Single-Factor Studies

- **ANOVA Model:**

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

where:

1. Y_{ij} is the value of the response variable in the jth trial for the ith treatment.

2. μ_i is the population mean (parameter) for the ith treatment.

3. ϵ_{ij} are independent $N(0, \sigma^2)$.

4. $i = 1, \ldots, r$; $j = 1, \ldots, n_i$.

- **Null and Alternative Hypotheses.**

H_0 : *The factor has NO effect.* vs. H_1 : *The factor has an effect.*

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ vs. H_1 : Not all are equal

ANOVA Table

Source	DF	Sum of Squares	Mean Square	F
Treatments (Between)	$r-1$	$SSTR = \sum_{i=1}^{r} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$MSTR = \dfrac{SSTR}{r-1}$	$F_{\text{obs}} = \dfrac{MSTR}{MSE}$
Error (Within)	$n_T - r$	$SSE = \sum_{i=1}^{r} (n_i - 1)s_i^2$	$MSE = \dfrac{SSE}{n_T - r}$	
Total	$n_T - 1$	$SSTO = \sum_{i=1}^{r}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$		

where,

1. $\bar{Y}_{i\cdot}$ is the average of the measurements from the ith treatment.

2. $\bar{Y}_{\cdot\cdot}$ is the average of all measurements.

3. s_i^2 is the sample variance of the ith treatment.

Note:

1. $SSTO = SSTR + SSE$.

2. Under H_0, $F_{\text{obs}} \sim f_{(r-1, n-r)}$.

3. $E(MSTR) = \sigma^2 + \frac{\sum n_i(\mu_i - \mu_\cdot)^2}{r-1}$.

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

- **Assumptions.**

1. The samples are independent SRS from each population.

2. The populations are assumed to be normal.

3. The standard deviations are equal. [Rule of thumb: *The largest standard deviation should not be more than twice the smallest standard deviation.*]

- **Idea.** ANOVA is based on separating the total variation observed in the data into two parts: variation *among the group means* and variation *within groups*. If the variation among groups is large relative to the variation within groups, we have evidence against the null hypothesis.

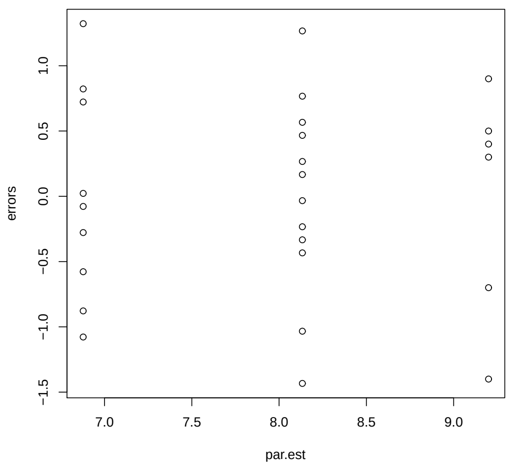
- **Example.**

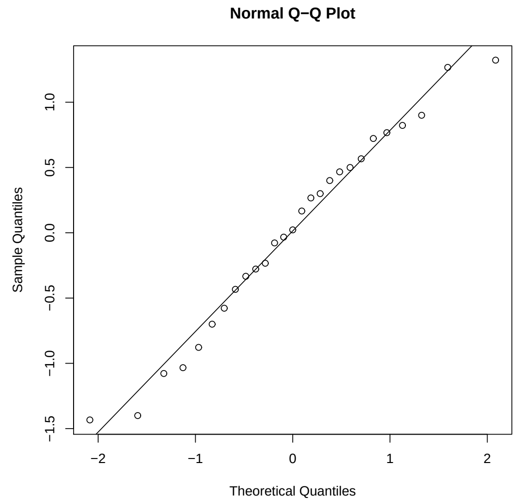
Productivity Improvement. An economist compiled data on productivity improvements last year for a sample of firms producing electronic computing equipment. The firms were classified according to the level of their average expenditures for research and development in the past three years (low, moderate, high). The results of the study is given in the table below (productivity improvement is measured on a scale from 0 to 100). Assuming that ANOVA model is appropriate, test the hypothesis that the level of expenditures for research and development has no effect on the productivity improvements. Use $\alpha = 0.05$.

	1	2	3	4	5	6	7	8	9	10	11	12
Low	7.6	8.2	6.8	5.8	6.9	6.6	6.3	7.7	6.0			
Moderate	6.7	8.1	9.4	8.6	7.8	7.7	8.9	7.9	8.3	8.7	7.1	8.4
High	8.5	9.7	10.1	7.8	9.6	9.5						

Sample R commands:

```
data=read.csv("ProductivityImprovement.csv", header=T)
head(data)
attach(data)        # This will attach the values to the column names.
results=aov(Improve ~ Budget)      # aov is the built-in ANOVA function in R.
summary(data)       # Note that you get the wrong d.f.

            Df Sum Sq Mean Sq F value    Pr(>F)
Budget       1 20.067 20.0668  32.532 6.111e-06 ***
Residuals   25 15.421  0.6168
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

attributes(Budget)
Budget=factor(Budget)
results=aov(Improve ~ Budget)
summary(data)
            Df Sum Sq Mean Sq F value    Pr(>F)
Budget       2 20.125 10.0626  15.720 4.331e-05 ***
Residuals   24 15.362  0.6401
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

plot(Budget,Improve)      # Creates boxplots for each factor level
plot(as.numeric(Budget),Improve)  # Creates a scatterplot
plot(as.numeric(Budget),Improve,xlab="Budget")

attributes(results)
par.est=results$fitted
points(as.numeric(Budget),par.est,pch=19,col="darkred")
cbind(Budget,Improve,par.est)
mean(Improve[Budget==1])      # Computes the mean of 'Improve' when 'Budget==1'
tapply(Improve,Budget,mean)   # Obtains the means of 'Improve' by 'Budget' level

errors=results$residuals
cbind(Budget,Improve,par.est,errors)

plot(par.est,errors)    # This scatterplot is useful for checking the
                        # constant variance assumption.
```






```
qqnorm(errors)      # This plot is useful for assessing the normality of the error terms.
qqline(errors)

shapiro.test(errors)

        Shapiro-Wilk normality test

data:  results$residuals
W = 0.9738, p-value = 0.7033
```

- **Homework.**

1. Rehabilitation Therapy. A rehabilitation center researcher was interested in examining the relationship between physical fitness prior to surgery of persons undergoing corrective knee surgery and time required in physical therapy until successful rehabilitation. Patient records in the rehabilitation center were examined, and 24 male subjects ranging in age from 18 to 30 years who had undergone similar corrective knee surgery during the past year were selected for the study. The number of days required for successful completion of physical therapy and the prior physical fitness status (below average, average, above average) for each patient follow.

	1	2	3	4	5	6	7	8	9	10
Below Average	29	42	38	40	43	40	30	42		
Average	30	35	39	28	31	31	29	35	29	33
Above Average	26	32	21	20	23	22				

a. Prepare aligned dot plots of the data.

b. Obtain the ANOVA table using the aov function.

c. Obtain the estimates for all the parameters μ_i (or the fitted values).

d. Obtain the residuals and verify that their sum is zero.

e. Create a scatter plot between the fitted values and the residuals. What can you say about the equal variance assumption?

f. Construct the qqplot for the residuals. What can you say about the normality assumption?

2. Do 16.5 on page 723.

3. Do 16.8 on page 723.