Chapter 8

Single–Factor Studies: One–Way ANOVA

8.1 Introduction

We look at testing whether or not the means from two or more populations are the same or different using the Analysis of Variance (ANOVA) method. From a "big picture" point of view, we continue to look at the statistical inference of (large n) multiple–sample mean problems.

		mean μ	variance σ^2	proportion π
	one	large n, 3.7, 3.8, 3.9, 3.10, 4.6 small n, 4.3, 4.6	4.4	6.2
sample	two	large n, 3.11 small n, 4.3	4.4	6.3
	multiple	**chapters** 7, **8**, 9	not done	6.2, 6.3

8.2 Completely Randomized Designs

We look at notation related to Completely Randomized Designed (CRD) experiments.

Exercise 8.1 (Notation For Completely Randomized Design (CRD))

1. **Drug Responses.**

drug 1	5.90	5.92	5.91	5.89	5.88	$\bar{y}_{1+} \approx 5.90$
drug 2	5.50	5.50				$\bar{y}_{2+} = 5.50$
drug 3	5.01	5.00	4.99	4.98	5.02	$\bar{y}_{3+} \approx 5.00$

(a) t = (circle one) **1** / **2** / **3** treatments

(b) N = (circle one) **2** / **5** / **12** responses

(c) n_1 = (circle one) **2** / **5** / **12** drug 1 responses

(d) n_2 = (circle one) **2** / **5** / **12** drug 2 responses

(e) drug 1, replication 1 response, y_{11} = (circle one) **5.50** / **5.90** / **5.92**

(f) drug 3, replication 4 response, y_{34} = (circle one) **4.98** / **5.90** / **5.92**

(g) sum of drug 1 responses, y_{1+} = (circle one) **5.50** / **5.90** / **29.5**

(h) sum of drug 2 responses, y_{2+} = (circle one) **4.98** / **5.90** / **11.00**

(i) average of drug 1 responses, $\bar{y}_{1+}$ = (circle one) **5.50** / **5.90** / **29.5**

(j) average of drug 3 responses, $\bar{y}_{3+}$ = (circle one) **4.98** / **5.00** / **11.00**

(k) variance of drug 1 responses, s_1^2 = (circle one) **0.00025** / **0.0158** / **29.5** (to get s_1^2, use calculator: type responses into L_1, then STAT CALC ENTER, then square S_x)

(l) variance of drug 3 responses, s_3^2 = (circle one) **0.00025** / **0.436** / **11.00**

(m) To say the average responses are *not* the same (circle one) **is** / **is not** the same thing as saying the average responses are *all* different from one another.

(n) *If* all the average responses to the three drugs are the same, then they (circle one) **are** / **are not** equal to the grand average.

(o) The *grand* average, $\bar{y}_{++} \approx 5.46$, is given by (circle none, one or more)

i. adding all twelve responses and dividing by 12.

ii. $\frac{\bar{y}_{1+}+\bar{y}_{2+}+\bar{y}_3}{12}$.

iii. $\frac{5\bar{y}_{1+}+2\bar{y}_{2+}+5\bar{y}_{3+}}{12}$.

(p) The *grand* variance, $s^2 \approx 0.0857$, is given by (circle none, one or more)

i. determining the variance of the twelve responses.

ii. $\frac{\bar{y}_{1+}+\bar{y}_{2+}+\bar{y}_3}{12}$.

iii. $\frac{(5-1)s_1^2+(2-1)s_2^2+(5-1)s_3^2}{12-3}$.

2. *Rats in New York.* Where are the rats in New York city? The rat count per square meter is recorded and shown in the following table.

sewers	parks	city hall
3	1	5
5	3	8
7	2	9
4		8
		10

(a) $t =$ (circle one) **1** / **2** / **3** treatments

(b) $N =$ (circle one) **2** / **4** / **12** responses

(c) $n_1 =$ (circle one) **3** / **4** / **12** responses

(d) $n_2 =$ (circle one) **2** / **3** / **12** responses

(e) sewer, replication 1 response, $y_{11} =$ (circle one) **3** / **5.90** / **5.92**

(f) city hall, replication 4 response, $y_{34} =$ (circle one) **8** / **5.90** / **5.92**

(g) sum of sewer responses, $y_{1+} =$ (circle one) **5.50** / **19** / **29.5**

(h) sum of parks responses, $y_{2+} =$ (circle one) **4.98** / **6** / **11.00**

(i) average of sewer responses, $\bar{y}_{1+} =$ (circle one) **4.75** / **5.90** / **29.5**

(j) average of city hall responses, $\bar{y}_{3+} =$ (circle one) **4.98** / **5.00** / **8**

(k) variance of sewer responses, $s_1^2 =$ (circle one) **0.00025** / **0.0158** / **2.92**

(l) variance of city hall responses, $s_3^2 =$ (circle one) **0.19025** / **0.436** / **3.5**

8.3 Analysis of Variance (ANOVA)

Exercise 8.2 (Analysis of Variance, Completely Randomized Design (CRD))

See Lab 10: One Way Analysis of Variance.

1. *Comparing Three Drugs.* Fifteen different patients are subjected to three drugs. A higher number indicates the patient is better off than a patient with a lower number.

drug 1	5.90	5.92	5.91	5.89	5.88	$\bar{y}_{1+} \approx 5.90$
drug 2	5.51	5.50	5.50	5.49	5.50	$\bar{y}_{2+} \approx 5.50$
drug 3	5.01	5.00	4.99	4.98	5.02	$\bar{y}_{3+} \approx 5.00$

Test if at least two of the three average patient responses to the drug are different at $\alpha = 0.05$.

(a) *Test Statistic Versus Critical Value.*

i. The statement of the test is (check none, one or more):

A. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 :$ $\mu_1 \neq \mu_2, \mu_1 = \mu_3$.

B. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 :$ $\mu_1 \neq \mu_3, \mu_1 \neq \mu_2$.

C. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus
$H_1 :$ at least one $\mu_i \neq \mu_j, i \neq j;$ $i, j = 1, 2, 3$.

D. H_0 : means the same versus H_1 : means different

ii. *Test.* After some effort, the ANOVA table[1] is given by,

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment (Drugs)	2	2.033	1.0165
Error	12	0.0022	0.00018
Total	14	2.035	

and so the test statistic[2] is

$$F \text{ test statistic } = \frac{1.0165}{0.00018} =$$

(circle one) **1.02** / **123** / **5647.2**.
The upper critical value at $\alpha = 0.05$, with $t - 1 = 3 - 1 = 2$ and $N - t = 15 - 3 = 12$ degrees of freedom, is
(circle one) **3.22** / **3.89** / **4.82**
(Use PRGM INVF ENTER 2 ENTER 12 ENTER 0.95 ENTER)

iii. *Conclusion.* Since the test statistic, 5647.2, is larger than the critical value, 3.89, we (circle one) **accept** / **reject** the null hypothesis that the average patient responses to the three drugs are the same.

(b) *P–Value Versus Level of Significance.*

i. The statement of the test is (check none, one or more):

A. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 : \ \mu_1 \neq \mu_2, \mu_1 = \mu_3$.

B. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 : \ \mu_1 \neq \mu_3, \mu_1 \neq \mu_2$.

C. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus
$H_1 :$ at least one $\mu_i \neq \mu_j, i \neq j; \ i, j = 1, 2, 3$.

D. H_0 : means the same versus H_1 : means different

ii. *Test.* Since the test statistic is $F = 5647.2$, the p–value, with $t - 1 = 3 - 1 = 2$ and $N - t = 15 - 3 = 12$ degrees of freedom, is given by

$$\text{p–value} = P(F \geq 5647.2)$$

which equals (circle one) **0.00** / **0.35** / **0.43**.
(Use 2nd DISTR 9:Fcdf(5647.2,E99,2,12).)
The level of significance is 0.05.

iii. *Conclusion.* Since the p–value, 0.00, is smaller than the level of significance, 0.05, we (circle one) **accept** / **reject** the null hypothesis that the average patient responses to the three drugs are the same.

[1]Type data into L_1, L_2 and L_3, then STAT TESTS ANOVA(L_1,L_2,L_3).

[2]Due to round–off error, the F calculated here is not the same as the one calculated by the calculator.

2. *Comparing Three Drugs Again.* Twelve different patients are subjected to three drugs.

drug 1	5.90	5.92	5.91	5.89	5.88	$\bar{y}_{1+} \approx 5.90$
drug 2	5.50	5.50				$\bar{y}_{2+} = 5.50$
drug 3	5.01	5.00	4.99	4.98	5.02	$\bar{y}_{3+} \approx 5.00$

Test if at least two of the three average patient responses to the drug are different at $\alpha = 0.05$.

(a) *Test Statistic Versus Critical Value.*

i. The statement of the test is (check none, one or more):

A. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 :$ $\mu_1 \neq \mu_2, \mu_1 = \mu_3$.

B. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 :$ $\mu_1 \neq \mu_3, \mu_1 \neq \mu_2$.

C. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus
$H_1 :$ at least one $\mu_i \neq \mu_j, i \neq j;$ $i, j = 1, 2, 3$.

D. H_0 : means the same versus H_1 : means different

ii. *Test.* After some effort, the ANOVA table is given by,

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment (Drugs)	2	2.029	1.0147
Error	9	0.002	0.00022
Total	11	2.031	

and so the test statistic is

$$F \text{ test statistic } = \frac{1.0147}{0.00022} =$$

(circle one) **1.02** / **123** / **4612.3**.
The upper critical value at $\alpha = 0.05$, with $t - 1 = 3 - 1 = 2$ and $N - t = 12 - 3 = 9$ degrees of freedom, is
(circle one) **3.22** / **3.89** / **4.26**
(Use PRGM INVF ENTER 2 ENTER 9 ENTER 0.95 ENTER)

iii. *Conclusion.* Since the test statistic, 4612.3, is larger than the critical value, 4.26, we (circle one) **accept** / **reject** the null hypothesis that the average patient responses to the three drugs are the same.

(b) *P–Value Versus Level of Significance.*

i. The statement of the test is (check none, one or more):

A. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 :$ $\mu_1 \neq \mu_2, \mu_1 = \mu_3$.

B. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 :$ $\mu_1 \neq \mu_3, \mu_1 \neq \mu_2$.

C. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus
H_1 : at least one $\mu_i \neq \mu_j, i \neq j;\ i, j = 1, 2, 3.$

D. H_0 : means the same versus H_1 : means different

ii. *Test.* Since the test statistic is $F = 4612.3$, the p–value, with $t - 1 = 3 - 1 = 2$ and $N - t = 12 - 3 = 9$ degrees of freedom, is given by

$$\text{p–value} = P(F \geq 4612.3)$$

which equals (circle one) **0.00** / **0.35** / **0.43**.
(Use 2nd DISTR 9:Fcdf(4612.3,E99,2,9).)
The level of significance is 0.05.

iii. *Conclusion.* Since the p–value, 0.00, is smaller than the level of significance, 0.05, we (circle one) **accept** / **reject** the null hypothesis that the average patient responses to the three drugs are the same.

3. *Rats in New York.* Where are the rats in New York city? The rat count per square meter is recorded and shown in the following table.

sewers	parks	city hall
3	1	5
5	3	8
7	2	9
4		8
		10

Test the claim that the average rat count per square meter in at least two of the three city areas is different at 5%.

(a) *Test Statistic Versus Critical Value.*

i. The statement of the test is (check none, one or more):

A. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 :\ \mu_1 \neq \mu_2, \mu_1 = \mu_3.$

B. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 :\ \mu_1 \neq \mu_3, \mu_1 \neq \mu_2.$

C. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus
$H_1 :$ at least one $\mu_i \neq \mu_j, i \neq j;\ i, j = 1, 2, 3.$

D. H_0 : means the same versus H_1 : means different

ii. *Test.* After some effort, the ANOVA table is given by,

Source	Sum Of Squares	Degrees of Freedom	Mean Squares
Between	70.2	2	35.1
Within	24.8	9	2.75
Total	95.0	11	

and so the test statistic is

$$F \text{ test statistic } = \frac{35.1}{2.75} =$$

(circle one) **12.76** / **123** / **4612.3**.
The upper critical value at $\alpha = 0.05$, with $k - 1 = 3 - 1 = 2$ and $N - k = 12 - 3 = 9$ degrees of freedom, is
(circle one) **3.22** / **3.89** / **4.26**
(Use PRGM INVF ENTER 2 ENTER 9 ENTER 0.95 ENTER)

iii. *Conclusion.* Since the test statistic, 12.76, is larger than the critical value, 4.26, we (circle one) **accept** / **reject** the null hypothesis that the average number of rats in the three locations are the same.

(b) *P–Value Versus Level of Significance.*

i. The statement of the test is (check none, one or more):

A. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 : \mu_1 \neq \mu_2, \mu_1 = \mu_3$.

B. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 : \mu_1 \neq \mu_3, \mu_1 \neq \mu_2$.

C. $H_0 : \mu_1 = \mu_2 = \mu_3$ versus
$H_1 :$ at least one $\mu_i \neq \mu_j, i \neq j;\ i, j = 1, 2, 3$.

D. H_0 : means the same versus H_1 : means different

ii. *Test.* Since the test statistic is $F = 12.76$, the p–value, with $k - 1 = 3 - 1 = 2$ and $N - k = 12 - 3 = 9$ degrees of freedom, is given by

$$\text{p–value} = P(F \geq 12.76)$$

which equals (circle one) **0.002** / **0.35** / **0.43**.
(Use 2nd DISTR 9:Fcdf(12.76,E99,2,9).)
The level of significance is 0.05.

iii. *Conclusion.* Since the p–value, 0.002, is smaller than the level of significance, 0.05, we (circle one) **accept** / **reject** the null hypothesis that the average number of rats in the three locations are the same.

8.4 Analysis of a CRD: Computational Details

The observed F is

$$F = \frac{MS[T]}{MS[E]}$$

where

$$MS[T] = \frac{SS[T]}{t - 1}, \quad MS[E] = \frac{SS[E]}{N - t}$$

where t is the number of treatments and N is the total number of replications, and where

$$\begin{aligned} SS[T] &= \sum_{i=1}^{t}[n_i(\overline{Y}_{i+} - \overline{Y}_{++})^2] \\ &= \sum_{i=1}^{t}\left(\frac{Y_{i+}^2}{n_i}\right) - \frac{Y_{++}^2}{N} \end{aligned}$$

where $CM = \frac{Y_{++}^2}{N}$ is called the correction for the mean and n_i is the number of replications in each treatment and

$$\begin{aligned} SS[E] &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 \\ &= \sum_{i=1}^{t}\left(\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{i+}^2}{n_i}\right) \\ &= \sum_{i=1}^{t}\left((n_i - 1)S_i^2\right) \\ &= SS[TOT] - SS[T] \end{aligned}$$

where

$$SS[TOT] = \sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij}^2 - CM$$

A summary of this information[3] can be displayed in the following ANOVA table.

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment	$t-1$	SS[T]	MS[T]
Error	$N-t$	SS[E]	MS[E]
Total	$N-1$	SS[TOT]	

Exercise 8.3 (Notation, the ANOVA table and F)

1. *Calculating ANOVA and F Using The Formulas.*
   If $y_{1+} = \sum_{j=1}^{4} y_{1j} = 23$, $\sum_{j=1}^{4} y_{1j}^2 = 2300$, $n_1 = 4$,
   $y_{2+} = \sum_{j=1}^{4} y_{2j} = 122$, $\sum_{j=1}^{4} y_{2j}^2 = 23040$, $n_2 = 6$,
   $y_{3+} = \sum_{j=1}^{4} y_{3j} = 32$, $\sum_{j=1}^{4} y_{3j}^2 = 4300$, and $n_3 = 6$, then
   calculate F.

   (a) $y_{++} = 23 + 122 + 32 =$ (circle one) **145** / **177** / **256**

[3]The TI–83 calculator is able to determine the ANOVA table from "raw" data, but is not able to determine the ANOVA table from "summarized" data, as given in this section.

(b) $SS[T] = \sum_{i=1}^{3} \left(\frac{y_{i+}^2}{n_i}\right) - \frac{y_{++}^2}{N} = \frac{(23)^2}{4} + \frac{(122)^2}{6} + \frac{(32)^2}{6} - \frac{(177)^2}{16} =$
(circle one) **145.7** / **825.5** / **1034.5**

(c) $SS[E] = \sum_{i=1}^{3} \left(\sum_{j=1}^{i} y_{ij}^2 - \frac{y_{i+}^2}{n_i}\right) =$
$\left(2300 - \frac{(23)^2}{4}\right) + \left(23040 - \frac{(122)^2}{6}\right) + \left(4300 - \frac{(32)^2}{6}\right) =$
(circle one) **145.7** / **26, 856.42** / **111, 034.5**

(d) $MS[T] = \frac{SS[T]}{t-1} = \frac{825.5}{3-1} =$ (circle one) **145.7** / **412.76** / **1034.5**

(e) $MS[E] = \frac{SS[E]}{N-t} = \frac{26{,}856.42}{16-3} =$ (circle one) **145.7** / **2, 065.87** / **11, 034.5**

(f) $F = \frac{MS[T]}{MS[E]} = \frac{412.76}{2{,}065.87} =$ (circle one) **0.13** / **0.20** / **1.34**
and the ANOVA table is

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment	2	825.52	412.76
Error	13	26,856.42	2,065.87
Total	15	27,681	

2. *Calculating F From a Data Set.* Calculate F from the following data set,

drug 1	5.90	5.92	5.91	5.89	5.88	$\bar{y}_{1+} \approx 5.90$
drug 2	5.51	5.50	5.50	5.49	5.50	$\bar{y}_{2+} \approx 5.50$
drug 3	5.01	5.00	4.99	4.98	5.02	$\bar{y}_{3+} \approx 5.00$

where $y_{1+} = \sum_{j=1}^{5} y_{1j} = 29.5$, $\sum_{j=1}^{5} y_{1j}^2 = 174.051$, $n_1 = 5$,
$y_{2+} = \sum_{j=1}^{5} y_{2j} = 27.5$, $\sum_{j=1}^{5} y_{2j}^2 = 151.2502$, $n_2 = 5$,
$y_{3+} = \sum_{j=1}^{5} y_{3j} = 25$, $\sum_{j=1}^{5} y_{3j}^2 = 125.001$, and $n_3 = 5$
(Type data into L_1, L_2, L_3; then use STAT CALC 1:1–Var Stats.)

(a) $y_{++} = 29.5 + 27.5 + 25 =$
(circle one) **82** / **176** / **256**

(b) $SS[T] = \sum_{i=1}^{3} \left(\frac{y_{i+}^2}{n_i}\right) - \frac{y_{++}^2}{N} = \frac{(29.5)^2}{5} + \frac{(27.5)^2}{5} + \frac{(25)^2}{5} - \frac{(82)^2}{15} =$
(circle one) **2.033** / **847.6** / **1034.5**

(c) $SS[E] = \sum_{i=1}^{3} \left(\sum_{j=1}^{5} y_{ij}^2 - \frac{y_{i+}^2}{n_i}\right) =$
$\left(174.051 - \frac{(29.5)^2}{5}\right) + \left(151.2502 - \frac{(27.5)^2}{5}\right) + \left(125.001 - \frac{(25)^2}{5}\right) =$
(circle one) **0.0022** / **23267.7** / **111034.5**

(d) $MS[T] = \frac{SS[T]}{t-1} = \frac{2.033}{3-1} =$ (circle one) **1.0165** / **423.8** / **1034.5**

(e) $MS[E] = \frac{SS[E]}{N-t} = \frac{0.0022}{15-3} =$ (circle one) **0.00018** / **1789.8** / **11034.5**

(f) $F = \frac{MS[T]}{MS[E]} = \frac{1.0165}{0.00018} =$ (circle one) **5647.2** / **1024.45** / **1346.73**
and the ANOVA table is

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment (Drugs)	2	2.033	1.0165
Error	12	0.0022	0.00018
Total	14	2.035	

(Confirm this by typing the three drug response data into L_1, L_2, and L_3 respectively then type STAT TESTS F:ANOVA(L_1, L_2, L_3) ENTER.)

3. *Chlorophyll In Leaves.* Calculate F using the following results obtained from a study measuring the amount of chlorophyll in three different types of leaves.

Treatment	Oak	Birch	Maple
replications, n_i	5	3	7
mean, $\bar{y}_i$	4.67	3.53	1.40
standard deviation, s_i	1.06	0.27	0.27

(a) Since $\bar{y}_{++} = \frac{n_1\bar{y}_{1+}+n_2\bar{y}_{2+}+n_3\bar{y}_{3+}}{N} = \frac{5(4.67)+3(3.53)+7(1.40)}{15} =$
(circle one) **1.916** / **2.916** / **3.916**.

(b) $SS[T] = n_1(\bar{y}_{1+} - \bar{y}_{++})^2 + n_2(\bar{y}_{2+} - \bar{y}_{++})^2 + n_3(\bar{y}_{3+} - \bar{y}_{++})^2 =$
$5(4.67 - 2.916)^2 + 3(3.53 - 2.916)^2 + 7(1.40 - 2.916)^2 =$
(circle one) **31.60** / **32.60** / **33.60**.

(c) $SS[E] = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 = (5-1)1.06^2 + (3-1)0.27^2 +$
$(7-1)1.4^2 =$
(circle one) **4.17** / **6.21** / **16.40**.

(d) $MS[T] = \frac{SS[T]}{t-1} = \frac{32.60}{3-1} =$ (circle one) **15.3** / **16.3** / **17.3**

(e) $MS[E] = \frac{SS[E]}{N-t} = \frac{16.40}{15-3} =$ (circle one) **0.417** / **0.517** / **6.21**

(f) $F = \frac{MS[T]}{MS[E]} = \frac{16.3}{6.21} =$ (circle one) **2.62** / **32.5** / **33.5**
and the ANOVA table is

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment (Leaves)	2	32.6	16.3
Error	12	16.40	6.21
Total	14	49.0	

8.5 The One–Way Classification Model

Not covered.

8.6 Checking For Violations of Assumptions

The ANOVA assumptions for a completely randomized design (CRD) are:

- populations have a normal distribution
- population variances are equal (are all equal to a constant)
- observed responses are independent random samples from t populations

If these conditions are satisfied, an ANOVA procedure will give correct results. We will find out how to determine if the data satisfies the first two conditions using *q–q plots* and $e \vee p$ plots, respectively[4]. More than just telling us that the data does not satisfy the ANOVA assumptions, the plots can often tell us in exactly what way the data does not satisfy the ANOVA assumptions and, in fact, tell us how to transform the data so that it does satisfy the ANOVA assumptions.

See Lab 10: Q–Q Plots and $e \vee p$ Plots For ANOVA.

Exercise 8.4 (ANOVA Assumptions For Completely Randomized Design (CRD))

1. *Storing and Retrieving Data On the TI–83.* It will be convenient, for later on, to be able to easily store and retrieve this stored data from the TI–83. Consider the following data set.

drug 1	5.90	5.92	5.91	5.89	5.88
drug 2	5.51	5.50	5.50	5.49	5.50
drug 3	5.01	5.00	4.99	4.98	5.02

   (a) *Store This Data in a file called DRUG.*
      i. Type data into L_1, L_2 and L_3
      ii. 2nd MEM 8:Group ENTER 1:Create New
      iii. type in file name, such as "DRUG", say, using *green* letters above the keys, then ENTER
      iv. arrow down to 4:List... ENTER and select L_1 ENTER, arrow down, L_2 ENTER, arrow down and finally L_3, then DONE ENTER

   (b) *Retrieving Stored Data from DRUG.*
      i. 2nd MEM 8:Group ENTER

[4]Both plots can also, often, be used to detect for dependence. Dependence may occur in different ways; for example, if there are confounding factors (dealt with by using randomization) or if the data is serially correlated.

ii. UNGROUP DRUG

iii. 3:Overwrite All (lists L_1, L_2, L_3 with new data)

2. *Checking ANOVA Assumptions: Three Drugs.* Fifteen different patients are subjected to three drugs.

drug 1	5.90	5.92	5.91	5.89	5.88
drug 2	5.51	5.50	5.50	5.49	5.50
drug 3	5.01	5.00	4.99	4.98	5.02

(a) *Normality?* Various q–q plots are given below.

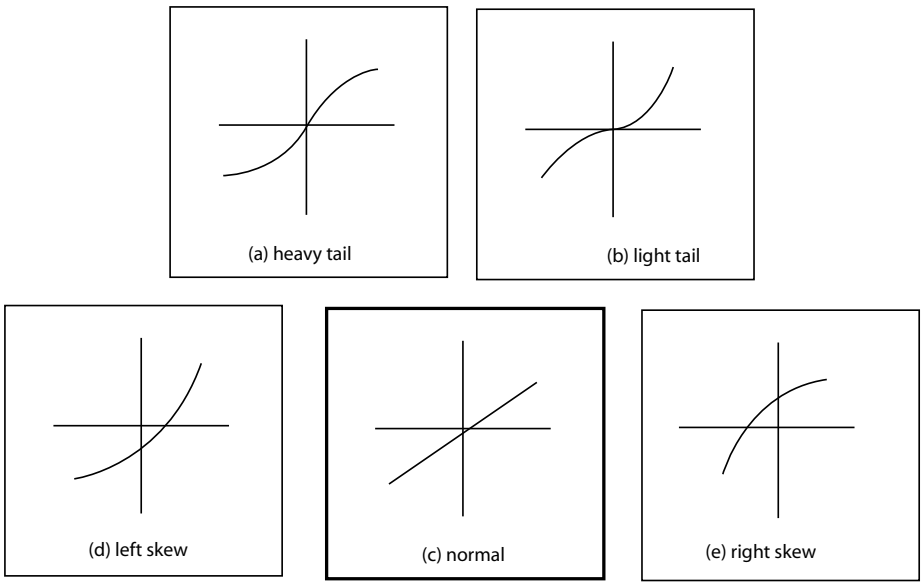

Figure 8.1 (Q–Q Plots)

The q–q plot for the data given above indicates (circle one)
heavy tail / light tail / normality
left skew / right skew / none of these
(Type drug 1, 2 and 3 responses in L_1, L_2, L_3 respectively, then PRGM QQPLTANV ENTER 3 ENTER)

(b) *Equal Variance?* Various $e \vee p$ plots are given below.

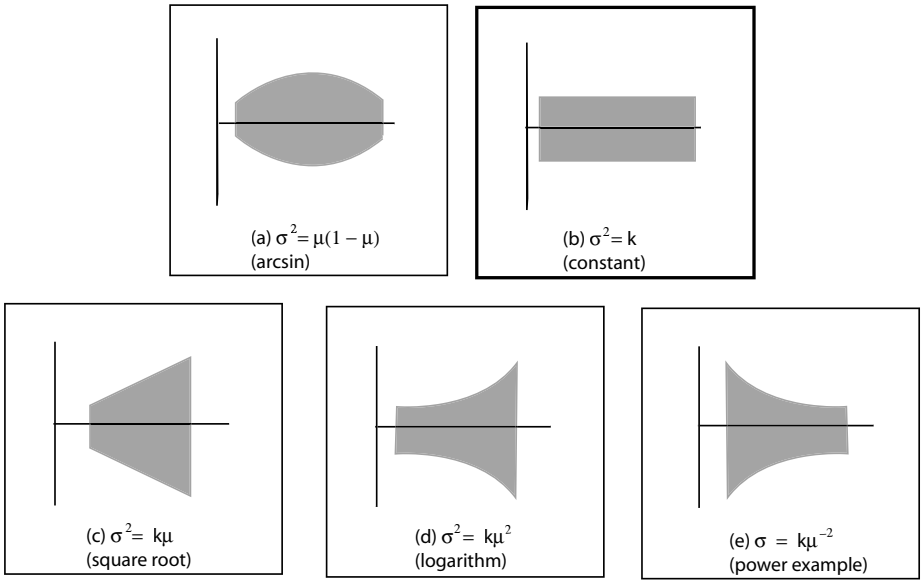

Figure 8.2 ($e \vee p$ Plots)

The $e \vee p$ plot for the data given above indicates variance is *not* not constant with respect to the mean μ, but varies in the following way: (circle one)
$\boldsymbol{\sigma^2 = \mu(1-\mu)}$ / $\boldsymbol{\sigma^2 = k}$ / $\boldsymbol{\sigma^2 = k\mu}$
$\boldsymbol{\sigma^2 = k\mu^2}$ / $\boldsymbol{\sigma^2 = k\mu^{-2}}$ / **none of these**.
(Type drug 1, 2 and 3 responses in L_1, L_2, L_3 respectively, then PRGM EVPPLOT ENTER 3 ENTER)

(c) It (circle one) **is** / **is not** possible to use ANOVA on this data because the variance is not constant.

(d) **True** / **False** In fact, since the $e \vee p$ plot does not appear to show any of the specified $e \vee p$ plot patterns, it is not clear how to transform the data to satisfy the ANOVA assumptions.

3. *Logarithm Transformation Required: Soil–Water Fluxes.* Soil–water fluxes were measured for different grades of soil.

grade 1 soil	0.306	0.363	0.437	
grade 2 soil	0.787	0.899	1.272	1.424
grade 3 soil	1.634	1.682	5.128	

(a) *Normality?* The q–q plot for the data given above indicates (circle one)
heavy tail / **light tail** / **normality**
left skew / **right skew** / **none of these**

(b) *Equal Variance?* The $e \vee p$ plot for the data given above indicates variance is related to the mean μ, in the following way: (circle one)
$\boldsymbol{\sigma^2 = \mu(1-\mu)}$ / $\boldsymbol{\sigma^2 = k}$ / $\boldsymbol{\sigma^2 = k\mu}$
$\boldsymbol{\sigma^2 = k\mu^2}$ / $\boldsymbol{\sigma^2 = k\mu^{-2}}$ / **none of these**.

(c) It (circle one) **is / is not** possible to use ANOVA on this data because of the left skew and the variance is not constant.

(d) **True / False** Since the $e \vee p$ plot appears to show $\sigma^2 = k\mu^2$, we can use the logarithm transformation, $g(y) = \ln y$, on the data to force the variance to become constant, $\sigma^2 = k$, to satisfy one of the ANOVA assumptions. Also, fortunately, this transformation also tends to make the data more closely follow a normal, another of the required ANOVA assumptions.

4. *Binomial (Arcsin Transformation Required): Pest–Free Apples.* The number of pest–free apples (out of 100 per tree) produced by apple plants in five differently applied pesticide plots of land is tabulated below. (The data has been sorted to make it easier to type into your calculators.)

pesticide 1	12	13	15	15	15	16	16	17	18	22
pesticide 2	28	30	31	32	33	33	37	37	40	41
pesticide 3	42	48	48	54	54	55	56	60	63	64
pesticide 4	70	71	72	73	75	75	77	79	79	80
pesticide 5	92	94	95	96	96	97	97	98	98	99

(a) *Normality?* The q–q plot for the data given above indicates (circle one)
heavy tail / light tail / normality
left skew / right skew / none of these

(b) *Equal Variance?* The $e \vee p$ plot for the data given above indicates variance is related to the mean μ, in the following way: (circle one)
$\boldsymbol{\sigma^2 = \mu(1-\mu)}$ / $\boldsymbol{\sigma^2 = k}$ / $\boldsymbol{\sigma^2 = k\mu}$
$\boldsymbol{\sigma^2 = k\mu^2}$ / $\boldsymbol{\sigma^2 = k\mu^{-2}}$ / **none of these**.

(c) It (circle one) **is / is not** possible to use ANOVA on this data because the variance is not constant.

(d) **True / False** Since the $e \vee p$ plot appears to show $\sigma^2 = k\mu(1-\mu)$, we can use the arcsin transformation, $g(y) = sin^{-1}\sqrt{y}$, on the data to force the variance to become constant, $\sigma^2 = k$, to satisfy one of the ANOVA assumptions. Also, fortunately, this transformation also tends to make the data more closely follow a normal, another of the required ANOVA assumptions.

In fact, the shape of the $e \vee p$ plot indicates the data from each of the five plots has been sampled from five different binomial distributions[5]

5. *Poisson (Square–Root Transformation Required): Barley Plant Mass.* Barley plant mass was measured for different amounts of soil drainage.

[5]In fact, I happen to know (but you would not know in general) the five binomial distributions have parameters $\pi_1 = 0.15$, $\pi_2 = 0.35$, $\pi_3 = 0.55$, $\pi_4 = 0.75$ and $\pi_5 = 0.95$.

soil drainage 1	7.6	8.4	8.9	8.9	8.5	7.7	6.7
soil drainage 2	9.5	10.2	10.2	9.6	8.5	7.1	5.6
soil drainage 3	6.3	9.0	11.3	12.5	12.5	11.3	9.5
soil drainage 4	3.3	8.4	14.0	17.5	17.5	14.6	10.4

(a) *Normality?* The q–q plot for the data given above indicates (circle one)
heavy tail / light tail / normality
left skew / right skew / none of these

(b) *Equal Variance?* The $e \vee p$ plot for the data given above indicates variance is related to the mean μ, in the following way: (circle one)
$\boldsymbol{\sigma^2 = k\mu(1-\mu)}$ / $\boldsymbol{\sigma^2 = k}$ / $\boldsymbol{\sigma^2 = k\mu}$
$\boldsymbol{\sigma^2 = k\mu^2}$ / $\boldsymbol{\sigma^2 = k\mu^{-2}}$ / **none of these**.

(c) It (circle one) **is** / **is not** possible to use ANOVA on this data because of the right skew and the variance is not constant.

(d) **True** / **False** Since the $e \vee p$ plot appears to show $\sigma^2 = k\mu$, we can use the square–root transformation, $g(y) = \sqrt{y}$, on the data to force the variance to become constant, $\sigma^2 = k$, to satisfy one of the ANOVA assumptions. Also, fortunately, this transformation also tends to make the data more closely follow a normal, another of the required ANOVA assumptions.

In fact, the shape of the $e \vee p$ plot indicates the data from each of the five plots has been sampled from four different Poisson distributions[6].

8.7 Analysis of Transformed Data

We use the $e \vee p$ plots to tell us how to transform the data so that it does satisfy the ANOVA assumptions and then perform an ANOVA procedure on the transformed data. Even though the ANOVA procedure is conducted on the *transformed* (rather than original) data, the results of this procedure typically still apply to the original data because the ANOVA procedure simply determines whether multiple means are the same or different. That is, the ANOVA procedure is a *relative*, rather than *absolute* procedure which checks to see if the means, whatever transformed scale they are on, are close to one another or not.

Exercise 8.5 (Transforming Data)

1. *Logarithm Transformation Required: Soil–Water Fluxes.* Soil–water fluxes were measured for different grades of soil.

[6]In fact, I happen to know (but you would not know in general) the four Poisson distributions have parameters $\lambda_1 = 20$ $(i = 17, \ldots, 23)$, $\lambda_2 = 15$ $(i = 13, \ldots, 19)$, $\lambda_3 = 10$ $(i = 6, \ldots, 14)$, $\lambda_4 = 5$ $(i = 1, \ldots, 8)$ where each are multiplied by 100.

grade 1 soil	0.306	0.363	0.437	
grade 2 soil	0.787	0.899	1.272	1.424
grade 3 soil	1.634	1.682	5.128	

(a) Use the logarithm transformation ($g(y) = \ln y$) to convert this data into one with equal variance (fill in the blanks):

grade 1 soil	-1.184	-1.103	-0.8278	
grade 2 soil	-0.2395	_____	0.24059	0.35347
grade 3 soil	0.49103	0.51998	_____	

(b) *Normality?* The q–q plot for the data given above indicates (circle one)
heavy tail / light tail / normality
left skew / right skew / none of these

(c) *Equal Variance?* The $e \vee p$ plot for the data given above indicates variance is related to the mean μ, in the following way: (circle one)
$\boldsymbol{\sigma^2 = k\mu(1-\mu)}$ / $\boldsymbol{\sigma^2 = k}$ / $\boldsymbol{\sigma^2 = k\mu}$
$\boldsymbol{\sigma^2 = k\mu^2}$ / $\boldsymbol{\sigma^2 = k\mu^{-2}}$ / **none of these**.

(d) Even though there still appears to be a left skew, complete the ANOVA table (fill in the blanks)

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment (Soil)	_____	_____	_____
Error	_____	_____	_____
Total	_____	_____	

where $F = 16.43$ and the p–value is essentially zero (in other words, the soil–water fluxes are different for different grades of soil). Although this data is *not* the same as the original data, it is safe to say the soil–water fluxes are different.
(Type the data into L_1, L_2, L_3, then type STAT TESTS F:ANOVA(L_1, L_2, L_3) ENTER.)

2. *Arcsin Transformation ($g(y) = sin^{-1}\sqrt{y}$): Pest–Free Apples.* The number of pest–free apples (out of 100 per tree) produced by apple plants in five differently applied pesticide plots of land is tabulated below. The $e \vee p$ plot tells us the data from each of the five plots has been sampled from five different binomial distributions.

pesticide 1	12	13	15	15	15	16	16	17	18	22
pesticide 2	28	30	31	32	33	33	37	37	40	41
pesticide 3	42	48	48	54	54	55	56	60	63	64
pesticide 4	70	71	72	73	75	75	77	79	79	80
pesticide 5	92	94	95	96	96	97	97	98	98	99

(a) Use the arcsin transformation ($g(y) = sin^{-1}\sqrt{y/100}$, where MODE is set to DEGREE) to convert this data into one with equal variance (fill in the blanks):

pesticide 1	20.2	21.1	22.8	22.8	22.8	23.6	23.6	24.4	25.1	27.9
pesticide 2	31.9	33.2	_____	34.5	35.1	35.1	37.5	37.5	39.2	39.8
pesticide 3	40.4	43.9	43.9	47.3	47.3	47.9	48.4	50.8	52.5	53.1
pesticide 4	56.8	57.4	58.1	58.7	_____	60	61.3	62.7	62.7	63.4
pesticide 5	73.6	75.8	77.1	78.5	78.5	80.0	80.0	81.9	81.9	84.3

(b) *Normality?* The q–q plot for the data given above indicates (circle one)
heavy tail / light tail / normality
left skew / right skew / none of these

(c) *Equal Variance?* The $e \vee p$ plot for the data given above indicates variance is related to the mean μ, in the following way: (circle one)
$\boldsymbol{\sigma^2 = k\mu(1-\mu)}$ / $\boldsymbol{\sigma^2 = k}$ / $\boldsymbol{\sigma^2 = k\mu}$
$\boldsymbol{\sigma^2 = k\mu^2}$ / $\boldsymbol{\sigma^2 = k\mu^{-2}}$ / **none of these**.

(d) Complete the ANOVA table (fill in the blanks)

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment (Pesticides)	_____	_____	_____
Error	_____	_____	_____
Total	_____	_____	

where $F = 538.7$ and the p–value is essentially zero (in other words, the pesticides are acting differently on the apple trees). Although this data is *not* the same as the original data, it is safe to say the pesticides are acting differently.
(Type the data into L_1, ..., L_5, then type STAT TESTS F:ANOVA(L_1, L_2, L_3, L_4, L_5) ENTER.)

3. *Poisson (Square–Root Transformation Required): Barley Plant Mass.* Barley plant mass was measured for different amounts of soil drainage.

soil drainage 1	7.6	8.4	8.9	8.9	8.5	7.7	6.7
soil drainage 2	9.5	10.2	10.2	9.6	8.5	7.1	5.6
soil drainage 3	6.3	9.0	11.3	12.5	12.5	11.3	9.5
soil drainage 4	3.3	8.4	14.0	17.5	17.5	14.6	10.4

(a) Use the square root transformation ($g(y) = \sqrt{y}$) to convert this data into one with equal variance (fill in the blanks):

soil drainage 1	_____	_____	2.98	2.98	2.90	2.77	2.59
soil drainage 2	3.09	3.20	3.20	3.10	2.91	2.66	2.36
soil drainage 3	2.51	3.00	3.36	3.54	3.54	3.37	3.08
soil drainage 4	1.84	2.90	3.75	4.19	4.19	3.82	3.23

(b) *Normality?* The q–q plot for the data given above indicates (circle one)
heavy tail / light tail / normality
left skew / right skew / none of these

(c) *Equal Variance?* The $e \vee p$ plot for the data given above indicates variance is related to the mean μ, in the following way: (circle one)
$\boldsymbol{\sigma^2 = k\mu(1-\mu)}$ / $\boldsymbol{\sigma^2 = k}$ / $\boldsymbol{\sigma^2 = k\mu}$
$\boldsymbol{\sigma^2 = k\mu^2}$ / $\boldsymbol{\sigma^2 = k\mu^{-2}}$ / **none of these**.

(d) Even though the ANOVA assumptions still do not appear to be satisfied, complete the ANOVA table (fill in the blanks)

Source	Degrees of Freedom	Sum Of Squares	Mean Squares
Treatment (Drainage)	_____	_____	_____
Error	_____	_____	_____
Total	_____	_____	

where $F = 1.98$ and the p–value is 0.143. Although this data is *not* the same as the original data, it appears the barley masses are the same for different soil drainage[7].
(Type the data into L_1, ..., L_4, then type STAT TESTS F:ANOVA(L_1, L_2, L_3, L_4) ENTER.)

8.8 One–Way Classified Ordinal Data

Not covered.

[7]This result is interesting since the averages are, in fact, different. This incorrect result may be due to the non–constant variance violation.