ANOVA

APPLIED STATISTICS (STAT200)

INTRODUCTION

Analysis of variance (ANOVA) is a term describing a large collection of statistical procedures.

The simplest ANOVA problem is called single-factor ANOVA; it involves the analysis of data sampled from two or more numerical populations (samples).

- The characteristic that labels the populations is called factor under study
- The different populations are referred to as the levels of the factor

Example 1: An experiment to study the effects of five different brands of gasoline on automobile engine operating efficiency (mpg).
Here, the factor is “gasoline brand” and it possesses five levels.

INTRODUCTION

Example 2: An experiment to study the effects of four different sugar solutions on bacterial growth.
Here, sugar is the factor, and it has has four levels.

Note:

- The factor is qualitative and nominal in both examples, hence the levels are categories
- An ANOVA may also be applied when the factor is qualitative and has ordinal structure. However, other approaches may be more beneficial in such a setting

SINGLE-FACTOR ANOVA

SECTIONS 11.1, 11.2, AND 11.3 (DEVORE & BERK 2012)

DEFINITION

Single-factor ANOVA (or: one-way ANOVA) focuses on a comparison of two or more populations – similar to a t-test, which analyzes only one or two populations. Let:

- I = the number of treatments being compared (= the number of levels)
- μ_1 = the mean of population 1 (or the true average response when treatment 1 is applied)

  $\vdots$

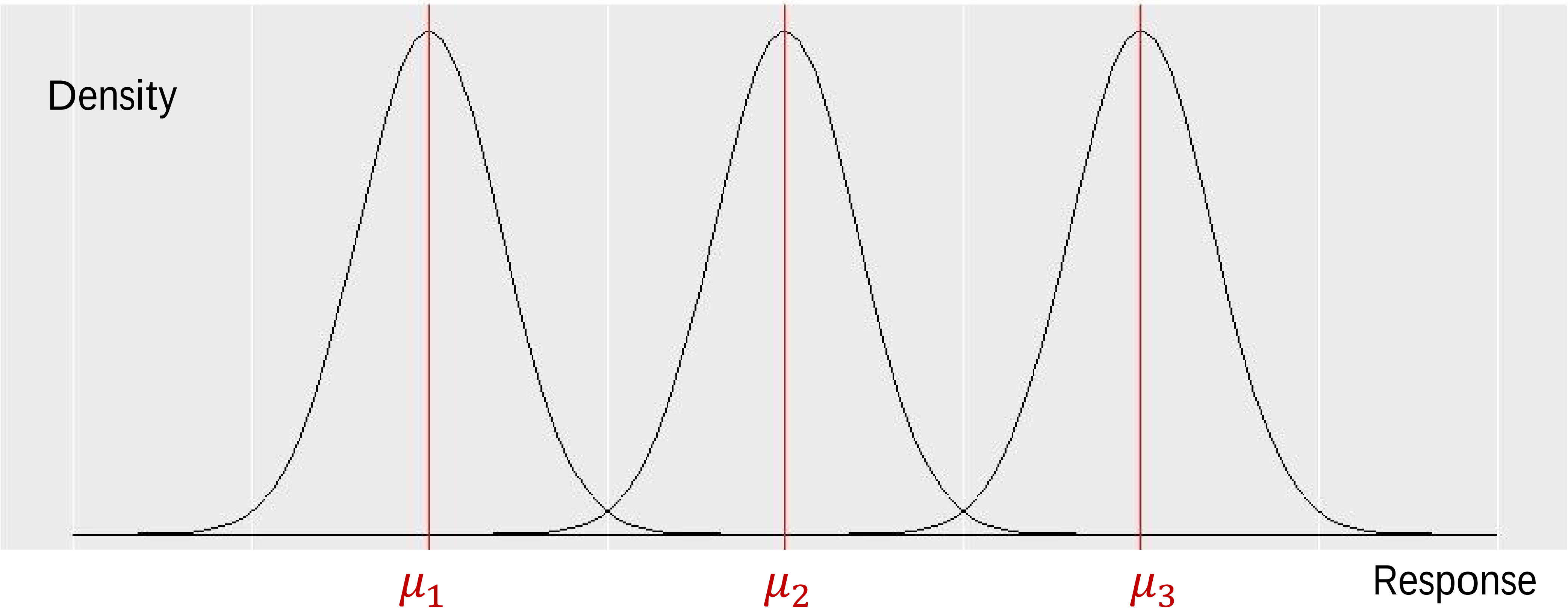
  μ_I = the mean of population I

Then we test the hypotheses

H_0: $\mu = \mu_1 = \mu_2 = \cdots = \mu_I$ vs.

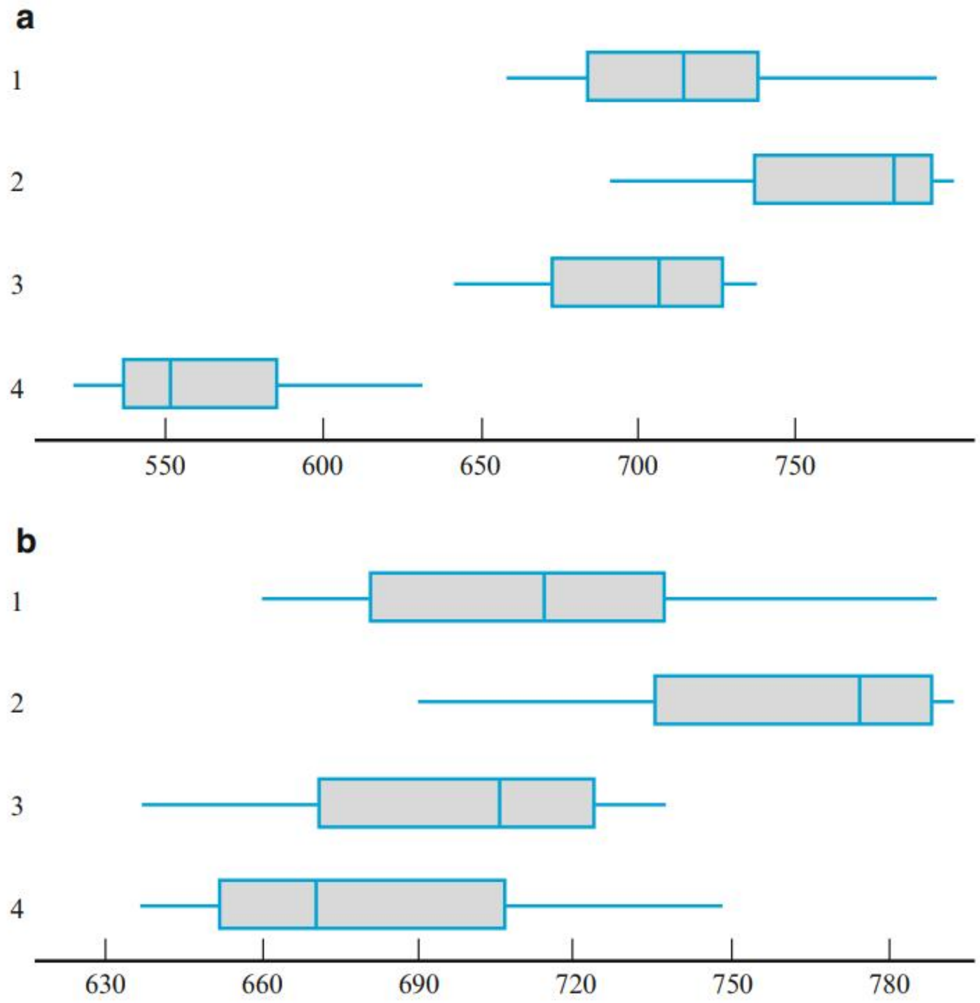
H_a: at least two of the μ_i's are different.

EXAMPLE WITH $I = 4$



Mean of the observations belonging to factor level 1

EXAMPLES WITH $I = 4$



ASSUMPTIONS

In order to carry out an ANOVA, various assumptions need to be met.

1. Observations within any particular sample are independent and that different samples are independent of each other
2. The sample sizes are equal for each factor level (we will discuss different sample sizes later on)
3. The *I* population distributions all follow a normal distribution with the same variance σ^2.
   This means that each X_{ij} is normally distributed with $E\left(X_{ij}\right) = \mu_i, V\left(X_{ij}\right) = \sigma^2$

???

NOTATION

We denote the number of different samples (=factor levels) with *I*; we denote the size of each sample with *J* (under the assumption that the sample sizes are equal).

We use the symbol X_{ij} to denote a random variable:

- the first subscript, *i*, identifies the sample number (factor level), or the population/treatment that's being sampled
- the second subscript, *j*, denotes the position/number of the observation within that sample

X_{ij} or $X_{i,j}$ = the random variable denoting the j^{th} measurement from the i^{th} population.

x_{ij} or $x_{i,j}$ = the observed value of X_{ij} when the experiment is performed.

NOTATION

The data set in this setting consists of IJ observations.

- The individual sample means are denoted with$\bar{X}_{1}.$, $\bar{X}_{2}.$, … , $\bar{X}_{I}.$, where

$$\bar{X}_{i\cdot} = \frac{\sum_{j=1}^{J} X_{ij}}{J}, i = 1,2, \ldots, I.$$

- Similarly, the average of all IJ observations is called grand mean, and is formulated as

$$\bar{X}_{\cdot\cdot} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij}}{IJ}.$$

- We denote the sample variances with $S_1^2, S_2^2, \ldots, S_I^2$, where

$$S_i^2 = \frac{\sum_{j=1}^{J} \left(X_{ij} - \bar{X}_{i\cdot}\right)^2}{J-1}$$

DEFINITIONS

The treatment sum of squares SSTr is a measure of differences among the sample means (so called “between-samples” variation). The formula is

$$SSTr = J\sum_{i}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 = J[(\bar{X}_{1\cdot} - \bar{X}_{\cdot\cdot})^2 + \cdots + (\bar{X}_{I\cdot} - \bar{X}_{\cdot\cdot})^2]$$

The error sum of squares SSE is a measure of variation calculated from within each sample.The formula is

$$SSE = \sum_{i}\sum_{j}(X_{ij} - \bar{X}_{i\cdot})^2 = \sum_{j}(X_{1j} - \bar{X}_{1\cdot})^2 + \ldots + (X_{Ij} - \bar{X}_{I\cdot})^2$$

$$= (J-1)S_1^2 + (J-1)S_2^2 + \cdots + (J-1)S_I^2 = (J-1)[S_1^2 + \cdots + S_I^2]$$

The total sum of squares SST is defined as

$$SST = \sum_{i}\sum_{j}(X_{ij} - \bar{X}_{\cdot\cdot})^2$$

THEOREM

Observe that the following holds:

H_0 is true $\Rightarrow \frac{STTr}{\sigma^2} \sim \chi^2$ with $I - 1$ degree of freedom, thus

H_0 is true $\Rightarrow E\left(\frac{STTr}{I-1}\right) = E\,(MSTr) = \sigma^2$,

where $MSTr$ is the mean square for treatments.

This means:

- $MSTr$ is an unbiased estimator of σ^2 if H_0 is true
- If H_0 is false $MSTr$ tends to overestimate σ^2

THEOREM

On the other hand holds:

$\frac{SSE}{\sigma^2} \sim \chi^2$ with $I(J-1)$ degree of freedom - even when H_0 is not true!

Thus:

$E\left(\frac{SSE}{I(J-1)}\right) = E\ (MSE) = \sigma^2$, where MSE is called mean square for error.

- This means that MSE always is an unbiased estimator of σ^2
- Note: SSE and $SSTr$ are independent variables

THE F-TEST

General idea of an F-test:

- Let $Y_1 \sim \chi^2$ with ν_1 df and $Y_2 \sim \chi^2$ with ν_2 df be independent random variables
- Then the ratio $F = \frac{Y1/\nu_1}{Y_2/\nu_2}$ follows an F distribution with ν_1 and ν_2 degrees of freedom df

Consequently, for an ANOVA we define the test ratio $F = \frac{MSTr}{MSE}$.

- F takes values around 1 if H_0 is true
- F takes values larger than 1 if H_0 is not true

THE F-TEST

In the case of single-factor ANOVA, we can write

$$F = \frac{MSTr}{MSE} = \frac{\left[\frac{SSTr}{\sigma^2}\right] \Big/ (I-1)}{\left[\frac{SSE}{\sigma^2}\right] \Big/ I(J-1)}$$

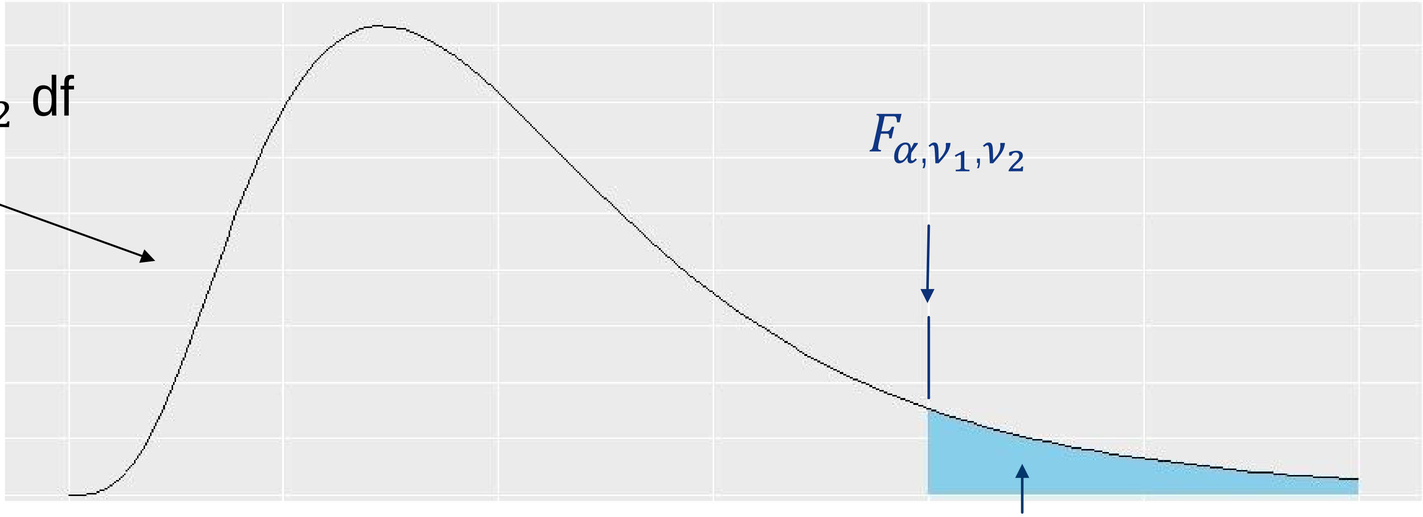
When H_0 is true:

- The numerator and denominator of F are independent chi-squared variables divided by their df's
- The df's are $I-1$ for the numerator and $I(J-1)$ for the denominator

The rejection region $f \geq F_{\alpha,I-1,I(J-1)}$ then specifies an upper-tailed test with significance level α.

THE F-TEST

Example: for given values of ν_1 and ν_2, the resulting density of the F statistic could look as follows



F curve for ν_1 and ν_2 df

F_{α,ν_1,ν_2}

Region / zone of rejection with α (%)

SUMMARY TABLE

Sum of squares	Df	Definition	Computing formula
Total = SST	$IJ - 1$	$\sum_i \sum_j \left(x_{ij} - \bar{x}_{..}\right)^2$	$\sum_i \sum_j x_{ij}^2 - x_{..}^2 \,/\, IJ$
Treatment = SSTr	$I - 1$	$J \sum_j (x_{i.} - \bar{x}_{..})^2$	$\frac{\sum_i x_{i.}^2}{J} - \frac{x_{..}^2}{IJ}$
Error = SSE	$I(J - 1)$	$\sum_i \sum_j \left(x_{ij} - \bar{x}_{i.}\right)^2$	SST - SSTr

FUNDAMENTAL ANOVA IDENTITY

The computational formula for SSE is a consequence of the fundamental ANOVA identity

$$\text{SST} = \text{SSTr} + \text{SSE}$$

Proof: homework

Hint: square both sides of the equation

$$x_{ij} - \bar{x}_{..} = \left(x_{ij} - \bar{x}_{i.}\right) + \left(\bar{x}_{i.} - \bar{x}_{..}\right)$$

and sum over all i’s and j’s.

FUNDAMENTAL ANOVA IDENTITY

Interpretation

- SST is a measure of total variation in the data. The identity says that this total variation can be partitioned into two parts, SSE and SSTr
- SSTr is the part of the total variation that can be explained by differences among μ_i's
- SSE represents the part of the total variation that is unexplained by the state of H_0 (in fact, it does not depend on whether H_0 is true or not)

Commonly used terms are

- SSE: within-sample(s) variation
- SSTr: between-sample(s) variation

ANOVA TABLE

The computations necessary for an ANOVA-type analysis are often summarized in a tabular format.

This is the so-called ANOVA table, and looks (approximately) like this:

Source of Variation	Df	Sum of Squares	Mean Square	F
Treatments	$I-1$	SSTr	MSTr=SSTr/$(I-1)$	MSTr/MSE
Error	$I(J-1)$	SSE	MSTr=SSTr/$I(J-1)$	
Total	$IJ-1$	SST		

SINGLE-FACTOR ANOVA – IMPLEMENTATION IN R

Several procedures for carrying out an ANOVA are available in R. For the beginning, we focus on the simplest one

- Use the function `aov`
- One argument necessary: the model specification having the form

  ```
  response ~ predictor
  ```
- `aov` creates an object of class "`aov`" for summarizing the results. Use the `summary` function to show the results

As before: if you are not sure how a test works, try a simple example.

`- script on screen -`

TUKEY'S PROCEDURE FOR POST-HOC ANALYSIS

How to interpret the F-statistic, and what to do further?

1. If the F-statistic isn't significant
   Conclusion: no effect of the factor, stop the analysis
2. If the F-statistic is significant,
   Conclusion: the factor has an effect, and at least one mean μ_i differs from the others

If 2. is the case, we proceed with a so-called post-hoc analysis. A post-hoc analysis allows to find out which pairs of μ_i's are different.

TUKEY'S PROCEDURE FOR POST-HOC ANALYSIS

Attention: comparing all pairs with each other leads to so-called multiple comparisons.
If this is not accounted for, one risks to detect differences between samples although there are none present in reality – they just occur randomly.

For the ANOVA, Tukey's procedure is a good approach to avoid problems related to multiple comparisons:

- The principle idea is to adjust the p-values of all the tests comparing μ_i and μ_j (for all *i* and *j*)
- More specifically, it control the simultaneous confidence level for all $I(I-1)/2$ intervals of the differences $\mu_i - \mu_j$

TUKEY'S PROCEDURE – DEFINITIONS

Let $Z_1, Z_2, \ldots, Z_m$ be m independent standard normal random variables and let $W \sim \chi^2$, independent of the Z_i's with ν df. Then the distribution of

$$Q = \frac{\max|Z_i - Z_j|}{\sqrt{W/_\nu}} = \frac{\max(Z_1, \ldots, Z_m) - \min(Z_1, \ldots, Z_m)}{\sqrt{W/_\nu}}$$

is called the studentized range distribution. The distribution has two parameters, *m* = the number of Z_i's, and ν = denominator df. We denote the critical value that captures upper-tail area α under the density curve Q by $Q_{\alpha,m,\nu}$.

Table A.9 in the appendix of Devore & Berk contains a table of these critical values.

TUKEY'S PROCEDURE

For a one-way ANOVA one proceeds as follows:

1. Select α and extract $Q_{\alpha,I,I(J-1)}$ from Appendix Table A.9, and calculate
$$w = Q_{\alpha,I,I(J-1)} \cdot \sqrt{MSE/J}$$
2. List the sample means in increasing order and underline those pairs that differ by less than w
3. Any pair of sample means not underscored by the same line corresponds to a pair of population means that are judged significantly different.

The quantity w is sometimes referred to as Tukey's honestly significantly difference (HSD).

TUKEY'S PROCEDURE – EXAMPLES

Example: Suppose that $I = 5$, and that $\bar{x}_{2.} < \bar{x}_{5.} < \bar{x}_{4.} < \bar{x}_{1.} < \bar{x}_{3.}$. Then

1. Consider first the smallest mean $\bar{x}_{2.}$. If $\bar{x}_{5.} - \bar{x}_{2.} \geq w$, we move to step 2. If $\bar{x}_{5.} - \bar{x}_{2.} < w$, connect these two means with a line segment. Then, if possible, extend this segment even further to the right to the largest $\bar{x}_{i.}$ that differs from $\bar{x}_{2.}$ by less than w.
2. Move to $\bar{x}_{5.}$, and extend a line segment to the largest $\bar{x}_{i.}$ to its right that differs from $\bar{x}_{5.}$ by less than w (it might not be possible to draw this line!)
3. Continue by moving to $\bar{x}_{4.}$ and repeating step 1. and 2. We then finally move to $\bar{x}_{1.}$

TUKEY'S PROCEDURE – INTERPRETATION OF α

What do we mean with simultaneous confidence level in this case (but also in other settings)?

Example: Consider calculating a 95% CI for a population mean μ based on a sample from a population. Then, you also calculate a 95% CI for a population proportion p based on another sample selected independently of the first sample.

- Prior to obtaining data, the probability that the first interval will include μ is .95, and this is also the probability that the second interval will include p
- The two samples are selected independently of each other, the probability that both intervals will include μ and p respectively is $(.95)(.95) \approx 0.90$. Thus the simultaneous or joint confidence level for the two intervals is roughly 90%

TUKEY'S PROCEDURE – INTERPRETATION OF α

Following the same argumentation, it is easy to see that:

- If three CIs are calculated based on independent samples, the simultaneous confidence level will be $100(.95)^3 \approx 86\%$
- Increase in number of intervals ⇒ decrease of simultaneous confidence level

In practice, we usually want to maintain the simultaneous confidence level at 95% (or at 99%,…). Then:

- For two independent samples: the individual confidence level for each test would have to be $100\sqrt{.95}\,\% = 97.5\%$
- Increase in number of intervals (comparisons) ⇒ increase of the individual CI to maintain the 95% simultaneous level
- Tukey's and related procedure(s) increase the intervals in "an intelligent" way

TUKEY'S PROCEDURE – IMPLEMENTATION IN R

The Tukey's procedure is available directly in R and easy to use

- Use the function `TukeyHSD`
- Only one argument necessary: an object of type `aov`
- Additional arguments allow to change the simultaneous / family-wise confidence level (`conf.level`) and to order the factor levels (`ordered`)

Example:

```
TukeyHSD(mod1)

TukeyHSD(mod2)
```

ALTERNATIVE DESCRIPTION OF THE ANOVA MODEL

The assumptions of single-factor ANOVA can be described succinctly by means of the model equation

$$X_{ij} = \mu_i + \varepsilon_{ij},$$

where

- ε_{ij} represents a Gaussian random deviation from the population mean μ_i
- $E(\varepsilon_{ij}) = 0, V(\varepsilon_{ij}) = \sigma^2$

Hence, $E(X_{ij}) = \mu_i$ and $V(X_{ij}) = \sigma^2$ for every i, j. We then define

$$\mu = \frac{1}{I}\sum_{i=1}^{I} \mu_i$$

(this works because J is the same for every sample!) and the parameters

$$\alpha_i = \mu_i - \mu \text{ for } i = 1,2, \ldots, I.$$

ALTERNATIVE DESCRIPTION OF THE ANOVA MODEL

Note that

- Now we have $I + 1$ coefficients $\mu, \alpha_1, \dots, \alpha_I$
- However, since we know that $\sum_i \alpha_i = 0$ (the average departure from the overall mean response is zero), only I parameters are independently determined

The model equation then becomes

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

for $i = 1, \dots, I, j = 1, \dots, J$.

Remark: This type of equation is useful for describing all kinds of statistical models and will be used very regularly.

ALTERNATIVE DESCRIPTION OF THE ANOVA MODEL

Using this alternative description,

- The claim that all μ_i's are identical is equivalent to the equality of all α_i's
- Since $\sum_i \alpha_i = 0$, the null hypothesis becomes $H_0: \alpha_1 = \cdots = \alpha_I = 0$.

Recall: if H_0 is false, MSTr tends to overestimate σ^2. More precisely, one can show that

$$E(MSTr) = \sigma^2 + \frac{J}{I-1}\sum_i \alpha_i^2$$

- When H_0 is true, $\mu_1 = \cdots = \mu_I \Rightarrow \alpha_1 = \cdots = \alpha_I = 0 \Rightarrow E(MSTr) = \sigma^2$
- A larger value of $\sum_i \alpha_i^2$ will result in a greater tendency for MSTr to overestimate σ^2

F-TEST AND T-TEST

If $I = 2$, the F-test of the ANOVA is testing $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$. We could also use a two-tailed, two-sample t test instead!

It can be shown that the single factor ANOVA F-test and the two-tailed pooled t test are equivalent; the p-values for the two tests will be identical!

Note:The two-sample t-test is more flexible than the F-test when $I = 2$ for two reasons. First, we do not need the assumption $\sigma_1 = \sigma_2$ (in principle). Second, we can perform one-tailed t-tests ($H_a: \mu_1 > \mu_2$ or $H_a: \mu_1 < \mu_2$) if this is required.

WHAT IF SAMPLE SIZES ARE UNEQUAL?

Let $J_1, J_2, \dots, J_I$ denote the I sample sizes. Let $n = \sum_i J_i$ denote the total number of observations. The previous formulae for this case then become:

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{ij} - \bar{X}_{..}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{ij}^2 - \frac{1}{n}\bar{X}_{..}^2 \text{ , } df = n - 1$$

$$SSTr = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{I}\frac{1}{J_i}\bar{X}_{i.} - \frac{1}{n}\bar{X}_{..}^2 \text{ , } \quad df = I - 1$$

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{ij} - \bar{X}_{i.}\right)^2 = SST - SSTr \text{ , } \qquad df = \sum_i J_i - 1 = n - I$$

Test statistic value: $f = \frac{MSTr}{MSE}$, where $MSTr = \frac{SSTr}{I-1}$ and $MSE = \frac{SSE}{n-I}$

Rejection region: $f \geq F_{\alpha, I-1, n-I}$

TESTING FOR HOMOSCEDASTICITY

A rough rule of thumb works as follows. Estimate the smallest standard deviation of all factor levels (s_{min}) and the largest one (s_{max}). If

$$\frac{s_{max}}{s_{min}} \leq 2,$$

it is reasonable to assume equal variances. The value “2” may vary, depending on the literature consulted.

However, several tests are also available for testing for the assumption of equal variances / homoscedasticity (variance stability):

1. Bartlett's test: is relatively sensitive to departures from normality, therefore not preferable – but still being used quite often

TESTING FOR HOMOSCEDASTICITY

2. Levene's test: more robust to departures from normality. Basically, the test consists of a 1-way-ANOVA itself with response variable

   $$Z_{ij} = \left|X_{ij} - \bar{X}_{i.}\right|$$

   Levene’s tests produces a statistic *W*, which is similar the common *F* statistic from an ANOVA

3. Brown–Forsythe test: also termed Brown-Forsythe modification of Levene’s test. Even more robust to outliers / departures from normality, $\bar{X}_{i.}$ is replaced by the median $\tilde{X}_{i.}$

4. Fligner-Killeen test: non-parametric test, highly robust against departures from normality. Excellent choice when a) data are non-normally distributed or b) outliers are present and cannot be removed / replaced. Bases on a χ^2-type statistic

The null hypothesis of all tests is homoscedasticity!

fligner.test
fligner.test

TESTING FOR HOMOSCEDASTICITY – IMPLEMENTATION IN R

All four tests for homoscedasticity are available in R:

- Bartlett's test: use the function `bartlett.test`
- Levene's test: use the function `leveneTest`
- Brown–Forsythe test: via the function `leveneTest` as well
- Fligner-Killeen test: use the function `fligner.test`

As before: if you are not sure how a test works, try a simple example.

`- script on screen -`

TESTING FOR THE NORMALITY ASSUMPTION

As before (see section on *t*-tests), you may use:

- The test of Shapiro-Wilk for sample sizes up to 5000 (currently)
- The Anderson–Darling test (or others) for greater sample sizes

There are two principle approaches for testing normality:

1. Test each sample for normality. Attention: here you might quickly run into the common problems related to multiple comparisons
2. Test the residuals (estimated true errors) of the model for normality. These are given by subtracting predicted from observed values:

$$\hat{\varepsilon}_{ij}(= e_{ij}) = x_{ij} - \hat{x}_{ij} = x_{ij} - \hat{\mu} - \hat{\alpha}_i$$

TESTING FOR THE NORMALITY ASSUMPTION

Remarks concerning Approach 2.:

- this testing principle also works well for ANOVAs with more than one factor, interactions, repeated measures,…
- the underlying idea is simple: suppose that the errors belonging to (at least) one factor level are not Gaussian. Then, joining them with more errors from other factor levels that are Gaussian does not make the resulting total errors Gaussian
- There are counterexamples showing that the previous statement is not true in general. However, this is rare in practical situation

Example:

`- script on screen – or exercise for students -`

DATA TRANSFORMATION

The use of ANOVA-type methods can be invalidated by

1. substantial differences in the variances $\sigma_1^2, \ldots, \sigma_I^2$ or
2. strong departures from normality

In many books for practitioner, one finds statements such as “the ANOVA is robust to moderate departures from normality. This is true and has been confirmed by simulation studies. However, what is “moderate”?

If you are in doubt, consider alternatives such as, e.g.,

- non-parametric approaches
- GLMs / GAMLSS / …
- data transformations

DATA TRANSFORMATION

Example: Sometimes it happens that $V\left(X_{ij}\right) = \sigma_i^2 = g(\mu_i)$, a known function of the means; hence if H_0 is not true, the variances can be different.
If X_{ij} has a Poisson distribution with parameter λ_i, then $\mu_i = \lambda_i$ and $\sigma_i^2 = \lambda_i$, so $g(\mu_i) = \mu_i$.

In general:

- One may try to use a data transformation to stabilize the variances, remove skewness,… . For example, the log-transformation often helps
- Visual inspection of the data may help to determine an appropriate transformation