly the second layout has a much larger typing speed than the other two. Again, are those statistically significantly different? We will have to perform a hypothesis testing procedure in order to address this question.

2.4. Test for Equal Means

The topic of this lesson is hypothesis testing for equal means. You will learn about how to perform the statistical test, particularly how to derive the test statistic and how to decide based on a test statistic.

Slide 3:

Using the **hypothesis testing procedure for equal means**, we test: The null hypothesis, which that the means are all equal ($\mu_1 = \mu_2 \dots = \mu_k$) versus the alternative hypothesis, that some means are different. Not all means have to be different for the alternative hypothesis to be true -- at least one pair of the means needs to be different.

Slide 4:

Under the null hypothesis, we can combine all k samples into one big large sample because null hypothesis of equal means translates into that observations have a normal distribution with a common mean μ and a common variance σ^2 . We can estimate this common mean by pooling all the response samples into one sample and estimating the mean with the sample mean of this combined response samples:

$$\bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$$

If we want to estimate the variance, similarly we're going to use the sample variance of the entire combined response samples. The difference between this variance estimator and the pool variance estimator is that now we are replacing the mean with the overall mean, not with the individual sample means. Because we are replacing only one parameter, the overall mean, we're now only losing one degree of freedom.

We can rewrite this estimate as sum of square total or abbreviated SST divided by $N - 1$. Again, N is the sum of all samples. Because we only have to estimate one mean, we only lose 1 degree of freedom above, and thus the denominator is $N-1$, not $N-k$ as in the pooled variance estimator.

$$S_0^2 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2}{N-1} = \frac{SST}{N-1}$$

• **SST = Sum of Squares Total**

For this variance estimator obtained under the null hypothesis, **the sample distribution of the variance estimator** is the chi-square distribution with $N - 1$ degrees of freedom:

$$\frac{(N-1)S_0^2}{\sigma^2} = \frac{SST}{\sigma^2} \sim \chi_{N-1}^2$$

Slide 5:

The sum of squares total can be decomposed into two components, the sum of squared error or SSE plus the sum of square of treatments, or SSTr:

SST = SSE + SSTr

Recall, the sum of square of errors (SSE) is the sum of square differences between the observations and the individual sample means. On the other hand, the sum of square treatment (SSTr) is the sum of the square difference between the sample means of the individual samples minus the overall mean:

where $SSTr = \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$ and $\bar{Y}_{i.}$ is the i^{th} sample mean

The interpretation of the decomposition is as follows:

1. $MSE = SSE/N-k = \text{within-group variability}$
2. $MSSTr = SSTr/k-1 = \text{between-group variability}$
3. ANOVA: comparing between to within variability
4. $F = \text{between-group variability/within-group variability}$

1. The ratio between the sum of squared errors (SSE) divided by $N - k$ is called **mean sum of squared errors (MSE)**. It's a measure of the **within-group variability**. Remember that we used this to estimate the pool variance estimator.

2. The **mean sum of square treatments (MSSTr)** is the sum of square treatment (SSTr) divided by $k - 1$, where k is the number of samples. This is a measure of the *between-group variability*.
3. One of the main purposes of ANOVA is to compare the variability between samples to the variability within a sample. The **F-test** is the ratio of between-group variability and within-group variability

Slide 6-7:

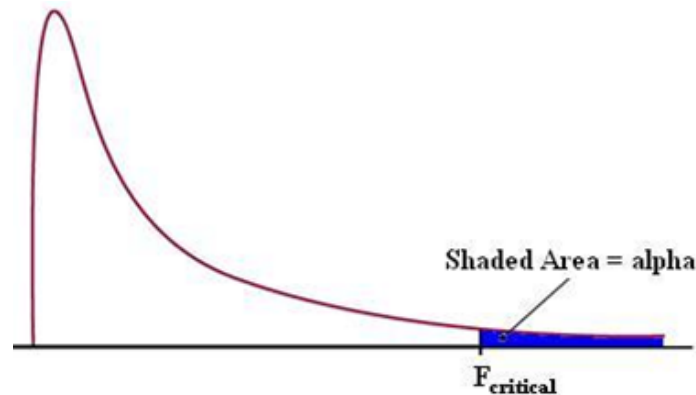
The **F-test** is the ratio between the sum of square treatments divided by $k - 1$ and the sum of squared errors divided by $N - k$. This is equivalent to the ratio between mean sum of square treatments (MSSTr) divided by the mean sum of squared error MSE. The F-test compares variability *between* groups against variability within groups. Under the null hypothesis that all means are equal, this ratio has an F distribution with $k-1$ and $N-k$ degrees of freedom.

$$\frac{SST_R / (k-1)}{SSE / (N-k)} \equiv \frac{MST_R}{MSE} = F_0 \sim F_{k-1, N-k}$$

If the F-test value is large, variability between groups is larger than variability within groups, and thus we reject the null hypothesis that the means are equal. We make this determination by comparing the F-test with the **F critical point** where the F critical point is for F distribution, with $k - 1$ and $N - k$ degrees of freedom.

Reject H_0 if $F_0 > F_{\alpha}(k-1, N-k)$, which is the upper α^{th} quantile of the F distribution.

We can also make a decision based on the p value, computed as the area under the **right tail** of the F distribution with $k-1$ and $N-k$ degrees of freedom, shown here in blue. Generally, we perform this test with statistical software as I will illustrate in the next slides with a data example.



Slide 8:

Let's go back to the first example where we interested in comparing the suicide rates by country region.

Are the mean rates statistically different?

Slide 9:

To perform the testing procedure for equal means, we can use the **aov()** command in **R**. The ANOVA table, we can use the summary() command. What I'm showing here in the output is what we call the ANOVA table. In the ANOVA table, we have two important rows, one corresponding to the treatments (country region) and the one corresponding to the residuals, which again are proxies of the errors.

- The first column provides the degrees of freedom for the treatment versus the residuals, there are 9 degrees of freedom for treatments because we have 10 groups. Thus $k - 1$ is going to be 9. We have 77 degrees of freedom for residuals, which corresponds to $N - k$.
- The second column provides the values for the sum of squares. The sum of square treatment is 1548 and the sum of squared error is 2779.
- The third column provides the mean sum of squares. For example, for the MSTr, we take the sum of square treatment and divide it by the corresponding degrees of freedom. Thus we take 1548 and divide it by 9, and we're going to get 172.06.
- The fourth column gives us the F value which is the ratio between the mean sum of square treatments divided by mean sum squared errors.

- The last column provides the P value of the F test. What we learn from here is that because the P value is approximately equal to zero, we reject the null hypothesis of equal mean for the suicide rates across regions.

Slide 10:

In the second example, we're interested to compare the means of the typing speed between the three keyboard layouts.

Are the mean typing times for the three keyboard layouts statistically different?

Slide 10:

Similar to the previous example, we perform an ANOVA using R. The summary is the ANOVA table provided on the slide. We have two degrees of freedom for the treatment because we have three groups. We have $N - k$ equal to 30, the sum of square of treatments is 121.24, the sum of square of error is 34.42, the F-Value is 52.84. Because the P-Value is approximately equal to zero we conclude that we reject the null hypothesis of equal mean typing speeds.

2.5. Comparing Pairs of Means

In this lesson, I'll continue with statistical inference on comparing pairs of means.

Slide 3:

One primary goal of ANOVA might be to determine which treatment means are bigger or smaller. One way to do this is to compare all possible pairs; there are $k(k-1)/2$ unique pairs of treatments. This is called **pairwise comparison**. To begin, we estimate the difference in the means for a pair μ_i and μ_j , specifically $\hat{\mu}_i$ and $\hat{\mu}_j$. However, whether the difference is smaller or larger than zero is not telling us much since the variability in the difference between the estimated means is important also. Thus we use confidence intervals for differences in means to make statements about pairs of means.

The **confidence interval** is going to be **centered at the estimated mean difference**, plus or minus a critical point times the standard deviation of the estimated difference in means. Here, the critical point is now the **alpha percentile of the studentized range distribution**, **not** the **critical point of the t-distribution**, which would be suggested by the sampling distribution of the difference in the estimated means.

$$\left(\hat{\mu}_i - \hat{\mu}_j \right) \pm q_{\alpha, k, N-k} \sqrt{\frac{\text{MSE}}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

The diagram shows the formula above with three blue arrows pointing to its components from below:

- The first arrow points to $(\hat{\mu}_i - \hat{\mu}_j)$ with the label: "The estimate of the difference in means"
- The second arrow points to $q_{\alpha, k, N-k}$ with the label: "The α percentile of the 'studentized range' distribution."
- The third arrow points to the standard deviation term $\sqrt{\frac{\text{MSE}}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$ with the label: "The standard deviation of the estimator"

Slide 4:

why do we use the q critical point rather than the t critical point? The reason is that we want to correct for simultaneous or joint inference. **The q critical point allows for correcting for multiplicity, which means that the intervals are wider, to compensate for the fact that we're making simultaneous comparisons.**

When we perform as a joint inference, confidence will decrease if we do not perform a correction. Consider calculating a 95% confidence interval for two population means based on two independent samples. Here, we're not interested in confidence intervals for each individual populations separately; we are interested on a simultaneous inferences on both, together. This means that their significance level is $(.95)(.95)$, or about 90%, not 95% as we initially wanted. Now imagine we are analyzing three populations and we want 95% confidence intervals. If we multiply 0.95, three times, we get 0.86, not 95%.

The q critical points for the pair comparison in ANOVA are not easy to compute. There are some tables in some of the textbooks, but I would suggest you use a statistical software to get the confidence intervals. The R statistical software does have a command that you can use, *called the **Tukey method***, developed by the statistician John Tukey.

Slide 5:

Let's illustrate the implementation of the comparison with the first example.

Which suicide rates across the country regions are statistically different?

Slide 6:

Recall that we rejected the hypothesis that the means are equal, meaning that at least two of the means are statistically different. The R command you can use to obtain the pairwise confidence intervals is called **TukeyHSD()**; for this you need to input the fitted ANOVA model. If you want to use a confidence level different than 95% which is the default, you need to specify that, as well. The partial output of this command is on the slide. I didn't include all the comparisons since there are 45 unique pairwise comparisons for 10 categories if we were to consider all; although some of the categories only include one observations and hence I discarded from this analysis since we cannot provide comparisons with regions including only one country or one observation.

The way you interpret this output is as follows:

- The first column provides the pairs of the means that we compare. For example, the first pair is the mean suicide rate between Eastern Europe and Asia.
- The second column will provide the differences in the estimated means.
- The next columns provide the lower and upper bounds of the confidence intervals for the differences in the means.

- The last column is the **adjusted p-value**, which gives us information whether we reject or not the null hypothesis for when we test whether the means of the pairs are equal or not.

So how can I use this output? First, you must look at the lower and upper bounds, and identify the **confidence intervals that include the zero values**. **For those confidence intervals, it's plausible for the difference to be zero.**

So for example, if you look at the difference between the mean suicide rate for the first pair, the lower bound is -0.86, the upper bound is 15.11. This confidence interval includes zero, meaning that the means of those two groups could be possibly equal to zero. You can also see the probability of p-value adjusted is 0.12, which indicates that the means of suicide rates between these two regions could be equal. For this p-value, we do not reject the null hypothesis of equal means.

Another aspect that we'll look into when evaluating the pairwise comparison is identifying the **confidence levels that have only positive values or only negative values**. For example, the confidence interval comparing Latin America vs Eastern Europe only includes negative values, suggesting that the mean of the suicide rates for countries in Latin America is lower than the mean of the suicide rates for countries in Eastern Europe.

Slide 7:

So here's a conclusion based on this output. We have a large number of categories to compare. It is particularly important in such examples to correct to multiplicity, hence we need to use the appropriate statistical inference approach. There are several regions with too few observations, which I recommend excluding from this comparison. Last, only three pairs have an adjusted p-value smaller than 0.05: Latin America vs Eastern Europe, Middle East vs Eastern Europe and Middle East vs Global West.

Slide 8:

Let's go back to the second example where we want to compare the means of the typing speed of the three keyboard layouts.

Which mean typing speeds for the three keyboard layouts are statistically different?

Slide 9:

Again, we can perform a pairwise comparison using the Tukey method. We have only three pairs: 2 and 1, 3 and 1, and 3 and 2. For the pair 3 and 1, the third and the first

layout, the confidence interval includes zero, which means that the means of the typing speed for the two layouts is plausibly similar. When we compare layouts 1 and 2 and layouts 3 and 2 respectively, we can see statistically significant differences.

In conclusion what we learn from this example is that the keyboard type 2 has statistically significantly higher typing time than keyboard layouts 1 and 3 on average. It's plausible that keyboard layout 1 and 3 have similar typing speed in average.

2.6. Model Fit Assessment

The topic of this lesson is analysis of variance with a focus on model fit assessment. Particularly, we're going to overview the ANOVA model assumptions and we'll learn about simple ways to diagnose these assumptions using graphical displays.

Slide 3:

the data in the ANOVA model consist of a response variable of interest observed for multiple populations differentiated by a categorical variable or a label. For example, for the analysis of suicide rates, the categorization could be instead age group, year, weather type etc.

Let's begin with the model. In our notation, Y_{ij} are the response data differentiated across the k categories. j is the index within group and i is the index across groups. The model is: $Y_{ij} = \mu_i$ (mean of the group i) + ϵ_{ij} (the error term epsilon_{ij}).

Data: Y_{ij} for $j = 1, \dots, n_i; i = 1, \dots, k$

Model: $Y_{ij} = \mu_i + \epsilon_{ij}$ where ϵ_{ij} = error term

The assumptions on ANOVA are with respect to the error term:

- **Constant variance assumption.** Recall that we assume the response data has constant variance across all groups, and thus the variance of the error terms is constant, equal to sigma square: $\text{Var}(\epsilon_{ij}) = \sigma^2$
- **Independence assumption.** Response data and error terms (" ϵ ") are independent.
- **Normality assumption.** Error terms are normal, and thus the y_{ij} response data are normal as well.

Slide 4:

To diagnose these assumptions, we do not diagnose assumptions of error terms because we do not know the means. Instead, we're going to **diagnose the assumptions on the residuals**. The residuals are the difference between the responses minus the estimated means of the individual samples:

$$\hat{\epsilon}_{ij} = Y_{ij} - \hat{\mu}_i$$

If the model fit is a good fit, then the residuals should be scattered around zero (randomly) in the plot of the residuals against either fitted or just their order.

Slide 5:

When we evaluate the assumptions in ANOVA, here are the type of residual plots we consider:

- We can plot the **residuals for each treatment group** to evaluate whether there is a **different variability across the groups**.
- We can plot the **quantile-quantile normal plot** to evaluate **normality**
- We can also plot a **histogram of the residuals**, similarly used to evaluate **normality**.

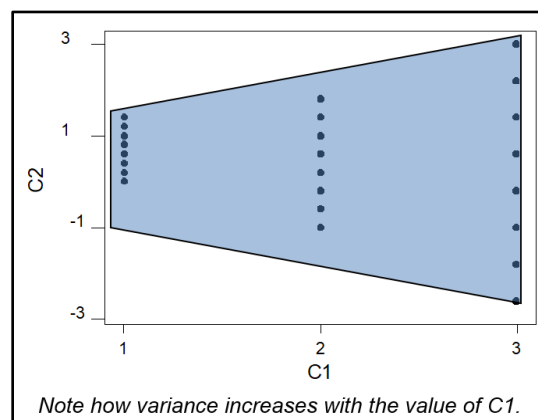
If the **scatter plot of the residuals (epsilon ij) is NOT random**:

- The sample responses are **not independent**, or
- the **variances** of responses are **not equal**

If the quantile-quantile normal plot and the histogram show departure from normality, you may consider a transformation in order to normalize the data.

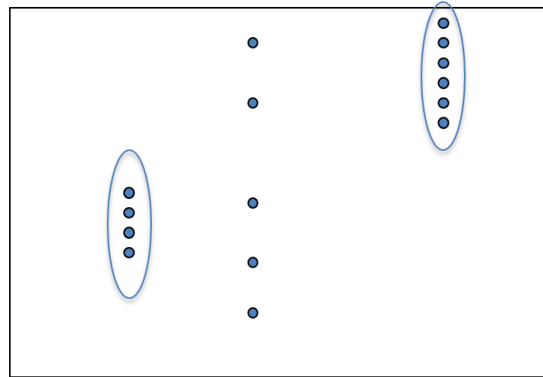
Slide 6:

Here is one example for departure from the assumption of *non-constant variance*. You can see here that the variability of the residuals changes from group one to group three. We have a much smaller variability for group one than for group three.



Slide 7:

This is another example of departure from one of the assumptions - we can see the residual clusters, which means that we may have *correlated errors*:



Clusters of residuals: correlated errors

Slide 8:

So let's go back to the first ANOVA example, where we're comparing the suicide rates across different country regions.

Are the inferences on the difference in height means reliable? Is the model a good fit?

Slide 9:

In order to address this question, we will perform a residual analysis using the three residual plots. The R code is here.

Slide 10:

The residual plots we are interested are here. The first one plots residuals against the fitted means. We're going to use this plot in order to evaluate whether the residuals are scattered around the zero line, and whether the variances are different across the groups or whether there are clusters in the residuals.

The second plot is the quantile-quantile normal plot. Here, we expect the residuals to line up on the line, meaning that the residuals will have a similar distribution to the normal distribution.

The third plot is a histogram of the residuals. We expect to see an approximately symmetric distribution.

For this example, all three assumptions hold.

2.7. ANOVA vs. Simple Linear Regression

The topic of this lesson is the comparison between the two models we've learned so far, analysis of variance and simple linear regression. Particularly, we'll learn that ANOVA is a particular case of linear regression.

Slide 3:

In the simple linear regression model in Unit 1, both the response and predictive variables have been quantitative variables. Can this be generalized to analyzing the variability in a response variable with different groups of predicting variables? For example:

- Does knowing the education level of a person, say high school, college, have predictive power for their annual salary?
- Is a return on a stock related to the industry group of the company?

This is a special kind of regression question in this context: if group membership has predictive power for the response, then the average mean of the response variable is different for different groups. This is actually a comparison of means, as we learned in analysis of variance. Thus ANOVA is a linear regression model where the predicting factor is a categorical or qualitative variable.

Let's look closer at ANOVA as a linear regression model.

Slide 4:

The data in the ANOVA consist of a response variable Y_{ij} observed across multiple categories with i from 1 to k ; but we can write the response Y_{ij} as the sum between the mean of the group i plus an error term. Furthermore, we can write μ_i as the sum between μ and τ_i , where μ is the overall mean and τ_i are the so-called treatment effects, where the sum of τ_i is equal to 0:

ANOVA:

Data: $Y_{ij} = 1, \dots, n_i; i = 1, \dots, k$

Model: $Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$ where $\sum_{i=1}^k \tau_i = 0$

$\mu_i = i\text{-th group mean decomposed into } \mu_i = \mu + \tau_i$

So let's see how we can write this model into a regression model. Now I take all the Y_{ij} 's and I stack them up into one vector of responses. The first n_1 values correspond to

the first group, the first data sample, the next n_2 values correspond to the second group. The last k values in this vector will correspond to the k th data sample or group. This is going to be the Y response variable:

Define Y be the response variable:

$$Y = (Y_{11}, \dots, Y_{1n_1}, Y_{21}, \dots, Y_{2n_2}, \dots, Y_{k1}, \dots, Y_{kn_k})$$

Now, I define another variable. I call it a label or categorical variable. Again, I stack those values, which are the labels. The first n_1 values in this vector corresponds to label 1, the label of the first category. The next n_2 values correspond to the label of the second category. The last n_k values correspond to the label of the k th group.

Define L be the label/categorical variable:

$$L = (l_{11}, \dots, l_{1n_1}, l_{21}, \dots, l_{2n_2}, \dots, l_{k1}, \dots, l_{kn_k}) \text{ Where } l_{ij} = i$$

Now ANOVA is a linear regression where we regress the label L onto the response Y .

Linear Regression: $Y \sim L$

Slide 5:

When we have a categorical or qualitative predicting variable with k different labels, we then convert those into what we call **dummy variables**, labeled x_1 to x_k . The x_1 variable for example has one on the first n_1 values and 0 for the rest of the values. The k th dummy variable has 0s at the beginning but the last n_k values are 1's. All of those dummy variables have the same length, and the length is N , which is the sum of all sample sizes across the k samples:

Categorical Variables in Linear Regression:

• **Transform categories into dummy variables**

$$x_1 = (1, \dots, 1, 0, 0, \dots, 0)^T; \dots; x_k = (0, 0, 0, \dots, 1, \dots, 1)^T$$

When we model a regression analysis with those variables, now those k dummy variables become the predicting variables, where Y (presented in a previous slide) is the response variable. If the model has an intercept we only include $k-1$ dummy variables because of the linear dependence between the x 's:

•If intercept in the model, only $k-1$ dummy variables because of linear dependence: $(1,1,\dots,1)^T = x_1 + x_2 + \dots + x_k$

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{k-1} x_{ik-1} + \varepsilon_i, i=1,\dots,n$

If the model does not have an intercept, then we'll include all k dummy variables in the model:

•If no intercept in the model, all k dummy variables

Model: $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i, i=1,\dots,n$

But most importantly, what you can see here is that I wrote the **ANOVA as a multiple linear regression model where the predicting variables are the dummy variables and the response variable is the response data stacked into one vector.**

Thus ANOVA is a particular case of a multiple regression model. In fact, β_1, β_2 and β_k in a model with no intercept correspond to the mean parameters, μ_1, μ_2 and μ_k .

I will expand on the linear models with multiple predicting variables, including both quantitative and qualitative variables in Unit 3 of this course.

2.8. Data Example

In the data example introduced in this last lesson of Unit 2 I will demonstrate how to implement ANOVA in R and how to draw inferences based on the R output.

Slide 3:

In this example, we're interested in the number of survival days for patients with different types of cancers that were treated with ascorbate.

Slide 4:

The response variable is the number of survival days for each patient for different types of cancer; 'j' again is the index for a patient and 'I' is the index for the type of cancer or the group. We have five different groups ($k = 5$). Each group corresponds to a cancer of a different organ (stomach, colon, ovary, breast). There are different sample sizes across the five groups. In fact, the group with the ovarian cancer has only six observations.

Slide 5:

A first step in any data analysis is to perform an exploratory data analysis. First, we're going to read the data in R using the **read.table()** command, where we need to specify the name of the file and whether the columns have a header or not. Then we need to extract the individual variables of interest. The response variable is 'survival' days. We can extract it from the data by specifying the name of the matrix, cancer_data and the name of the column with a dollar sign in-between.

Our first step is to analyze whether the response variable has an approximately normal distribution using a histogram with the R command hist(). Do not forget to specify the X axis and the Y axis labels. The histogram of the survival response data is on the slide. What we see here is that the response variable has a skewed distribution, an indication that the normality assumption does not hold. We will need to transform the data to normalize it. A common transformation to normalize when there is a strong skewness is to use a **log() transformation**. After taking this transformation, you can see that we don't have that long tail on the right anymore and the distribution looks symmetric. We're not looking for a perfect normal, symmetric distribution here. In fact, if you remember from the data structure in the ANOVA model, the data come from multiple normal distributions. What we should expect is to see a multi-modal distribution where

the modes correspond to each individual group. After transformation, we do see two modes in this distribution, but the shape of the distribution is symmetric.

Slide 6:

For this study, we're going to use the (log) transformed data, not the original data. For the ANOVA implementation, we need two variables. One is going to be the response variable, in this case, it is the log of the survival number of days, and the second variable is the label corresponding to the type of cancer. Recall, the response variable consists of all the stacked response data and now we have to specify which response observation corresponds to what category or cancer type. However, we need to ensure that R identifies this as a categorical variable. For this, we need to convert into a factor using the **as.factor()** command. By doing so, R will know that this is a categorical variable and when you apply the side-by-side boxplot in R, you will plot the survival for each category separately.

The side by side boxplot of the log of survival data versus the cancer type is on the slide. You see here that there are differences in the means between the groups. For example if you compare those patients with breast cancer versus other cancers, the number survival days will be much larger for those with breast cancer compared to those not with breast cancer. We can also see that the variability within each group is slightly different. But that may be also because some groups have more observations than others. For example, the group of patients that have ovary cancer has a very small sample size, and thus will have much larger variability also.

When we study the side-by-side boxplot, we compared the within-variability to the between-variability. I will remind you that the within-variability is the variability within the group. The between-variability is the variability of the means between the groups.

Slide 7:

We'll next perform an **Analysis of Variance** on the log survival rate versus the cancer type. The output of the ANOVA (AOV command in R) consists of the ANOVA table, which provides information about the sum of squared errors or residuals as well as sum of square of treatments, the degrees of freedom, mean sum of squares, the F test for and the p-value of the test of equal means.

The first column in the ANOVA table are the degrees of freedom for each source of variability, the treatment (cancertype) and the residuals. There are four degrees of freedom for the cancer type, because there are five different cancer types. The next column corresponds to the sum of squares. The sum of square treatments is 24.49 and the sum of square of residuals is 84.27. The F-value is used for the F-test for performing the test for equality of the means. The F-value is 4.286 and the P-value of the F-test is 0.004.

The other set of output is provided by the **model.tables command** where we input the fitted model. This output provides us the overall mean across all groups and the means of the log survival rate for the individual groups. For example, the estimated mean for the group that has breast cancer, the average log survival rate is 6.56 and the number of patients in this group is 11.

Recall that we are performing ANOVA in order to test whether the means are equal across the groups, across the cancer types. **Since we reject the null hypothesis of equal means, we can ask the next question, which means are statistically significantly different?**

Slide 8:

To answer this question, we perform a pairwise comparison. We'll use the **TukeyHSD command** in R which will provide the confidence intervals for the difference in means for all possible combinations of the pairs of the means.

We see is that there are only two pairs of means that have statistically significantly different means. The first one, when we compare the means of the log of the number survival of days between patients with a Bronchus cancer versus with those with Breast cancer. The other one is the pair of means between Stomach cancer and Breast cancer. In both cases, the confidence intervals include only negative values, which means that the log mean of the number of survival days for those with breast cancer is significantly larger than the log mean of the number of survival days for those with Bronchus or Stomach cancer.

All the other pairs are not statistically significantly different, which means that it's possible that the number of survival days across all the other cancers is similar.

Slide 9:

It's important also to evaluate the assumptions (via residual analysis), meaning evaluating the goodness of fit of the model. This is important since without a good fit, we cannot rely on the statistical inference we made using the hypothesis testing on the pairwise comparison, for example.

I'll remind you that the three assumptions in ANOVA are the constant variance, independence, and normality. The first two plots, the quantile normal plot and the histogram are used to evaluate normality. The next two plots can be used to evaluate the other two assumptions, constant variance and uncorrelated. The normality plot and the histogram look reasonably well, which is an indication that the distribution of the residuals is approximately normal. Also, we don't see any pattern in the residuals.

Slide 10:

Using the Analysis of Variance, we've learned from this example that:

- There is strong evidence for differences in the survival days across the five types of cancer
- The survival time is statistically different for those patients with Breast cancer versus those with Bronchus or Stomach cancer