Chapter 11
The Analysis of Variance

11.1 One Factor Analysis of Variance

11.1 One Factor Analysis of Variance
11.1.1 One Factor Layouts

- Suppose that an experimenter is interested in **k populations** with unknown population means $\mu_1, \mu_2, \ldots, \mu_k$
- The one factor analysis of variance methodology is appropriate for comparing three or more populations.
- The observation x_{ij} represents the j-th observation from the *i*-th population.
- The sample from population *i* consists of the n_i observations, $\mathrm{x_{i1}}, \mathrm{x_{i2}}, \cdots, x_{in_i}$.
- If the sample sizes, $n_1, n_2, \cdots, n_k,$ are all equal, then the data set is said to be balanced; otherwise, the data set is said to be unbalanced.

- The total sample size of the data set is $n_T = n_1 + \cdots + n_k$.
- A data set of this kind is called a one-way or **one factor** layout.
- The single factor is said to have *k* **levels** corresponding to the *k* populations under consideration.
- Completely randomized designs : the experiment is performed by randomly allocating a total of $\mathrm{n_T}$ "units" among the *k* populations.
- Modeling assumption: $x_{ij} = \mu_i + \epsilon_{ij}$
  where the error terms ϵ_{ij} follow $\mathrm{N}(0, \sigma^2)$.
- Equivalently,

$$\mathrm{x_{ij}} \ (iid) \sim N(\mu_i, \sigma^2)$$

- Point estimates of the unknown population means

$$\hat{\mu}_i = \overline{x}_{i\cdot} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, 1 \leq i \leq k$$

- Consider testing H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ vs H_A: $not\ H_0$.

- Acceptance of the null hypothesis indicates that there is no evidence that any of the population means are unequal.
- Rejection of the null hypothesis implies that there is evidence that at least some of the population means are unequal.

Example 62 : Collapse of Blocked Arteries

level 1 : stenosis = 0.78
level 2 : stenosis = 0.71
level 3 : stenosis = 0.65

$$\hat{\mu}_1 = \overline{x}_{1.} = 11.209$$
$$\hat{\mu}_2 = \overline{x}_{2.} = 15.086$$
$$\hat{\mu}_3 = \overline{x}_{3.} = 17.330$$

11.1.2 Partitioning the Total Sum of Squares

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{..}\right)^2$$

$$= \sum_{i=1}^{k} n_i (\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i.}\right)^2$$

$$= SSTr + SSE$$

- P-value considerations

  The plausibility of the null hypothesis that the factor level means are all equal depends upon the **relative size** of the sum of squares for treatments, SSTr, to the sum of squares for error, SSE.

Example 62 : Collapse of Blocked Arteries

$\overline{x}_{1.} = 11.209, \quad \overline{x}_{2.} = 15.086, \quad \overline{x}_{3.} = 17.330$

$\overline{x}_{..} = 14.509$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 = 7710.39$$

SST$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - n_T\overline{x}_{..}^2 = 342.5$

SSTr$= \sum_{i}^{k} n_i\overline{x}_{i.}^2 - n_T\overline{x}_{..}^2 = 204.0$

$SSE = SST - SSTr = 138.5$

11.1.3 The Analysis of Variance Table

- Mean square error (MSE)

$$MSE = \frac{SSE}{n_T - k}$$

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n_T - k}$$

$$E(MSE) = \sigma^2$$

- Mean squares for treatments (MSTr)

$$MSTr = \frac{SSTr}{k-1}$$

- If the factor level means μ_i are all equal (H_0), then $E(MSTr) = \sigma^2$ and $\frac{SSTr}{\sigma^2} \sim \chi^2_{k-1}$
- Under H_0,

$$F = \frac{MSTr}{MSE} \sim F_{k-1, n_T - k}$$

Analysis of variance table for one factor layout

Source	d.f.	Sum of squares	Mean squares	F−statistic	P−value
Treatments	$\boldsymbol{k-1}$	SSTr	$\boldsymbol{MSTr = \frac{SSTr}{k-1}}$	$\boldsymbol{F = \frac{MSTr}{MSE}}$	$\boldsymbol{P(F_{k-1,n_T-k} \geq obs(F))}$
Error	$\boldsymbol{n_T-k}$	SSE	$\boldsymbol{MSE = \frac{SSE}{n_T-k}}$		
total	$\boldsymbol{n_T-1}$	SST			

Example 62 : Collapse of Blocked Arteries

MSTr=102.0 with df $k-1=3-1=2.$

MSE=4.33 with df $n_T-k=35-3=32$

Obs(F)=$\frac{102.0}{4.33}=23.6.$

P-value=$P(F\geq 23.6)\approx 0$ where $F\sim F_{k-1,n_T-k}.$

Consequently, the null hypothesis that the average flowrate at collapse is the same for all three amounts of stenosis is rejected at the sig. level 0.01.

Python codes for ANOVA with ‘airquality’ data

```
import pandas as pd
from statsmodels.stats.anova import anova_lm
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt
aq = pd.read_csv('data/airquality.csv', index_col=0)
print(aq)
model = ols('Wind ~ C(Month)', aq).fit()
print(anova_lm(model))
```

Python output

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41.0	190.0	7.4	67	5	1
2	36.0	118.0	8.0	72	5	2
3	12.0	149.0	12.6	74	5	3
4	18.0	313.0	11.5	62	5	4
5	NaN	NaN	14.3	56	5	5
..	...	...	...	...	...	...
149	30.0	193.0	6.9	70	9	26
150	NaN	145.0	13.2	77	9	27
151	14.0	191.0	14.3	75	9	28
152	18.0	131.0	8.0	76	9	29
153	20.0	223.0	11.5	68	9	30

[153 rows x 6 columns]

	df	sum_sq	mean_sq	F	PR(>F)
C(Month)	4.0	164.270802	41.067701	3.529048	0.00879
Residual	148.0	1722.283054	11.637048	NaN	NaN

- Decision:

H_0 is rejected if the sig. level α is larger than

11.1.4 Pairwise Comparisons of the Factor Level Means (T-Method: Tukey's Multiple Comparisons Procedure)

- When the null hypothesis is rejected, the experimenter can follow up the analysis with pairwise comparisons of the factor level means to discover which ones have been shown to be different and by how much.

- With k factor levels there are k(k-1)/2 pairwise differences.

- A set of $1-\alpha$ confidence level simultaneous confidence intervals for the differences, $\mu_{i_1}-\mu_{i_2}$, are

$$\left(\overline{x}_{i_1\cdot}-\overline{x}_{i_2\cdot}-\hat{\sigma}\frac{q_{\alpha,k,v}}{\sqrt{2}}\sqrt{\frac{1}{n_{i_1}}+\frac{1}{n_{i_2}}},\overline{x}_{i_1\cdot}-\overline{x}_{i_2\cdot}\right.$$

- These confidence intervals are similar to the t-intervals
  - Difference : $q_{\alpha,k,v}/\sqrt{2}$ is used instead of $t_{\alpha/2,v}$.
  - T-intervals have an individual confidence level whereas this set of simultaneous confidence intervals has an **overall** confidence level
  - All of the k(k-1)/2 confidence intervals contain their respective parameter value $\mu_{i_1} - \mu_{i_2}$
  - $q_{\alpha,k,v}/\sqrt{2}$ is larger than $t_{\alpha/2,v}$.
- If the confidence interval for the difference $\mu_{i_1} - \mu_{i_2}$ contains zero, then there is no evidence that the means at factor levels i_1 and i_2 are different.

Example 62 : Collapse of Blocked Arteries

- $\hat{\sigma} = \sqrt{4.33} = 2.080$
- With 32 degrees of freedom for error, the critical point is $q_{0.05,3,32} = 3.48$
- The overall confidence level is 1-0.05=0.95
- The 95% simultaneous confidence interval (SCI) for $\mu_1 - \mu_2$:

$$\left(11.209 - 15.086 - 2.080 \times \frac{3.48}{\sqrt{2}} \sqrt{\frac{1}{11} + \frac{1}{14}} \, , 11.209 - 15.086 \right.$$

- None of these three confidence intervals contains zero, and so the experiment has established that each of the three stenosis levels results in a different average flow rate at collapse.

Example 64. Comparison of Wear of Carpet Fiber Blends(p.513-4)

FIGURE 11.18

Pairwise confidence intervals for the carpet fiber blends experiment

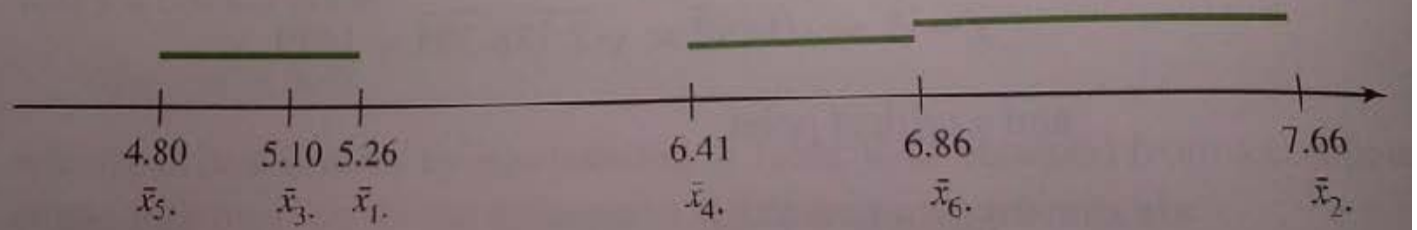
The critical point is $q_{0.05,6,84} = 4.12$ and $s = \hat{\sigma} = \sqrt{MSE} = 1.004$ so

$$\mu_i - \mu_j \in \bar{x}_{i\cdot} - \bar{x}_{j\cdot} \pm \frac{1.004 \times 4.12}{\sqrt{2}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$\mu_1 - \mu_2$ (−3.44, −1.37)				
$\mu_1 - \mu_3$ (−0.94, 1.25) **contains 0**	$\mu_2 - \mu_3$ (1.47, 3.66)			
$\mu_1 - \mu_4$ (−2.19, −0.12)	$\mu_2 - \mu_4$ (0.22, 2.28)	$\mu_3 - \mu_4$ (−2.41, −0.22)		
$\mu_1 - \mu_5$ (−0.61, 1.53) **contains 0**	$\mu_2 - \mu_5$ (1.79, 3.93)	$\mu_3 - \mu_5$ (−0.83, 1.43) **contains 0**	$\mu_4 - \mu_5$ (0.54, 2.68)	
$\mu_1 - \mu_6$ (−2.66, −0.55)	$\mu_2 - \mu_6$ (−0.25, 1.85) **contains 0**	$\mu_3 - \mu_6$ (−2.87, −0.65)	$\mu_4 - \mu_6$ (−1.50, 0.60) **contains 0**	$\mu_5 - \mu_6$ (−3.15, −0.97)

FIGURE 11.19

Schematic presentation of the results of the carpet fiber blends experiment



Python codes for multiple comparison

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
    # `tukeyhsd' stands for 'Tukey's Honestly Significant Difference'.
import matplotlib.pyplot as plt
comp = pairwise_tukeyhsd(aq['Wind'], aq['Month'], alpha=0.05)
print(comp)  =➔ next sheet
aq.boxplot('Wind', by='Month', grid=False, figsize=(12,6))
plt.show()     =➔ 2 sheets after
```

Python output: Multiple comparison of means

Multiple Comparison of Means - Tukey HSD, FWER=0.05

==========

group1 group2 meandiff p-adj lower upper reject

---

5 6 -1.3559 0.5264 -3.7688 1.0569 False

5 7 -2.6806 0.0197 -5.0736 -0.2876 True

5 8 -2.829 0.0117 -5.222 -0.436 True

5 9 -1.4426 0.4685 -3.8554 0.9703 False

6 7 -1.3247 0.5465 -3.7376 1.0881 False

6 8 -1.4731 0.4472 -3.886 0.9397 False

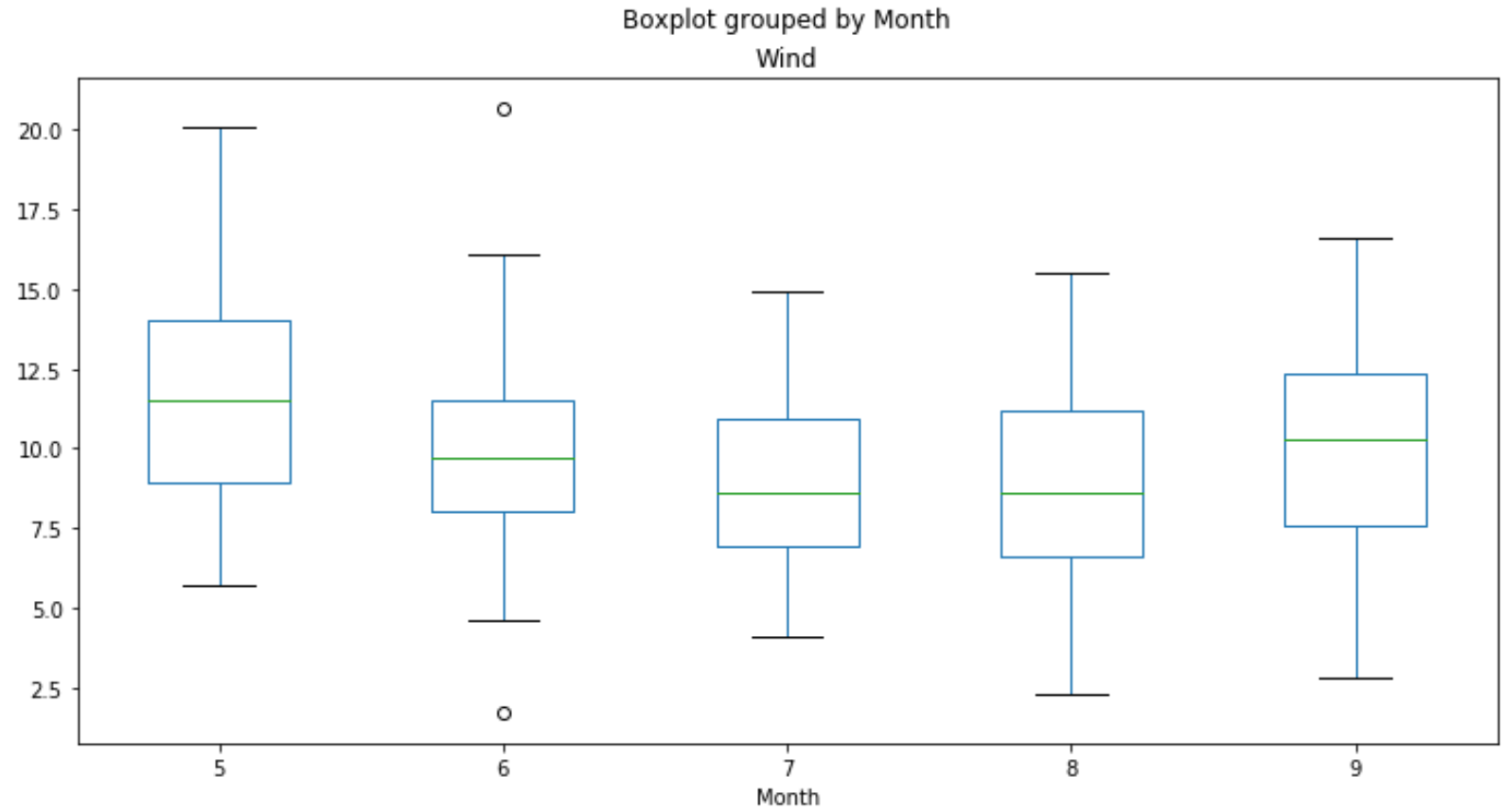
6 9 -0.0867 0.9 -2.5192 2.3459 False

7 8 -0.1484 0.9 -2.5414 2.2446 False

7 9 1.2381 0.6024 -1.1748 3.6509 False

8 9 1.3865 0.5067 -1.0264 3.7993 False

Python output: Box plots of wind speeds across the 5 months



11.1.5 Sample Size Determination

- The sensitivity afforded by a one factor analysis of variance depends upon the k sample sizes
- The power of the test of the null hypothesis that the factor level means are all equal increase as the sample sizes increase.
- An increase in the sample size results in a decrease in the lengths of the pairwise confidence intervals.
- If the sample sizes n_i are unequal,

$$L = \sqrt{2}\,\hat{\sigma} q_{\alpha,k,v} \sqrt{\frac{1}{n_{i_1}} + \frac{1}{n_{i_2}}}$$

- The critical point $q_{\alpha,k,v}$ gets larger as the number of factor levels increases.

11.1.6 Model Assumptions

Modeling assumption of the analysis of variance

- Observations are distributed independently with normal distribution that has a common variance
- The ANOVA is fairly robust to the distribution of data, so that it provides fairly accurate results as long as the distribution is not very far from a normal distribution.
- The equality of the variances across the k factor levels can be judged from a comparison of the sample variances or from a visual comparison of the lengths of boxplots of the observations at each factor level.

Summary of Section 11.1

11.1 One Factor Analysis of Variance

SST=SSTr+SSE

Distribution of MSTr/MSE

Simultaneous confidence intervals