

The idea of ANOVA

Reminders: A **factor** is a variable that can take one of several **levels** used to differentiate one group from another.

An experiment has a **one-way**, or **completely randomized, design** if several levels of one factor are being studied and the individuals are randomly assigned to its levels. (There is only one way to group the data.)

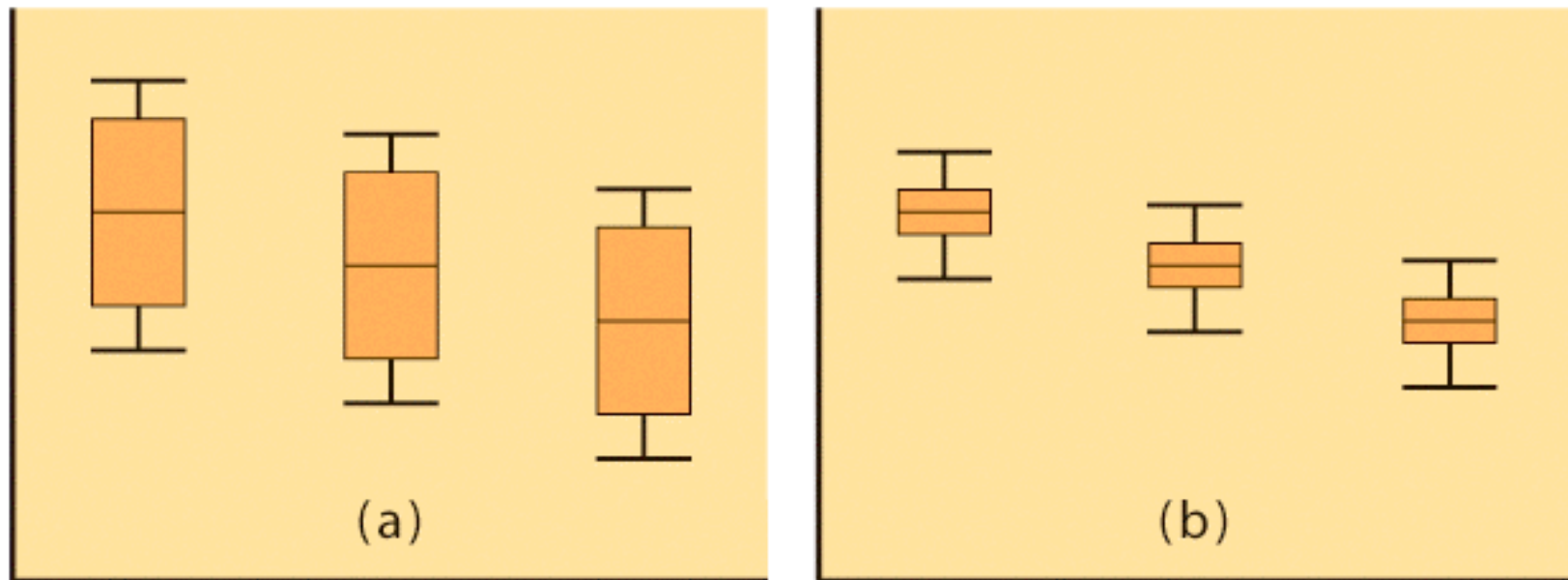
- Example: Four levels of herbicide strength in an experiment on dry weight of treated plants.
- Two plant species and four levels of herbicide would be a two-way design.

Analysis of variance (ANOVA) is the technique used to determine whether more than two population means are equal.

One-way ANOVA is used for completely randomized, one-way designs.

Comparing means

We want to know if the observed differences in sample means are likely to have occurred by chance just because of random sampling.



This will likely depend on both the difference between the sample means and how much variability there is within each sample.

Two-sample t statistic

A two sample t -test assuming equal variance and an ANOVA comparing only two groups will give you the same p -value (for a two-sided hypothesis).

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

One-way ANOVA

F-statistic

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

t -test assuming equal variance

t -statistic

$F = t^2$ and both p -values are the same.

But the t -test is more flexible: You may choose a one-sided alternative instead, or you may want to run a t -test assuming unequal variance if you are not sure that your two populations have the same standard deviation σ .

An Overview of ANOVA

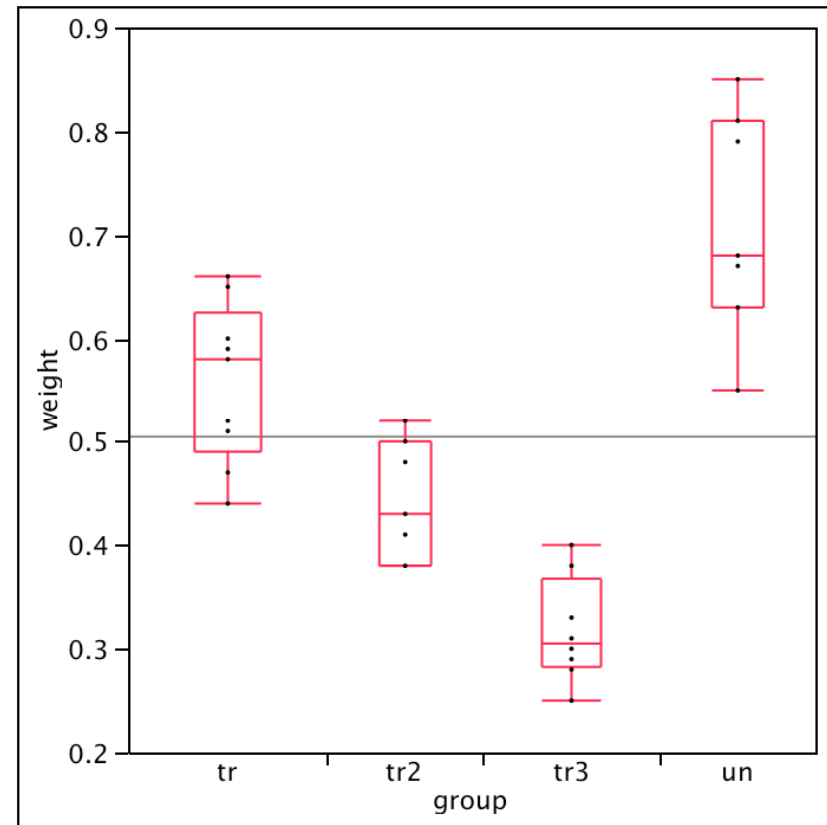
- We **first** examine the multiple populations or multiple treatments to test for overall statistical significance as evidence of any difference among the parameters we want to compare. → **ANOVA F-test**
- If that overall test showed statistical significance, then a detailed follow-up analysis is legitimate.
 - If we planned our experiment with specific alternative hypotheses in mind (before gathering the data), we can test them using **contrasts**.
 - If we do not have specific alternatives, we can examine all pair-wise parameter comparisons to define which parameters differ from which, using **multiple comparisons procedures**.

Herbicides and plant weight after treatment

Does the amount of herbicide on a plant affect plant weight? A completely randomized design is constructed to test this: how might this be done?

Hypotheses: All μ_i are the same (H_0)

versus not All μ_i are the same (H_a)
(at least one of the mean weights is different)

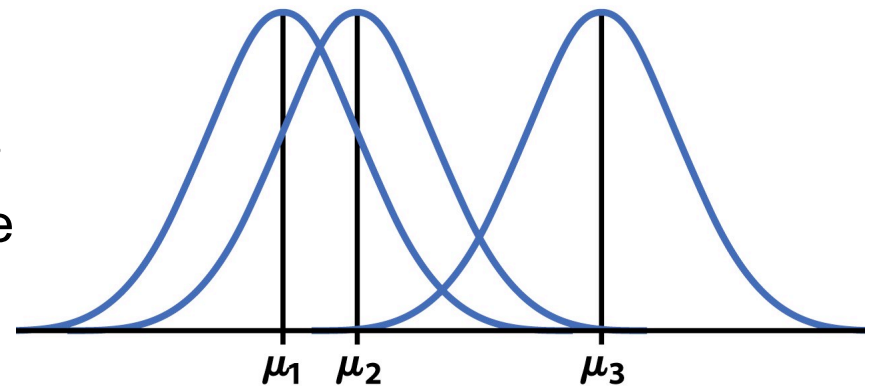


The ANOVA model

Random sampling always produces chance variations. Any “factor effect” would thus show up in our data as the factor-driven differences plus chance variations (“error”):

Data = fit (“factor/groups”) + residual (“error”)

The one-way ANOVA model analyses situations where chance variations are normally distributed $N(0, \sigma)$ so that:



$$X_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, \dots, I$ and $j = 1, \dots, n_i$. The ϵ_{ij} are assumed to be from an $N(0, \sigma)$ distribution. The **parameters of the model** are the population means $\mu_1, \mu_2, \dots, \mu_I$ and the common standard deviation σ .

Testing hypotheses in one-way ANOVA

We have **I independent SRSs**, from I populations or treatments.

The i^{th} population has a **normal distribution** with unknown mean μ_i .

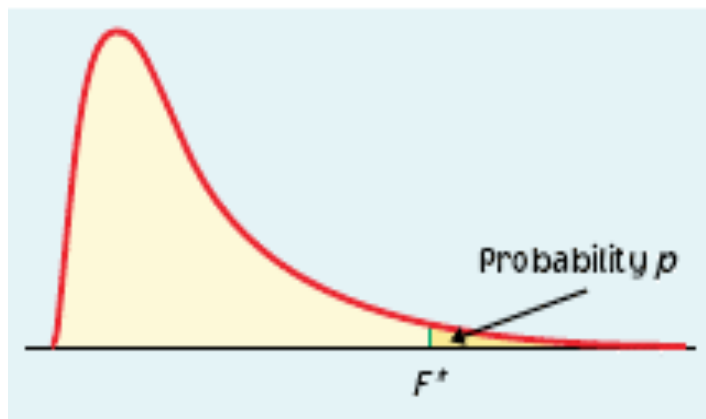
All I populations have the **same standard deviation σ** , unknown.

The ANOVA F statistic tests:

$$F = \frac{\text{SSG}/(I - 1)}{\text{SSE}/(N - I)}$$

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

$$H_a: \text{not all the } \mu_i \text{ are equal.}$$



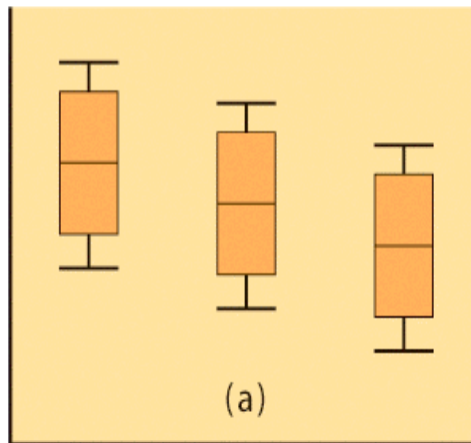
When H_0 is true, F has the **F distribution** with $I - 1$ (*numerator*) and $N - I$ (*denominator*) degrees of freedom. Here, N =total sample size, I =#levels.

The ANOVA F -test

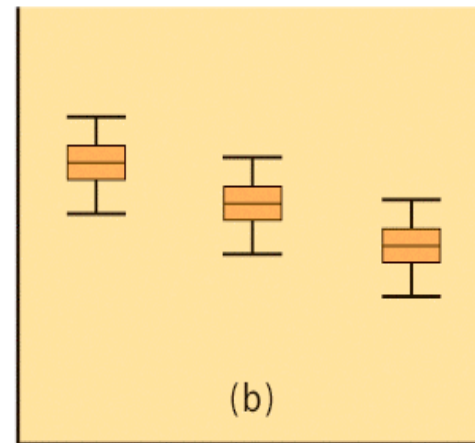
The **ANOVA F -statistic** compares variation due to specific sources (levels of the factor) with variation among individuals who should be similar (individuals in the same sample).

$$F = \frac{\text{variation among sample means}}{\text{variation among individuals in same sample}}$$

Difference in means small relative to overall variability



→ F tends to be small



→ F tends to be large

Difference in means large relative to overall variability

Larger F -values typically yield more significant results. How large depends on the degrees of freedom ($I - 1$ and $N - I$).

Checking our assumptions

Each of the k populations must be **normally distributed** (histograms or normal quantile plots). But the test is robust to deviations from normality for large enough sample sizes.

The ANOVA F-test requires that all populations have the **same standard deviation σ** . Since σ is unknown, this can be hard to check.

Practically: The results of the ANOVA F-test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample standard deviation.

(Equal sample sizes also make ANOVA more robust to deviations from the equal σ rule)

Do herbicides affect plant weight?

group	n_j	Mean(weight)	Std Dev(weight)
tr	9	0.55777778	0.07710022
tr2	7	0.44285714	0.05736267
tr3	8	0.3175	0.05063878
un	7	0.71142857	0.10838643

Conditions required:

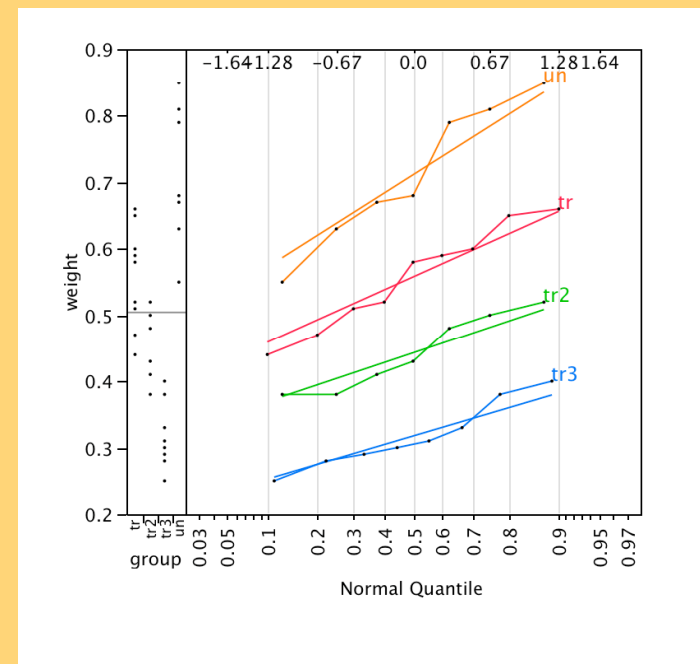
- equal variances: checking that largest s_i is only slightly more than twice the smallest s_i : Largest $s_i = .1084$; smallest $s_i = .0506$. Note that the reference lines are ~parallel...

- Independent SRSs

Four groups are independent

- Distributions “roughly” normal

It is hard to assess normality with such small samples, but the normal quantile plots are OK... points follow the reference lines...



The ANOVA table

Source of variation	Sum of squares SS	DF	Mean square MS	F	P value	F crit
Among or between “groups”	$\sum n_i (\bar{x}_i - \bar{x})^2$	$I - 1$	MSG = SSG/DFG	MSG/MSE	Tail area above F	Value of F for α
Within groups or “error”	$\sum (n_i - 1) s_i^2$	$N - I$	MSE = SSE/DFE			
Total	SST=SSG+SSE ! $\sum (x_{ij} - \bar{x})^2$	$N - 1$				
$R^2 = \text{SSG}/\text{SST}$ Coefficient of determination		$\sqrt{\text{MSE}} = s_p$ Pooled standard deviation				

The sum of squares represents variation in the data: $\text{SST} = \text{SSG} + \text{SSE}$.

The degrees of freedom likewise reflect the ANOVA model: $\text{DFT} = \text{DFG} + \text{DFE}$.

Data (“Total”) = fit (“Groups”) + residual (“Error”)

SAS output for the one-way ANOVA

The GLM Procedure

Dependent Variable: weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.63163361	0.21054454	36.50	<.0001
Error	27	0.15573413	0.00576793		
Corrected Total	30	0.78736774			

R-Square	Coeff Var	Root MSE	weight Mean
0.802209	15.05341	0.075947	0.504516

Here, the calculated F-value (36.50) is larger than F_{critical} for $\alpha = 0.05$.

(check this value in an F-Table (3,27) or use technology...).

Thus, the test is significant at $\alpha = .05 \rightarrow$ Not all mean plant weights are the same; the treatment amount of herbicide is an influential factor.

The **boxplots** on slide #5 above and the normal quantile plots on slide #10 both suggest that increasing amounts of herbicide yield lower plant weights.

Computation details

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\text{SSG}/(I - 1)}{\text{SSE}/(N - I)}$$

MSG, the mean square for groups, measures how different the individual means are from the overall mean (~ weighted average of square distances of sample averages to the overall mean). SSG is the sum of squares for groups.

$$\text{MSG} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \dots + n_I(\bar{x}_I - \bar{x})^2}{I - 1}$$

MSE, the mean square for error is the **pooled sample variance s_p^2** and estimates the common variance σ^2 of the I populations (~ weighted average of the variances from each of the I samples). SSE is the sum of squares for error.

$$\text{MSE} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_I - 1)s_I^2}{N - I}$$



You have calculated a p-value for your ANOVA test. Now what?

If you found a significant result, you still need to determine which treatments were different from which.

- ▣ You can gain insight by looking back at your plots (boxplot, mean \pm s).
- ▣ There are several tests of statistical significance designed specifically for multiple tests. You can choose *apriori* **contrasts**, or *aposteriori* **multiple comparisons**.
- ▣ You can find the confidence interval for each mean μ_i shown to be significantly different from the others.

Multiple comparisons

Multiple comparison tests are variants on the two-sample t -test.

- They use the pooled standard deviation $s_p = \sqrt{\text{MSE}}$.
- The pooled degrees of freedom **DFE**.
- And they compensate for the multiple comparisons.

We compute the t -statistic for all pairs of means:

$$t_{ij} = \frac{\bar{X}_i - \bar{X}_j}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

A given test is significant (μ_i and μ_j significantly different), when

$$|t_{ij}| \geq t^{**} \text{ (df = DFE).}$$

The value of t^{**} depends on which procedure you choose to use.

The Bonferroni procedure

The **Bonferroni procedure** performs a number of pair-wise comparisons with *t*-tests and then multiplies each *p*-value by the number of comparisons made. This ensures that the probability of making *any* false rejection among all comparisons made is no greater than the chosen significance level α .

As a consequence, the higher the number of pair-wise comparisons you make, the more difficult it will be to show statistical significance for each test. But the chance of committing a type I error also increases with the number of tests made. The Bonferroni procedure lowers the working significance level of each test to compensate for the increased chance of type I errors among all tests performed ($\alpha/n \dots$)

Simultaneous confidence intervals

We can also calculate simultaneous level C **confidence intervals for all pair-wise differences** $(\mu_i - \mu_j)$ between population means:

$$CI : (\bar{x}_i - \bar{x}_j) \pm t^{**} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

- s_p^2 is the pooled variance, MSE.
- t^{**} is the t critical with degrees of freedom $DFE = N - I$, adjusted for multiple, simultaneous comparisons (e.g., Bonferroni procedure).