Ch. 10 - Analysis of Variance (ANOVA)

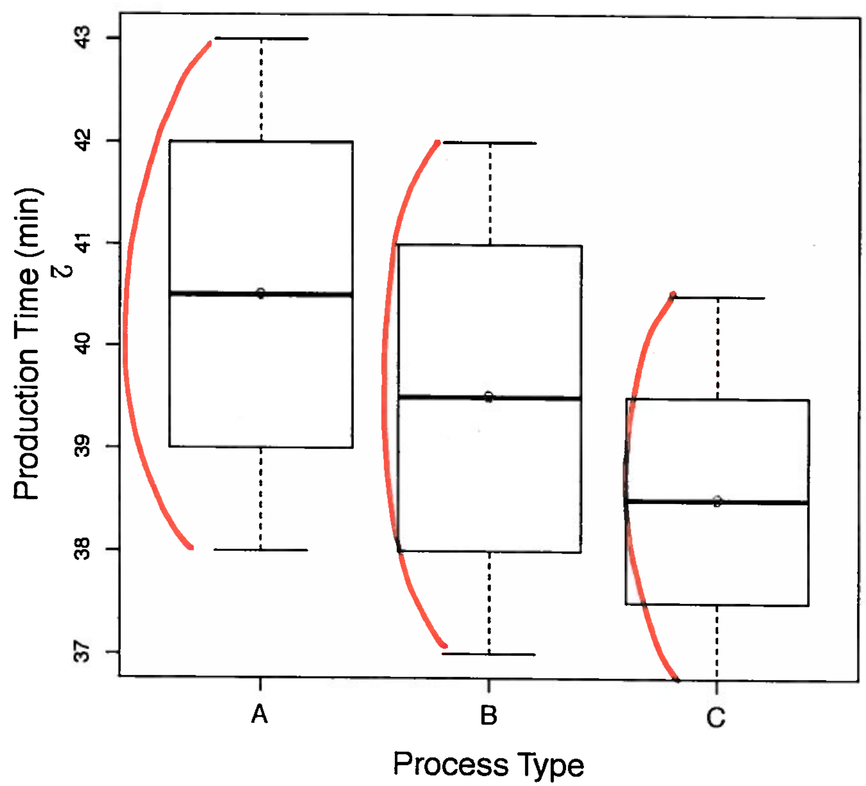
Objective: to compare the means of 3 or more independent populations
e.g.

- Comparison of the mean reduction in blood pressure for patients randomized to three different treatment groups: drugs A, B, and C
- Compare the mean production time for 3 different process types.

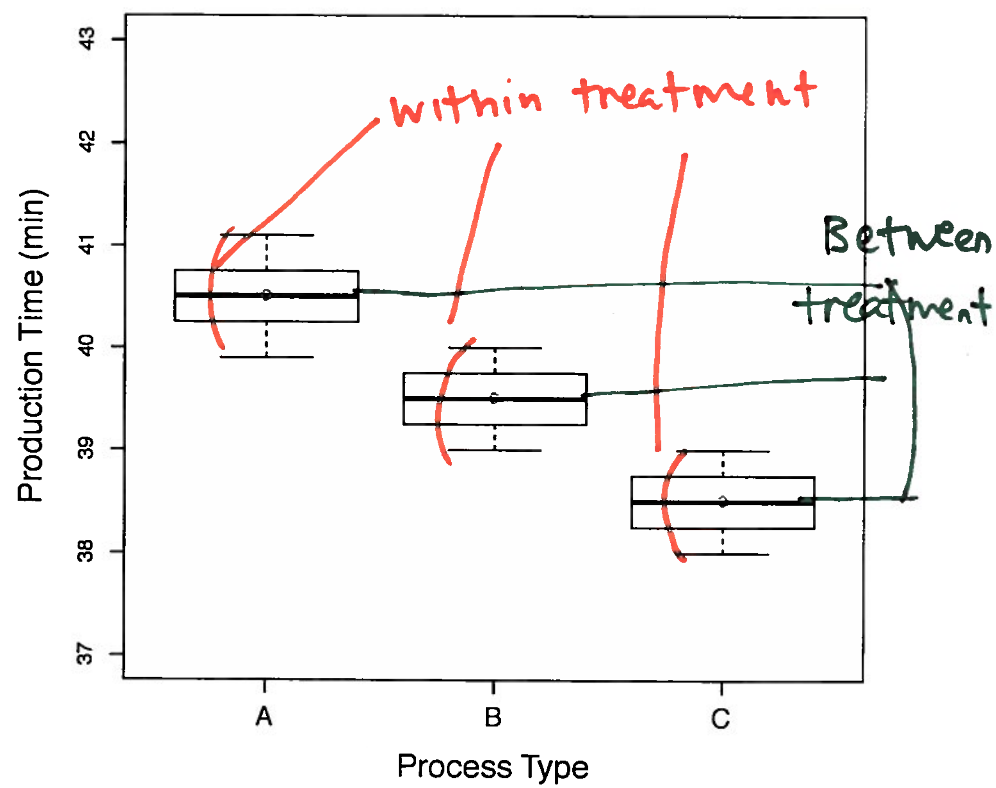
ANOVA is a statistical method that tests the equality of three or more population means by analyzing the variation in the data.

Let's compare the mean production time for three different process types, A, B and C. Which of the two plots below show stronger evidence that the mean production time differs significantly among the three processes?

Case A:



Case B:



ANOVA Notation:

k = the number of populations under investigation

x_{ij} = measurement on the ith individual taken from the jth sample

Population	Population mean	Population SD	Sample size	Sample mean	Sample SD
1	μ_1	σ_1	n_1	$\bar{x}_1$	s_1
2	μ_2	σ_2	n_2	$\bar{x}_2$	s_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	μ_k	σ_k	n_k	$\bar{x}_k$	s_k
			Total sample size	Grand Mean	
			$N = n_1 + n_2 + ... + n_k$	$\bar{\bar{x}} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}$	

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_A : \mu_i \neq \mu_j \text{ for some } i \neq j$$

Assumptions:

Rule of thumb: Ratio between largest and smallest sd < 2

1. The k population distributions are normal
2. The k population variances are equal ($\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$)
3. The samples are selected randomly and independently from their respective populations

The total variation in the data ($SSTotal$ total sum of squares) comes from two sources:

1. variation between groups/ treatments (SST Treatment sum of squares)
2. variation within groups/treatments (SSE Error sum of squares)

$$SSTotal = \quad SST \quad + \quad SSE$$

obs

jth group mean

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{\overline{x}})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{x}_j - \overline{\overline{x}})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_j)^2$$

$$= \sum_{j=1}^{k} n_j(\overline{x}_j - \overline{\overline{x}})^2 + \sum_{j=1}^{k}(n_j - 1)s_j^2$$

each obs

grand mean

jth group mean

grand mean

$$\text{Mean Squares } (MS) = \frac{\text{Sum of Squares}}{\text{degrees of freedom}}$$

1. $MST = \frac{SST}{k-1}$ (Mean Square Treatment)
2. $MSE = \frac{SSE}{N-k}$ (Mean Square Error)

<u>ANOVA test procedure</u>

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_A : \mu_i \neq \mu_j \text{ for some } i \neq j$$

Test statistic:

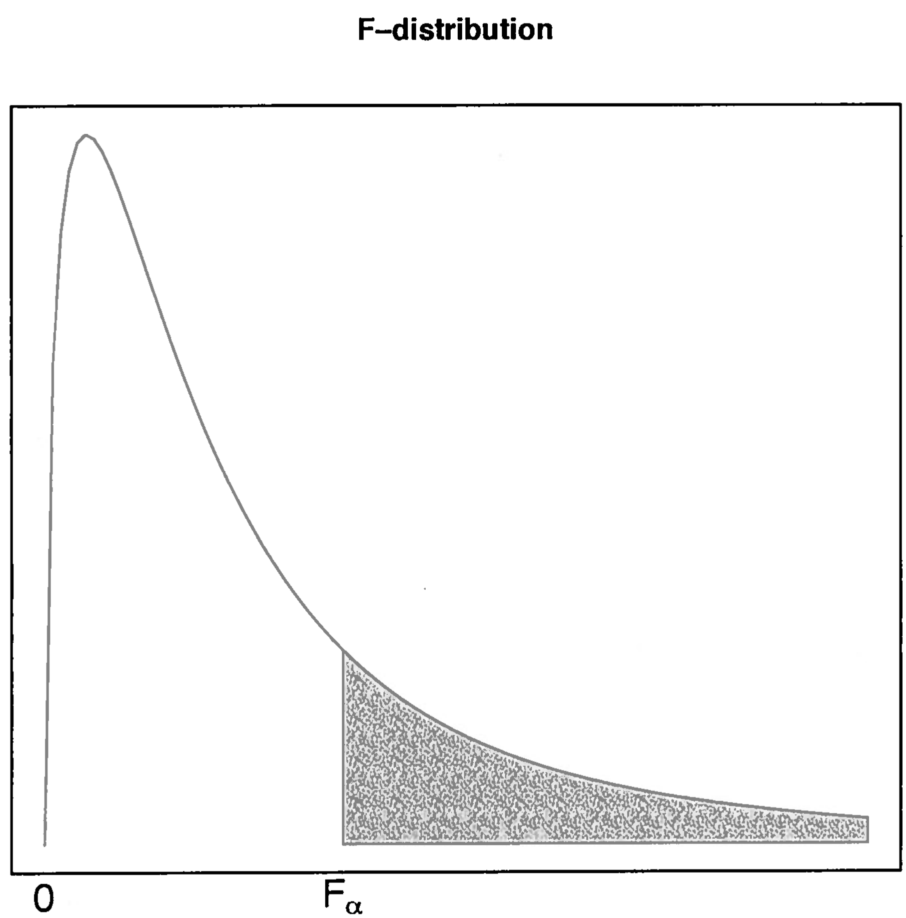
$$F_{obs} = \frac{MST}{MSE}$$

variation btw samples

variation among individuals within same sample

Under H_0, $F_{obs} = \frac{MST}{MSE}$ follows the F-distribution with

- ν_1 (numerator degrees of freedom $= df(SST) = k - 1$) and
- ν_2 (denominator degrees of freedom $= df(SSE) = N - k$)



Rejection Region:
If $F_{obs} \geq F_{\alpha}$ we reject H_0
If $F_{obs} < F_{\alpha}$ we do not reject H_0

When H_0 is rejected, we conclude that at least one pair of population means are significantly different from each another.

The ANOVA table:

Source of Variation	df	Sum of Squares	Mean Squares	F-ratio
Treatments	$k-1$	SS_T	$MS_T = \frac{SS_T}{k-1}$	$\frac{MS_T}{MS_E}$
Error	$N-k$	SS_E	$MS_E = \frac{SS_E}{N-k}$	
Total	$N-1$	SS_{Total}		

$k=3$

$N = 26 + 30 + 32 = 88$

Example 1. Resting pulse rates for a random sample of 26 heavy smokers had a mean of 80 beats per minute (bpm) and a standard deviation of 5 bpm. Among 32 randomly chosen light smokers the mean and standard deviation were 77 and 6 bpm. Among 30 non-smokers the mean was 74 bpm with a standard deviation of 5bpm. All three sets of data were roughly symmetric and had no outliers. The ANOVA table shows:

Analysis of Variance Table

Source	df	Sum of Squares	Mean Square	F Ratio
Smoking	2	502.364	251.182	8.66
Error	85	2466	29.01	
Total	87	2968.364		

Fill in the ANOVA table. Is there evidence of a difference in mean pulse rate between these groups at $\alpha = 0.05$?

Solution:

μ_i: population mean resting pulse rate for $i = 1, 2, 3$

heavy light non.

$H_0: \mu_1 = \mu_2 = \mu_3$

H_A: not all means are equal

Assumptions:

- independent random samples
- Spreads of 3 groups appear to be about the same.
- we are told data appear to be approximately Normal.

$F_{obs} = 8.66$

Look up F-table

$F_{2, 85, \alpha=0.05} \approx F_{2, 60, \alpha=0.05} = 3.15$

numerator df

denominator df

use smaller df

Since $F_{obs} = 8.66 > 3.15$

we reject Ho. Conclude there are statistically significant differences between 3 groups

Last Class:

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

$H_A: \mu_i \neq \mu_j$ for some $i \neq j$

Test stat,

$$F_{obs} = \frac{MST}{MSE} \geqslant F_\alpha \quad \text{Reject } H_0.$$

$\nu_1 = k-1$

$\nu_2 = N-K$

Bonferroni's Multiple Comparison Procedure

- If we do not reject H_0, we stop - there is no point in further testing.
- When we reject H_0, we conclude that at least one pair of population means are significantly different from each another.
  What if we want to know which means are different? We can test whether any pairs differ.

We know how to compare two means with a t-test, but now we want to do several tests and each test poses the risk of a Type I error. With each additional test, the risk of making Type I error grows bigger than the α level of each individual test.

The **Bonferroni method** allows us to examine many pairs simultaneously while the overall Type I error rate stays at (or below) α. It adjusts the critical value so that the resulting confidence intervals are wider and the corresponding Type I error rates are lower for each test.

- In general, if we have k groups/treatments, there are $K = \binom{k}{2}$ comparisons.
  e.g. if $k = 3$ groups, there are $K = \binom{3}{2} = 3$ comparisons
- If we wish to have an overall confidence level of $(1-\alpha) \times 100$, then we use the adjusted alpha level:

$$\alpha^* = \frac{\alpha}{K}$$

- The K simultaneous Bonferroni confidence intervals have as endpoints the values:

$$(\bar{x}_A - \bar{x}_B) \pm t^{\alpha^*}_{N-k} \times \sqrt{MSE}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

MSE estimates the common variance of the populations.

Recall:

- If the confidence interval contains 0, there is no evidence to suggest that the population means are different.

- If 0 is not included in the CI, there is evidence to suggest that the two population means are different.

Example 2. Continued from example 1 (smoking example) above. Find a 95% confidence interval for the difference between two groups, light smokers and non-smokers. $N = 88, \quad k = 3$

$$K = \binom{3}{2} = 3 \text{ comparisons}$$

$$\alpha^* = \frac{\alpha}{K} = \frac{0.05}{3} = 0.0167$$

$$(\bar{x}_2 - \bar{x}_3) \pm t_{N-k}^{\alpha^*} \sqrt{MSE} \sqrt{\frac{1}{n_2} + \frac{1}{n_3}}$$

$$(77 - 74) \pm t_{85, \frac{0.0167}{2}} \sqrt{29.01} \sqrt{\frac{1}{32} + \frac{1}{30}}$$

split α^* into 2 because its a confidence interval.

$$3 \pm 2.44 \times 5.39 \sqrt{\frac{1}{32} + \frac{1}{30}}$$

$t_{\frac{0.0167}{2}}$ not available in t-table

use $qt(1 - \frac{0.0167}{2}, 85)$ in R if you want to try it.

$3 \pm 3.34 = (-0.34, 6.34)$ bpm.

We cannot conclude there is a significant difference in mean pulse rates between light + non smokers since 0 is in the interval.

You could do same process for smokers + non and smokers + light.

$$MSE = \frac{SSE}{N-K}$$

$$\sqrt{MSE} = \sqrt{\sum_{j=1}^{K} \frac{(n_j - 1)\, s_j^2}{N-K}}$$

$$= \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + \ldots + (n_K - 1)\, s_K^2}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_K - 1)}}$$