

Topic 3. Single factor ANOVA [ST&D Ch. 7]

"The analysis of variance is more than a technique for statistical analysis. Once it is understood, ANOVA is a tool that can provide an insight into the nature of variation of natural events" Sokal & Rohlf (1995), BIOMETRY.

3. 1. The F distribution [ST&D p. 99]

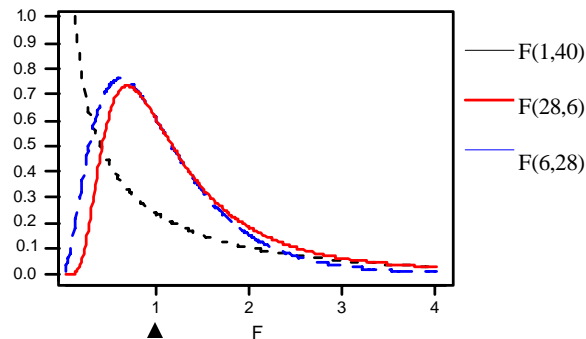


Fig 1. Three representative F-distributions. (Note $F_{(1,40)}$ similar χ^2_1)

Assume that you are sampling at random from two different populations with **equal variance**:

- 1) Sample n_1 items and calculate s^2_1 (df: $n_1 - 1$)
- 2) Sample n_2 items and calculate s^2_2 (df: $n_2 - 2$)
- 3) Calculate for each sample

$$F_s = s^2_1 / s^2_2 \quad \text{This ratio will be close to 1.}$$

The expected distribution of this statistic is called **F-distribution**. Is determined by **two** values for df (indicated by Nu **v**).

Values in the table represent the proportion of the F-distribution to the right of the given- F-value (in one tail) and v_1, v_2 are the degrees of freedom for the numerator and denominator of the variance ratio, respectively.

If we are testing $H_0 = s^2_1 = s^2_2$ vs. $H_a = s^2_1 \neq s^2_2$ (two tail test)

And we find $F_{\alpha/2=0.025 [v_1=9, v_2=9]} = 4.03$

Interpretation: The ratio s^2_1 / s^2_2 , from samples of ten individuals from normally distributed populations with equal variance, is expected to be larger than 4.03 ($F_{\alpha/2=0.025}$) or lower than 0.24 ($F_{1-\alpha/2=0.975}$ not in Table) by chance in only 5% of the experiments.

3. 2. Testing the hypothesis of equality of variances [ST&D 116-118]

The F statistic can be used as a test for the hypothesis

$$H_0: s_u^2 = s_v^2 \text{ vs. } H_A: s_u^2 \neq s_v^2$$

H_0 is rejected at the α level of significance if the ratio s_v^2 / s_u^2 is $\geq F_{\alpha/2, m-1, n-1}$

This test is rarely used because it is **very sensitive to departures from normality**.

3. 3. Testing the hypothesis of equality of two means [ST&D p. 98-112]

The ratio between two estimates of σ^2 can be used to test differences between means:

$$H_0: \mu_1 - \mu_2 = 0 \text{ versus } H_A: \mu_1 - \mu_2 \neq 0$$

How can we use variances to test differences between means?

$$F = \frac{\text{estimate of } \sigma^2 \text{ from means}}{\text{estimate of } \sigma^2 \text{ from individuals}}$$

The denominator is an estimate of σ^2 from the individuals in each sample. It is a **weighted average** of the sample variances.

The numerator is an estimate of σ^2 from the means.

$$\sigma_y^2 = \sigma^2/n$$

so

$$\sigma^2 = \sigma_y^2 * n$$

This implies that means may be used to estimate σ^2 by multiplying the variance of sample means σ^2/n by n .

When the two populations have **different means** (but same variance), the estimate of σ^2 based on sample means will include a contribution attributable to the difference between population means and **F will be higher than expected by chance**.

Example

Table 1. Yields (100 lb./acre) of wheat varieties 1 and 2.

Varieties	Replications	Y_i	\bar{Y}_i	s^2_i
1	19 14 15 17 20	85	$\bar{Y}_{1.} = 17$	6.5
2	23 19 19 21 18	100	$\bar{Y}_{2.} = 20$	4.0
		$Y_{..} = 185$	$\bar{Y}_{..} = 18.5$	

t = 2 treatments and r = 5 replications,

We will *assume* that the two populations have the same variance s^2

1) Estimate the sample variance *within samples* or *experimental error*

$$s_w^2 = \frac{(r_1 - 1)s_1^2 + (r_2 - 1)s_2^2}{(r_1 - 1) + (r_2 - 1)} = 4 * 6.5 + 4 * 4.0 / (4 + 4) = \mathbf{5.25}$$

2) Estimate the *between samples* variability. Under H_0 , $\bar{Y}_{1.}$ and $\bar{Y}_{2.}$ estimate the same population mean. To estimate the variance of means:

$$s_Y^2 = \frac{\sum_{i=1}^t (\bar{Y}_i - \bar{Y}_{..})^2}{t - 1} = [(17 - 18.5)^2 + (20 - 18.5)^2] / (2 - 1) = 4.5$$

Therefore the *between samples* estimate is

$$s_b^2 = r s_Y^2 = 5 * 4.5 = 22.5$$

These two variances are used in the F test as follows:

$$\mathbf{F} = s_b^2 / s_w^2 = 22.5 / 5.25 = 4.29$$

under our assumptions (normality, equal variance, etc.), this ratio is distributed according to an $F_{(t-1, t(r-1))}$ distribution.

This result indicates that the variability *between* the samples is 4.29 times larger than the variability *within* the samples.

$$s_b^2 \mathbf{v} = 1 \text{ (2 treatments - 1)} \quad s_w^2 \mathbf{v} = t(r-1) = 4 + 4 = 8.$$

From Table A.6, p. 614 of ST&D, $F_{0.05, 1, 8} = 5.32$. Since $4.29 < 5.32$ we fail to reject H_0 at $\alpha=0.05$ significance level. An F value of 4.29 or larger happens just by chance about 7% of the times for these degrees of freedom.

skip

3. 3. 1 Relationship between F and t for two treatments

$$F_{(1, df), 1-\alpha} = t_{df, 1-\alpha/2}^2 \quad \text{or} \quad t = \sqrt{\frac{s_b^2}{s_w^2}}$$

Example: $t_{0.025, 8} = 2.306$ and $t^2 = 2.306^2 = 5.32 = F_{0.05, 1, 8}$

3. 4. The linear additive model

3. 4. 1. One population: Applicable to the problem of estimating or making inferences about population means and variances.

$$Y_i = \mu + \epsilon_i \text{ (epsilon sub i)}$$

- This model attempts to explain an observation as a mean μ plus a random element of variation ϵ_i .
- The ϵ_i 's are assumed to be from a population of uncorrelated ϵ 's with mean zero. Independence among ϵ 's is assured by random sampling.

3. 4. 2. Two populations: This model is more general because it describes two populations simultaneously

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \text{Where } \mu = (\mu_1 + \mu_2)/2$$

Y_{ij} is composed of the grand mean μ of the population plus an effect for the treatment τ_i plus a random deviation ϵ_{ij} .

The data represent the model as:

$$Y_{ij} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.})$$

3. 4. 3. More than 2 populations: Assumptions of the *Model I ANOVA or fixed model* (Model II: Topic 10).

1. **Treatment effects are additive and fixed by the researcher**

$$\sum \tau_i = 0 \quad \text{and} \quad H_0: \tau_1 = \dots = \tau_t = 0 \quad \text{vs.} \quad H_A: \text{some } \tau_i \neq 0$$

2. Experimental errors are **random, independent** and **normally** distributed about a zero mean, and with a **common variance**.

3. When H_0 is false there will be an additional component in the variance between treatments = $r \sum \tau_i^2 / (t-1)$ (* by r because variation of means)

3. 5. 1. The Completely Random Design CRD

- **CRD** is the basic ANOVA design.
- A single factor is varied to form the different treatments.
- These treatments are applied to **t** independent random samples of size **r**.
- The total sample size is $n = rt$.
- $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_t$ against **H_1 : not all the μ_i 's are equal.**

The results are usually summarized in an ANOVA table:

Source	df	Definition	SS	MS	F
Treatments	t - 1	$r \sum_i (Y_{i.} - \bar{Y}_{..})^2$	SST	SST/(t-1)	MST/ MSE
Error	t(r-1) = n - t	$\sum_{i=1}^t \sum_{j=1}^r (Y_{ij} - \bar{Y}_{i.})^2$	SS - SST	SSE/(n-t)	
Total	n - 1	$\sum_{i,j} (Y_{ij} - \bar{Y}_{..})^2$	SS		

SST = sum of squares for treatments.

SSE = sum of squares for error.

SST definition formula: the * **r** results in **MST** estimating σ^2 rather than σ^2/r

$$\mathbf{SS} = \mathbf{SST} + \mathbf{SSE}.$$

We can decompose the **total SS** into a portion due to **variation among groups** and another portion due to **variation within groups**.

The v are also additive.

MST = SST/(t-1), the *mean square for treatments*

It is an independent estimate of σ^2 when H_0 is true.

MSE = SSE/(n-t) is the *mean square for error*

- It gives the average dispersion of the items around the group means.
- It is an estimate of a common σ^2
- Is the **within variation** or variation among observations treated alike.
- MSE estimates the common σ^2 if the treatments have equal variances.

F = **MST/MSE**. We expect **F** approximately equal to 1 if no treatment effect.

$$\text{Expected } \frac{MST}{MSE} = \frac{\sigma^2 + r \sum \tau_i^2 / (t-1)}{\sigma^2}$$

3. 5. 1. 2. Testing the assumptions associated with ANOVA

Independence of error: Guaranteed by the random allocation of e.u.

Normal distribution: Shapiro and Wilk test statistics W (ST&D p.567, and SAS PROC UNIVARIATE NORMAL). Normality is rejected if W is sufficiently smaller than 1.

Homogeneity of variances: to determine if the variance is the same within each of the groups defined by the independent variable.

- **Bartlett's test:** (ST&D 481) inaccurate if the underlying distribution is even slightly nonnormal.
- **Levene's test:** is more robust to deviations from normality.

Levene's test is an ANOVA of the squares of the residuals of each observation from the treatment means.

Original data				Residuals		
T1	T2	T3		T1	T2	T3
3	8	5		-1	2	0
4	6	6		0	0	1
5	4	4		1	-2	-1
4	6	5	Treatment means			

Advantages of CRD (ST&D 140)

- Simple design
- Can accommodate well unequal number of replications per treatment
- Loss of information due to missing data is smaller than in other designs
- The number of d.f. for estimating experimental error is maximum
- Can accommodate unequal variances, using a **Welch's variance-weighted ANOVA** (Biometrika 1951 v38, 330, SAS ANALYST).

Disadvantages of CRD (ST&D 141)

- The experimental error includes the entire variation among e.u. except that due to treatments.

3.5.1.4.1. Power

- The power of a test is the probability of detecting a real treatment effect.
- To calculate the ANOVA power, we 1st calculate the critical ϕ value (a standardized measure of the expected differences among means in σ units).

ϕ depends on:

- The number of treatments (k)
- The number of replications (r) -> **When r \uparrow \Rightarrow Power \uparrow**
- The **treatment difference** we want to detect (**d**) -> **When d \uparrow \Rightarrow Power \uparrow**
- An estimate of the population variance ($\sigma^2 = MS_{error}$)
- The probability of rejecting a true null hypothesis (α).

$$\phi = \sqrt{\frac{r}{MSE} \sum \frac{\tau_i^2}{k}} \quad \text{With } \tau_i = \mu_i - \mu$$

In a CRD we can **simplify** this general formula if we assume all $\tau_i = 0$ except the extreme treatment effects $\mu_{(k)} - \mu_{(1)}$. If **d** = $\mu_{(k)} - \mu_{(1)} \Rightarrow \tau_i = d/2$

$$\sum \frac{\tau_i^2}{k} = \frac{(d/2)^2 + (d/2)^2}{k} = \frac{d^2/4 + d^2/4}{k} = \frac{d^2/2}{k} = \frac{d^2}{2k}$$

And the ϕ formula simplifies to
$$\phi = \sqrt{\frac{d^2 * r}{2k * MS_{error}}}$$

Entering the chart for $v_1 = df = k-1$ and α (0.05 or 0.01) \Rightarrow the interception of ϕ and $v_2 = df = k(n-1)$ gives the power of the test.

Example: Suppose that one experiment has k=6 treatments with r=2 replications each. The difference between the extreme means was 10 units, MSE= 5.46, and the required $\alpha = 5\%$. To calculate the power:

$$\phi = \text{SQR}(2 * 10^2 / 2 * 6 * 5.46) = 1.75$$

- Use Chart $v_1 = k-1 = 5$.
- Use the set of curves to the left ($\alpha = 5\%$).
- Select curve $v_2 = k(n-1) = 6$.
- The height of this curve corresponding to the abscissa of $\phi = 1.75$ is the power of the test.
- In this case the power is slightly greater than 0.55.

3. 5. 1. 4. 2. Sample size

To calculate the number of replications 'n' for a given α and power:

- a) Specify the constants,
- b) Start with an arbitrary number of 'n' to compute ϕ ,
- c) Use Pearson and Hartley's charts to find the power,
- d) Iterate the process until a minimum 'n' value which satisfies a required power for a given α level is found.

Example:

Suppose that 6 treatments will be involved in a study and the anticipated difference between the extreme means is 15 units. What is the required sample size so that this difference will be detected at $\alpha = 1\%$ and power = 90%, knowing that $\sigma^2 = 12$?

(note, $k = 6$, $\alpha = 1\%$, $\beta = 10\%$, $d = 15$ and $MSE = 12$).

$$\phi = \sqrt{\frac{d^2 * r}{2k * MS_{error}}}$$

n	df	ϕ	(1- β) for $\alpha=1\%$
2	6(2-1)= 6	1.77	0.22
3	6(3-1)= 12	2.17	0.71
4	6(4-1)= 18	2.50	0.93

Thus 4 replications are required for each treatment to satisfy the required conditions.

3. 5. 2. Subsampling: the nested design

If measures of the same *experimental unit* vary too much, the experimenter can make several observations within each *experimental unit*.

Such observations are made on subsamples or *sampling units*.

Examples

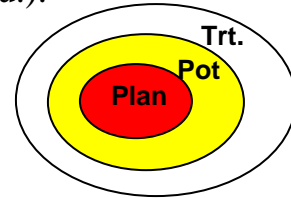
- Sampling individual plants within pots where the pots (e.u.).
- Sampling individual trees within an orchard plot (e.u.).

Hierarchical way or **nested analysis of variance**.

Can have multiple levels

Applications of nested ANOVA:

- Ascertain the magnitude of error at various stages of an experiment or an industrial process.
- Estimate the magnitude of the variance attributable to various levels of variation in a study (e.g. quantitative genetics).
- Discover sources of variation in natural population or systematic studies.



3. 5. 2. 1. Linear model for subsampling:

$$Y_{ijk} = \mu + \tau_i + \varepsilon_{j(i)} + \delta_{k(ij)}$$

- Two random elements are obtained with each observation: ε and δ .
- The $\delta_{k(ij)}$ are the errors associated with each subsample.
- The $\delta_{k(ij)}$ are assumed normal with mean 0 and variance s^2 .

The subscript $\varepsilon_{j(i)}$ indicates that the j^{th} level of factor B (pot) is **nested** under the i^{th} level of factor A (treatment). This means that other measures below that level are not real replications.

This is represented in the data as

$$Y_{ijk} = \bar{Y} \dots + (\bar{Y}_{i..} - \bar{Y} \dots) + (\bar{Y}_{ij.} - \bar{Y}_{i..}) + (Y_{ijk} - \bar{Y}_{ij.}).$$

The dot notation: the dot replaces a subscript and indicates that all values covered by that subscript have been added

3. 5. 2. 2. Nested ANOVA with equal subsample numbers: computation

Example: ST&D p. 159: **6 treatments** and **3 pots** nested under each level of treatment (replications), and **4 subsamples**. Variable: stem growth

Note that the Pot number is just an ID (not a treatment)



	Treatment 1			Treatment 2...		
	T₁ : low T/ 8 h	T₂ :low T/ 12 h	T₃ :low T/ 16 h	T₄ :high T/ 8 h	T₅ :high T/ 12 h	T₆ :high T/ 16 h
Plant N _o	Pot number 1 2 3	Pot number 1 2 3	Pot number 1 2 3	Pot number 1 2 3	Pot number 1 2 3	Pot number 1 2 3
1	3.5 2.5 3.0	5.0 3.5 4.5	5.0 5.5 5.5	8.5 6.5 7.0	6.0 6.0 6.5	7.0 6.0 11.0
2	4.0 4.5 3.0	5.5 3.5 4.0	4.5 6.0 4.5	6.0 7.0 7.0	5.5 8.5 6.5	9.0 7.0 7.0
3	3.0 5.5 2.5	4.0 3.0 4.0	5.0 5.0 6.5	9.0 8.0 7.0	3.5 4.5 8.5	8.5 7.0 9.0
4	4.5 5.0 3.0	3.5 4.0 5.0	4.5 5.0 5.5	8.5 6.5 7.0	7.0 7.5 7.5	8.5 7.0 8.0
Pot totals = Y _{ij.}	15 17.5 11.5	18 14 17.5	19 21.5 22	32 28 28	22 26.5 29	33 27 35
Treatment totals = Y _{i..}	44.0	49.5	62.5	88.0	77.5	95.0
Treatment means = $\bar{Y}_{i..}$	3.7	4.1	5.2	7.3	6.5	7.9

In this example $t = 6$, $r = 3$, and $s = 4$, and $n = trs = 72$

CRD

$$\sum_{i=1}^t \sum_{j=1}^r (Y_{ij} - \bar{Y}_{..})^2 = r \sum_{i=1}^t (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^t \sum_{j=1}^r (Y_{ij} - \bar{Y}_{i.})^2 \text{ or } \mathbf{SS} = \mathbf{SST} + \mathbf{SSE}.$$

Degrees of freedom: **SSv** = $n-1$, **SSTv** = $t-1$, and **SSEv** = $n-t$, respectively.

In the nested design the SST is unchanged but the SSE is further partitioned into two components. The resulting equation can be written

Nested CRD

$$\sum_{i=1}^t \sum_{j=1}^r \sum_{k=1}^s (Y_{ijk} - \bar{Y}_{...})^2 = rs \sum_{i=1}^t (\bar{Y}_{i..} - \bar{Y}_{...})^2 + s \sum_{i=1}^t \sum_{j=1}^r (\bar{Y}_{ij.} - \bar{Y}_{i..})^2 + \sum_{k=1}^s (Y_{ijk} - \bar{Y}_{ij.})^2$$

or **SS** = **SST** + **SSEE** + **SSSE**

SSE= SSEE + SSSE

SSEE: Sum of squares due to *experimental error*: variation among plots (pots) within treatments.

SSSE: Sum of squares due to *sampling error*: variation among subsamples (plants) within plots (pots).

ANOVA table:

Source of variation	df	SS	MS	F	Expected MS
Treatments (t _i)	t - 1 = 5	SST	SST/ 5	MST/MSEE	$\sigma_{\delta}^2 + 4\sigma_{\epsilon}^2 + 12\Sigma\tau^2/5$
Exp. Error (e _{j(i)})	t (r - 1) = 12	SSEE	SSEE/ 12	MSEE/MSSE	$\sigma_{\delta}^2 + 4\sigma_{\epsilon}^2$
Samp. Error (d _{k(ij)})	rt (s - 1) = 54	SSSE	SSSE/ 54		σ_{δ}^2
Total	trs - 1 = 71	SS			

In testing a hypothesis about population treatment means, the appropriate divisor for *F* is the experimental error MS since it includes variation from all sources that contribute to the variability of treatment means except treatments.

To estimate the different components of variance in the pot experiment:

Variance Source	df	Sum of Squares	Mean Squares	Variance component	Percent of total
Total	71	255.91	3.60	4.05	100.0 %
trtmt	5	179.64	35.92	2.81	69.4 %
pot	12	25.83	2.15	0.30	7.5 %
error	54	40.43	0.93	0.93	23.0 %

MSSE= 0.93

MSEE= $\sigma_{\delta}^2 + 4\sigma_{\epsilon}^2$ $\sigma_{\epsilon}^2 = (MSEE - \sigma_{\delta}^2)/4 = (2.15-0.93)/4 = 0.30$

MST= $\sigma_{\delta}^2 + 4\sigma_{\epsilon}^2 + Tr$ $Tr = (MST - MSEE)/12 = (35.92-2.15)/12 = 2.81$

SAS CODE

proc GLM;

```
class trtmt pot;
model growth= trtmnt pot (trtmt);
random pot (trmt);
test h=trmt e=pot(trtmt);
```

proc VARCOMP;

```
class trtmt pot;
model growth= trtmt pot (trtmt);
```

Pot(trtmt): Pot only has meaning within treatment (is an ID variable).
Pot 1 in treatment 1 is not related to Pot 1 in treatment 2!

3.5.2.3. The optimal allocation of resources

ST&D 173 or Biometry Sokal & Rohlf pg. 309)

- One of the main reasons to do a nested design is to investigate how the variation is distributed between experimental units and subsamples.
- Once the variance of the experimental units ($s^2_{e.u.}$) and the variance of the subsamples s^2_{SUB} are known, the variance of the means can be calculated as

$$s^2_{\bar{Y}} = \frac{s^2_{eu}}{N_s * N_r} + \frac{s^2_{SUB}}{N_r} \quad N_s = N^{\circ} \text{ of subsamples per e.u. } \& \quad N_r = N^{\circ} \text{ of replications.}$$

- This formula can be used to test the relative efficiency of designs with \neq number or replicates and subsamples.

The relative efficiency is not meaningful, unless the **relative cost of obtaining the two designs are taken into consideration.**

- If one design is twice as efficient as another but at the same time is ten times as expensive we might not choose it.
- Cost function.

$$C = N_s * N_r(C_{SUB}) + N_r(C_{eu})$$

- To calculate the number of subsamples (N_s) per experimental unit that will result in simultaneous minimal cost and minimal variance the following formula may be used:

$$N_s = \sqrt{\frac{C_{e.u.} * s^2_{SUB}}{C_{SUB} * s^2_{e.u.}}}$$

The optimum number of subsamples will increase when the relative cost of the subsamples is low and the variance within the experimental unit is high (S^2_{SUB}).

Example Pot: \$3, plant \$1 then: $N_s = \sqrt{\frac{3 * 0.93}{1 * 0.30}} = 3$ optimum 3 plants per pot

If $C_{eu} = C_{SUB} \Rightarrow N_s = \frac{s_{SUB}}{s_{e.u.}}$ subsamples are valuable only if

s between subsamples > s between e.u.