

Data Analysis in Sociology

Lecture 4. One-way analysis of variance.

February 2023

Some organizational things

- Data clinic on t-test this Wednesday at 18:10.
- Project 2 is approaching

Previously in this course

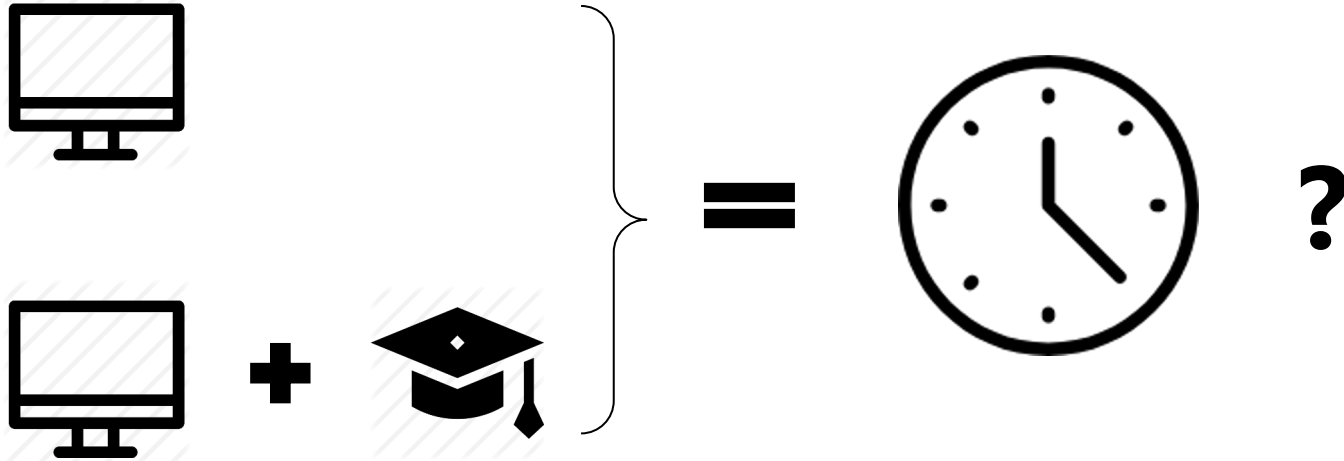
- Research hypothesis vs. statistical hypothesis
- Central tendency measures, measures of variability
- Basic graphs
- Chi-square test of independence
- The t-test

Previously in this course

	Binary	Categorical (3+ cats)	Numeric
Binary	Chi-square test (Yates correction)	Chi-square test	T-test MW U-test
Categorical (3+ cats)	Chi-square test	Chi-square test	
Numeric	T-test MW U-test		

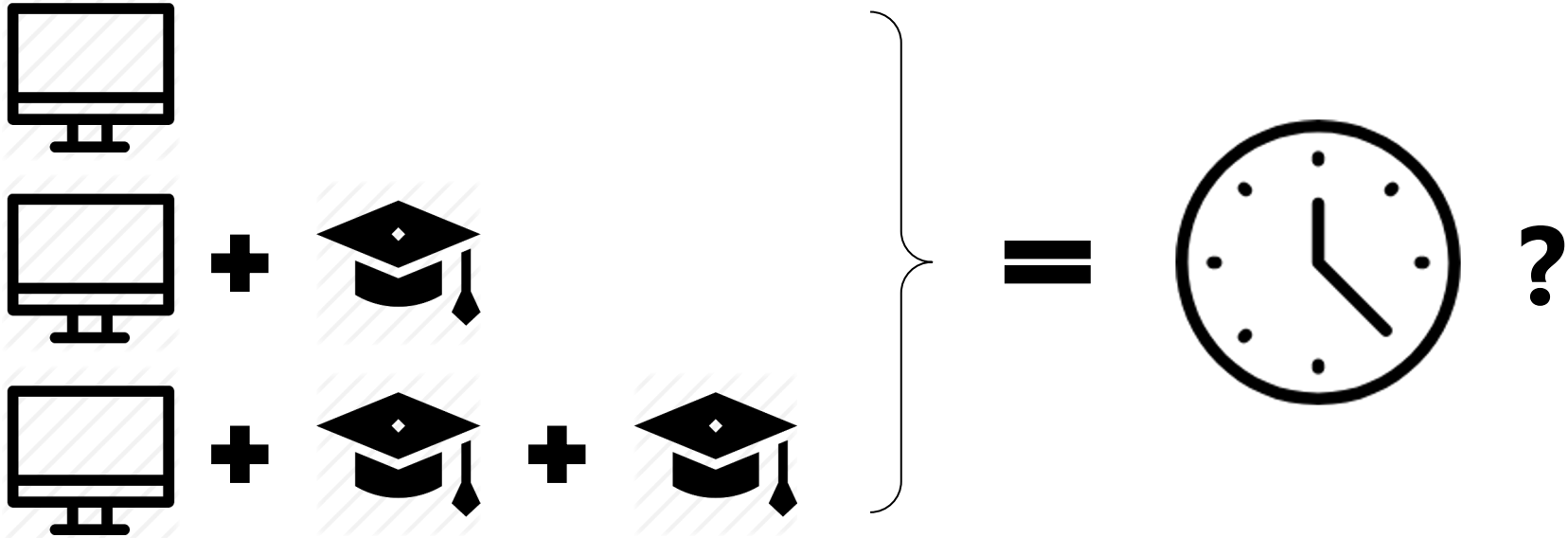
Can we solve all the questions of comparison now?

Ex. 1: Do programmers with higher education work the same hours per week as those without higher education?



Can we solve all the questions of comparison now?

Ex. 2: Do programmers with education - less than higher, higher, and postgraduate - work the same hours?



Ex. 2: Do programmers with education - less than higher, higher, and postgraduate - work the same hours?

! This problem should not be solved using the t-test, to avoid *familywise error* (FWER) resulting from multiple comparisons (=multiple null hypotheses) tested on the same data.

When multiple statistical tests are performed on the same data, some of them will have p-values less than α purely by chance, even if all null hypotheses are true (= *false-positive results*).

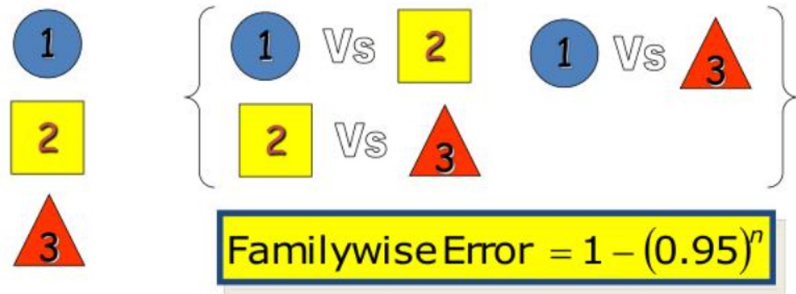
#familywise_error #multiple_comparison_problem

https://en.wikipedia.org/wiki/Family-wise_error_rate

https://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method

Ex. 2: Do programmers with education - less than higher, higher, and postgraduate - work the same hours?

Imagine three groups ($\alpha=0.05$) :



The probability of observing at least one significant result (at least one p -value < 0.05) just due to chance is: $1 - p(\text{no. of significant results}) = 1 - (1 - 0.05)^3 = 0.143$

So, **with as few as 3 tests being considered, we already have a 14.3% chance of observing at least one false-positive result**, even if all of the tests are actually not significant.

with 10 groups, 45 comparisons, this probability of false-positives raises to 90% (!)

Picture: Andy Field <https://www.youtube.com/watch?v=SULO2-gjZoY>

Ex. 2: Do programmers with education - less than higher, higher, and postgraduate - work the same hours?

So, what to do? **ANOVA** generalises beyond the two groups.

ANOVA is used to assess whether the mean of the outcome variable is different for different (3+) levels of a categorical variable (called a 'factor'):

H_0 : $\mu_{no_edu} = \mu_{higher} = \mu_{postgrad}$ (\Rightarrow the 3 education levels are equal in terms of mean working hours).

H_1 : **at least one** mean is different (\Rightarrow at least one education level is different from at least another one in terms of working hours)

This lecture

- Idea of one-way analysis of variance (ANOVA)
- Logic of the test
 - Visualization
 - F-test
 - Post hoc comparisons
- Assumptions of ANOVA
- Non-parametric Kruskal-Wallis ANOVA
- Effect size

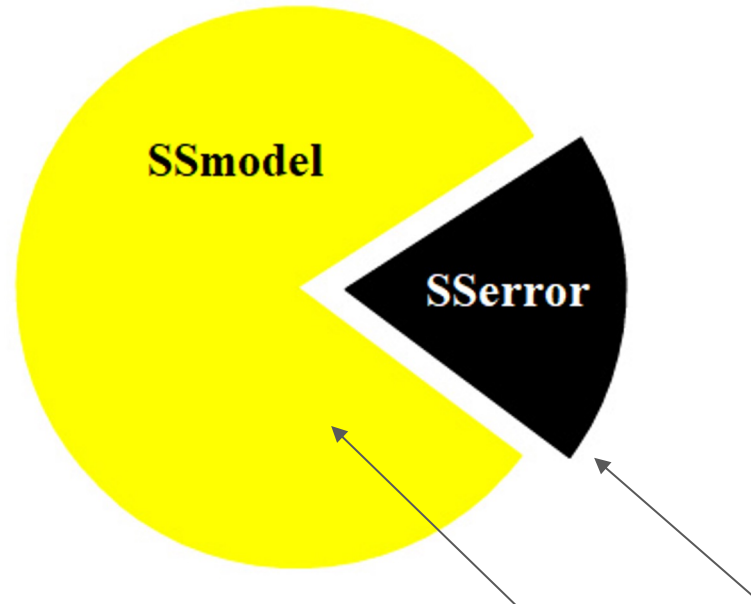
Idea of ANOVA

Analysis of variance is a widespread technique rooted in experimental research: control group vs. treatment 1, control vs. treatment 2, etc.

Factors are categorical, the **outcome** is always a continuous variable.

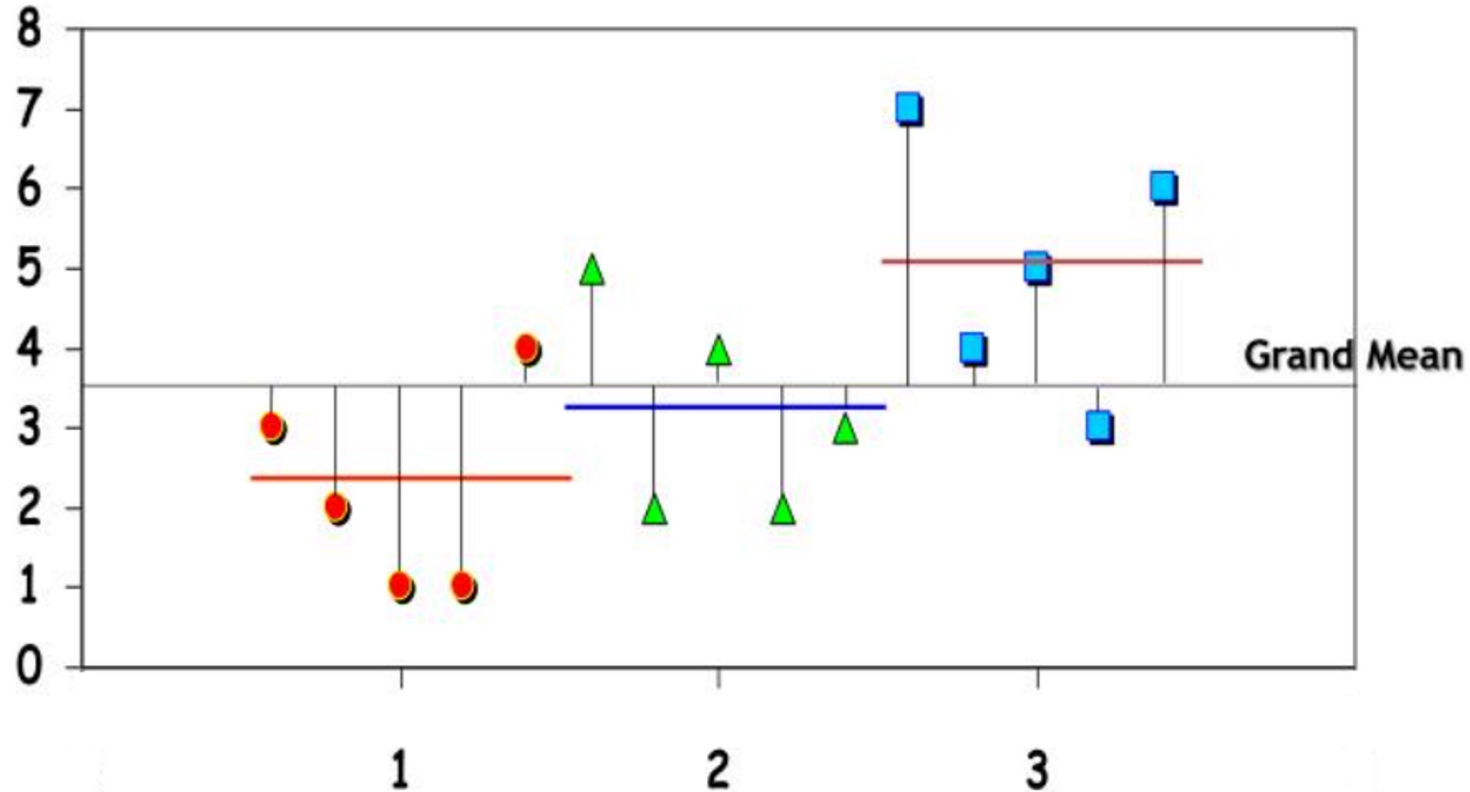
Ex.: Do bachelor students of all education programmes have the same ideal salary?
factor = Bachelor's education programme,
outcome = ideal salary

Analysis of variance: **total variance** = model + error



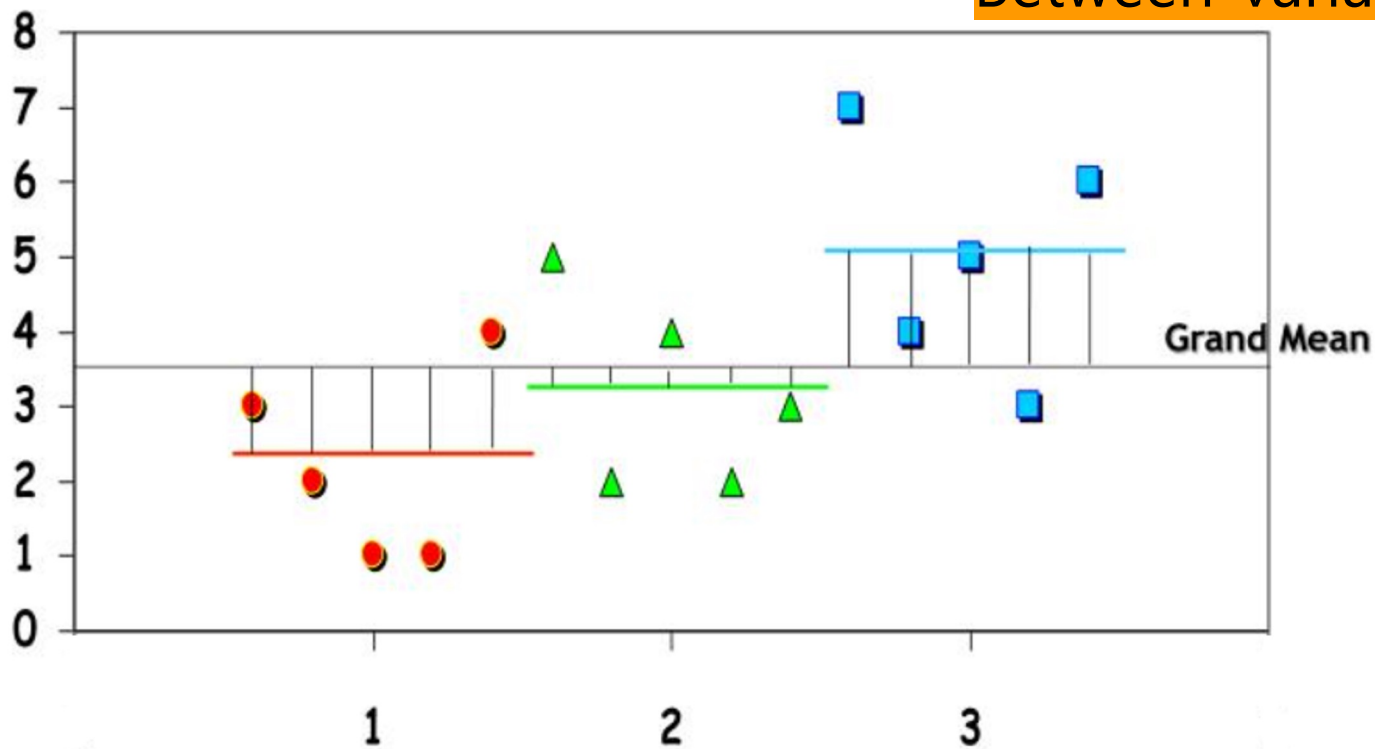
Variance of the outcome = group-related + individual

The distance from data points to the **grand mean** shows the *total variance*:



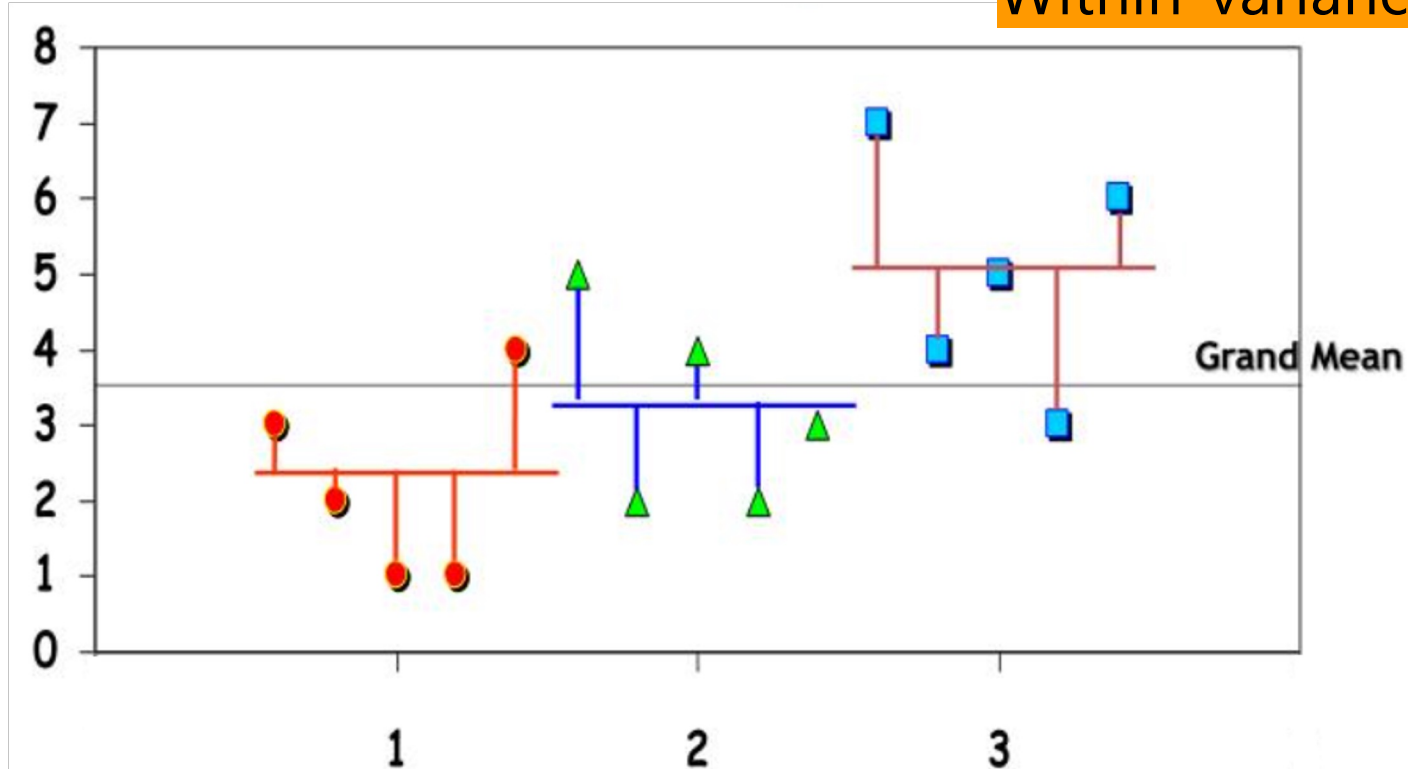
If replaced with **group means**, the distances are the *model variance*:

Between-variance



The difference between individual values and group means is the *error (residual) variance*:

Within-variance



One-way analysis of variance (one-way ANOVA):

'One-way' means there is only one factor, i.e., a grouping variable with more than 2 levels

The statistic is called the **F-ratio**: it is the ratio of variance between the groups to the variance within the groups.

If the groups are consistently different, F would be large.

A small F-value means we cannot reject H_0 , but we do not prove that it is true or that H_1 is true.

Purpose of ANOVA testing

...is to compare the means from two or more groups to see whether they are so far apart that the observed differences cannot all reasonably be attributed to sampling variability:

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

Formulas

Data = Model + Error

Model here is the differences between the group means and the grand mean (SSm).

Error corresponds to the differences between the *observed* values and the *group* means (SSe).

Mean sum of squares = SS / df. There are two degrees of freedom: group + individual

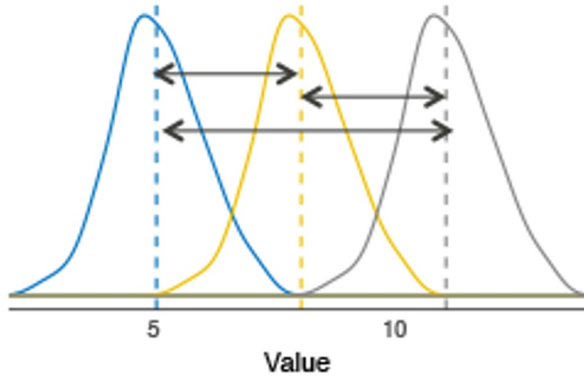
- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

$$F = \frac{MSG}{MSE}$$

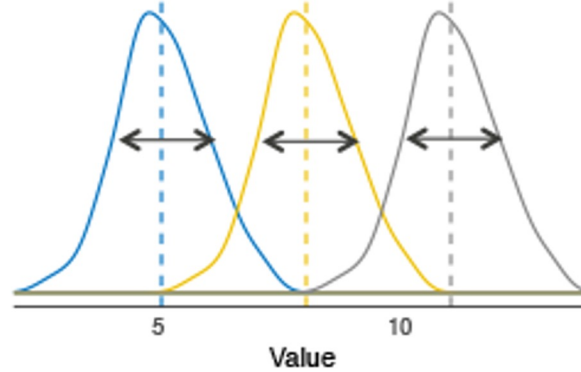
F-ratio = Mean sum of squares between groups / Mean sum of squares within groups

$F = MSm / MSe$

A
Between-group variation
(i.e. Differences among group means)



B
Within-group variation
(i.e. Variability within each group)



If the distance **between groups** is much larger than the distances **within groups**, then **the samples come from populations with different means.**

Fig: <https://www.datanovia.com/en/lessons/anova-in-r/>

What F-ratio shows

The F-test shows the ratio between how much variability is due to the group or experimental manipulation vs. due to individual difference.

Experiment tests whether some treatment works better than the control group.

Group comparison tests whether the groups are more different between than within.

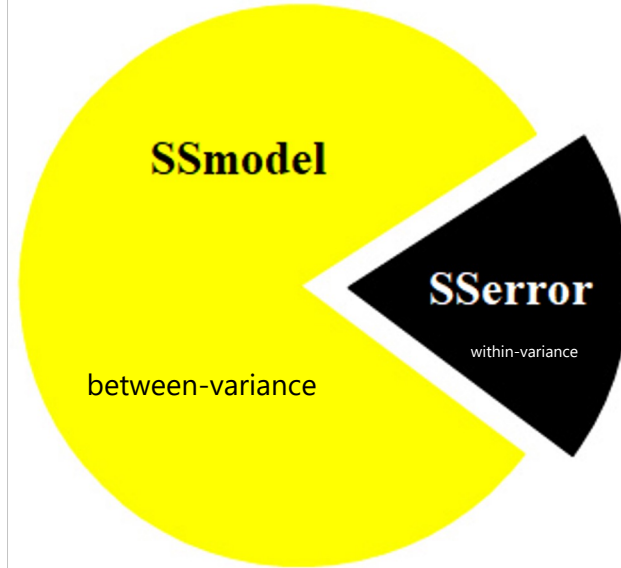
To sum up, $F = \text{systematic difference} / \text{unsystematic difference}$

! ANOVA is called 'an *omnibus* (=global) test': it tests for overall differences between groups, but it does not show which pairs are exactly different.

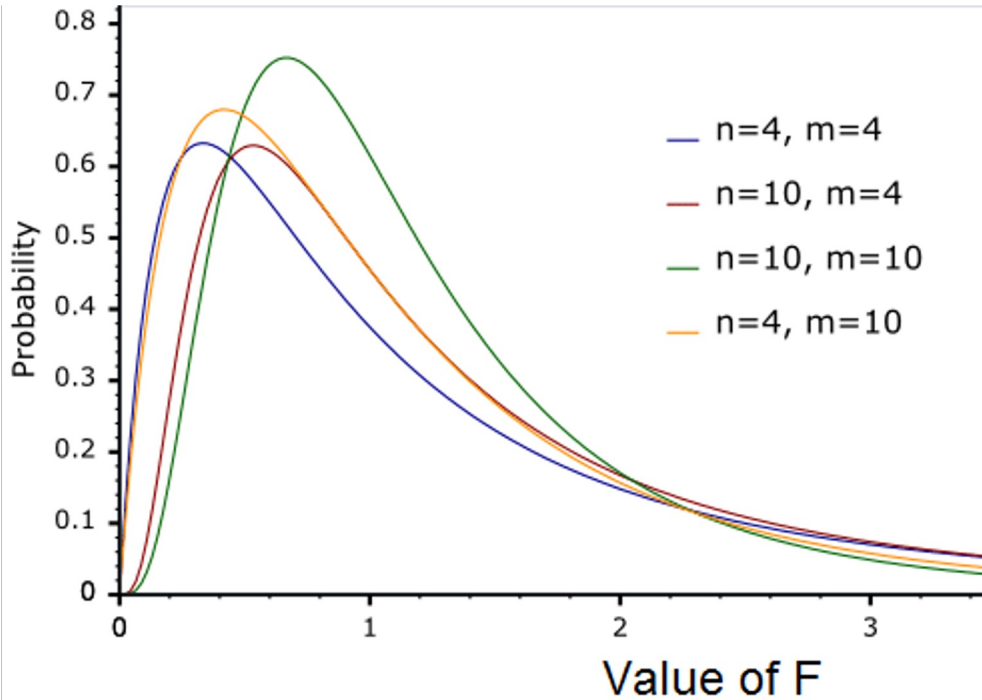
What F-ratio shows

If an F-ratio is **significant**, *the model explains more variance than it can't explain.*

If the between-variance is significantly larger than the within-variance, the group means are declared to be different.



F-distribution: one tail and two df's

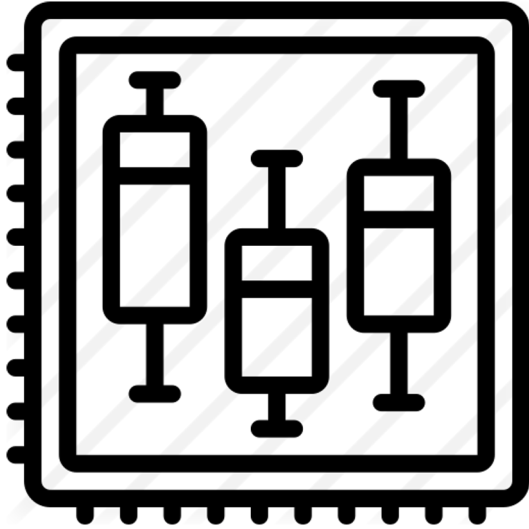


p-value in the F-test is the probability of at least as large a ratio between the “between group” and “within group” variability, if in fact the means of all groups are equal ($H_0 = \text{true}$)

Logic of the Test

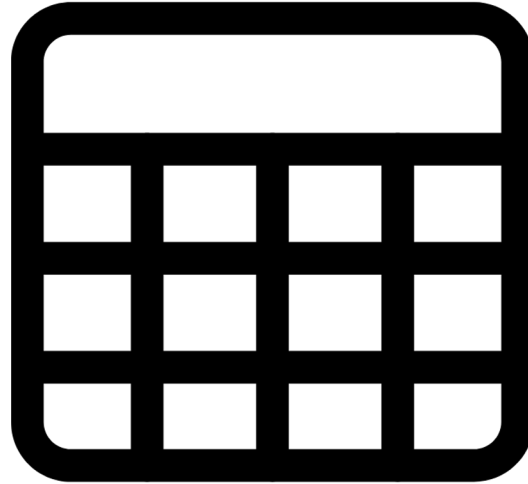
Visualization

Show a boxplot



+

Examine group sizes, means and SDs



F-test

After visualization,

Test for the assumptions to choose a proper F-test formula:

- Independence of observations (one unit can count once)
- Normality of the continuous variable
- Equality of variances

F-test functions:

```
oneway.test(y ~ x, var.equal=T/F) # Welch's F-test
```

```
aov.out <- aov(y ~ x); summary(aov.out) # has a full summary
```

What's after? Which group is unlike the others?

If the **null hypothesis is not rejected** ($p\text{-value} \geq 0.05$), it means that we fail to reject H_0 that the observed differences in sample means are attributable to sampling variability (or chance).

Write up the results.

If the **null hypothesis is rejected** ($p\text{-value} < 0.05$), the data provide convincing evidence that at least one mean is different (but we can't tell which one).

Use a follow-up, *post hoc* test (*Latin* "after this"), which are tests with a correction for multiple pairwise-comparison.

Post hoc comparisons

Post-hoc tests take into account that multiple tests are done and deal with the problem by *adjusting alpha (α) in some way*. There are dozens of them, all good for their purposes (see the literature).

When running an experiment, use planned comparisons.

When comparing group means, use a post hoc test:

- Tukey Honestly Signif. Differences when variances are equal
- Games-Howell when variances are unequal
- Bonferroni as a safe option (efficient but Type II raises) $\alpha^* = \alpha/K$
- Dunnett's t (compares all treatment groups vs. control)
- Benjamini-Hochberg for false discoveries, etc.

Assumptions of ANOVA in detail

One-way ANOVA assumptions

1. Independence:
 - a. each item is randomly selected
 - b. there is no relationship between the groups (e.g. not the same people across groups)
2. Variable types (continuous outcome, categorical predictor)
3. Groups have equal (homogeneous) variances
4. Normality: experimental errors are normally distributed OR the continuous variable is normally distributed by each group.

Independence

Make sure you deal with one-off observations of the subject.

One-way ANOVA is a 'between-subject' design, comparing different units.

! No 'before-after' design (there are other types of ANOVA for that).

Variable types

The continuous variable is the 'dependent variable' (in experiment) or the 'outcome' (in observational data).

The continuous variable always goes first in the ANOVA formula: $y \sim x$.

It must be of numeric or double class in R.

The categorical variable is the 'grouping' factor.

It must be read by R as a factor.

Pay attention to your interpretations and conclusions as the outcome is the continuous variable.

Equal variances

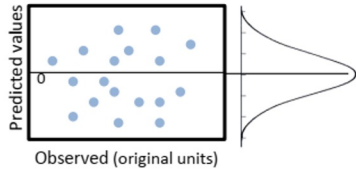
Can be tested with:

- Levene's test (like in the t-test assumptions check)
- Bartlett's test, when normality holds (like in the t-test)
- Residual plot: the **residuals versus fitted values plot**

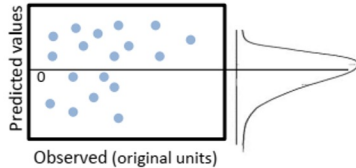
Equal variances

Testing for Equal Variances – Residual Plots

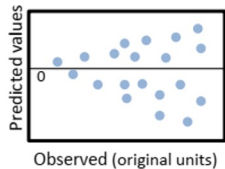
Residual plots in R (multiple plots):
`plot(lm(YIELD~VARIETY))` (2nd plot)



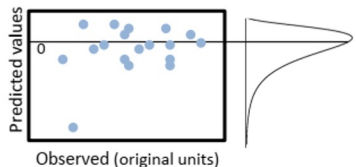
- **NORMAL distribution:** equal number of points along observed
- **EQUAL variances:** equal spread on either side of the mean_{predicted value=0}
- **Good to go!**



- **NON-NORMAL distribution:** unequal number of points along observed
- **EQUAL variances:** equal spread on either side of the mean_{predicted value=0}
- **Optional to fix**



- **NORMAL/NON NORMAL:** look at histogram or test
- **UNEQUAL variances:** cone shape – away from or towards zero
- **This needs to be fixed for ANOVA** (transformations)



- **OUTLIERS:** points that deviate from the majority of data points
- **This needs to be fixed for ANOVA** (transformations or removal)

More relevant
when group sizes
are different

Equal variances

If variances are equal and normality holds, use `aov()`

If variances are unequal with normality, use `oneway.test()`

If there are many outliers and normality is out of question, use Kruskal-Wallis ANOVA `kruskal.test()` or `kruskal_test()`

Normality

Two ways to go:

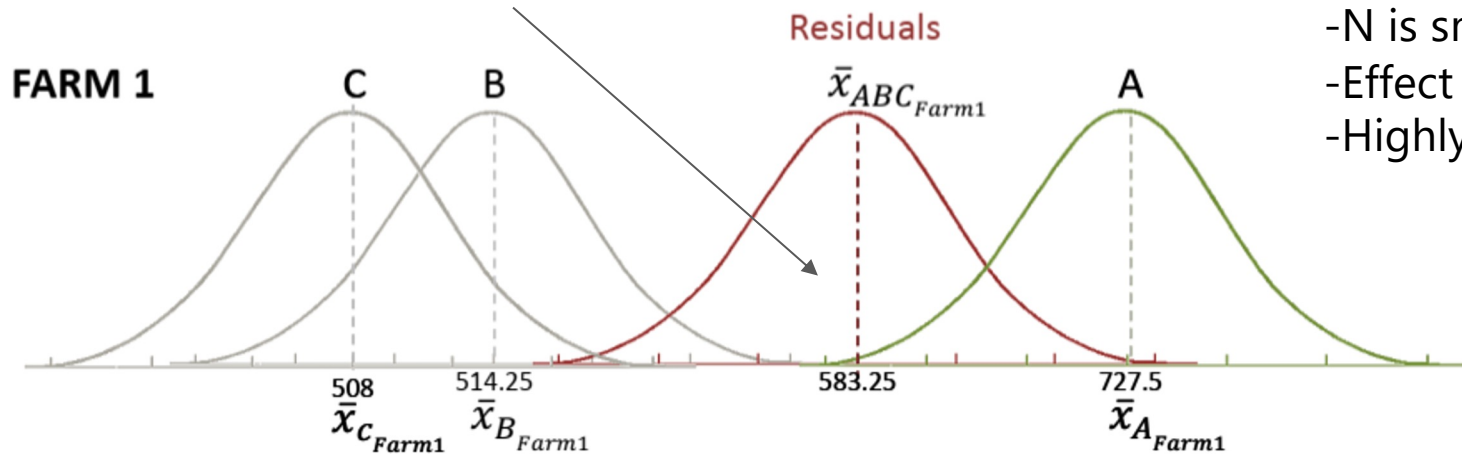
- 1) either check normality by group — when groups are few and group sizes are big:
 - a) Use histograms,
 - b) QQ-plots, or
 - c) formal tests (better for smaller samples, e.g., <50 but there is no strict rule)
- 2) or analyse the residuals of the outcome — when there are many small groups:

```
aov.out <- aov(y ~ x)  
hist(residuals(aov.out)); shapiro.test(residuals(aov.out))
```

Normally distributed residuals:

Determine this by looking at the residuals of your sample:

residuals : subtract overall mean from the sample means



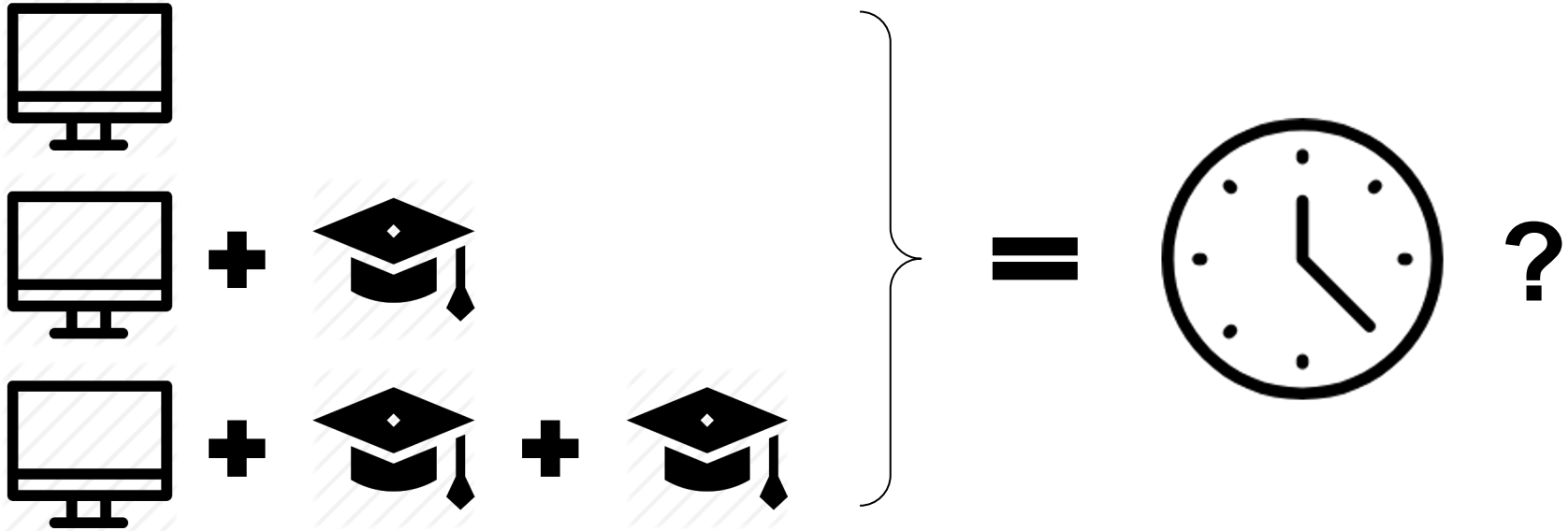
More relevant for
small groups

Important when:

- N is small,
- Effect size is small
- Highly skewed Y

How can we answer that question now?

Ex. 2: Do programmers with education - less than higher, higher, and postgraduate - work the same hours?



Overview of one-way ANOVA

Procedure:

- 1) Visualise the differences with a boxplot
- 2) Analyze descriptives by group
- 3) Check for normality
- 4) Check for equality of variances
- 5) Run the proper F-test
- 6) If statistically significant, run a post hoc test (Tukey HSD/Games-Howell/Bonferroni)
- 7) Estimate the effect size
- 8) Report and rejoice!

Non-parametric Kruskal-Wallis ANOVA

Non-parametric Kruskal-Wallis ANOVA: use when assumptions do not hold

Replaces absolute values with ranks. Has a different null hypothesis:

In the general case, the KW ANOVA tests whether 'at least one sample stochastically dominates one other sample' (compares their variances).

However, if the distributions have identical shapes and scales, except for the differences in medians, then H_0 is that the medians of all groups are equal, and H_1 is that at least one population median of one group is different from the population median of at least one other group.

Games-Howell or Dunn post hoc test can be used as a follow-up.

Effect size

Effect size is the share of variance explained by the model

- The partial ***Eta-squared*** shows the proportion of variance that is attributable to the group:
 - Sums of the partial Eta-squared (η^2) values can be greater than 1.00.
 - It estimates the effect size for the sample and always overestimates it for the population (the bias is smaller if the sample is large).
- An estimate of the amount of variance accounted for in the population is the ***omega-squared***:
 - The formula is: $\omega^2 = (SS_{\text{effect}} - (df_{\text{effect}} * (MS_{\text{error}})) / MS_{\text{error}} + SS_{\text{total}}$
 - Because eta-squared is a sample estimate and omega-squared is a population estimate, omega-squared is always going to be smaller than eta-squared or partial eta-squared.
 - Omega-squared is an unbiased effect size for small sample sizes.
- Rules of thumb for the interpretation of both: 0.01-0.06 = 'small'; 0.06-0.14 = 'medium'; > 0.14 = 'large'.
 - `library(sjstats); anova_stats(aov.out)`

Linear model framework*

Assumptions of the following tests are similar as they belong to the same framework:

A framework means a common approach, a set of guiding principles.

t-test assumptions:

- measurement level, normality (no outliers), homogeneity of variances

ANOVA assumptions:

- measurement level, normality (no outliers), homogeneity of variances

General linear model is a common framework

Data = model + error:

“Data” here is any continuous variable

“Model” is the explaining variable(s)

ANOVA can be thought of *as a special case of linear regression*.

Linear regression and correlation are also part of this framework.







Additionally, they share the assumption of linear relationship (will be discussed later).

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying

notebook: <https://lindeloev.github.io/tests-as-linear>

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: $\text{lm}(y \sim 1 + x)$	y is independent of x P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	$\text{lm}(y \sim 1)$ $\text{lm}(\text{signed_rank}(y) \sim 1)$	✓ for N > 14	One number (intercept, i.e., the mean) predicts y . - (Same, but it predicts the <i>signed rank</i> of y .)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y1, y2, paired=TRUE) wilcox.test(y1, y2, paired=TRUE)	$\text{lm}(y_2 - y_1 \sim 1)$ $\text{lm}(\text{signed_rank}(y_2 - y_1) \sim 1)$	✓ for N > 14	One intercept predicts the pairwise y ₂ - y ₁ differences. - (Same, but it predicts the <i>signed rank</i> of y ₂ - y ₁ .)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	$\text{lm}(y \sim 1 + x)$ $\text{lm}(\text{rank}(y) \sim 1 + \text{rank}(x))$	✓ for N > 10	One intercept plus x multiplied by a number (slope) predicts y . - (Same, but with <i>ranked x</i> and y)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y1, y2, var.equal=TRUE) t.test(y1, y2, var.equal=FALSE) wilcox.test(y1, y2)	$\text{lm}(y \sim 1 + G_2)^A$ $\text{gls}(y \sim 1 + G_2, \text{weights} = \dots^B)^A$ $\text{lm}(\text{signed_rank}(y) \sim 1 + G_2)^A$	✓ ✓ for N > 11	An intercept for group 1 (plus a difference if group 2) predicts y . - (Same, but with one variance <i>per group</i> instead of one common.) - (Same, but it predicts the <i>signed rank</i> of y .)	
Multiple regression: $\text{lm}(y \sim 1 + x_1 + x_2 + \dots)$	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_N)^A$ $\text{lm}(\text{rank}(y) \sim 1 + G_2 + G_3 + \dots + G_N)^A$	✓ for N > 11	An intercept for group 1 (plus a difference if group ≠ 1) predicts y . - (Same, but it predicts the <i>rank</i> of y .)	
	P: One-way ANCOVA	aov(y ~ group + x)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_N + x)^A$	✓	- (Same, but plus a slope on x .) <i>Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.</i>	
	P: Two-way ANOVA	aov(y ~ group * sex)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_N + S_2 + S_3 + \dots + S_K + G_2^*S_2 + G_3^*S_3 + \dots + G_N^*S_K)$	✓	Interaction term: changing sex changes the y ~ group parameters. <i>Note: $G_{2:10}$ is an indicator (0 or 1) for each non-intercept levels of the group variable. Similarly for $S_{2:10}$ for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the group * sex interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.</i>	[Coming]
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	Equivalent log-linear model $\text{glm}(y \sim 1 + G_2 + G_3 + \dots + G_N + S_2 + S_3 + \dots + S_K + G_2^*S_2 + G_3^*S_3 + \dots + G_N^*S_K, \text{family} = \dots)^A$	✓	Interaction term: (Same as Two-way ANOVA.) <i>Note: Run glm using the following arguments: glm(model, family=poisson()) As linear-model, the Chi-square test is $\log(y) = \log(N) + \log(a_i) + \log(\beta_j) + \log(a_i\beta_j)$ where a_i and β_j are proportions. See more info in the accompanying notebook.</i>	Same as Two-way ANOVA
	N: Goodness of fit	chisq.test(y)	$\text{glm}(y \sim 1 + G_2 + G_3 + \dots + G_N, \text{family} = \dots)^A$	✓	(Same as One-way ANOVA and see Chi-Square note.)	1W-ANOVA

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation $y \sim 1 + x$ is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are **"dummy coded" indicator variables** (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. `lm` requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at <https://lindeloev.github.io/tests-as-linear>.

<https://lindeloev.github.io/tests-as-linear/>

^A See the note to the two-way ANOVA for explanation of the notation.

^B Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.



Jonas Kristoffer Lindeløv

<https://lindeloev.net>

Summary

1. When we run too many t-tests on the same data, Type I Error rate increases ("familywise error"). This issue is resolved by using a modified significance level ("corrections").
2. ANOVA allows testing the mean differences between 3+ groups. It is an omnibus test. If significant, use post hoc tests for pairs.
3. When variances are unequal, use Welch's ANOVA. When the normality assumption does not hold, use Kruskal-Wallis ANOVA.
4. Both the t-test and ANOVA (and linear regression as well) are cases of the linear model.

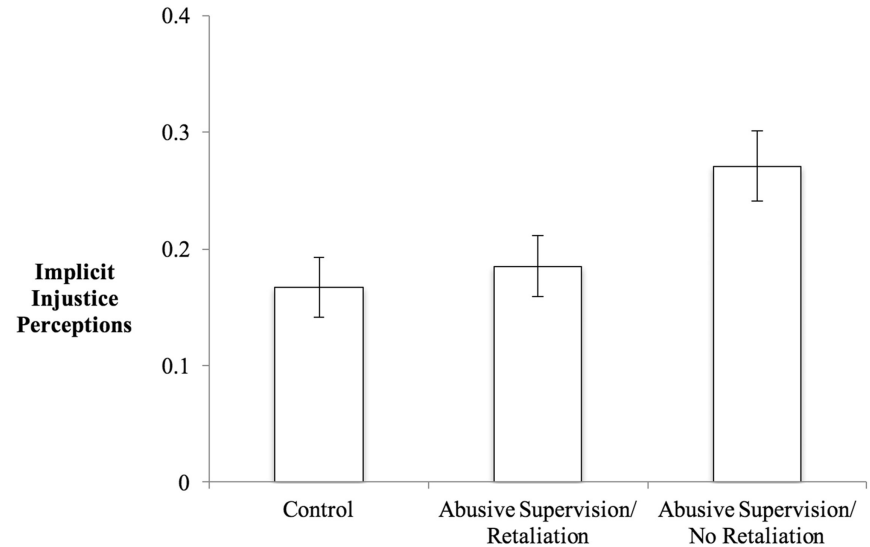
Summary

	Binary	Categorical (3+ cats)	Numeric
Binary	Chi-square test (Yates correction)	Chi-square test	T-test MW U-test
Categorical (3+ cats)	Chi-square test	Chi-square test	Anova K-W test
Numeric	T-test MW U-test	Anova K-W test	

Anova in IgNobel prize winning articles

Liang, Lindie & Brown, Douglas & Lian, Huiwen & Hanig, Samuel & Ferris, D. & Keeping, Lisa. (2018). Righting a wrong: Retaliation on a voodoo doll symbolizing an abusive supervisor restores justice. *The Leadership Quarterly*. 29. 10.1016/j.leaqua.2018.01.004.

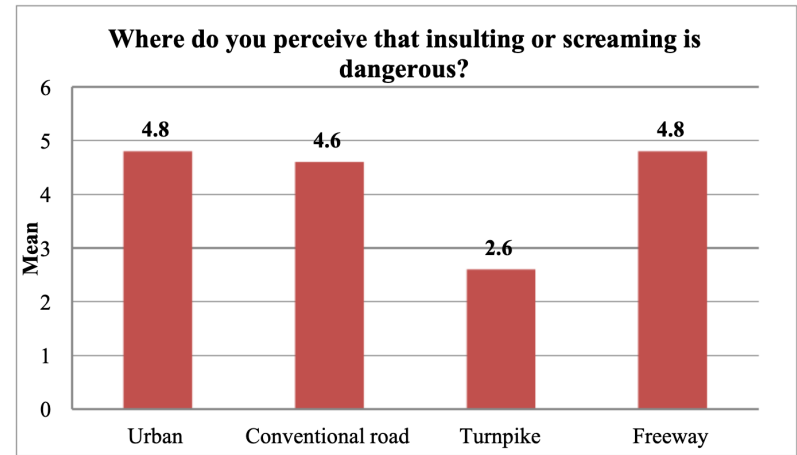
"To test the hypothesis that not retaliating following abusive supervision engenders greater perceptions of injustice compared to having the opportunity to retaliate ([Hypothesis 1](#)), we conducted a One-Way Analysis of Variance (ANOVA). In support of [Hypothesis 1](#), there was a significant effect of condition (i.e., control condition, abusive supervision/retaliation, and abusive supervision/no retaliation) on participants' implicit injustice perceptions [$F(2, 192) = 3.81, p = .02, \eta^2 = 0.04$]."



Anova in IgNobel prize winning articles

Alonso, F., Esteban, C., Serge, A., & Ballestar, M. L. (2017). Shouting and cursing while driving: Frequency, reasons, perceived risk and punishment. *Journal of Sociology and Anthropology*, 1(1), 1-7.

"If we analyze the relationship between the risk perceived by drivers in each one of the studied behaviors as a cause of accidents and in the type of road they use for their trips, those who mainly do urban journeys attributed, on average, higher scores to all the behavior, while the average scores of other participants only differ significantly in the case of shouting or insulting while driving $F(3,1086)=7.29$; $p<.001$ (see [Figure 4](#) and [Table 2](#))."



Really Helpful Reading

1. **OpenIntro Stats - ANOVA**

<https://docs.google.com/presentation/d/1zFclgaHK588gAo9qCla1tyiLlkDzvyf0XqsVnWD0po/edit>

2. **ANOVA in R** <https://towardsdatascience.com/anova-in-r-4a4a4edc9448> -> check out the 'report' library mentioned there

3. **More about the Kruskal-Wallis ANOVA**

https://influentialpoints.com/Training/Kruskal-Wallis_ANOVA_use_and_misuse.htm

4. **Overview of post hoc test procedures**

http://web.pdx.edu/~newsomj/uvclass/ho_posthoc.pdf

5. **Video by Andy Field, 18+** <https://www.youtube.com/watch?v=SULO2-gjZoY>

6. **2 Anova examples (in Russian)** <https://rpubs.com/ovolchenko/anovaexamples>

Sample R output for `aov()`

Analysis of Variance Table

$df_G = k - 1$

Response: Values

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	1	60	60.000	64.444	5.503e-11 ***
Residuals	58	54	0.931		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$df_E = n - 1 - df_G$

MSmodel

MSerror