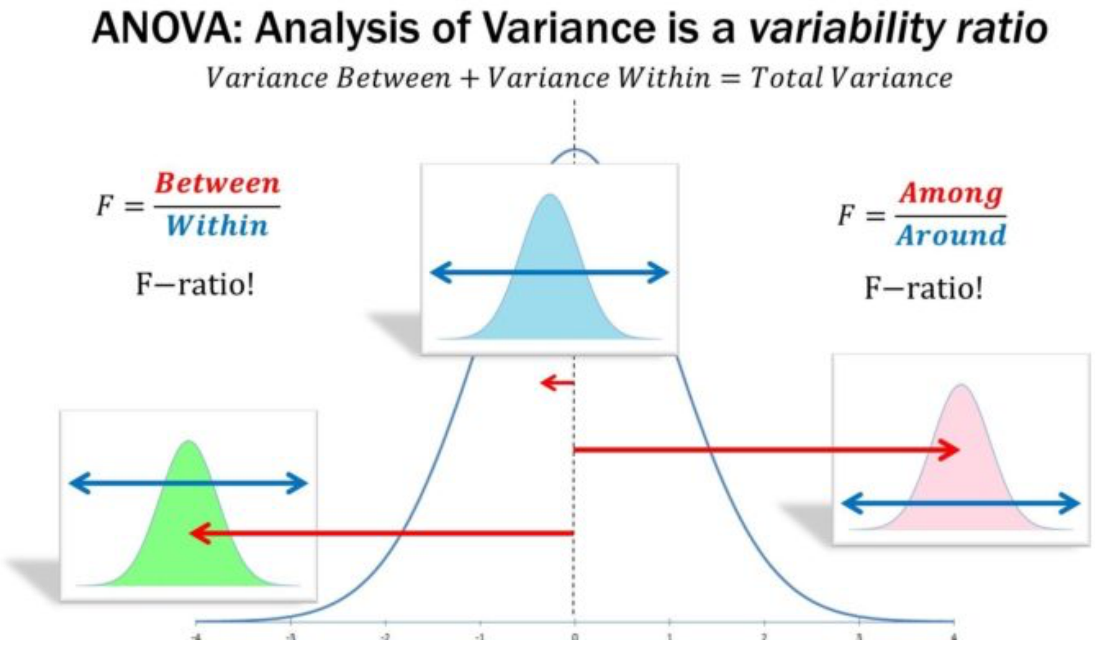
Comparing Centers of Several Independent Groups

EN5423 | Spring 2024

w13_several_independent_02.pdf
(Week 13)

Contents


ANOVA: Analysis of Variance is a *variability ratio*
Variance Between + Variance Within = Total Variance
F = Between / Within
F−ratio!
F = Among / Around
F−ratio!
-4
-3
-2
-1
0
1
2
3
4

1 Analysis of Variance (One-factor)

Example Scenario: The Effect of Soil Type on Plant Growth

Objective: Determine if different types of soil affect the growth rate of a specific plant species.

Factor (Grouping Variable): **Type of Soil**

Levels of the Factor:

1) Sandy Soil, 2) Clay Soil, 3) Loamy Soil

Table 1. Hypothesis tests with ***one factor*** and their characteristics.
[ANOVA, analysis of variance; BDM, Brunner-Dette-Munk test; MCT, multiple comparison test. HA is the alternative hypothesis, the signal to be found if it is present]

	Objective of test (H_A)				
	Data from at least one group is frequently higher than the other groups		**Mean of at least one group is higher than the mean of the other groups**		
Test	Kruskal-Wallis test	BDM test	ANOVA	Welch's adjusted ANOVA	Permutation test on group means
Class of test	Nonparametric	Nonparametric	Parametric	Parametric	Permutation
Distributional assumption for group data	None	None	Normal distribution; equal variances	Normal distribution	Exchangeable
Multiple comparison test	Pairwise rank-sum tests or Dunn’s test	Pairwise rank-sum tests or Dunn’s test	Tukey’s MCT	Tukey’s MCT	Tukey’s MCT

• Analysis of variance (ANOVA) determines whether the mean of ***at least one group*** differs from the means for other groups.

• If the group means are dissimilar, some of them will differ from the overall mean, as in **Figure 1**. If the group means are similar, they will also be similar to the overall mean, as in **Figure 2**.

• Why should a test of differences between means be named ***analysis of variance***?
>> In order to determine if the differences between group means (***the signal***) can be seen above the variation within groups (***the noise***), ***the total noise in the data as measured by the total sum of squares is split into two parts***:

Total sum of squares (Overall variation) = Factor sum of sqaures (Group means - overall mean) + Residual sum of sqaures (Variation within groups)

$$\sum_{j=1}^{k}\sum_{i=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2 = \sum_{j=1}^{k} n_j\left(\bar{y}_j-\bar{y}\right)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_i}\left(y_{ij}-\bar{y}_j\right)^2$$

Therefore,

$$\sum_{j=1}^{k}\sum_{i=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2=\sum_{j=1}^{k}n_j\left(\bar{y}_j-\bar{y}\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_i}\left(y_{ij}-\bar{y}_j\right)^2 \qquad \text{Eq. (1)}$$

Where y_{ij} is the ith observation in the jth group, there are k groups, and n_j designates that sample sizes within the jth group may or may not be equal to those in other groups.

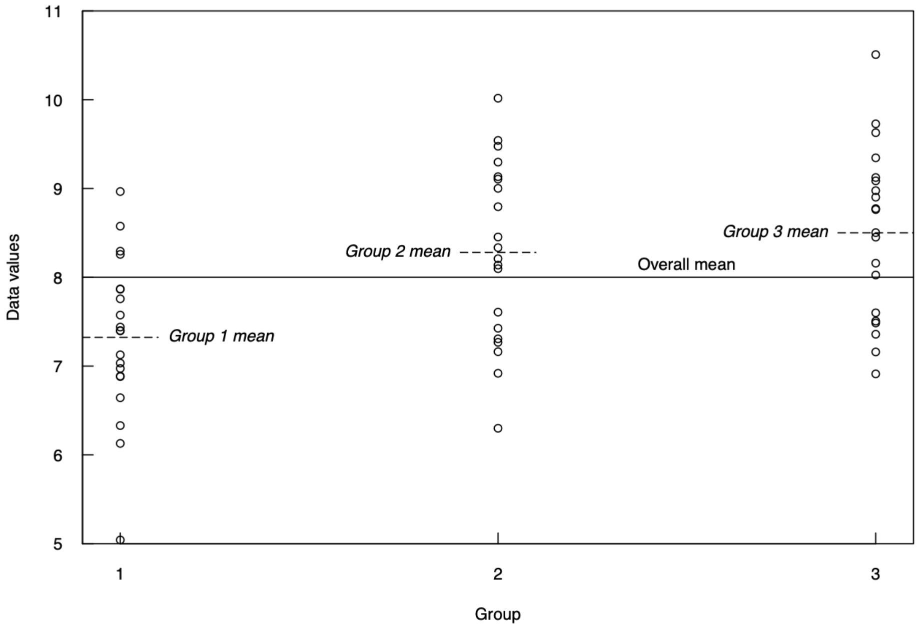


Figure 1. Hypothetical data for three groups. Factor mean square > residual mean square, and group means are found to differ.

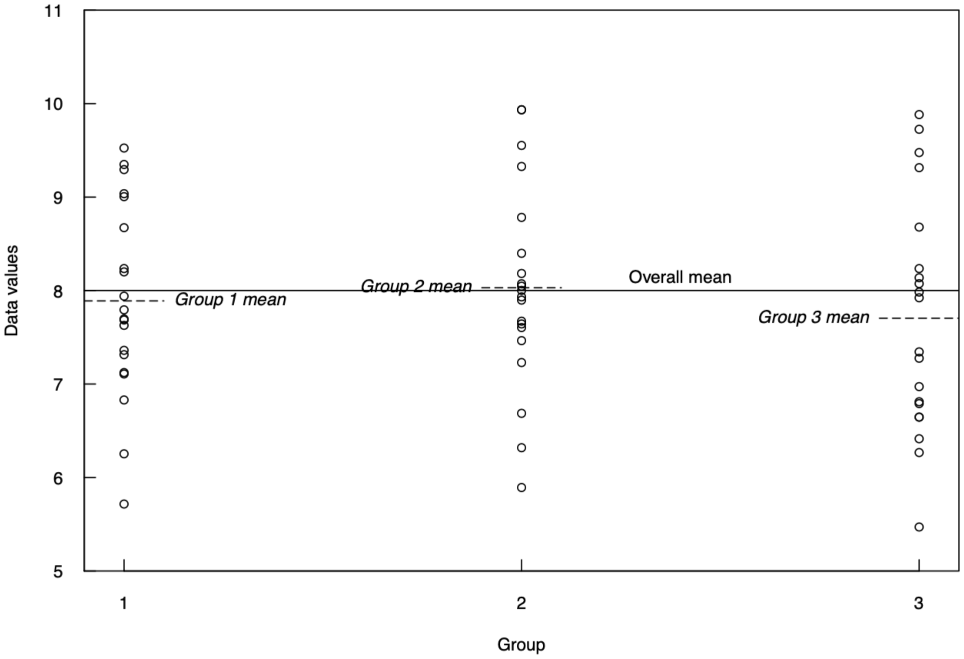


Figure 2. Hypothetical data for three groups. Factor mean square ≅ residual mean square, and group means do not significantly differ.

• If the total sum of squares is divided by $N - 1$, where N is the total number of observations, it equals the variance of the y_{ij}s.

• Thus, ANOVA ***partitions*** the ***variance of the data into two parts***, one measuring the signal (factor mean square, representing differences between groups) and the other measuring the noise (residual mean square, representing differences within groups).

• ***If the signal (factor mean square) is large compared to the noise (residual mean square), the means are found to be significantly different.***

Important:

ANOVA decomposes the total variability (SST; $\sum_{j=1}^{k} \sum_{i=1}^{n_i} \left(y_{ij} - \bar{y}\right)^2$) into two components: the variability due to ***differences among group means*** (SSF; $\sum_{j=1}^{k} n_j\left(\bar{y}_j - \bar{y}\right)^2$) and the variability due to ***differences within the groups themselves*** (SSE; $\sum_{j=1}^{k} \sum_{i=1}^{n_i} \left(y_{ij} - \bar{y}_j\right)^2$).

1.1 Null and Alternate Hypotheses for Analysis of Variance

The null and alternate hypotheses for the analysis of variance are

H_0: The group means are identical $\mu_1 = \mu_2 = \cdots = \mu_k$

H_A: At least one mean is different.

This is always a two-sided test.

1.2 Assumptions of the Analysis of Variance Test

ANOVA extends the *t*-test to more than two groups. It is not surprising then, that the same assumptions

apply to both tests:

1. All samples are ***random samples*** from their respective populations.

2. All samples are ***independent*** of one another.

3. Departures from the ***group mean*** $(y_{ij} - \bar{y}_j)$ are ***normally distributed*** for all j groups.

4. All groups ***have equal population variance***, σ^2, estimated for each group by s_j^2.

$$s_j^2 = \frac{\sum_{i=1}^{n_i} \left(y_{ij} - \bar{y}_j\right)^2}{n_j - 1} \qquad \text{Eq. (2)}$$

1.3 Computation of Classic ANOVA

Each observation, y_{ij}, can be written as

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij} \qquad \text{Eq. (3)}$$

Where

y_{ij} is the ith individual observation in group j, $j = 1, 2, \dots, k$;
μ is the overall mean (overall all groups);
α_j is the group effect, or $(\mu_j - \mu)$
ϵ_{ij} are the residuals or error within groups.

• If H_0 is true, all j groups have the same mean equal to the overall mean, μ, and thus $\alpha_j = 0$ for all j.

• If group means differ, $\alpha_j \neq 0$ for some j. To detect a difference between means, ***the variation within a group around its mean must be sufficiently small in comparison to the difference between group means*** so that the group means may be seen as different (see fig. 1).

• The noise ***within groups*** is estimated by the ***residual*** or ***error mean square (MSE)***, and the signal between group means is estimated by the ***factor*** or ***treatment mean square*** (***MSF***). Their computation is shown below.

The ***residua***l or ***error sum of squares***

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_i} \left(y_{ij} - \bar{y}_j\right)^2 \qquad \text{Eq. (4)}$$

estimates the total ***within-group*** noise using departures from the sample group mean, $\bar{y}_j$. Error in this context refers not to a mistake, but to the inherent noise within a group.

The ***factor*** or ***treatment sum of squares***

$$SSF = \sum_{j=1}^{k} n_j\left(\bar{y}_j - \bar{y}\right) \qquad \text{Eq. (5)}$$

estimates the factor effect using differences between group means, $\bar{y}_j$, and the overall mean, $\bar{y}$, weighted by sample size.

Each ***sum of squares*** has an associated number of degrees of freedom, or the ***number of independent pieces of information used to calculate the statistic***. For the ***factor sum of squares*** this equals k−1, as when k−1 of the group means are known, the kth group mean can be calculated. The ***total sum of squares*** has N−1 degrees of freedom. The ***residual sum of squares*** has degrees of freedom equal to the difference between the above two, or N−k.

Note:

1. **Degrees of Freedom (DF):** This refers to the number of independent values or quantities that can vary in an analysis without violating any constraints imposed on them. It essentially represents the amount of "free" data that is available to estimate another piece of data.
2. **Factor Sum of Squares (SS Between Groups):** This measures the variability due to the differences between the group means. The formula $k-1$ for its degrees of freedom comes from the fact that if you have k groups, knowing the means of any $k-1$ groups allows you to compute the mean of the kth group, assuming the overall mean is known. Here, k represents the number of groups.
3. **Total Sum of Squares (SS Total):** This measures the total variability in the data without taking into account how the data is grouped. The degrees of freedom here is $N-1$, where N is the total number of observations. This is because with N observations, knowing $N-1$ values allows you to compute the Nth value, assuming you know the overall mean.
4. **Residual Sum of Squares (SS Within Groups):** This represents the variability within each group, essentially the noise or error that is not explained by the group differences. Its degrees of freedom is calculated as $N-k$, which is the total number of observations minus the number of groups. This reflects the leftover degrees of freedom after accounting for the variability between the groups.

• Dividing the sums of squares by their degrees of freedom produces variance estimates: the total variance, the ***MSF***, and the ***MSE***.

• The ***MSF*** estimates the variance as a result of any signal between groups.

$$MSF = \frac{SSF}{DF} = \frac{\sum_{j=1}^{k} n_j\left(\bar{y}_j - \bar{y}\right)}{k-1} \qquad \text{Eq. (6)}$$

• If the ***MSF*** is similar to the ***MSE***, there is not much signal and H_0 is not rejected (fig. 2).

$$MSE = \frac{SSE}{DFE} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_i} \left(y_{ij} - \bar{y}_j\right)^2}{N-k} \qquad \text{Eq. (7)}$$

• If the ***MSF*** is sufficiently larger than the ***MSE***, the null hypothesis will be rejected and at least one group has a mean different from the others (fig. 1).

• The test to compare the two estimates of variance, ***MSF*** and ***MSE***, is whether their ratio equals 1.

$$F = \frac{MSF}{MSE} \qquad \text{Eq. (8)}$$

Important:

If the variability BETWEEN the means (distance from overall mean) in the numerator is relatively large compared to the variance WITHIN the samples (internal spread) in the denominator, the ratio will be much larger than 1. The samples then most likely do NOT come from a common population: REJECT NULL HYPOTHESIS that means are equal.

$\frac{LARGE}{small} = \boldsymbol{Reject\ H_0}$: At least one mean is an outlier and each distribution is narrow; distinct from each other.

$\frac{similar}{similar} = \boldsymbol{Fail\ to\ Reject\ H_0}$: Means are fairly close to overall mean and/or distributions overlap a bit; hard to distinguish.

$\frac{small}{Large} = \boldsymbol{Fail\ to\ Reject\ H_0}$: The means are every close to overall mean and/or distributions “melt” together.

• The test statistic F is compared to quantiles of an F-distribution and H_0 is rejected for large F. Equivalently, reject H_0 if the p-value for the test < α.

• The computations and results of an ANOVA are organized into an ANOVA table. Items usually provided in a one-way ANOVA table are shown in table 2. Note that the R summary command for analysis of variance does not display the Total row.

Table 2. Schematic of a one-factor ANOVA table.

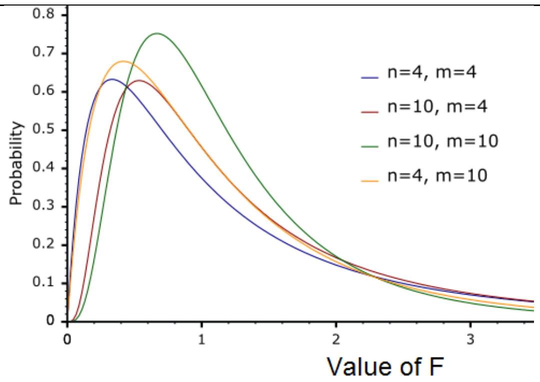
[df, degrees of freedom; *k*, number of groups; *N*, number of observations in all groups together; *SS*, sum of squares; *SSF*, *SS* for factor; *SSE*, *SS* for error; *MS*, mean square; *MSF*, *MS* for factor; *MSE*, *MS* for error; *F*, *F* test statistic; -, not applicable]

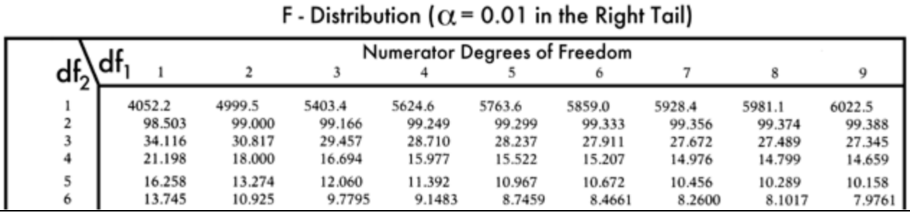
Source	df	SS	MS	F	*p*-value
Factor/Treatment	$(k-1)$	SSF	MSF	MSF/MSE	p
Residual error	$(N-k)$	SSE	MSE	-	-
Total	$N-1$	Total SS	-	-	-

Example 1: Specific capacity—Classic ANOVA

Properties of the F-distribution: The F-distribution is the theoretical distribution that the test statistic F follows under the null hypothesis that all group means are equal (i.e., the factor has no effect). The F-distribution is used because it is specifically designed for comparing two estimates of variance (two chi-squared distributions divided by their respective degrees of freedom). The *MSF* is essentially a scaled variance estimate among groups, and the *MSE* is a scaled variance estimate within groups. Each of these follows a *chi-squared distribution* under certain assumptions (normality, independent samples).

$$F_{n,m} = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}}$$




F - Distribution (α = 0.01 in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom 1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761

Example 1: Specific capacity—Classic ANOVA

```
# Fit the model
model = ols('spcap ~ rock', data=data).fit()

# Perform ANOVA
anova_results = sm.stats.anova_lm(model, typ=1)  # typ=1 for Type I
SS

# Print the ANOVA table
print(anova_results)
```

```
             df         sum_sq     mean_sq        F    PR(>F)
rock        3.0    6475.532023 2158.510674  2.51153  0.059859
Residual  196.0  168450.315468  859.440385      NaN       NaN
```

Classic ANOVA is run on the specific capacity data of last week’s example (Knopman, 1990) to illustrate the effects of its non-normality and unequal variance. There are 50 observations in each of the four groups.

The *F*-statistic of 2.51 is not significant (*p*=0.0599) at an *α* of 0.05.

The temptation is to declare that no difference has been found. However, neither of the requirements of ***normality*** or ***equal variance*** was met.

The data analyst should be worried about the effects of failing to meet these assumptions when using ANOVA, even with a dataset of 50 observations per group. Using the preferred Welch adjustment instead would ***address the issue of unequal variance, but not non-normality.***

Important:
MSF measures variance due to the main effect of the groups (the signal), while MSE measures the variance within the groups (the noise). The comparison of these two, typically through an *F*-test, helps determine if the factor significantly impacts the response variable across the groups.

1.4 Welch's Adjusted ANOVA

• Recognition of the loss of power for ANOVA as ***a result of heteroscedasticity*** has slowly made its way into statistics software. This mirrors standard practice for the *t*-test, where Welch's adjustment is the default (because heteroscedasticity is common.)

• Welch's *F*-statistic is computed by weighting ***each group's contribution*** to the ***MSF*** by $\frac{n}{s^2}$, so that ***groups with greater variability have lower weight***. The ***MSE*** (or residual mean square) is computed using an adjusted degrees of freedom whose value decreases from the ANOVA residual degrees of freedom as group variances become dissimilar.

• The resulting *F*-test is more accurate for heteroscedastic data. ***There is little disadvantage to using the Welch adjustment as its correction to the classic ANOVA F-statistic is negligible when heteroscedasticity is not present.***

Example 2: Specific capacity—Welch's adjusted ANOVA

```
# Fligner-Killeen test of homogeneity of variances
grouped_data = [data[data['rock'] == rock]['spcap'] for rock in
data['rock'].unique()]
stat, p = fligner(*grouped_data)

print("Fligner-Killeen test:")
print("Chi-squared:", stat, ", p-value:", p)

# One-way ANOVA not assuming equal variances
anova_results = anova_oneway(data['spcap'], groups=data['rock'],
use_var='unequal')

print("\nOne-way ANOVA (not assuming equal variances):")
print(anova_results)
```

```
Fligner-Killeen test:
Chi-squared: 39.45805263692878 , p-value: 1.3880856753298259e-08

One-way ANOVA (not assuming equal variances):
statistic = 3.439684687787948
pvalue = 0.02052032650226662
df = (3.0, 82.54121109385693)
df_num = 3.0
df_denom = 82.54121109385693
nobs_t = 200.0
n_groups = 4
means = [15.49219977  9.83265055  3.15103934  1.07100378]
nobs = [50. 50. 50. 50.]
vars_ = [1546.74565809 1759.18092118  129.99549886    1.83946202]
use_var = unequal
welch_correction = True
```

```
tuple = (3.439684687787948, 0.020520326502226662)
```

Welch’s adjustment increases the F-statistic signal to 3.4 from the classic ANOVA’s 2.5 by reducing the residual MS originally inflated by unequal variance.

The cost of adjustment is that the denominator (residual) degrees of freedom decreases from 198 to 82.

$$F = \frac{\sum_{j=1}^{k} \frac{\left(\bar{y}_j - \bar{y}\right)^2}{s_j^2/n_j}}{\sum_{j=1}^{k} \frac{s_j^2/n_j^2}{n_j - 1}} \qquad df = \frac{\left(\sum_{j=1}^{k} \frac{s_j^2}{n_j}\right)^2}{\sum_{j=1}^{k} \frac{\left(s_j^2/n_j\right)^2}{n_j - 1}}$$

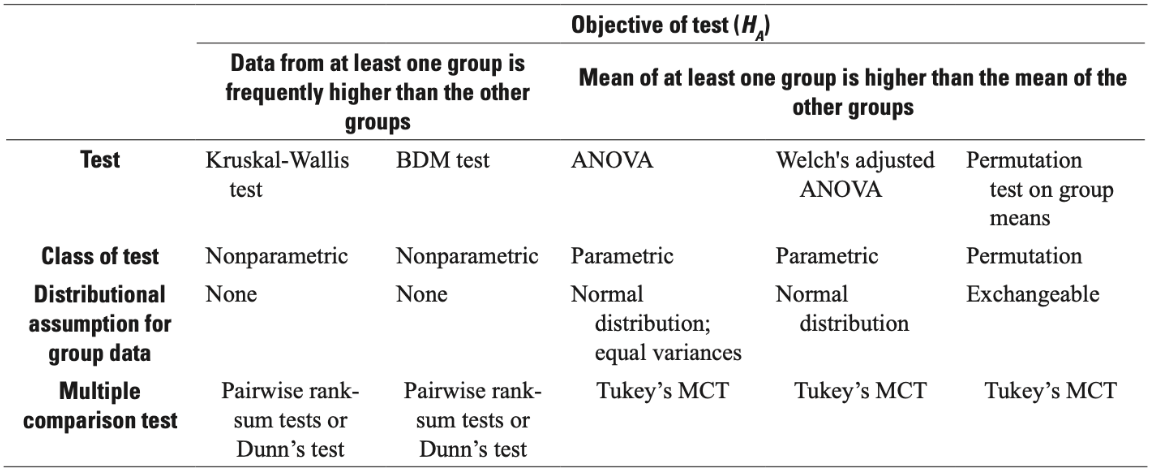
The cost is small compared to the benefit, as the adjusted *p*-value of 0.02 is sufficiently lower than the unadjusted *p*-value to reject the null hypothesis, finding a difference between group means.

This illustrates the power loss for this dataset by using classic ANOVA on data with unequal group variances. Even with the relatively large sample size of 50 observations per group, violation of the two primary assumptions of classical ANOVA can lead to a loss of power.

Water resources and other environmental data are known for their strong skewness, leading to violations of both normality and constant variance. If a test on means is the appropriate objective, ***use the Welch’s adjusted ANOVA to correct for violation of equal variance***, or use a permutation test (see later week’s pdf file) t***o avoid the interferences of both unequal variance and non-normality***. If the objective is to determine whether at least one group has higher values than another, the Kruskal-Wallis test (see last week’s pdf file) addresses that frequency objective directly.

2 Permutation Test for Difference in Means (One-factor)

Table 1. Hypothesis tests with ***one factor*** and their characteristics.
[ANOVA, analysis of variance; BDM, Brunner-Dette-Munk test; MCT, multiple comparison test. HA is the alternative hypothesis, the signal to be found if it is present]



	Objective of test (H_A)				
	Data from at least one group is frequently higher than the other groups		Mean of at least one group is higher than the mean of the other groups		
Test	Kruskal-Wallis test	BDM test	ANOVA	Welch's adjusted ANOVA	Permutation test on group means
Class of test	Nonparametric	Nonparametric	Parametric	Parametric	Permutation
Distributional assumption for group data	None	None	Normal distribution; equal variances	Normal distribution	Exchangeable
Multiple comparison test	Pairwise rank-sum tests or Dunn’s test	Pairwise rank-sum tests or Dunn’s test	Tukey’s MCT	Tukey’s MCT	Tukey’s MCT

• Permutation tests ***allow the means of skewed datasets to be tested without the loss of power inherent in parametric ANOVA procedures***.

• Permutation tests determine whether group means differ ***without requiring the assumption of normality and without suffering the same consequences of unequal variance that parametric tests have***.

• Permutation tests require exchangeability—any observation found in one group could have come from another group because they originate from the same population. This is simply a restatement of the null hypothesis—***there is only*** *one population* ***from which all groups are sampled.***

• Permutation tests on means do ***not require a mathematical adjustment for unequal variance as do parametric tests because they do not use a standard deviation or variance parameter to compute the test statistic***.

• However, if the observed variances are unequal, the larger variance spills over into all groups during the permutation process.

• In short, permutation tests are ***less susceptible*** ***to loss of power caused by unequal variance than are*** *parametric tests****, but the observed variability of a single group will be spread to all groups***. There is really no way around this if the goal is to perform a test of means.

2.1 Computation of the Permutation Test of Means

• Permutation tests compute either all test results possible for rearrangements of the observed data (exact test), or thousands of test results for a large random selection of possible rearrangements.

• The proportion of computed results equal to, or more extreme than, the one result obtained from the original data is the p-value of the test.

• For a one-factor permutation test, the simplest visualization is that the column of group assignments is randomly reordered thousands of times, and the ANOVA F-test statistic computed for each reordering.

• By reordering group assignments the number of observations per group stays the same, but the observations assigned to groups differ for each randomization.

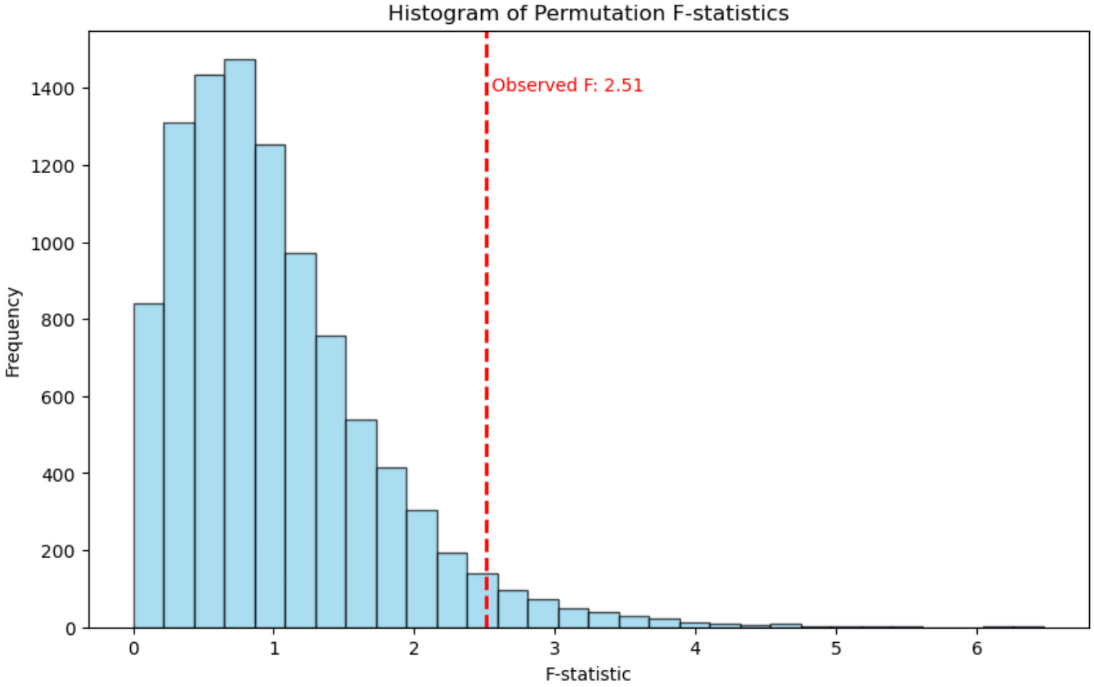
• If the null hypothesis is true, each group has the same data distribution; each has the same mean and variance.

• Therefore, for the null hypothesis the group assignment is basically random—any observation from one group could have just as easily come from another group.

• The F-statistics from the thousands of random group assignments represent the distribution of F-statistics expected when the null hypothesis is true.

• This distribution may or may not resemble any specific shape, the data determine the shape of the test statistic distribution.

• A histogram of 10,000 F-statistics from permutations of the specific capacity data from previous example representing the null hypothesis of no group differences are shown in figure 3.



```
Observed F-statistic: 2.511530423479519
Permutation p-value: 0.0409 (based on 10000 permutations)
```

Figure 3. Histogram of F-statistics for 10,000 permutations of the specific capacity group assignments from previous example. The vertical dashed line at 2.51 is the F-statistic of the analysis of variance for the original data.

• The p-value of 0.0429 states that 4.29 percent of the permutation F-test statistics equaled or exceeded the original observed F of 2.51.

• Therefore, the mean specific capacity is declared ***significantly different for the four rock types*** in previous example. Running the procedure again will produce a slightly different p-value, but using 10,000 rearrangements ensures that the variation in p-values will be small.

• A permutation test can always be used instead of classic (or Welch's) ANOVA when violations of ***normality or equal variance assumptions occur***. Here the similarity between permutation and Welch's adjustment results indicates that ***unequal variance was the more important violation of assumptions for this dataset***, pushing the classic ANOVA's p-value above 0.05.

Example 3: Specific capacity—Welch's adjusted ANOVA

```
# Fit ANOVA model to get the F-statistic for the original data
model = ols('spcap ~ rock', data=data).fit()
aov_table = sm.stats.anova_lm(model, typ=2)
f_observed = aov_table['F'].iloc[0]  # Using .iloc for proper indexing

# Permutation test setup
```

```
n_permutations = 10000
f_permutations = []

for _ in range(n_permutations):
    # Shuffle the 'rock' labels
    shuffled_rock = np.random.permutation(data['rock'])
    data['shuffled_rock'] = shuffled_rock

    # Fit model with shuffled labels
    model_shuffled = ols('spcap ~ shuffled_rock', data=data).fit()
    aov_table_shuffled = sm.stats.anova_lm(model_shuffled, typ=2)
    f_permutations.append(aov_table_shuffled['F'].iloc[0])  #
Corrected to use .iloc

# Calculate p-value
p_value = np.mean([f >= f_observed for f in f_permutations])
print(f"Observed F-statistic: {f_observed}")
print(f"Permutation p-value: {p_value} (based on {n_permutations}
permutations)")
```

```
Observed F-statistic: 2.511530423479519
Permutation p-value: 0.0409 (based on 10000 permutations)
```

3 Two-factor Analysis of Variance

• Often, multiple factors might affect the outcomes we observe in a study. Multi-factor tests help us understand the ***impact of each factor simultaneously***, similar to how multiple regression analysis works. ***This means we can see the effect of one factor while accounting for the influence of the others***. This approach is used in ***factorial analysis of variance***, which is a more complex version of ANOVA.

• ***Factorial ANOVA*** is used when the factors being studied are independent of each other, meaning ***no factor is just a part of another***. Sometimes, studies might involve factors that are subsets of each other, ***known as nested factors***, and these require different methods to analyze. (For more details on nested ANOVA, you can look up resources like Aho (2016))

(E.g., **Factor 1**: School (School A, School B, School C); **Factor 2**: Classroom (Classrooms are nested within Schools: Classroom 1A, 2A, 3A are in School A; Classroom 1B, 2B, 3B are in School B; Classroom 1C, 2C, 3C are in School C))

• Here, we focus on discussing ANOVA with two factors, but it is possible to include more than two; however, that is more complex and is not covered in this explanation.

3.1 Null and Alternate Hypotheses for Analysis of Variance

• Last week, we read a ***two-factor*** example, the determination of chemical concentrations among stream basins at low flow. The objective was to determine **whether concentrations**

differed as a function of mining history (whether or not each basin was mined, and if so, whether it was reclaimed) and **of rock type**.

• Call the two factors A and B. There are *i* = 1 to *α* ≥ 2 categories of factor A, and *j* = 1 to b ≥2 categories of factor B. Treatment groups are defined as all the possible combinations of factors A and B, so there are a · b treatment groups. Within each treatment group there are n_{ij} observations. The test determines whether mean concentrations are identical among all the a · b treatment groups, or whether at least one differs.

*H*0: All treatment group means, μ_{ij}, are equal. $\mu_{11} = \mu_{12} =, \dots, = \mu_{ab}$
HA: At leat one μ_{ij} differs from the rest.

For the $k = 1, 2, \dots, n_{ij}$ observtions in treatment group *ij*, the magnitude of any observation, y_{ijk}, differs from the overall mean, μ, by being affected by several possible influences:

$$y_{ijk} = \mu + \gamma_i + \delta_j + \gamma\delta_{ij} + \epsilon_{ijk}$$

where

γ_i is the influence of the *i*th category of factor A;
δ_j is the influence of the *j*th category of factor B;
$\gamma\delta_{ij}$ is the **interaction effect** between factors A and B beyond those of γ_i and δ_j individually for the *ij*th treatment group; and
ϵ_{ijk} is the residual error, the difference between the *k*th observation (k=1,2,...,n_{ij}) and the treatment group mean $\mu_{ij} = \mu + \gamma_i + \delta_i + \gamma\delta_{ij}$.

The null hypothesis states that treatment group means μ_{ij} all equal the overall mean, μ. Therefore γ_i, δ_i, and $\gamma\delta_{ij}$ effects are ***sufficiently nonzero***, the ***null hypothesis is rejected*** and at least one treament group mean significantly differs from the others.

3.2 Assumptions of Two-factor ANOVA

• In two-factor ANOVA, the residuals ϵ_{ij} from each treatment group mean μ_{ij} (each combination of factors A and B) are ***assumed to be normally distributed*** with identical variance σ^2.

• The normality and constant variance assumptions could be checked by inspecting separate boxplots of data for each treatment group, but more powerfully by testing the ANOVA residuals from all groups together, ϵ_{ij}, ***using the Shapiro-Wilk test*** and plotting them on one normal probability plot or boxplot.

• The effect of violating these assumptions is the same as for one-way ANOVA, a loss of power ***leading to higher p-values*** and failure to find significant differences that are there. ***No convenient version of a Welch's adjustment exists for two or more factor designs***.

• Brunner and others (1997) state that ***Welch adjustments to factorial ANOVA are "cumbersome." Instead, permutation tests are the primary method for computing factorial ANOVA on non-normal or heteroscedastic data.***

3.3 Computation of Two-factor ANOVA

• The influences of factors A, B, and their interaction are evaluated separately by partitioning the total sums of squares into component parts for each effect. After dividing by their respective degrees of freedom, the mean squares for factors A (***MSA***), B (***MSB***), and their interaction (mean square for interaction, ***MSI***) are produced. As with a one-way ANOVA, these are compared to the ***MSE*** using F-tests to determine their significance.

• The sums of squares for factor A (***SSA***), factor B (***SSB***), interaction (***SSI***), and error (***SSE***), assuming constant sample size $n_{ij} = n$ per treatment group, are presented in table 3.

• Dividing the sums of squares by their degrees of freedom produces the mean squares ***MSA***, ***MSB***, ***MSI***, and ***MSE*** as in table 4. If H_0 is true and γ_i, δ_j, and $\gamma\delta_{ij}$ ***all equal 0***, all variation is simply around the overall mean, μ.

• The ***MSA***, ***MSB***, and ***MSI*** will then all approximate the ***MSE***, and all three F-tests will have ratios similar to 1. However, when the alternate hypothesis H_A is true, ***at least one of the mean squares in the numerators will be significantly larger*** than the ***MSE***, and the ratio ***MSfactor***/***MSE*** (F-statistic) will be larger than the ***appropriate quantile of the F distribution***.

• H_0 is then rejected and that factor is considered significant at a risk level α. The formulae in the two-factor ANOVA table (table 4) are for ***an equal number of observations*** in each treatment group (all $n_{ij} = n$).

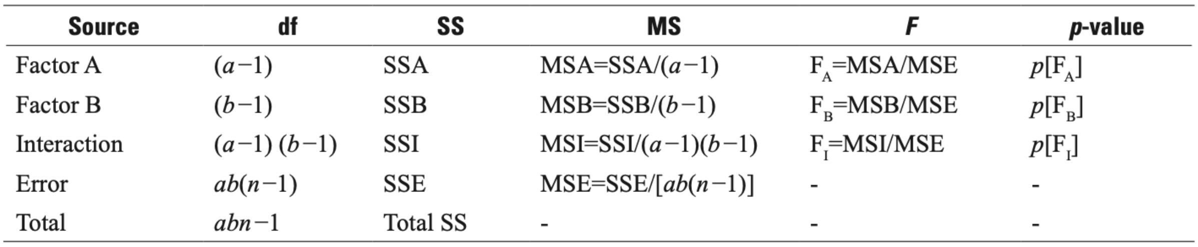
• More complex formulae are involved when there are unequal numbers of observations (an unbalanced design).

Table 3. Sums of squares definitions for two-factor ANOVA.
[SS, sum of squares; SSA, SS for factor A; SSB, SS for factor B; SSE, SS for error; SSI, SS for interaction]

Sums of squares formula	Effect
$SSA = \sum^{a} \frac{\left(\sum^{b}\sum^{n} y\right)^2}{bn} - \frac{\left(\sum^{a}\sum^{b}\sum^{n} y\right)^2}{abn}$	$\mu_i - \mu$
$SSB = \sum^{b} \frac{\left(\sum^{a}\sum^{n} y\right)^2}{an} - \frac{\left(\sum^{a}\sum^{b}\sum^{n} y\right)^2}{abn}$	$\mu_j - \mu$
$SSI = Total\ SS - SSA - SSB - SSE$	$\mu_{ij} - (\mu_i + \mu_j) + \mu$
$SSE = \sum^{a}\sum^{b}\sum^{n} (y)^2 - \sum^{a}\sum^{b} \frac{\left(\sum^{n} y\right)^2}{n}$	$y_{ijk} - \mu_{ij}$
$Total\ SS = \sum^{a}\sum^{b}\sum^{n} (y)^2 - \frac{\left(\sum^{a}\sum^{b}\sum^{n} y\right)^2}{abn}$	$y_{ijk} - \mu$

Table 4. Schematic for a two-factor ANOVA table.

[SS, sum of squares; SSA, SS for factor A; SSB, SS for factor B; SSE, SS for error; SSI, SS for interaction; MS, mean square; MSA, MS for factor A; MSB, MS for factor B; MSE, MS for error; MSI, MS for interaction; df, degrees of freedom; F, F-test statistic; -, not applicable]



Source	df	SS	MS	*F*	*p*-value
Factor A	$(a-1)$	SSA	MSA=SSA/$(a-1)$	F_A=MSA/MSE	$p[F_A]$
Factor B	$(b-1)$	SSB	MSB=SSB/$(b-1)$	F_B=MSB/MSE	$p[F_B]$
Interaction	$(a-1)\,(b-1)$	SSI	MSI=SSI/$(a-1)(b-1)$	F_I=MSI/MSE	$p[F_I]$
Error	$ab(n-1)$	SSE	MSE=SSE/$[ab(n-1)]$	-	-
Total	$abn-1$	Total SS	-	-	-

Example 4: Iron at low flows—Two-factor ANOVA

Iron concentrations were measured at low flow in numerous small streams in the coal-producing areas of eastern Ohio (Helsel, 1983).

Each stream drains either an ***unmined area***, ***a reclaimed coal mine***, or an ***abandoned coal mine***. Each site is also underlain by either a ***sandstone*** or ***limestone*** formation.

Are iron concentrations at low flow influenced by upstream mining history, by the underlying rock type, or by both?

Boxplots for total iron concentrations are shown in **Figure 4**, where three outliers greater than 100 milligrams per liter in the sandstone, abandoned (sand_ab) group are not shown.

Note the skewness evidenced by the larger upper portions of several boxes. Also note the differences in variance as depicted by differing box heights. This exercise illustrates the problems incurred when parametric ANOVA is applied to data with (commonly occurring) non-normal, heteroscedastic characteristics.

There are ***six treatment groups***, combining the three possible mining histories (unmined, abandoned mine, and reclaimed mine) and the two possible rock types (sandstone and limestone).

Subtracting the group mean from each group's data, the Q-Q plot of residuals clearly shows the three high outliers and several large negative residuals resulting from subtracting the large mean of the abandoned sandstone group from its lower concentration data (**Figure. 5**).

The pattern is not consistent with a normal distribution. The ANOVA table is shown below. Tests are computed for the factors of mining history alone, rock type alone, and their interaction.

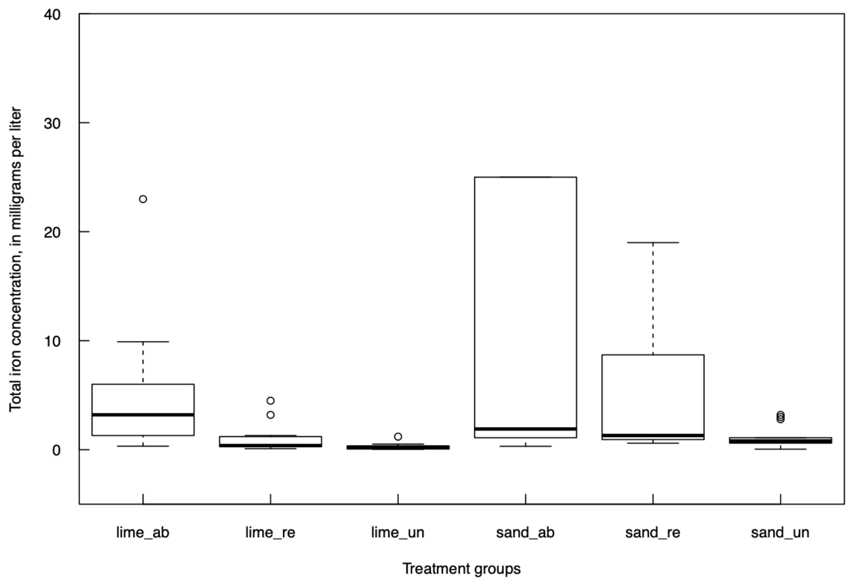


Figure 4. Boxplots of iron concentrations at low flow from Helsel (1983). Three outliers greater than 100 milligrams per liter are not shown. Lime_ab, limestone, abandoned mine; lime_re, limestone, reclaimed; lime_un, limestone, unmined; sand_ab, sandstone, abandoned mine; sand_re, sandstone, reclaimed; sand_un, sandstone, unmined.

```
# Fit ANOVA model
model = ols('Fe ~ C(Mining) * C(Rocktype)', data=data_iron).fit()

# Shapiro-Wilk test on residuals
shapiro_test = stats.shapiro(model.resid)
print("Shapiro-Wilk normality test")
print(f"W = {shapiro_test.statistic}, p-value =
{shapiro_test.pvalue}")

# Summary of the ANOVA model
anova_table = sm.stats.anova_lm(model, typ=2)  # Type II ANOVA
DataFrame
print("\nANOVA table:")
print(anova_table)

# Q-Q plot of residuals
plt.figure(figsize=(6, 6))
sm.qqplot(model.resid, line='s')
plt.title('Normal Q-Q plot of residuals')
plt.show()

# Fligner-Killeen test of homogeneity of variances
fligner_test = stats.fligner(*[group['Fe'].values for name, group in
data_iron.groupby('Group', observed=True)])
print("Fligner-Killeen test of homogeneity of variances")
print(f"Chi-squared = {fligner_test.statistic}, df =
{len(data_iron['Group'].unique()) - 1}, p-value =
{fligner_test.pvalue}")
```

```
Shapiro-Wilk normality test
```

```
W = 0.3350742649225724, p-value = 4.6104433466381143e-17
ANOVA table:
                              sum_sq    df         F    PR(>F)
C(Mining)               32282.304955   2.0  2.492636  0.089804
C(Rocktype)             15411.161730   1.0  2.379906  0.127289
C(Mining):C(Rocktype)   25868.829021   2.0  1.997428  0.143133
Residual               466238.520226  72.0       NaN       NaN
```

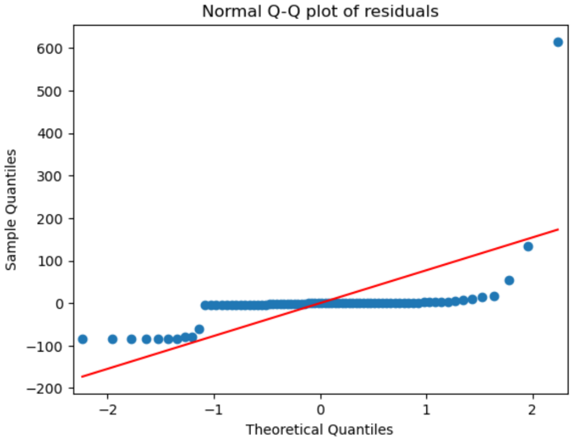


Figure 5. Q-Q plot showing the non-normality of the ANOVA residuals of the iron data from example 7.5.

Neither factor nor the interaction appears significant at the $\alpha = 0.05$ level, as their p-values are all larger than 0.05. However, the gross violation of the test's assumptions of normality, shown by the **Shapiro-Wilk test** and Q-Q plot (**Figure. 5**), and of equal variance, shown in the boxplots (**Figure. 4**) and by the ***Fligner-Killeen*** test, must not be ignored.

Perhaps the failure to reject H_0 is due not to a lack of an influence, but to the parametric test's lack of ***power to detect these influences because of the violation of test assumptions***. We will examine that possibility in later section using a permutation test.

3.4 Interaction Effects in Two-factor ANOVA

- ***Interaction*** is a synergistic or antagonistic change in the mean for a combination of the two factors, beyond what is seen from individual factor effects.

- Without interaction, the effect of factor B is identical for all groups of factor A, and the effect of factor A is identical for all groups of factor B.

- Plotting the means of all a·b groups, with factor A on the x axis and factor B represented by different connecting lines (an interaction plot—**Figure. 6**), ***the lines are parallel***, showing that there is no change in the effect of one factor based on the levels of the second factor and thus ***there is no interaction***.

- When interaction is present ($\gamma\delta_{ij} \neq 0$), the treatment group means are not determined solely ***by the additive effects of factors A and B alone***.

- Some of the groups will have mean values larger or smaller than those expected from the individual factors. The effect of factor A can no longer be discussed without reference to which group of factor B is of interest, and the effect of factor B can likewise not be stated apart from a knowledge of the group of factor A—***the lines are not parallel***.

- This is the pattern exhibited by the mining history and rock type effects of the example 4 data (**Figure. 7**).

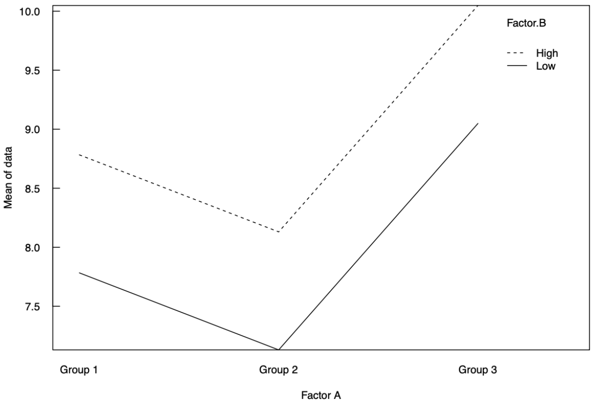


Figure 6. Interaction plot presenting the means of data in the six treatment groups from “Iron at low flows” data showing no interaction between the two factor effects.

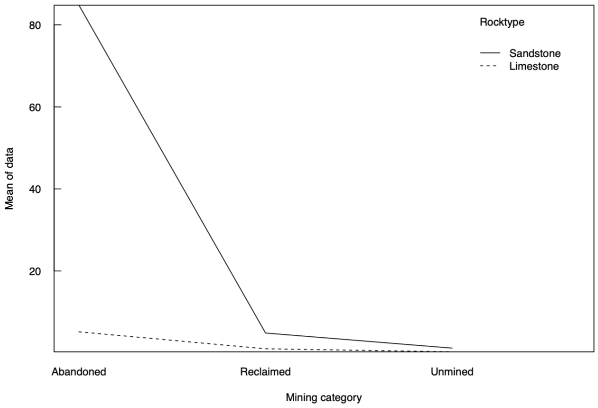


Figure 7. Interaction plot showing interaction by the large nonparallel increase in the mean for the combination of abandoned mining history and sandstone rock type.

Note:

Factors and Setting:
Imagine you are studying the growth of flowers and you want to understand how two factors, Soil Type (Factor A) and Amount of Fertilizer (Factor B), affect their growth. You have two types of soil (sandy and loamy) and two levels of fertilizer (low and high).

Without Interaction:
If there is ***no interaction*** between the soil type and the amount of fertilizer, then the effect of changing the fertilizer amount would be the same regardless of the soil type. Similarly, the effect of changing the soil type would be the same no matter how much fertilizer you use. In this scenario, if you plotted the average flower heights for each combination on a graph (soil type on the x-axis, different lines for fertilizer levels), the lines would be parallel. This means ***the increase or decrease in growth is consistent across different soil types when you change the fertilizer level, indicating no interaction***.

With Interaction:
However, suppose in your experiment, flowers grow ***exceptionally well in loamy soil*** with high fertilizer but ***poorly in sandy soil*** with the same amount of fertilizer. Meanwhile, flowers in sandy soil only show slight improvement with more fertilizer compared to loamy soil. ***This variation suggests an interaction between soil type and fertilizer amount because the effect of adding more fertilizer depends on the type of soil.***

When you plot this on a graph (soil type on the *x*-axis, different lines for fertilizer levels), the lines would not be parallel. They might cross or diverge, indicating that the combination of soil type and fertilizer amount has a unique effect that isn't just the sum of their separate effects.

3.5 Two-factor ANOVA on Logarithms or Other Transformations

• A common approach for analyzing two-way designs with ***non-normal and heteroscedastic*** data is to perform ANOVA on data transformed by a power transformation, such as logarithms.

• This transformation aims to create a dataset that is ***closer to normal with constant variance***. In logarithmic terms, the effects of each factor are additive, which models a multiplicative influence on the original scale.

• However, an ANOVA on transformed data tests for differences in ***geometric means***, essentially estimating the median of the data on the original scale, not the arithmetic means. This means that if the actual means are of interest, then using a transformation alters the focus of the test.

• While ANOVA on logarithms can be effective for assessing typical (median) differences between groups, the assumptions about residuals for the log-transformed data should still be verified.

• Natural logarithms of the low-flow iron concentrations from iron example are shown in **Figure 8**. After transformation, most treatment groups still display distinct right-skewness, except for the unmined limestone (lime_un) group, which appears less symmetric.

• The log transformation is ***not a universal solution***; other transformations may also fail to correct issues like positive skewness or unequal variance and could even create new problems, such as turning a symmetric distribution into a left-skewed one, as seen with the lime_un group.

• If the analysis's goal is to test differences in means, ***a permutation test is recommended over transforming the data scale***.

Example 5: Iron at low flows—Two-actor ANOVA using logarithms

```
# Fit ANOVA model
model = ols('np.log(Fe) ~ C(Mining) * C(Rocktype)',
data=data_iron).fit()

# Summary of the ANOVA model
anova_table = sm.stats.anova_lm(model, typ=2)  # Type II ANOVA
DataFrame
print("\nANOVA table:")
print(anova_table)
```

```
ANOVA table:
                          sum_sq   df          F    PR(>F)
C(Mining)              69.746910  2.0  15.890508  0.000002
C(Rocktype)            26.312445  1.0  11.989581  0.000904
C(Mining):C(Rocktype)   2.441812  2.0   0.556320  0.575759
```

```
Residual               158.011862  72.0        NaN       NaN
```

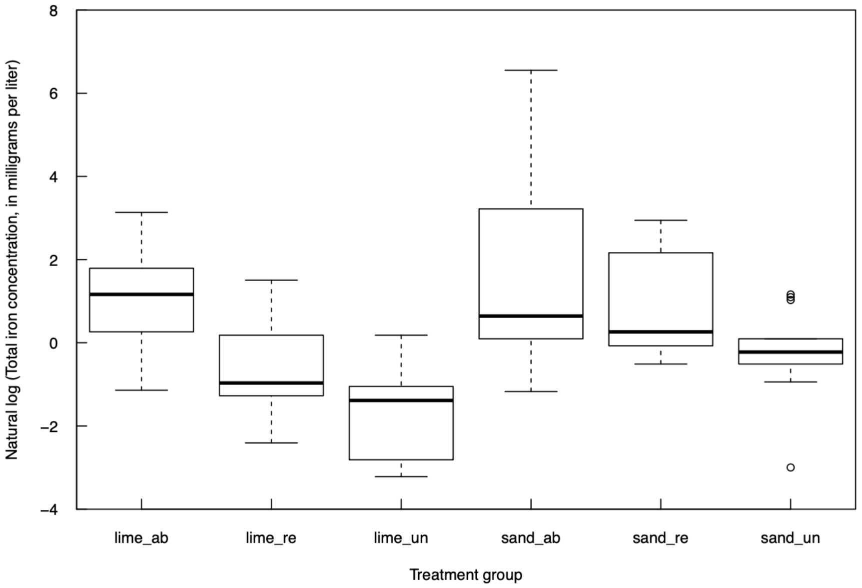


Figure 8. Boxplots of the natural logarithms of the iron data from iron example. Lime_ab, limestone, abandoned mine; lime_re, limestone, reclaimed; lime_un, limestone, unmined; sand_ab, sandstone, abandoned mine; sand_re, sandstone, reclaimed; sand_un, sandstone, unmined.

Both mining and rock type factors are significant, showing that some combinations of the two factors demonstrate higher geometric means, even though differences in means were not significant.

3.6 Fixed and Random Factors

• The previously mentioned F-test equations are based on the assumption that both factors involved are fixed.

• This means the conclusions drawn from the results apply only to the specific treatment groups studied. For instance, in the iron data analysis, we are looking at the chemical differences between three distinct mining histories. This is different from a random factor where groups are chosen randomly from a larger set, aiming to represent a general factor. Results from tests with random factors aim to understand a broad effect rather than pinpoint effects on specific groups.

• Consider an example where soil concentrations of a trace metal are compared across three different particle sizes statewide to identify the best size for reconnaissance. Here, ***particle size is a fixed effect*** because the interest lies in these specific sizes. Additionally, a random factor might be introduced if funding limitations restrict sampling.

• For example, suppose only seven counties can be sampled. These counties are chosen randomly, and the results from these locations will help determine not only which particle size is best but also if this finding holds true across the sampled counties. This random sampling approach assumes that the chosen counties represent spatial variability across the state.

- In cases where all factors are random, ***the mean square for interaction***, rather than ***the mean square for error***, is used as the denominator in F-tests. When dealing with designs that combine both random and fixed factors, known as mixed effects designs, the calculation gets more complex.

- Typically, fixed factors use the interaction mean squares for their calculations, while random factors use the error mean square. This might seem counterintuitive at first. The complexity of mixed effects F-tests increases with the number of factors involved. For a detailed exploration of various ANOVA designs, including those with mixed effects, the work by Aho (2016) provides a comprehensive discussion.