Statistics

Marco Caserta
marco.caserta@ie.edu

IE University

One-Way Analysis of Variance

ANOVA

- Design of Experiment *vs.* Observational Study (ANOVA *vs.* Regression)
- Test the null hypothesis that several independent population means are equal

Examples:

- You want to prescribe the smallest dosage of a drug that is effective for treating fever. You select five possible dosages and administer them to people with high fevers. You record the time in minutes until the temperature returns to normal. Do average times for fever reduction differ for the five dosages?
- How do promotions affect production sales? As a marketing manager, you assign one of three display types to each of 100 stores and record the change in euro sales for your product from the previous week. Are all three display types equally effective?
- You have four possible ways of filling containers on a production line. You randomly assign 50 workers to each process and determine the weight of the resulting boxes. You want to know whether there is a relation between the final weight of the boxes and the method used to fill them.

One-Way Analysis of Variance: Setting

- Response : Variable of interest of the experiment (dependent variable)
- Factor : Factor that might affect the response (independent variable)
- Factor Level : Possible values for the factor
- Treatments : Factor levels combinations (same as Factor Level if only one factor)
- Experimental Unit : Objects on which response and factors are observed

Example: Do TV shows with violence and sex impair memory for commercials?

A sample of 324 adults, after viewing a program, was scored on his/her recall of brand names in commercials.

- Response: Recall factor (0 to 9)
- Factor : Movie type
- Factor Level : Sex, violence, neutral
- Treatment : Assigning a factor level to a viewer
- Experimental Unit : Each viewer

One-Way Analysis of Variance: Key Idea

1. Compare the variance *between* different groups (movie types: sex, violence, neutral), believed to be due to the independent variable, with the variability *within* each of the groups, believed to be due to chance.
2. Compute the F ratio, as variance between groups divided by variance within group
3. A large F ratio indicates that there is more variability due to group type (type of movie watched) than due to error, *i.e.*, random within group variability
4. If F is large enough, we reject the null hypothesis, *i.e.*, that the population means are equal. However, we do not know which groups differ.
5. Post-hoc tests are used to determine which groups are actually different

Two types of one-way ANOVA

i. Between-groups ANOVA, when participants in each group are different (*completely randomized design*)

ii. Repeated-measures ANOVA, when the same participants are measures under different conditions (*randomized block design*)

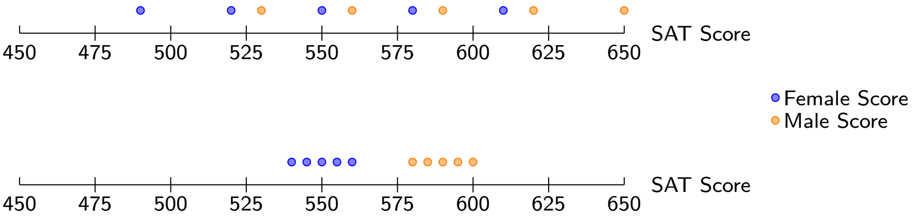
Completely Randomized Design: Single Factor

Independent random samples of experimental units are selected for each treatment. The objective is to compare the treatment means:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \text{at least two of the } k \text{ treatment means differ}$$

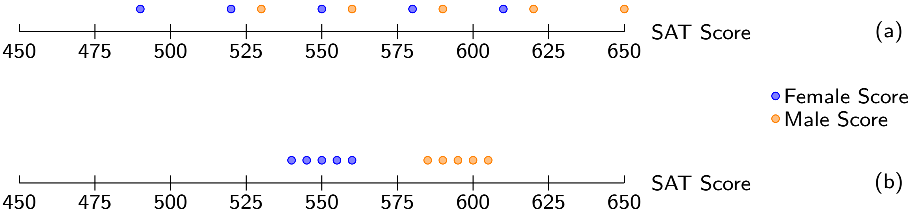
We use the unbiased estimators sample means $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$

Example



Intrablock Variability

Example



- $\overline{x}_M^a = 590, \overline{x}_F^a = 550$; $\overline{x}_M^b = 590, \overline{x}_F^b = 550$; Can we conclude that $\mu_M > \mu_F$?
- In case (b), intrablock variability is so small that it is quite likely that $\mu_M > \mu_F$
- In case (a), the variability within each factor (or treatment) is large and, therefore, we do not know whether $\overline{x}_M > \overline{x}_F$ because $\mu_M > \mu_F$ or just because of the "specific" sample we have selected
- The answer to the test depends on the ratio between the variability of each treatment (male, female) with respect to the total mean and the variability within each treatment

z/t TEST VS. ANOVA - PURPOSE

z/t test

Compare means from TWO groups to see whether they are so far apart that the observed difference cannot reasonably be attributed to sampling variability.

$$H_0 : \mu_1 = \mu_2$$

ANOVA

so far apart that the observed differences cannot all reasonably be Compare the means from TWO OR MORE groups to see whether they are attributed to sampling variability.

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

z/t test vs. ANOVA - Method

z/t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

ANOVA

Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

z/t test vs. ANOVA - Method

z/t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

ANOVA

Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

- Large test statistics lead to small p-values.
- If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal.

z/t test vs. ANOVA

- With only two groups t-test and ANOVA are equivalent, but only if we use a pooled standard variance in the denominator of the test statistic.

z/t TEST VS. ANOVA

- With only two groups t-test and ANOVA are equivalent, but only if we use a pooled standard variance in the denominator of the test statistic.
- With more than two groups, ANOVA compares the sample means $\bar{x}_i$ to an overall GRAND MEAN $\bar{x}$.

ANOVA: Intuition & Formulas

SST	Sum of Squares for Treatments	$(\overline{x}_i - \overline{x})^2$	variability of each treatment
SSE	Sum of Squares for Errors	$(x_{ij} - \overline{x}_i)^2$	variability within each treatment

- SST : variability of the observations with respect to the overall mean due to treatment (SAT score differences are due to different treatment levels, *i.e.*, gender).
- SSE : even within the same treatment level, there is variability, thus there must be more to the story. We call this "error".

$$\begin{aligned} SST &= \sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{x}\right)^2 \\ SSE &= \sum_{j=1}^{n_1} (x_{1j} - \overline{x}_1)^2 + \sum_{j=1}^{n_2} (x_{2j} - \overline{x}_2)^2 + \ldots + \sum_{j=1}^{n_k} \left(x_{kj} - \overline{x}_k\right)^2 \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2 \qquad \text{with} \quad s^2 = \frac{(x_i - \overline{x})^2}{n-1} \end{aligned}$$

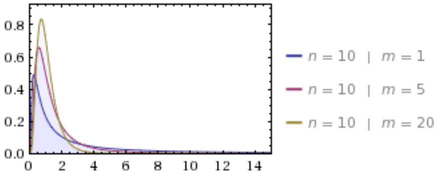
ANOVA: Intuition & Formulas

- To make the sum of squares comparable, we take the means by dividing them by the df:

$$MST = \frac{SST}{k-1}$$
$$MSE = \frac{SSE}{n-k}$$

F Statistic

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



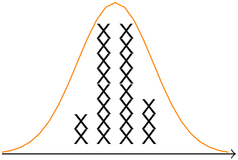
- The higher F, the better, since it increases the amount of variability captured/explained by the factor
- A high value of F indicates that other unknown factors are not too important in explaining the variability of the response
- High values of F lead to the rejection of H_0, when compared with $F_\alpha(k-1, n-k)$

ANOVA : Assumptions

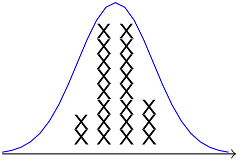
1. Probability distribution of population of response associated with each treatment should be approximately normal
2. Probability distribution of population of response associated with each treatment should have approximately the same variance HOMOSCEDASTICITY)
3. Samples of experimental units must be random and independent

ANOVA : Assumptions I

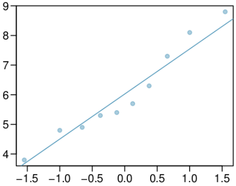
1. Normally distributed

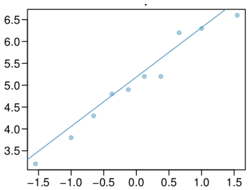


SAT Male



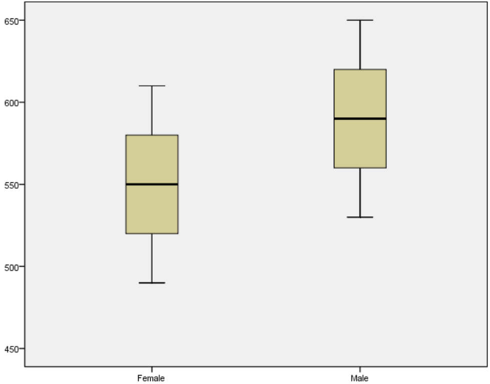
SAT Female





ANOVA : Assumptions II

2. Same variance (Homoscedasticity): Box Plot



3. Random and independent samples: It depends on the design of the experiment.

Steps of ANOVA for Completely Randomized Design

1. Verify that sampling is independent and randomized
2. Check assumptions of normality and equal variance
3. Create ANOVA table

Source	DF	SS	MS	F	p-value
Treatments	k-1	SST	MST	MST/MSE	$p(f > F)$
Error	n-k	SSE	MSE		
Total	n-1	SSTotal			

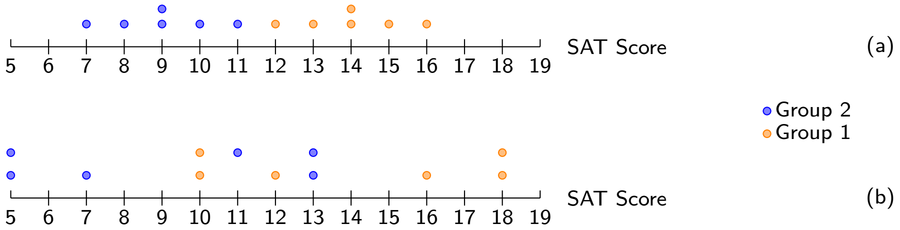
4. If F-test leads to conclusion that means differ (p-value below α):
   a. Compute Effect Size as proportion of variance of DV explained by the IV:

   $$\eta^2 = \frac{\text{SST}}{\text{SSTotal}}$$

   where $\eta^2 = 0.01$ implies small effect size, $\eta^2 = 0.06$ implies medium effect size, and $\eta^2 = 0.14$ implies large effect.
   b. Conduct a multiple comparison procedure to determine which treatments are statistically different
   c. Form confidence intervals for one or more treatment means
5. If F-test leads to non-rejection:
   a. We still cannot accept the null hypothesis, due to the "danger" of type II error
   b. Maybe the treatment means do differ, but other important factors account for the total variability. These factory might be inflating SSE (intrablock variability) and this are producing small F values. Either increase sample size or use a different experimental design.

ANOVA: Manual Computation

Example



SOLUTION:

- (a) SST = 75, SSE = 20, F = 75/2 = 37.5 $\Rightarrow$ Reject, since $F_{0.05}(1, 10) = 4.96)$
- (b) SST = 75, SSE = 144, F = 75/14.4 = 5.21 $\Rightarrow$ Reject since $F_{0.05}(1, 10) = 4.96)$

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is "which ones?"

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is "which ones?"
- We can do two sample t tests for differences in each possible pair of groups.

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is "which ones?"
- We can do two sample t tests for differences in each possible pair of groups.

Can you see any pitfalls with this approach?

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is "which ones?"
- We can do two sample t tests for differences in each possible pair of groups.

Can you see any pitfalls with this approach?

- When we run too many tests, the Type 1 Error rate increases.
- This issue is resolved by using a modified significance level.

Multiple comparisons

- The scenario of testing many pairs of groups is called multiple comparisons.

Multiple comparisons

- The scenario of testing many pairs of groups is called MULTIPLE COMPARISONS.
- The BONFERRONI CORRECTION suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^\star = \alpha/K$$

where K is the number of comparisons being considered.

Multiple comparisons

- The scenario of testing many pairs of groups is called multiple comparisons.
- The Bonferroni correction suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^{\star} = \alpha / K$$

  where K is the number of comparisons being considered.
- If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

Which means differ? (cont.)

If the ANOVA assumption of equal variability across groups is satisfied, we can use the data from all groups to estimate variability:

- Estimate any within-group standard deviation with $\sqrt{MSE}$, which is s_{pooled}
- Use the error degrees of freedom, $n - k$, for t-distributions

Difference in two means: after ANOVA

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

Anova Example

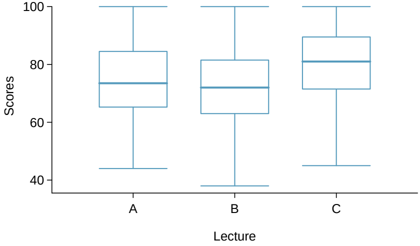
Statistics Courses

Consider a statistics dept that runs three lectures of a statistics course. We want to determine whether there are statistically significant differences in the exam scores of these three classes (A, B, and C).

Class i	A	B	C
n_i	58	55	51
$\bar{x}_i$	75.10	71.96	78.94
s_i	13.87	13.77	13.12

1. Write down the hypothesis test for this problem.
2. Check the assumptions of anova.
3. Run anova analysis.
4. Run post-hoc analysis.

Anova Example: Assumptions



Anova Example: Analysis

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture	2	1290.11	645.06	3.48	0.0330
Residuals	161	29810.13	185.16		

$s_{pooled} = 13.61$ on $df = 161$

Anova Example: Post-hoc

- We need to check which classes are different, running pairwise comparisons.
- Using Bonferroni correction, we set $\alpha' = \alpha/3$.
- Compute $s_p = \sqrt{MSE} = \sqrt{185.16} = 13.61$, with $df = 161$.
- Run the following two-independent sample t-tests:

$$\begin{cases} H_0: & \mu_A = \mu_B \\ H_A: & \mu_A \neq \mu_B \end{cases} \qquad \begin{cases} H_0: & \mu_A = \mu_C \\ H_A: & \mu_A \neq \mu_C \end{cases} \qquad \begin{cases} H_0: & \mu_B = \mu_C \\ H_A: & \mu_B \neq \mu_C \end{cases}$$

- Summarize your results.

Completely Randomized Design: An Example

Diamonds (`Diamonds.xlsx`)

Consider the `Diamonds` data set, where we have two quantitative variables:

- price (in USD)
- carats (in number of carats)

and three qualitative vararaiables:

- color (D, E, F, G, H, I),
- clarity (IF, VS1, VS2, VVS1, VVS2),
- certification (GIA, HRD, IGI)

Select price and dependent variable and one qualitative variable at a time as factor. Run ANOVA to determine whether price depends on the factor. In addition, run a post-hoc analysis to establish a ranking. Use $\alpha = 0.05$. Interpret your results.

Case Study : Anorexia Therapies

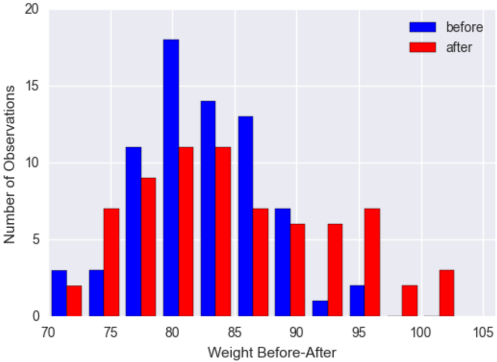
Anorexia Therapies (`anorexia.xlsx`)

We have data for three groups of anorexic girls who were treated with either cognitive behavioral therapy (group 1), family therapy (group 3), or no therapy at all (control group, group 2). For each girl, we have weights before and after the treatment and the change in weight.
The hypothesis that you want to test is that in the population, the three treatments produce different results in terms of weight change. In other words, is one treatment more effective than the others in dealing with anorexia?
Download file `anorexia.xlsx` and work through the case.

1. Visualize the data using histograms and draw some qualitative conclusions.
2. Is there any difference in the weights before and after the treatments?
3. Which treatment seems to be more effective?

Anorexia

Q_1 : Is there any difference in weights before and after the treatments?



Anorexia

Q_2 : Is there any difference among treatments?