

MATH 204: PRINCIPLES OF STATISTICS 2
WINTER 2026

HANDOUTS

MATH 204: PRINCIPLES OF STATISTICS 2

WINTER 2026

Instructor David A. Stephens (david.stephens@mcgill.ca, Burnside 1225)
Textbook *Statistics* (13th Edition)
J. T. McClave and T. Sincich
Purchasing the textbook is not mandatory; the course will follow the structure of Chapters 10–14, but the course materials will be self-contained.

Method of Evaluation: Evaluation for the course will be on the following basis

- **50 % Final Examination**
 - Examination will be in-person and held during the final exam period.
 - Please note that the Instructor has no jurisdiction over the date of the final examination.
- **20 % Midterm Examination**
 - Midterm will be held on **Thursday, 19th February**
 - Midterm examination will be a written exam covering material up to and including Lecture 12 material.
- **15 % Assignments**
 - There will be **two** assignments.
 - Assignments will be a combination of short and long answer questions, submitted electronically.
- **15 % Quizzes**
 - There will be **eight** quizzes, the best **five** will count for the assessment.
 - Quizzes will be open-book time-limited short answer myCourses questions.
 - Quiz topics will be closely related to topics discussed in lectures.

Please see the Course Schedule for the timing of the in-term assessments.

Note: The final grade will be determined by the aggregate score

$$70\% \text{ Final Exam} + 15\% \text{ Assignments} + 15\% \text{ Quizzes}$$

if this score exceeds the principal method of evaluation.

The instructor bears sole responsibility for grading issues. Please contact Prof. Stephens (by email at david.stephens@mcgill.ca) if you have any questions concerning grading of any assessed work.

TARGET COURSE SCHEDULE

Week		Lectures	Topics	Q	A
1	05/01	1–2	Outline. Recap of Hypothesis Testing Concepts		
2	12/01	3–4	Designed Experiments. The Analysis of Variance. The Completely Randomized Design. The ANOVA F-test.		
3	19/01	5–6	Completely Randomized Design Examples and Assumptions	Q1	
4	26/01	7–8	Blocking and Randomized Block Designs. Replication.	Q2	
5	02/02	9–10	Factorial Experiments and multi-factor ANOVA.	Q3	A1
6	09/02	11–12	Parameterization. Illegal Models.	Q4	
7	16/02	13–14	Linear regression models. Probabilistic models. Least-squares.		
8	23/02	15–16	Model assumptions. Hypothesis testing.		
9	02/03	-	Reading Break	-	-
10	09/03	17–18	Correlation. R^2 statistic. Prediction. ANOVA in regression. Residuals.	Q5	
11	16/03	19–20	Polynomial regression. The linear model. Factor predictors. Multiple linear regression. Model selection.	Q6	
12	23/03	21–22	ANOVA and F-tests in multiple regression. Unbalanced designs. Pitfalls.	Q7	A2
13	30/03	23–24	Chi-squared tests for count data. Exact tests. Non-parametric tests.	Q8	
14	06/04	25–26	Non-parametric tests based on ranks.		

- ‘Q’ refers to Quiz number; ‘A’ refers to Assignment number.
- Assessments will usually be released at the end of the week stated.
- Extra (non-assessed) practice quizzes and other materials may be provided.

Policy on Late Submission:

- Late submission of the Quizzes will not be possible due to the time-limited and machine-scored nature of the assessment.
- Late submission of the Assignments may be subject to 100% penalty.
- Requests for accommodation due to extenuating circumstances should be directed to Prof. Stephens, david.stephens@mcgill.ca.

Policy on the use of AI:

- You may use generative AI tools to assist with coding and with the production of text-based answers. However, anything that you submit for assessment must be **your own work**: for example, it must be the result of **your own use** of the AI tools.
- You are responsible for the work you submit: errors resulting from use of AI tools will be penalized in the usual way.

Key Academic Dates: <https://www.mcgill.ca/importantdates/key-dates>

- Add/Drop deadline: Tuesday, January 20th
- Winter Reading Break: March 2nd to March 6th inclusive
- Last lecture: Thursday, April 9th
- Classes end: Tuesday, April 14th

Tutorials & Office Hours:

- There will be two one hour optional tutorials per week; these will begin in Week 2.
- Instructor and Teaching Assistants will also hold weekly Office Hours.

Scheduling details will be announced on myCourses.

myCourses:

- Official communications will take place via myCourses.
- All course materials will be available via myCourses.

David A. Stephens
Department of Mathematics and Statistics
McGill University

UNDERSTANDING THE ANOVA F-STATISTIC

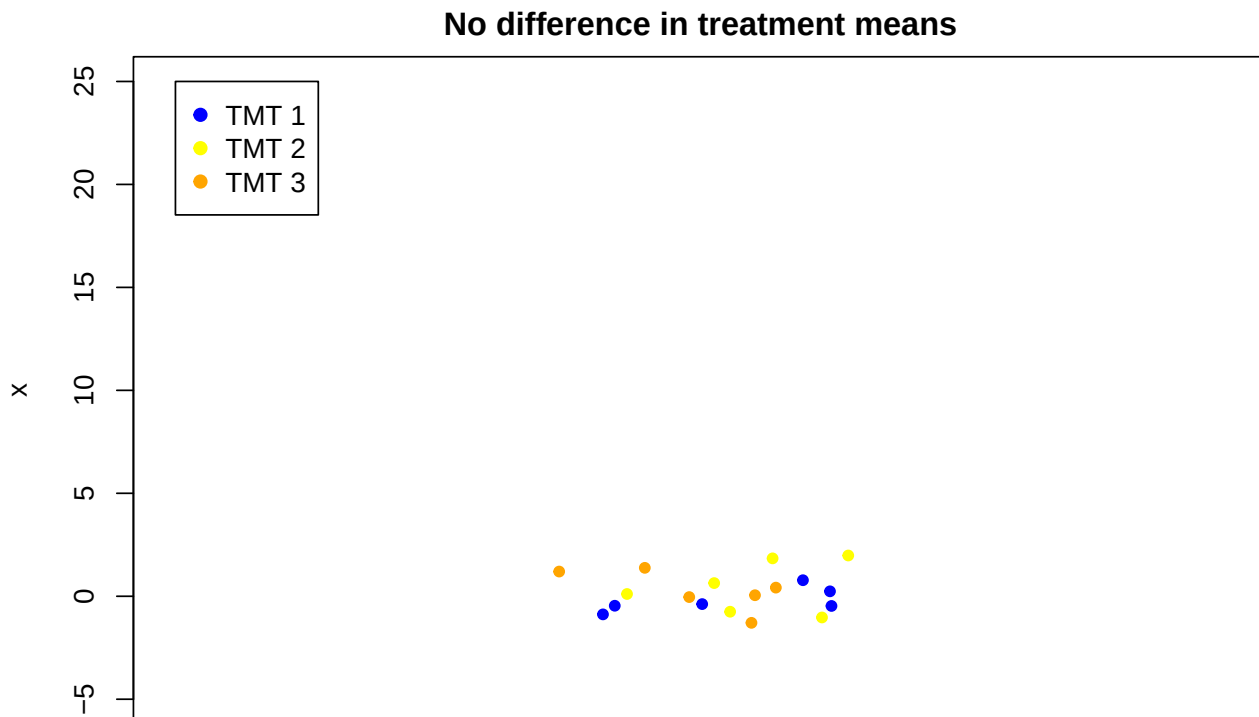
Suppose that we have $k = 3$ treatment groups in a Completely Randomized Design, with sample sizes $n_1 = n_2 = n_3 = 6$. Suppose first that the treatment means are all zero, that is

$$\mu_1 = \mu_2 = \mu_3 = 0$$

and that the treatment group variance parameter σ^2 is equal to 1. A typical data set is

							$\bar{x}_i$	s_i^2
TMT 1	-0.88	0.24	-0.46	0.78	-0.47	-0.38	-0.195	0.358
TMT 2	-0.75	0.11	0.64	1.98	-1.03	1.84	0.465	1.611
TMT 3	1.38	1.20	0.42	0.05	-1.29	-0.04	0.287	0.939

```
> x<-c(-0.88,0.24,-0.46,0.78,-0.47,-0.38,
+      -0.75,0.11,0.64,1.98,-1.03,1.84,
+      1.38,1.20,0.42,0.05,-1.29,-0.04)
> tmt<-rep(1:3, each=6)
> nvec<-rep(6,3)
> n<-sum(nvec)
> k<-3
> cvec<-c('blue','yellow','orange')
> par(mar=c(1,4,2,2))
> plot(runif(sum(nvec),-3,3),x,col=cvec[tmt],pch=19,cex=0.8,
+      ylim=range(-5,25),xlim=range(-10,10),xaxt='n')
> legend(-10,25,c('TMT 1','TMT 2','TMT 3'),pch=19,col=cvec)
> title('No difference in treatment means')
```



```

> #Summary statistics
> xbar.i<-tapply(x,tmt,mean)
> xvar.i<-tapply(x,tmt,var)
> cbind(1:3,xbar.i,xvar.i)

+           xbar.i    xvar.i
+ 1 1 -0.195000 0.357830
+ 2 2  0.465000 1.610750
+ 3 3  0.286667 0.939187

> xbar<-mean(x)
> xbar

+ [1] 0.185556

```

These data yield $\bar{x} = 0.185556$, and

$$s_P^2 = \frac{1}{n-k} \sum_{i=1}^k (n_i - 1) s_i^2 = 0.969.$$

```

> sp.sq<-sum((nvec-1)*xvar.i)/(n-k)
> sp.sq

+ [1] 0.969256

```

For these data, we have using the definitions from lectures

$$SST = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 = 1.399 \qquad SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = 14.539$$

and

$$SS = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = 15.938$$

so that the equation $SS = SST + SSE$ holds.

```

> SST<-sum(nvec*(xbar.i-xbar)^2)
> SSE<-sum((x-rep(xbar.i, nvec))^2)
> SS<-sum((x-xbar)^2)
> c(SST, SSE, SS)

+ [1] 1.39881 14.53883 15.93764

```

For the F-statistic, we have

$$F_{obs} = \frac{MST}{MSE} = \frac{SST/(k-1)}{SSE/(n-k)} = \frac{1.399/2}{14.539/15} = 0.722$$

```
> Fobs<-(SST/(k-1))/(SSE/(n-k))
> Fobs

+ [1] 0.72159
```

To complete the test, we compare this with the upper α probability point of the F-distribution with $(k-1, n-k) = (2, 15)$ degrees of freedom. With $\alpha = 0.05$, from the tables in McClave and Sincich, we see that

$$F_{\alpha,2,15} = 3.68$$

and we **do not reject** the ANOVA F-test null hypothesis

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

Here, this is the **correct** conclusion, as in fact all the true treatment means are zero. Thus a **small** value of the test statistic F supports H_0 .

```
> alpha<-0.05
> qf(1-alpha, k-1,n-k)

+ [1] 3.68232
```

Now suppose that, in fact,

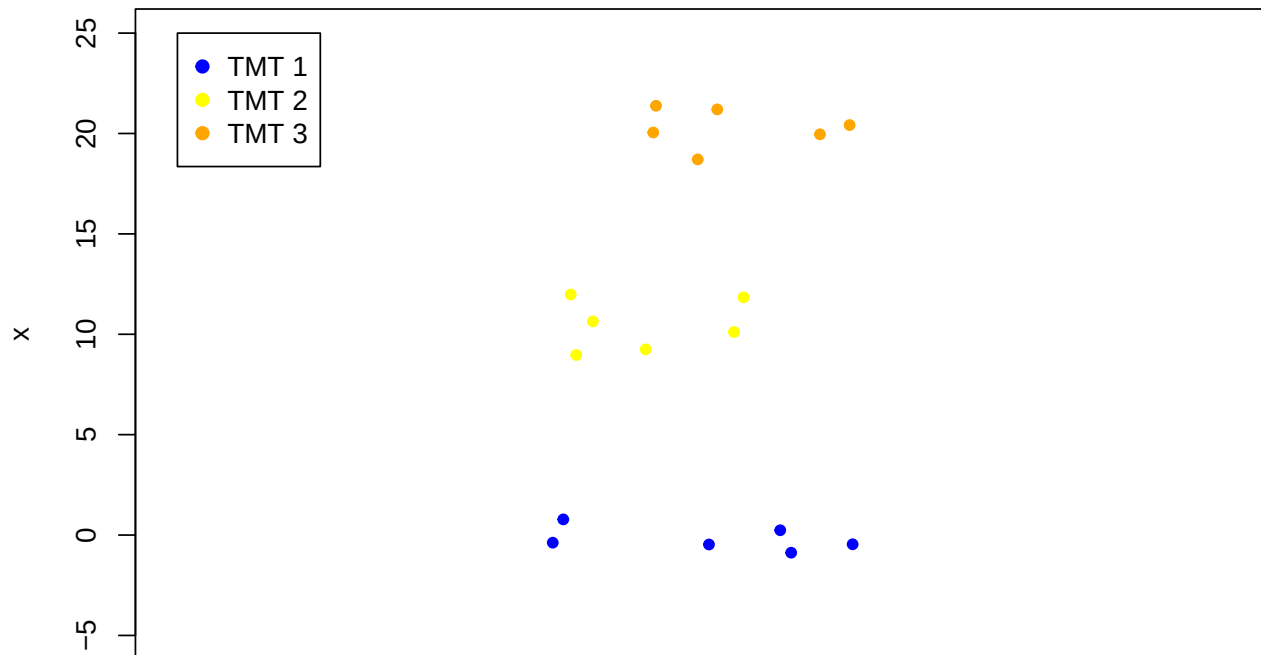
$$\mu_1 = 0 \quad \mu_2 = 10 \quad \mu_3 = 20.$$

The equivalent data set to the one above but with the treatment means changed in this way takes the form

							$\bar{x}_i$	s_i^2
TMT 1	-0.88	0.24	-0.46	0.78	-0.47	-0.38	-0.195	0.358
TMT 2	9.25	10.11	10.64	11.98	8.97	11.84	10.465	1.611
TMT 3	21.38	21.20	20.42	20.05	18.71	19.96	20.287	0.939

```
> x<-c(-0.88,0.24,-0.46,0.78,-0.47,-0.38,
+      -0.75,0.11,0.64,1.98,-1.03,1.84,
+      1.38,1.20,0.42,0.05,-1.29,-0.04)
> x[7:12]<-x[7:12]+10
> x[13:18]<-x[13:18]+20
> tmt<-rep(1:3, each=6)
> nvec<-rep(6,3)
> n<-sum(nvec)
> k<-3
> cvec<-c('blue','yellow','orange')
> par(mar=c(1,4,2,2))
> plot(runif(sum(nvec),-3,3),x,col=cvec[tmt],pch=19,cex=0.8,
+      ylim=range(-5,25),xlim=range(-10,10),xaxt='n')
> legend(-10,25,c('TMT 1','TMT 2','TMT 3'),pch=19,col=cvec)
> title('Difference in treatment means')
```

Difference in treatment means



```
> xbar.i<-tapply(x,tmt,mean)
> xvar.i<-tapply(x,tmt,var)
> cbind(1:3,xbar.i,xvar.i)

+      xbar.i    xvar.i
+ 1 1 -0.1950 0.357830
+ 2 2 10.4650 1.610750
+ 3 3 20.2867 0.939187

> xbar<-mean(x)
> xbar

+ [1] 10.1856
```

yielding $\bar{x} = 10.186$, and

$$s_P^2 = \frac{1}{n-k} \sum_{i=1}^k (n_i - 1) s_i^2 = 0.969.$$

Note that the sample means have changed accordingly, but that the sample variances **have not changed at all**.

```
> sp.sq<-sum((nvec-1)*xvar.i)/(n-k)
> sp.sq

+ [1] 0.969256
```

On further calculation, we have

$$SST = 1259.199$$

$$SSE = 14.539$$

$$SS = 1273.738$$

so that

$$MST = \frac{1259.199}{2} = 629.600 \quad MSE = \frac{14.539}{15} = 0.969$$

so that

$$F_{obs} = \frac{629.600}{0.969} = 649.570.$$

```
> SST<-sum(nvec*(xbar.i-xbar)^2)
> SSE<-sum((x-rep(xbar.i, nvec))^2)
> SS<-sum((x-xbar)^2)
> c(SST, SSE, SS)

+ [1] 1259.1988    14.5388 1273.7376

> Fobs<-(SST/(k-1))/(SSE/(n-k))
> Fobs

+ [1] 649.57
```

We again compare this with $F_{\alpha,2,15} = 3.68$ (the critical value, C_R), and notice that the F statistic is **much larger** than this critical value. The test statistic thus lies within the rejection region, and hence we **reject H_0** .

This example illustrates that SST measures the variability **between** means across the treatment groups, whereas SSE measures the variability **within** treatment groups, allowing for the possibility that the treatment means may be different. The quantity SS measures the total amount of variability; in the first example $SS = SST + SSE$ gives

$$15.938 = 1.399 + 14.539$$

so most of the variability is contributed by SSE, whereas in the second example, we have

$$1273.738 = 1259.199 + 14.539$$

and most of the variability is contributed by SST.

THE SUMS OF SQUARES DECOMPOSITION

We have seen the fundamental decomposition

$$SS = SST + SSE$$

borne out in numerical calculations. The decomposition is justified as follows:

$$\begin{aligned}
 SS &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 \\
 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 \\
 &= \sum_{i=1}^k \sum_{j=1}^{n_i} [(x_{ij} - \bar{x}_i)^2 + (\bar{x}_i - \bar{x})^2 + 2(x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x})] \\
 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})(x_{ij} - \bar{x}_i) \\
 &= \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + 0 \\
 &= SST + SSE
 \end{aligned}$$

- Line 2 follows from line 1 by adding $0 = \bar{x}_i - \bar{x}_i$ inside the bracket.
- Line 3 follows from line 2 by squaring out the bracket

$$(a + b)^2 = a^2 + b^2 + 2ab$$

with $a = (x_{ij} - \bar{x}_i)$ and $b = (\bar{x}_i - \bar{x})$.

- Line 4 follows from line 3 by properties of summation and rearrangement.
- Line 5 follows from line 4 as for the three separate sums, we have for the first term

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$$

as the summand $(\bar{x}_i - \bar{x})^2$ does not depend on j – it is **constant** in the inner sum, and gets counted n_i times. For the third term we have

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})(x_{ij} - \bar{x}_i) = \sum_{i=1}^k \left\{ (\bar{x}_i - \bar{x}) \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) \right\} = 0$$

as the term in the curly bracket is zero for all i as, by properties of summation,

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = \sum_{j=1}^{n_i} x_{ij} - n_i \bar{x}_i = \sum_{j=1}^{n_i} x_{ij} - \sum_{j=1}^{n_i} x_{ij} = 0.$$

USING THE F TABLES

Tables in the textbook contain information on the upper α points for the F-distribution for $\alpha = 0.1, 0.05, 0.025$ and 0.01 respectively, and for different values of the **degrees of freedom** parameters. The values in the body of the table are the numbers x which solve the equation

$$\Pr[F > x] = \alpha$$

when the statistic F has a F-distribution with ν_1 and ν_2 degrees of freedom, written

$$F \sim F_{\nu_1, \nu_2}$$

where ν_1 and ν_2 are whole numbers greater than zero.

The table below is the F table for $\alpha = 0.05$, equivalent to the textbook table.

Table 1: Upper $\alpha = 0.05$ points of F_{ν_1, ν_2} distribution

ν_2	ν_1								
	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65

From textbook, p 822. Reported to 2 dp.

We read ν_1 from the **column** and ν_2 from the **row**. For example,

- if $\nu_1 = 9$ and $\nu_2 = 4$, we know from the table that

$$\Pr[F > 6.00] = 0.05$$

- if $\nu_1 = 6$ and $\nu_2 = 3$, we know from the table that

$$\Pr[F > 8.94] = 0.05$$

- if $\nu_1 = 2$ and $\nu_2 = 14$, we know from the table that

$$\Pr[F > 3.74] = 0.05$$

Table 2: Upper $\alpha = 0.01$ points of F_{ν_1, ν_2} distribution

ν_2	ν_1								
	1	2	3	4	5	6	7	8	9
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03

From textbook, p 826. Reported to 2 dp.

The equivalent table for $\alpha = 0.01$ is produced above.

Note: The F-distribution is a non-symmetric probability distribution with a specific property that allows the tables to tabulate only the **right-hand tail** of the distribution. If we need to look up the left-hand tail, we can use the fact that if $F \sim F_{\nu_1, \nu_2}$, and $0 < p < 1$

$$\Pr[F > x] = p \quad \implies \quad \Pr[1/F \leq 1/x] = p \quad \implies \quad \Pr[1/F > 1/x] = 1 - p.$$

But it transpires that

$$F \sim F_{\nu_1, \nu_2} \quad \implies \quad \frac{1}{F} \sim F_{\nu_2, \nu_1}.$$

To look up the **left-tail** α probability point for the $F \sim F_{\nu_1, \nu_2}$ distribution, we look up the **right-tail** $1 - \alpha$ probability point for the F_{ν_2, ν_1} distribution, and then take the reciprocal. For example,

- if $F \sim F_{9,4}$, we use tables to discover that as $F_{0.05,4,9} = 3.63$, it follows that

$$\Pr[F \leq 1/3.63] = \Pr[F \leq 0.28] = 0.05$$

giving the 0.05 (left-tail) probability point of the $F_{9,4}$ distribution as 0.28.

```
> qf(0.05, 9, 4)
+ [1] 0.275248
> 1/qf(0.95, 4, 9)
+ [1] 0.275248
```

Table 3: Upper $\alpha = 0.05$ points of F_{ν_1, ν_2} distribution

	ν_2	ν_1																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	Inf
14	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
	120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
	Inf	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

From textbook, pp 822–823. Reported to 2 dp.

ONE-WAY ANOVA: WORKED EXAMPLE

A standard model of memory is that the degree to which the subject remembers verbal material is a function of the degree to which it was processed when it was initially presented.

Reference: Craik, F. I. M. and Lockhart, R. S. (1972). Levels of Processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.

Experiment: Fifty subjects aged between 55 and 65 years were randomly assigned to one of five groups which carried out different memory tasks. The five groups included

- The Counting group was asked to read through a list of words and simply count the number of letters in each word.
- The Rhyming group was asked to read each word and think of a word that rhymed with it.
- The Adjective group had to process the words to the extent of giving an adjective that could reasonably be used to modify each word on the list.
- The Imagery group was instructed to try to form vivid images of each word.
- The Intentional group was told to read through the list and to memorize the words for later recall.

After subjects had gone through the list of 27 items three times, they were given a sheet of paper and asked to write down all the words they could remember. The response data were the number of words recalled by each individual in each group, and are presented below:

Counting	Rhyming	Adjective	Imagery	Intentional
9	7	11	12	10
8	9	13	11	19
6	6	8	16	14
8	6	6	11	5
10	6	14	9	10
4	11	11	23	11
6	6	13	12	14
5	3	13	10	15
7	8	10	19	11
7	7	11	11	11

Research question: Does the level of processing required when material is processed affect how much material is remembered ?

```
> load("204DataSets.rda")
> head(Memory)

+           Task Memory
+ 1    Counting      9
+ 2    Rhyming       7
+ 3  Adjective     11
+ 4    Imagery     12
+ 5 Intentional    10
+ 6    Counting      8

> par(mar=c(4, 4, 2, 2))
> boxplot(Memory~Task, data=Memory)
```

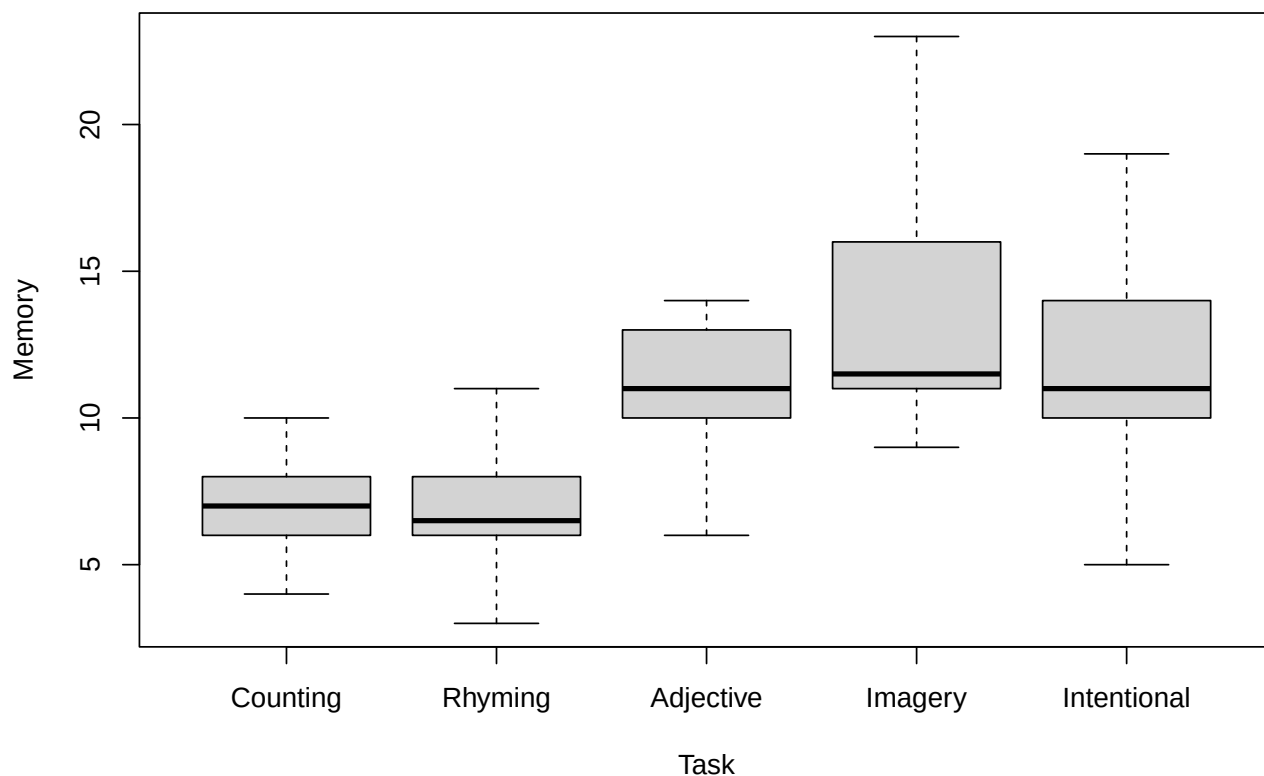


Figure 1: Boxplot of number of errors made by memory task

First we assess the equal variance assumption using Levene's Test from the `car` library.

```
> #Levenes Test
> car::leveneTest(Memory~Task, center='mean', data=Memory)

+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value Pr(>F)
+ group  4    2.529 0.0536 .
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> car::leveneTest(Memory~Task, center='median', data=Memory)

+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group  4    0.893 0.476
+ ---
```

The null hypothesis of equal group variances is not rejected, using either centering by the mean or centering by the median in the Levene test.

We may proceed with the ANOVA F-test.

```
> print(summary(aov(Memory~Task,data=Memory)), digits = 5)

+           Df Sum Sq Mean Sq F value    Pr(>F)
+ Task         4 351.52   87.880   9.0848 1.815e-05 ***
+ Residuals    45 435.30    9.673
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions: The F statistic in this ANOVA is 9.085. This is to be compared with the $F_{k-1, n-k}$ distribution; here $k - 1$ is the degrees of freedom for treatment (here for factor `Task`), and $n - k$ is the residual degrees of freedom. The values are given in the `Df` column as 4 and 45. The result of the test is given by the p-value which is 1.815e-05; hence we **reject** the null hypothesis that there is no difference between treatments at the $\alpha = 0.05$ significance level.

We can also use the `lm` function to carry out an analysis

```
> summary(lm(Memory~Task,data=Memory))

+
+ Call:
+ lm(formula = Memory ~ Task, data = Memory)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+  -7.00  -1.85  -0.45   2.00   9.60
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)    7.000     0.984    7.12  6.8e-09 ***
+ TaskRhyming   -0.100     1.391   -0.07  0.9430
+ TaskAdjective  4.000     1.391    2.88  0.0061 **
+ TaskImagery    6.400     1.391    4.60  3.4e-05 ***
+ TaskIntentional 5.000     1.391    3.59  0.0008 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 3.11 on 45 degrees of freedom
+ Multiple R-squared:  0.447, Adjusted R-squared:  0.398
+ F-statistic: 9.08 on 4 and 45 DF, p-value: 1.82e-05
```

The result of the ANOVA F-test is given in the final line of output.

ANOVA FOR THE COMPLETELY RANDOMIZED DESIGN: EXAMPLES

EXAMPLE Milk protein level by Diet: data set `Milk`.

```
> par(mar=c(4, 4, 2, 2))  
> boxplot(protein~Diet, data=Milk, xlab="Diet", ylab="Protein Level (%)")
```

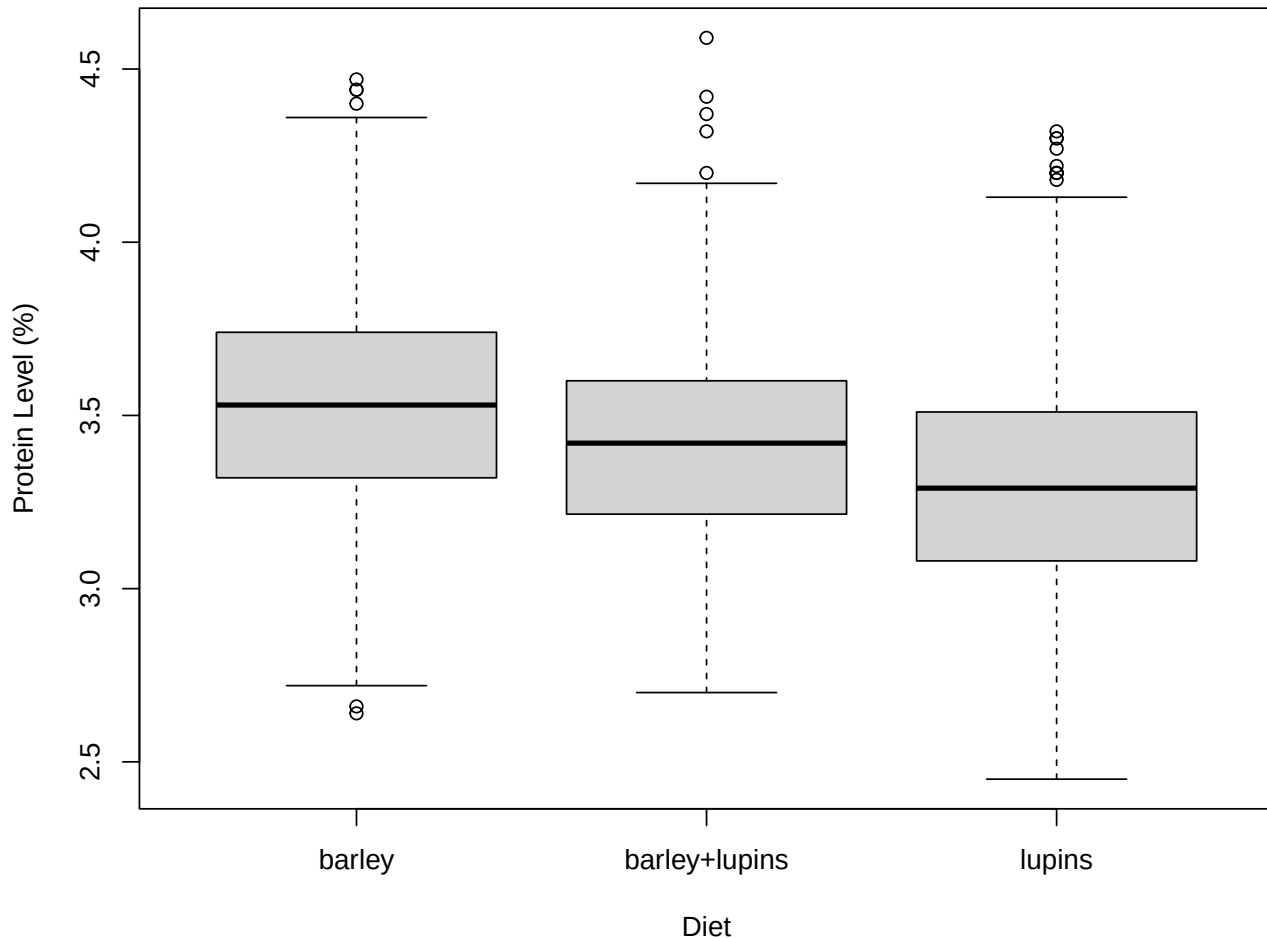


Figure 2: Boxplot of Protein level by Diet

We may write a function to compute basic summary details:

```
> #A summary function  
> my.summary<-function(x) {  
  
+     my.sum<-data.frame(n=length(x), Mean=mean(x),  
+                        StDev=sd(x), StdErr=sd(x)/sqrt(length(x)),  
+                        Lower=mean(x)-1.96*sd(x)/sqrt(length(x)),  
+                        Upper=mean(x)+1.96*sd(x)/sqrt(length(x)),  
+                        Min=min(x), Max=max(x))  
+     return(round(my.sum, 5))  
+ }
```

Applying this function to the data gives the summaries by Diet.

```
> tapply(Milk$protein, Milk$Diet, my.summary)
+ $barley
+      n      Mean    StDev  StdErr   Lower   Upper   Min   Max
+ 1 425 3.53193 0.31921 0.01548 3.50158 3.56228 2.64 4.47
+
+ $`barley+lupins`
+      n      Mean    StDev  StdErr   Lower   Upper   Min   Max
+ 1 459 3.42969 0.30234 0.01411 3.40204 3.45735 2.7 4.59
+
+ $lupins
+      n      Mean    StDev  StdErr   Lower   Upper   Min   Max
+ 1 453 3.31238 0.33709 0.01584 3.28134 3.34343 2.45 4.32
+
> my.summary(Milk$protein)
+      n      Mean    StDev  StdErr   Lower   Upper   Min   Max
+ 1 1337 3.42245 0.33175 0.00907 3.40466 3.44023 2.45 4.59
```

The Levene test results indicate that the equal variance assumption seems adequate.

```
> car::leveneTest(protein~Diet, center='mean', data=Milk)
+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value Pr(>F)
+ group  2    1.838  0.16
+      1334
+
> car::leveneTest(protein~Diet, center='median', data=Milk)
+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group  2    1.734  0.177
+      1334
```

We may now compute the sum of squares quantities for the ANOVA F-test using

- aov
- lm
- anova

functions.

```
> #Total
> summary(aov(protein~1, data=Milk))
+      Df Sum Sq Mean Sq F value Pr(>F)
+ Residuals 1336    147    0.11
+
> anova(lm(protein~1, data=Milk))
+ Analysis of Variance Table
+
+ Response: protein
+      Df Sum Sq Mean Sq F value Pr(>F)
+ Residuals 1336    147    0.1101
```

```
> #Fit and Error (Residual)
> summary(aov(protein~Diet,data=Milk))

+           Df Sum Sq Mean Sq F value Pr(>F)
+ Diet           2    10.6      5.3    51.9 <2e-16 ***
+ Residuals   1334   136.4      0.1
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(lm(protein~Diet,data=Milk))

+ Analysis of Variance Table
+
+ Response: protein
+           Df Sum Sq Mean Sq F value Pr(>F)
+ Diet           2   10.61    5.303   51.85 <2e-16 ***
+ Residuals 1334  136.43    0.102
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(lm(protein~Diet,data=Milk))

+
+ Call:
+ lm(formula = protein ~ Diet, data = Milk)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.8919 -0.2197 -0.0124  0.1981  1.1603
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)    3.5319     0.0155  227.68 < 2e-16 ***
+ Dietbarley+lupins -0.1022     0.0215   -4.75 2.3e-06 ***
+ Dietlupins      -0.2195     0.0216  -10.17 < 2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.32 on 1334 degrees of freedom
+ Multiple R-squared:  0.0721, Adjusted R-squared:  0.0707
+ F-statistic: 51.9 on 2 and 1334 DF, p-value: <2e-16
```

```
> par(mar=c(4,4,1,2))
> ggplots::plotmeans(protein~Diet,data=Milk, ylim=range(3.2,3.75))
```

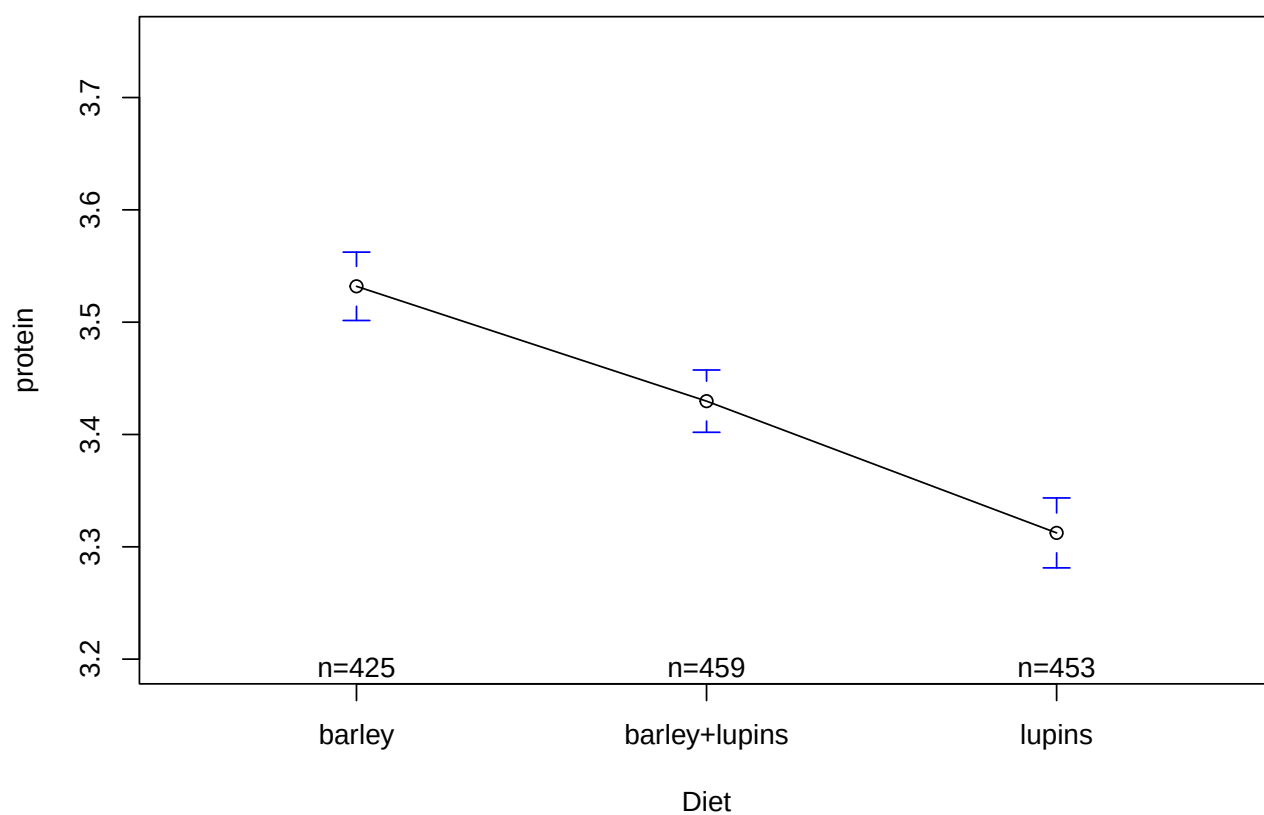


Figure 3: Means plot of Protein level by Diet, with 95 % confidence intervals for means

```
> confint(lm(protein~1+Diet, data=Milk))
+           2.5 %   97.5 %
+ Dietbarley      3.50150 3.56236
+ Dietbarley+lupins 3.40041 3.45898
+ Dietlupins      3.28291 3.34186
```

EXAMPLE Anxiety level by Dose Level: data set Pull

```
> par(mar=c(4, 4, 2, 2))
> boxplot(Pull~Dose, data=Pull, xlab="Dose", ylab="Pull strength")
```

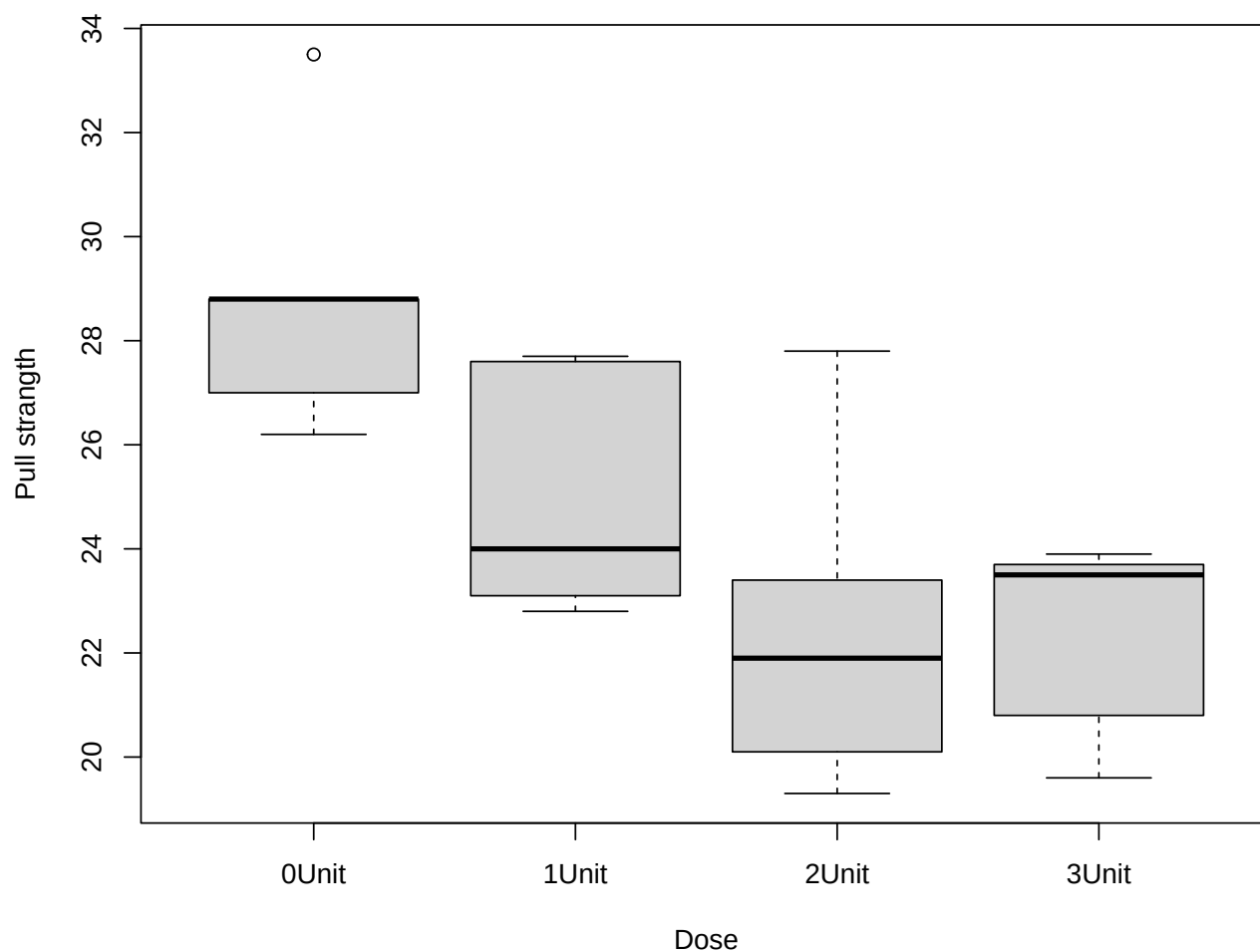


Figure 4: Boxplot of Protein level by Diet

```
> tapply(Pull$Pull, Pull$Dose, my.summary)
+ $`0Unit`
+   n Mean   StDev StdErr  Lower  Upper  Min  Max
+ 1  5 28.86 2.83161 1.26633 26.378 31.342 26.2 33.5
+
+ $`1Unit`
+   n Mean   StDev StdErr  Lower  Upper  Min  Max
+ 1  5 25.04 2.42343 1.08379 22.9158 27.1642 22.8 27.7
+
+ $`2Unit`
+   n Mean   StDev StdErr  Lower  Upper  Min  Max
+ 1  5 22.5 3.36378 1.50433 19.5515 25.4485 19.3 27.8
+
+ $`3Unit`
```

```
+   n Mean   StDev   StdErr   Lower   Upper   Min   Max
+ 1  5 22.3 1.9685 0.88034 20.5745 24.0255 19.6 23.9
```

```
> my.summary(Pull$Pull)
```

```
+   n   Mean   StDev   StdErr   Lower   Upper   Min   Max
+ 1 20 24.675 3.67364 0.82145 23.0649 26.285 19.3 33.5
```

```
> car::leveneTest(Pull~Dose,center='mean',data=Pull)
```

```
+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value Pr(>F)
+ group 3    0.295  0.829
+      16
```

```
> car::leveneTest(Pull~Dose,center='median',data=Pull)
```

```
+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group 3    0.196  0.898
+      16
```

```
> #Fit and Error (Residual)
```

```
> summary(aov(Pull~Dose,data=Pull))
```

```
+              Df Sum Sq Mean Sq F value Pr(>F)
+ Dose          3    140    46.7    6.42 0.0046 **
+ Residuals    16    116     7.3
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(lm(Pull~Dose,data=Pull))
```

```
+ Analysis of Variance Table
+
+ Response: Pull
+              Df Sum Sq Mean Sq F value  Pr(>F)
+ Dose          3  140.1    46.70    6.423 0.00462 **
+ Residuals    16  116.3     7.27
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(lm(Pull~Dose,data=Pull))
```

```
+
+ Call:
+ lm(formula = Pull ~ Dose, data = Pull)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -3.20  -2.02  -0.33   1.45   5.30
+
+ Coefficients:
```

```

+           Estimate Std. Error t value Pr(>|t|)
+ (Intercept)    28.86      1.21   23.93  5.9e-14 ***
+ Dose1Unit      -3.82      1.71   -2.24  0.0396 *
+ Dose2Unit      -6.36      1.71   -3.73  0.0018 **
+ Dose3Unit      -6.56      1.71   -3.85  0.0014 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 2.7 on 16 degrees of freedom
+ Multiple R-squared:  0.546, Adjusted R-squared:  0.461
+ F-statistic: 6.42 on 3 and 16 DF,  p-value: 0.00462

```

```

> par(mar=c(4,4,0,2))
> gplots::plotmeans(Pull~Dose,data=Pull, ylim=range(15,35))

```

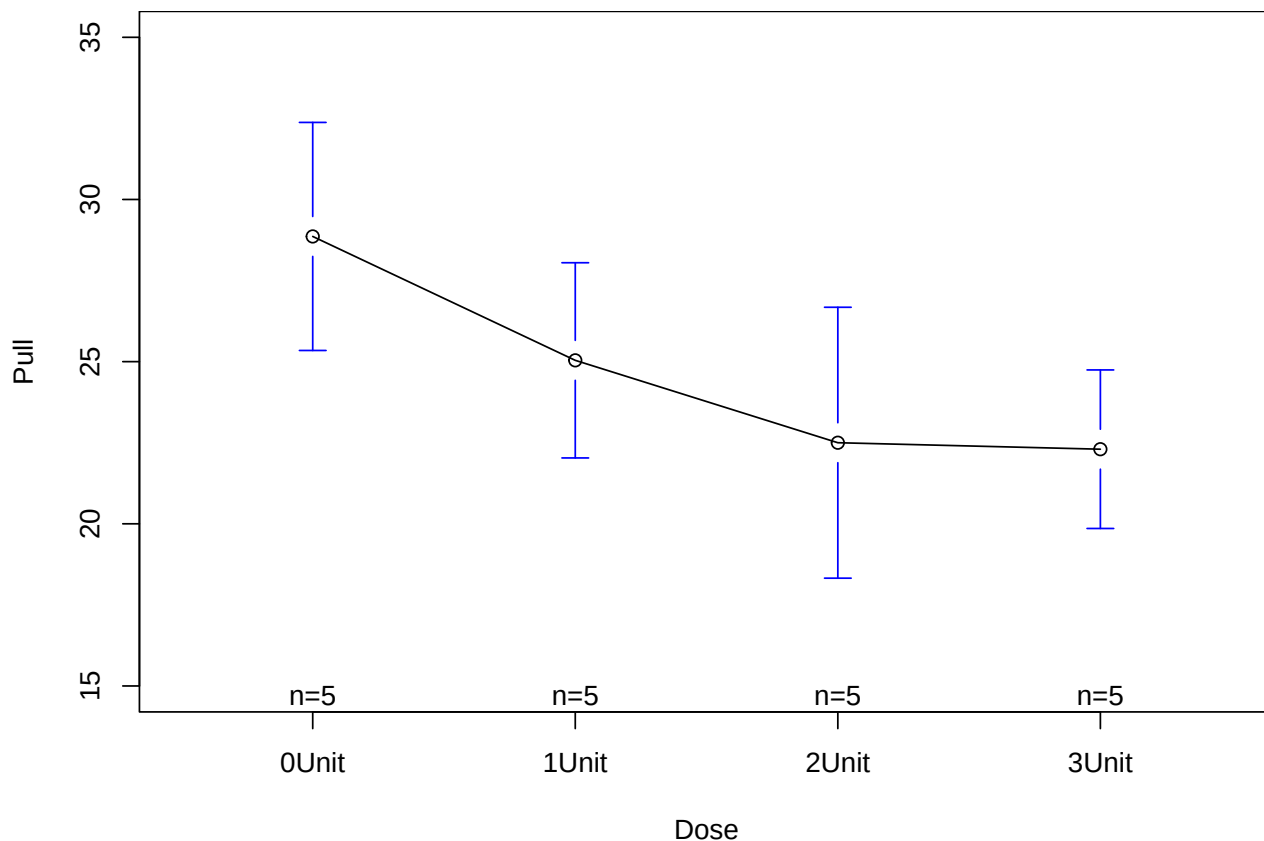


Figure 5: Means plot of pull strength by Dose level, with 95 % confidence intervals for means

```

> #To estimate the Dose-specific means
> summary(lm(Pull~-1+Dose,data=Pull))$coef
+           Estimate Std. Error t value    Pr(>|t|)
+ Dose0Unit    28.86    1.20584  23.9335 5.90944e-14
+ Dose1Unit    25.04    1.20584  20.7656 5.35775e-13
+ Dose2Unit    22.50    1.20584  18.6592 2.78178e-12
+ Dose3Unit    22.30    1.20584  18.4933 3.18983e-12

```

Post-hoc tests:

```
> summary(aov(Pull~Dose, data=Pull))

+               Df Sum Sq Mean Sq F value Pr(>F)
+ Dose           3    140    46.7    6.42 0.0046 **
+ Residuals     16    116     7.3
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> pairwise.t.test(Pull$Pull, Pull$Dose, p.adj = "none")

+
+ Pairwise comparisons using t tests with pooled SD
+
+ data:  Pull$Pull and Pull$Dose
+
+      0Unit 1Unit 2Unit
+ 1Unit 0.040 -      -
+ 2Unit 0.002 0.156 -
+ 3Unit 0.001 0.128 0.908
+
+ P value adjustment method: none

> pairwise.t.test(Pull$Pull, Pull$Dose, p.adj = "bonferroni")

+
+ Pairwise comparisons using t tests with pooled SD
+
+ data:  Pull$Pull and Pull$Dose
+
+      0Unit 1Unit 2Unit
+ 1Unit 0.238 -      -
+ 2Unit 0.011 0.935 -
+ 3Unit 0.009 0.766 1.000
+
+ P value adjustment method: bonferroni

> TukeyHSD(aov(Pull~Dose, data=Pull))

+ Tukey multiple comparisons of means
+ 95% family-wise confidence level
+
+ Fit: aov(formula = Pull ~ Dose, data = Pull)
+
+ $Dose
+      diff      lwr      upr      p adj
+ 1Unit-0Unit -3.82 -8.69894  1.05894 0.154638
+ 2Unit-0Unit -6.36 -11.23894 -1.48106 0.008884
+ 3Unit-0Unit -6.56 -11.43894 -1.68106 0.006999
+ 2Unit-1Unit -2.54 -7.41894  2.33894 0.466156
+ 3Unit-1Unit -2.74 -7.61894  2.13894 0.402693
+ 3Unit-2Unit -0.20 -5.07894  4.67894 0.999398
```

EXAMPLE Gut permeability by AIDS diagnosis: data set AIDS

```
> par(mar=c(4, 4, 2, 2))
> boxplot(mann~diag, data=AIDS, xlab="Diagnosis", ylab="Mannitol Level")
```

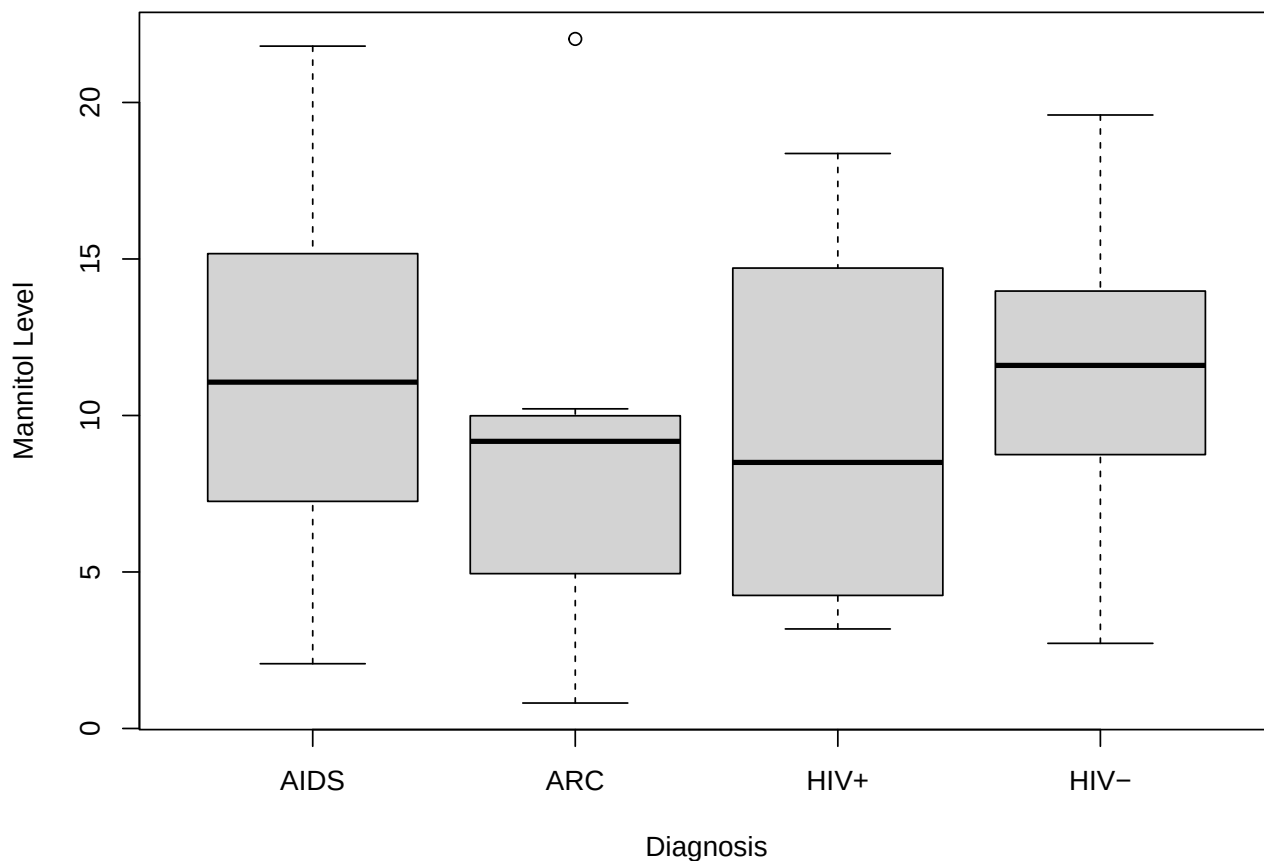


Figure 6: Boxplot of Mannitol level by diagnosis

```
> car::leveneTest(mann~diag, center='mean', data=AIDS)
+ Levene's Test for Homogeneity of Variance (center = "mean")
+   Df F value Pr(>F)
+ group 3    0.866  0.464
+   55

> car::leveneTest(mann~diag, center='median', data=AIDS)
+ Levene's Test for Homogeneity of Variance (center = "median")
+   Df F value Pr(>F)
+ group 3    0.733  0.537
+   55

> summary(aov(mann~diag, data=AIDS))
+           Df Sum Sq Mean Sq F value Pr(>F)
+ diag         3     49    16.3      0.6   0.62
+ Residuals   55    1510    27.4
```

```

> summary(lm(mann~diag,data=AIDS))
+
+ Call:
+ lm(formula = mann ~ diag, data = AIDS)
+
+ Residuals:
+     Min       1Q   Median       3Q      Max
+  -9.262  -3.992   0.169   3.336  13.188
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  11.3312     1.0276   11.03  1.5e-15 ***
+ diagARC      -2.4893     2.2312   -1.12    0.27
+ diagHIV+     -1.6208     2.2312   -0.73    0.47
+ diagHIV-      0.0658     1.5814    0.04    0.97
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 5.24 on 55 degrees of freedom
+ Multiple R-squared:  0.0314, Adjusted R-squared:  -0.0214
+ F-statistic: 0.595 on 3 and 55 DF,  p-value: 0.621

```

```

> par(mar=c(4,4,0,2))
> ggplots::plotmeans(mann~diag,data=AIDS,ylim=range(0,18))

```

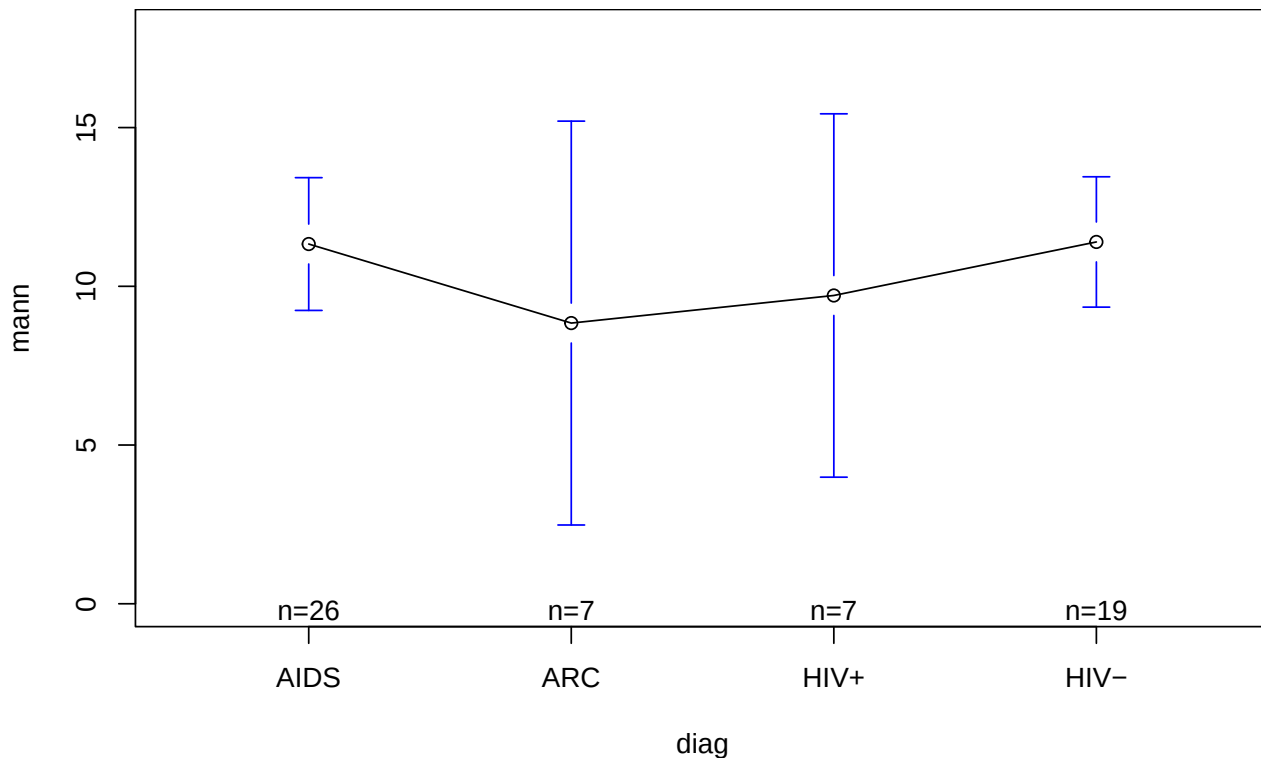


Figure 7: Means plot of gut permeability of mannitol by diagnosis, with 95 % confidence intervals for means

EXAMPLE Bacteria level in batches of milk: data set MilkBac

```
> car::leveneTest(BacLevel~Batch,center='mean',data=MilkBac)
+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value Pr(>F)
+ group  4    0.384  0.818
+      25

> car::leveneTest(BacLevel~Batch,center='median',data=MilkBac)
+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group  4    0.243  0.911
+      25

> summary(aov(BacLevel~Batch,data=MilkBac))
+      Df Sum Sq Mean Sq F value   Pr(>F)
+ Batch      4     803   200.8     9.01 0.00012 ***
+ Residuals  25     557    22.3
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(lm(BacLevel~Batch,data=MilkBac))
+
+ Call:
+ lm(formula = BacLevel ~ Batch, data = MilkBac)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -8.83  -2.54   0.25   2.54   9.17
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)    23.83      1.93    12.37 3.8e-12 ***
+ Batch2        -10.50      2.73    -3.85 0.00072 ***
+ Batch3        -12.17      2.73    -4.46 0.00015 ***
+ Batch4        -14.67      2.73    -5.38 1.4e-05 ***
+ Batch5         -6.00      2.73    -2.20 0.03716 *
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 4.72 on 25 degrees of freedom
+ Multiple R-squared:  0.59, Adjusted R-squared:  0.525
+ F-statistic: 9.01 on 4 and 25 DF, p-value: 0.00012

> par(mar=c(4,4,0,2))
> gplots::plotmeans(BacLevel~Batch,data=MilkBac, ylim=range(0,30))
```

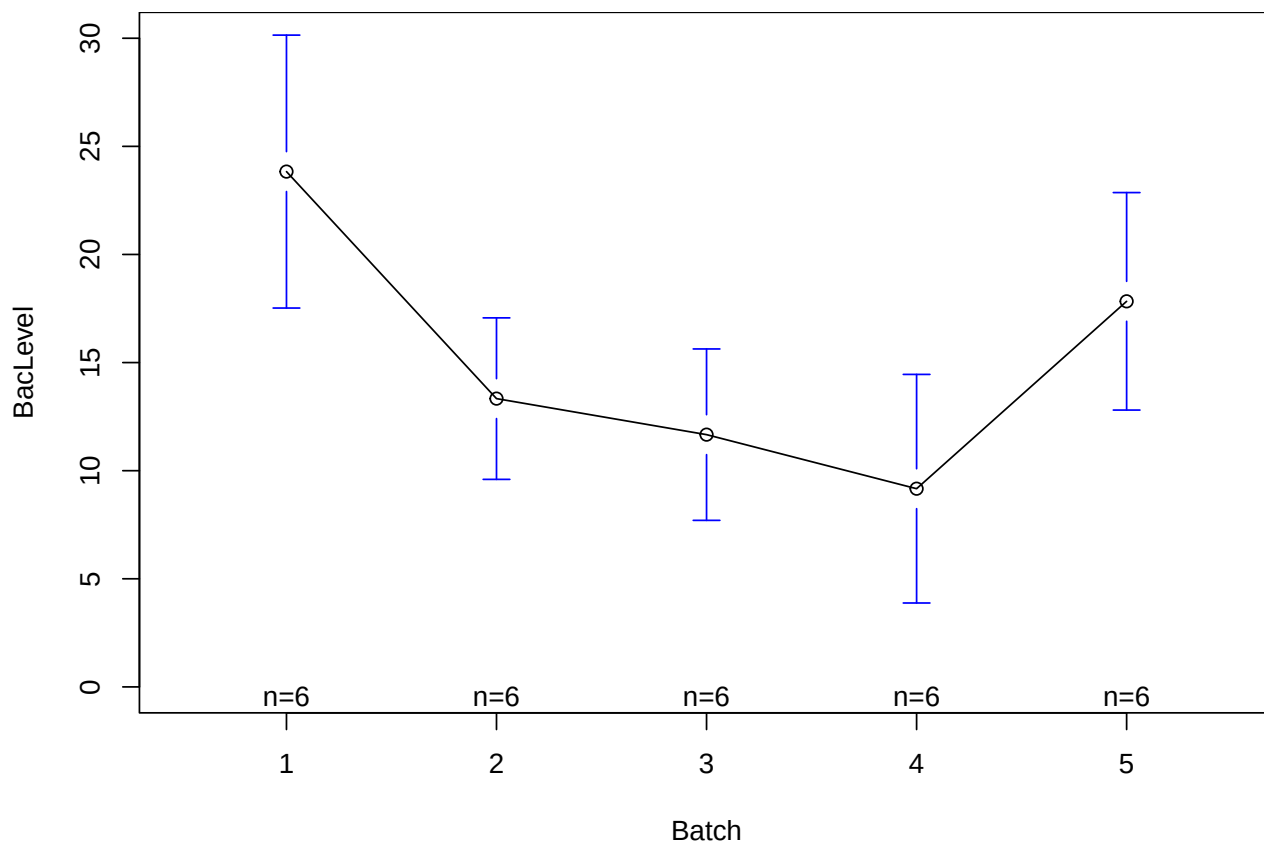


Figure 8: Means plot of bacteria level by batch, with 95 % confidence intervals for means

```
> summary(lm(BacLevel ~ 1 + Batch, data=MilkBac))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
+ Batch1	23.83333	1.92729	12.36625	3.76038e-12
+ Batch2	13.33333	1.92729	6.91818	2.98452e-07
+ Batch3	11.66667	1.92729	6.05341	2.52240e-06
+ Batch4	9.16667	1.92729	4.75625	7.00794e-05
+ Batch5	17.83333	1.92729	9.25306	1.49983e-09

EXAMPLE Post-traumatic stress disorder: data set PTSD

Data Source: Foa, E. B., Rothbaum, B. O., Riggs, D. S., & Murdock, T. B. (1991) Treatment of post traumatic stress disorder in rape victims: A comparison between cognitive-behavioral procedures and counseling. *Journal of Consulting and Clinical Psychology*, 59, 715-723.

```
> par(mar=c(4, 4, 2, 2))
> boxplot(PTSD~GROUP, data=PTSD, xlab="Diagnosis", ylab="Mannitol Level")
```

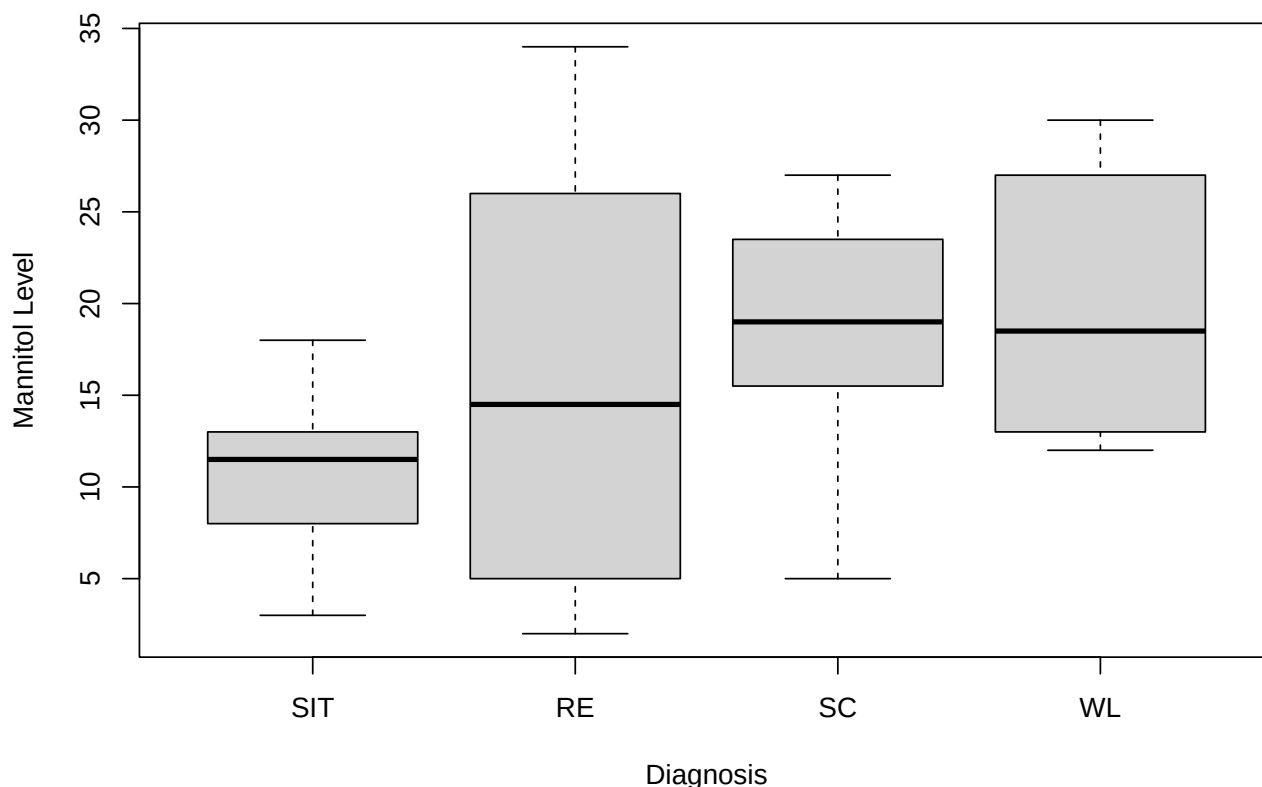


Figure 9: Boxplot of PTSD score by therapy

```
> car::leveneTest(PTSD~GROUP, center='mean', data=PTSD)
+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value    Pr(>F)
+ group  3    6.633 0.000934 ***
+      41
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> car::leveneTest(PTSD~GROUP, center='median', data=PTSD)
+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value    Pr(>F)
+ group  3    6.171 0.00147 **
+      41
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

> summary(aov(PTSD~GROUP,data=PTSD))

+           Df Sum Sq Mean Sq F value Pr(>F)
+ GROUP          3      508   169.3    3.05  0.039 *
+ Residuals    41     2279    55.6
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(lm(PTSD~GROUP,data=PTSD))

+
+ Call:
+ lm(formula = PTSD ~ GROUP, data = PTSD)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -13.40  -4.40   0.50   4.93  18.60
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)    11.07      1.99     5.56 1.8e-06 ***
+ GROUPRE         4.33      3.09     1.40  0.1683
+ GROUpsc         7.02      3.00     2.34  0.0244 *
+ GROUPL         8.43      3.09     2.73  0.0093 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 7.46 on 41 degrees of freedom
+ Multiple R-squared:  0.182, Adjusted R-squared:  0.122
+ F-statistic: 3.05 on 3 and 41 DF, p-value: 0.0394

> par(mar=c(4,4,1,2))
> ggplot2::plotmeans(PTSD~GROUP,data=PTSD, ylim=range(0,30))

```

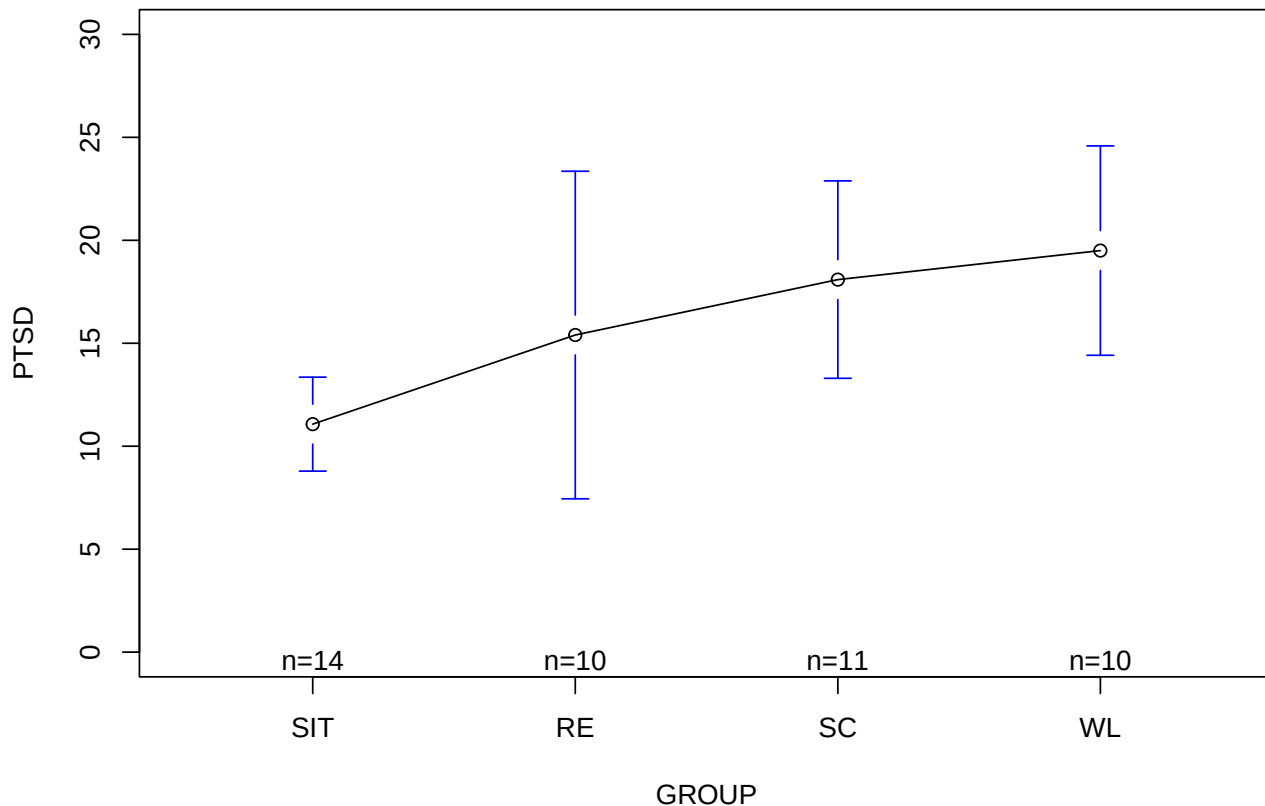


Figure 10: Means plot of PTSD score by therapy, with 95 % confidence intervals for means

```
> summary(lm(PTSD~1+GROUP, data=PTSD))$coef
```

```
+      Estimate Std. Error t value    Pr(>|t|)
+ GROUPSIT    11.0714    1.99247  5.55664 1.84459e-06
+ GROUPRE     15.4000    2.35752  6.53229 7.57159e-08
+ GROUPSC     18.0909    2.24781  8.04825 5.69479e-10
+ GROUPWL     19.5000    2.35752  8.27141 2.81770e-10
```

```
> summary(aov(PTSD~GROUP, data=PTSD))
```

```
+      Df Sum Sq Mean Sq F value Pr(>F)
+ GROUP      3      508    169.3    3.05  0.039 *
+ Residuals  41     2279     55.6
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> pairwise.t.test(PTSD$PTSD, PTSD$GROUP, p.adj = "none")
```

```
+
+ Pairwise comparisons using t tests with pooled SD
+
+ data: PTSD$PTSD and PTSD$GROUP
+
+      SIT  RE  SC
+ RE 0.168 -   -
+ SC 0.024 0.414 -
```

```

+ WL 0.009 0.226 0.668
+
+ P value adjustment method: none

> pairwise.t.test(PTSD$PTSD, PTSD$GROUP, p.adj = "bonferroni")

+
+ Pairwise comparisons using t tests with pooled SD
+
+ data: PTSD$PTSD and PTSD$GROUP
+
+ SIT RE SC
+ RE 1.00 - -
+ SC 0.15 1.00 -
+ WL 0.06 1.00 1.00
+
+ P value adjustment method: bonferroni

```

```

> TukeyHSD(aov(PTSD~GROUP, data=PTSD))

+ Tukey multiple comparisons of means
+ 95% family-wise confidence level
+
+ Fit: aov(formula = PTSD ~ GROUP, data = PTSD)
+
+ $GROUP
+ diff lwr upr p adj
+ RE-SIT 4.32857 -3.936483 12.5936 0.505162
+ SC-SIT 7.01948 -1.023440 15.0624 0.106248
+ WL-SIT 8.42857 0.163517 16.6936 0.044145
+ SC-RE 2.69091 -6.031118 11.4129 0.841777
+ WL-RE 4.10000 -4.827279 13.0273 0.611888
+ WL-SC 1.40909 -7.312936 10.1311 0.972524

```

RANDOMIZED BLOCK DESIGNS AND THE ANOVA F-TEST

Consider a **randomized block design** (RBD) with k treatments and b blocks. Assume that each block has k experimental units, and that one unit is assigned to each treatment. Let x_{ij} be the measured response for the experimental unit from block j in treatment i and

- sample mean for **treatment** i

$$\bar{x}_i = \frac{1}{b} \sum_{j=1}^b x_{ij} \quad i = 1, \dots, k$$

- sample mean for block j

$$\bar{x}_j^{(B)} = \frac{1}{k} \sum_{i=1}^k x_{ij} \quad j = 1, \dots, b$$

- overall sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^b x_{ij}$$

- Sum of Squares for Treatments (SST)

$$SST = \sum_{i=1}^k b(\bar{x}_i - \bar{x})^2$$

- Sum of Squares for Blocks (SSB)

$$SSB = \sum_{j=1}^b k(\bar{x}_j^{(B)} - \bar{x})^2$$

- Overall Sum of Squares (SS)

$$SS = \sum_{i=1}^k \sum_{j=1}^b (x_{ij} - \bar{x})^2$$

The following decomposition holds

$$SS = SST + SSB + SSE \quad \therefore \quad SSE = SS - SST - SSB$$

For testing

$$H_0 : \mu_1 = \dots = \mu_k$$

$$H_a : \text{At least two treatment means different}$$

in an RBD, the test statistic is

$$F = \frac{MST}{MSE}$$

where

$$MST = \frac{SST}{k-1} \quad MSE = \frac{SSE}{n-b-k+1}$$

If H_0 is **true**, then $F \sim F_{k-1, n-b-k+1}$, and the rejection region for the test with significance level α is

$$F > F_{\alpha, k-1, n-b-k+1}$$

where F_{α, ν_1, ν_2} is the upper α point of the F-distribution with ν_1 and ν_2 degrees of freedom.

Example: Measurements were made on the amount of sulphur (in parts per million) in soil samples using four different solvents. The soil samples were collected from five different geographical locations in Florida, USA, and represented different soil types.

The **response variable** sulphur level. The single **factor** is the *solvent* and there are $k = 4$ factor levels:

1. Calcium Chloride (CaCl_2)
2. Ammonium Acetate (NH_4OAc)
3. Mono-Calcium Phosphate ($\text{Ca}(\text{H}_2\text{P O}_4)_3$)
4. Water (H_2O)

The *soil* types determine the $b = 5$ **blocks**

1. Troup, Jackson Co. (*Paleudults* soil)
2. Lakeland, Walton Co. (*Quartzipsamments* soil)
3. Leon, Duval Co. (*Haplaquads* soil)
4. Chipley, Jackson Co. (*Quartzipsamments* soil)
5. Norfolk, Alachua Co. (*Paleudults* soil)

The data observed in the study were as follows:

Treatment	Block				
	Troup	Lakeland	Leon	Chipley	Norfolk
CaCl_2	5.07	3.31	2.54	2.34	4.71
NH_4OAc	4.43	2.74	2.09	2.07	5.29
$\text{Ca}(\text{H}_2\text{P O}_4)_3$	7.09	2.32	1.09	4.38	5.70
H_2O	4.48	2.35	2.70	3.85	4.98

```
> print(summary(aov(sulphur~solvent+soil,data=Soil)), digits = 5)

+               Df Sum Sq Mean Sq F value    Pr(>F)
+ solvent         3   1.621   0.5404   0.6726 0.5851298
+ soil           4  33.965   8.4912  10.5683 0.0006629 ***
+ Residuals     12   9.642   0.8035
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The this is the standard ANOVA table for the randomized block design. As expected, there is a significant difference between **blocks** ($F_{obs} = 10.568$, p-value=0.001), but **no significant difference** between **treatments** ($F_{obs} = 0.673$, p-value=0.585).

THE NEED FOR BLOCKING IN AN RBD ANALYSIS

Consider the following response data: five measurements collected in three treatment groups:

	Block				
—	1	2	3	4	5
Group 1	-20.88	-4.76	-0.46	10.78	-10.47
Group 2	-15.75	0.11	5.64	16.98	-6.03
Group 3	-8.62	6.20	10.42	20.05	-1.29

Raw data broken down by block

```
> Y<-c(-20.88,-4.76,-0.46,10.78,-10.47,
+      -15.75,0.11,5.64,16.98,-6.03,
+      -8.62,6.20,10.42,20.05,-1.29)
>
> Group<-gl(3,5,labels=c("Group1","Group2","Group3"))
> Block<-gl(5,1,15,labels=c(1:5))
> par(mar=c(4,4,1,2))
> gplots::plotmeans(Y~Group,ylim=range(-40,40))
> stripchart(Y~Group,vertical=T,pch=4,cex=0.75,lwd=2,add=T)
```

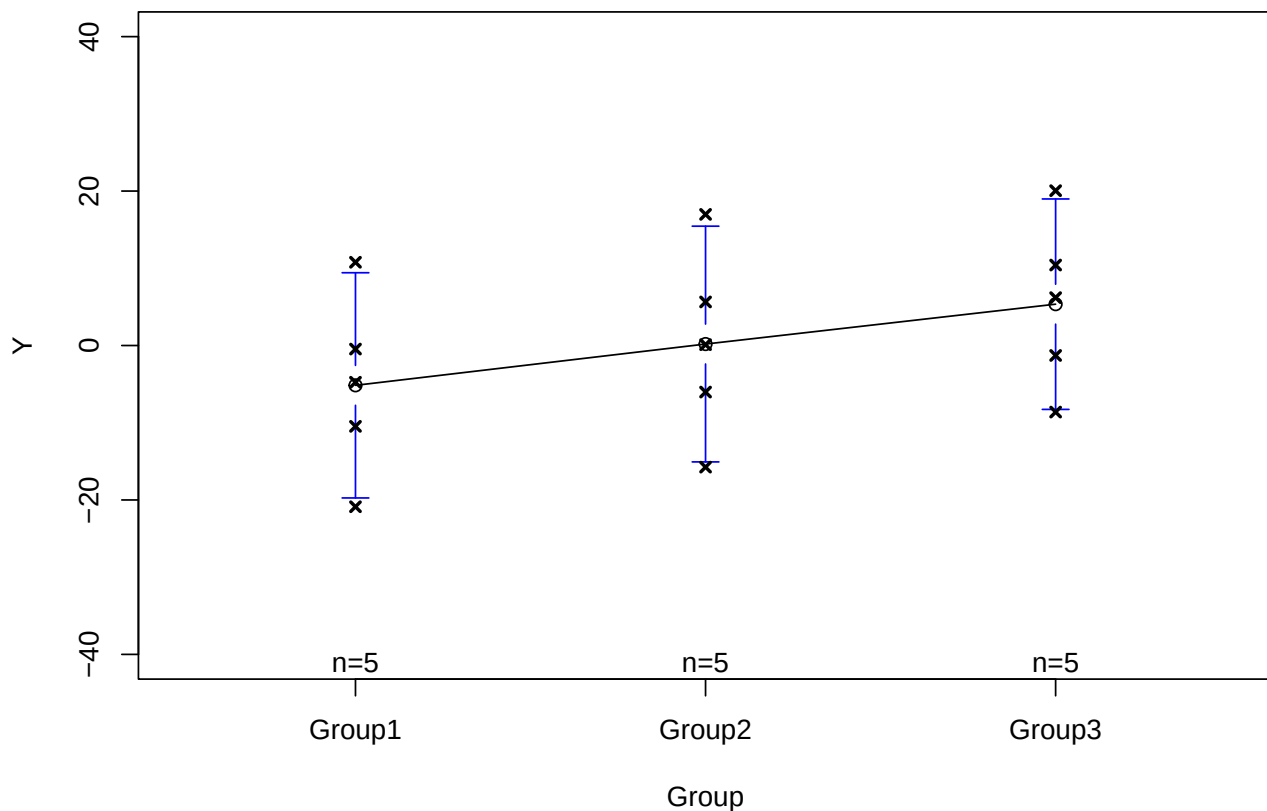


Figure 11: Raw data with block ignored

```
> print(summary(aov(Y~Group)), digits = 5)

+           Df  Sum Sq Mean Sq F value Pr(>F)
+ Group       2   276.18   138.09   1.0118 0.3926
+ Residuals   12  1637.80   136.48
```

It seems from this analysis that there is no difference between groups; however, this is not the case if the block structure is taken into account.

```
> blocklabel<-paste(rep("Block", 5), c(1:5))
> par(mar=c(4, 4, 1, 2))
> gplots::plotmeans(Y~Group, ylim=range(-40, 40))
> points(rep(0.9, 5), Y[Group=="Group1"], pch=c(1:5))
> points(rep(1.9, 5), Y[Group=="Group2"], pch=c(1:5))
> points(rep(2.9, 5), Y[Group=="Group3"], pch=c(1:5))
> legend(0.5, 40, blocklabel, pch=(1:5))
```

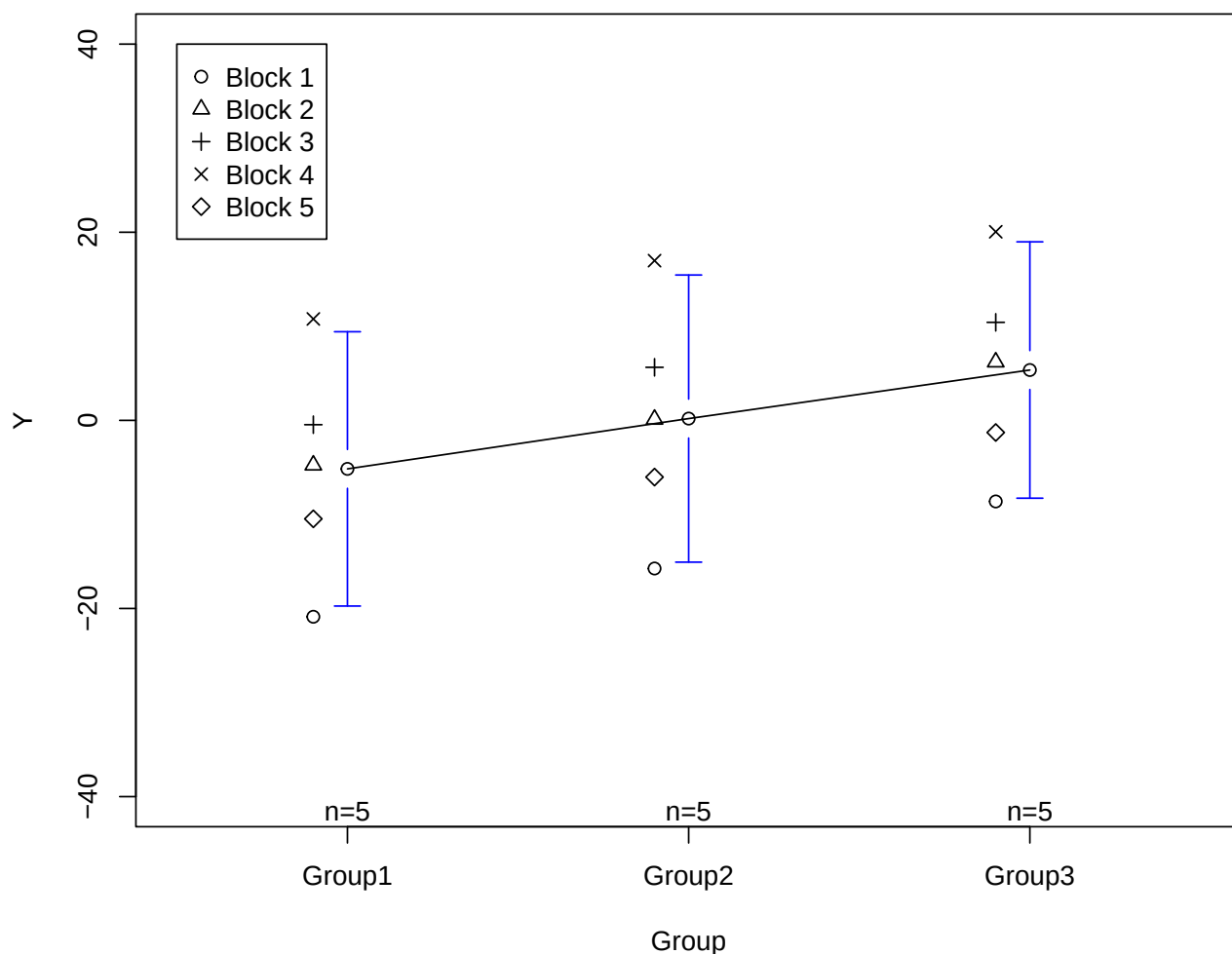


Figure 12: Raw data broken down by block

```
> print(summary(aov(Y~Group+Block)), digits = 5)

+           Df   Sum Sq Mean Sq F value    Pr(>F)
+ Group         2    276.18   138.09   178.81 2.292e-07 ***
+ Block         4   1631.62   407.91   528.18 1.009e-09 ***
+ Residuals     8     6.18     0.77
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In fact, there is hidden structure in the data. If this structure is taken into account, evidence that the treatment means are significantly different is uncovered. The reason that the CRD and one-way ANOVA do not discover this is that they assume that the variability can be decomposed as

$$SS = SST + SSE$$

whereas in fact

$$SS = SST + SSB + SSE$$

that is, the CRD assumes that the random variability that is observed is MUCH LARGER than it actually is. Once the variation due to BLOCKS is taken into account, the ANOVA-F test result for TREATMENTS becomes significant.

RANDOMIZED COMPLETE BLOCK DESIGNS WITH BALANCED REPLICATION

Consider a **randomized block design** (RBD) with k treatments and b blocks, and r **replications**, giving $n = rbk$ observations in total. Let x_{ijt} be the t th replicated observation in the (i, j) th treatment/block combination.

- sample mean for **treatment** i

$$\bar{x}_i = \frac{1}{br} \sum_{j=1}^b \sum_{t=1}^r x_{ijt} \quad i = 1, \dots, k$$

- sample mean for **block** j

$$\bar{x}_j^{(B)} = \frac{1}{kr} \sum_{i=1}^k \sum_{t=1}^r x_{ijt} \quad j = 1, \dots, b$$

- sample mean for replicates in (i, j) th **treatment/block** combination

$$\bar{x}_{ij} = \frac{1}{r} \sum_{t=1}^r x_{ijt} \quad i = 1, \dots, k, j = 1, \dots, b$$

- overall sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^b \sum_{t=1}^r x_{ijt}$$

- Sum of Squares for Treatments (SST)

$$SST = \sum_{i=1}^k br(\bar{x}_i - \bar{x})^2$$

- Sum of Squares for Blocks (SSB)

$$SSB = \sum_{j=1}^b kr(\bar{x}_j^{(B)} - \bar{x})^2$$

- Sum of Squares for Interaction (SSI)

$$SSI = \sum_{i=1}^k \sum_{j=1}^b r(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j^{(B)} + \bar{x})^2$$

- Overall Sum of Squares (SS)

$$SS = \sum_{i=1}^k \sum_{j=1}^b \sum_{t=1}^r (x_{ijt} - \bar{x})^2$$

The following decomposition holds

$$SS = SST + SSB + SSI + SSE \quad \therefore \quad SSE = SS - SST - SSB - SSI$$

Define

$$MST = \frac{SST}{k-1} \quad MSB = \frac{SSB}{b-1} \quad MSI = \frac{SSI}{(k-1)(b-1)}$$

and

$$MSE = \frac{SSE}{n - bk}$$

- For testing for a **TREATMENT** effect, use

$$F = \frac{MST}{MSE}$$

Under the assumption of **NO TREATMENT EFFECT**, then

$$F \sim F_{k-1, n-bk}$$

which defines the rejection region and p-value in the usual way.

- For testing for a **BLOCK** effect, use

$$F = \frac{MSB}{MSE}$$

Under the assumption of **NO BLOCK EFFECT**, then

$$F \sim F_{b-1, n-bk}$$

- For testing for an **INTERACTION**, use

$$F = \frac{MSI}{MSE}$$

Under the assumption of **NO INTERACTION**, then

$$F \sim F_{(k-1)(b-1), n-bk}$$

Example: Measurements were made on the lifetimes of batteries (in hours) for three battery types constructed from different materials, to investigate the effect of operating temperature on lifetime. It was believed before the experiment that the battery types were likely to behave differently in the experiment.

The **response variable** is lifetime. The single **factor** is the *temperature* and there are $k = 3$ factor levels:

1. 15 Celsius
2. 70 Celsius
3. 125 Celsius

Treatment	Block		
	Lead	Acetate	Nickel Cadmium
15	130,155,74,180	150,188,159,126	138,119,168,160
70	34,40,80,75	126,122,106,115	174,120,150,139
120	20,70,82,58	25,70,58,45	96,104,82,60

The *material* types determine the $b = 3$ **blocks**

1. Lead
2. Acetate
3. Nickel Cadmium

$r = 4$ replicate measurements were made, so that $n = 3 \times 3 \times 4 = 36$ data were obtained in total.

The data observed in the study were as follows:

```
> print(summary(aov(life~temp*material,data=Batteries)), digits = 5)

+               Df Sum Sq Mean Sq F value    Pr(>F)
+ temp              2   39083  19541.6  29.3438 1.694e-07 ***
+ material           2   10633   5316.6   7.9834 0.001888 **
+ temp:material      4    9438   2359.4   3.5429 0.018973 *
+ Residuals         27   17981    666.0
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> car::leveneTest(life~temp*material,data=Batteries,center='mean')

+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value Pr(>F)
+ group 8    1.059  0.42
+      27

> car::leveneTest(life~temp*material,data=Batteries,center='median')

+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group 8    0.936  0.504
+      27
```

There is a

- **significant difference between blocks** (material, $F_{obs} = 7.983$, p-value 0.00189),
- **significant difference between treatments** (temp, $F_{obs} = 29.344$, p-value $1.69e - 07$),
- **significant interaction** (temp:material, $F_{obs} = 3.543$, p-value 0.01897).

Levene's test reveals that there is no evidence to suspect that the population variances are different:

```

> par(mar=c(4,4,1,2))
> gplots::plotmeans(life~temp, subset=(material=='Lead'),
+   n.label=FALSE, col='blue', barcol='blue', data=Batteries,
+   ylim=range(0,250))
> suppressWarnings(gplots::plotmeans(life~temp,
+   subset=(material=='Acetate'), n.label=FALSE,
+   col='orange', barcol='orange', data=Batteries, add=TRUE))
> suppressWarnings(gplots::plotmeans(life~temp,
+   subset=(material=='Nickel Cadmium'), n.label=FALSE,
+   col='cyan', barcol='cyan', data=Batteries, add=TRUE))
> legend(2.5,250,c('Lead','Acetate','Nickel Cadmium'),
+   col=c('blue','cyan','orange'), lty=1)

```

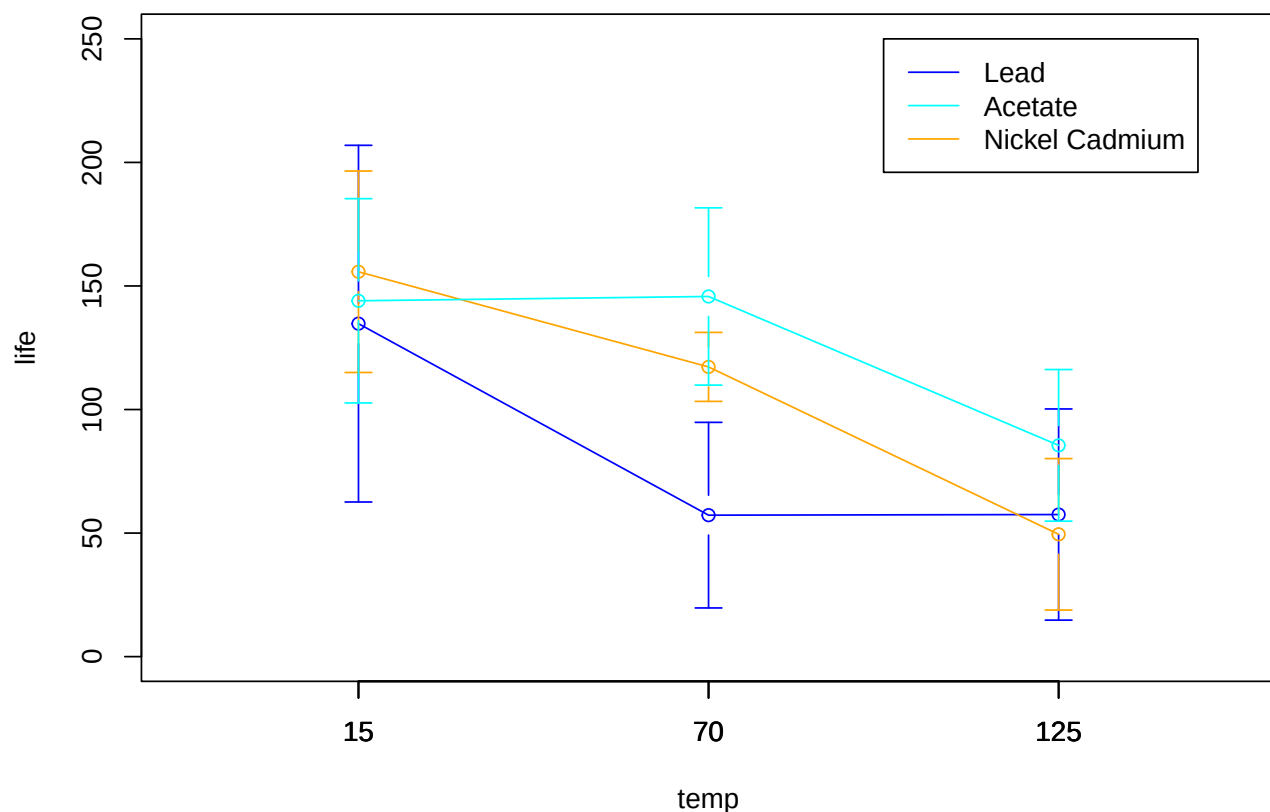


Figure 13: Means Plot by Temperature

```

> par(mar=c(4,4,1,2))
> gplots::plotmeans(life~material, subset=(temp=='15'),
+   n.label=FALSE,
+   col='blue', barcol='blue', data=Batteries,ylim=range(0,250))
> suppressWarnings(
+   gplots::plotmeans(life~material,subset=(temp=='70'),
+   n.label=FALSE, col='orange', barcol='orange', data=Batteries,
+   add=TRUE))
> suppressWarnings(
+   gplots::plotmeans(life~material, subset=(temp=='125'),

```

```

+   n.label=FALSE, col='cyan', barcol='cyan', data=Batteries,
+   add=TRUE) )
> legend(2.5, 250, c('15', '70', '125'),
+   col=c('blue', 'cyan', 'orange'), lty=1)

```

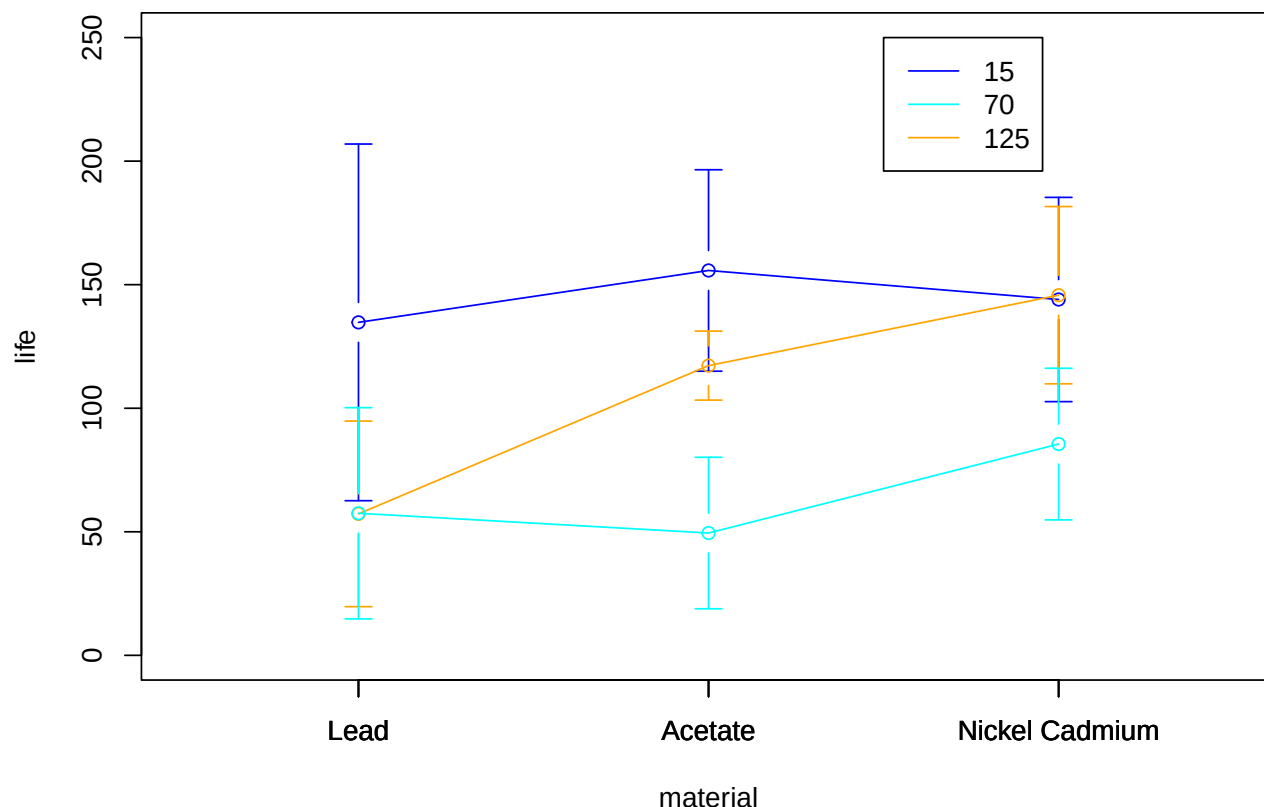


Figure 14: Means Plot by Material

To produce a means plot, we also use the full ggplot functionality. First we compute a by-treatment summary using the `summarySE` function from the `Rmisc` library.

```

> par(mar=c(4, 4, 1, 2))
> batsum <- Rmisc::summarySE(Batteries, measurevar="life",
+   groupvars=c("temp", "material"))
> batsum

```

	temp	material	N	life	sd	se	ci
1	15	Lead	4	134.75	45.35324	22.67662	72.1671
2	15	Acetate	4	155.75	25.61738	12.80869	40.7630
3	15	Nickel Cadmium	4	144.00	25.97435	12.98717	41.3310
4	70	Lead	4	57.25	23.59908	11.79954	37.5514
5	70	Acetate	4	117.25	8.77021	4.38511	13.9554
6	70	Nickel Cadmium	4	145.75	22.54440	11.27220	35.8732
7	125	Lead	4	57.50	26.85144	13.42572	42.7266
8	125	Acetate	4	49.50	19.26136	9.63068	30.6491
9	125	Nickel Cadmium	4	85.50	19.27866	9.63933	30.6766

```

> pd <- position_dodge(0.2)
> ggplot(batsum,
+   aes(x=temp, y=life, colour=material, group=material)) +
+   geom_errorbar(aes(ymin=life-ci, ymax=life+ci),
+     width=0.25, position=pd) +
+   geom_line(position=pd) +
+   geom_point(position=pd, size=3) +
+   theme(legend.position="bottom")

```

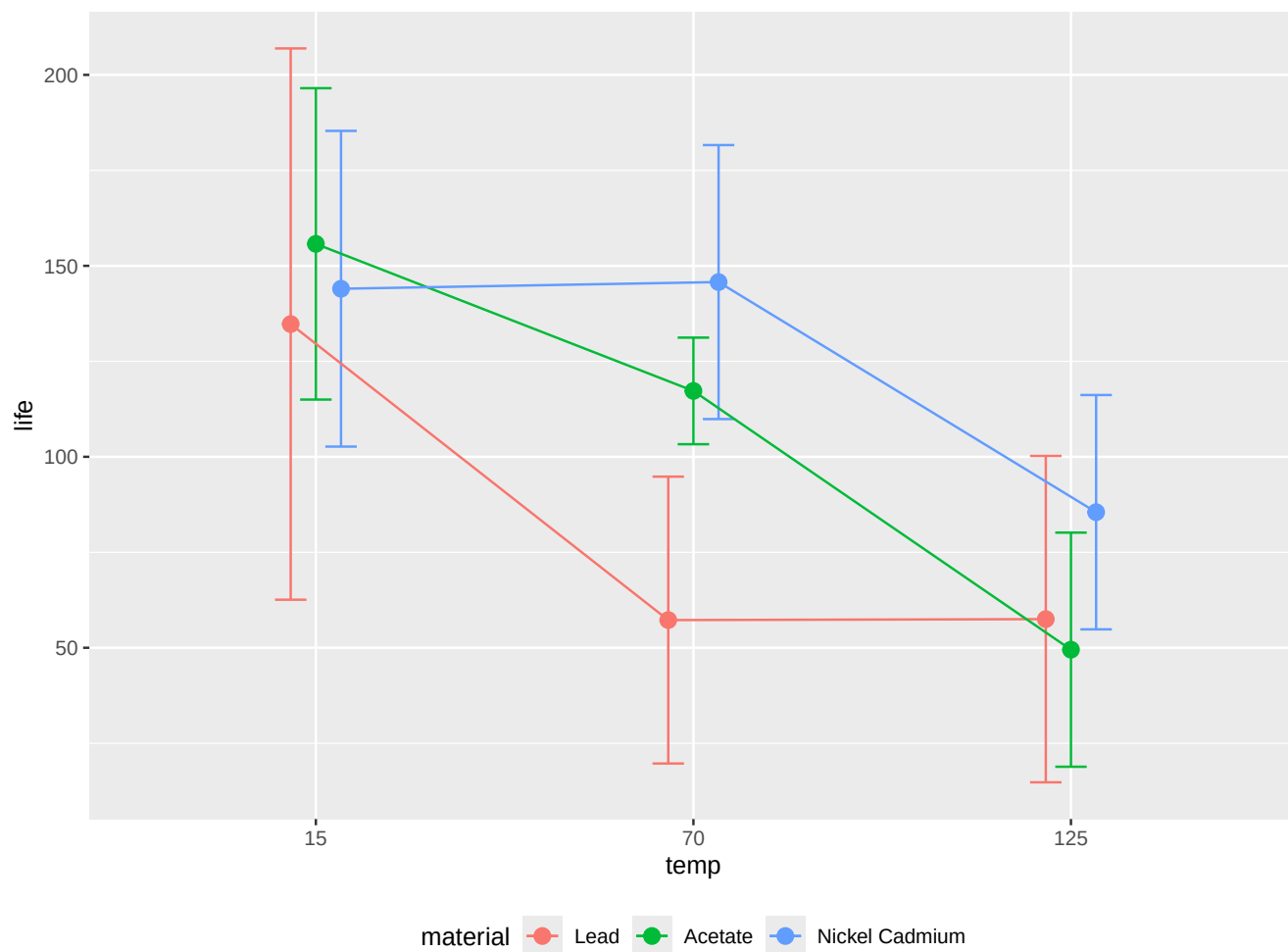


Figure 15: Means Plot for Batteries data

```

> pd <- position_dodge(0.2)
> ggplot(batsum,
+   aes(x=material, y=life, colour=temp, group=temp)) +
+   geom_errorbar(aes(ymin=life-ci, ymax=life+ci),
+     width=0.25, position=pd) +
+   geom_line(position=pd) +
+   geom_point(position=pd, size=3) +
+   theme(legend.position="bottom")

```

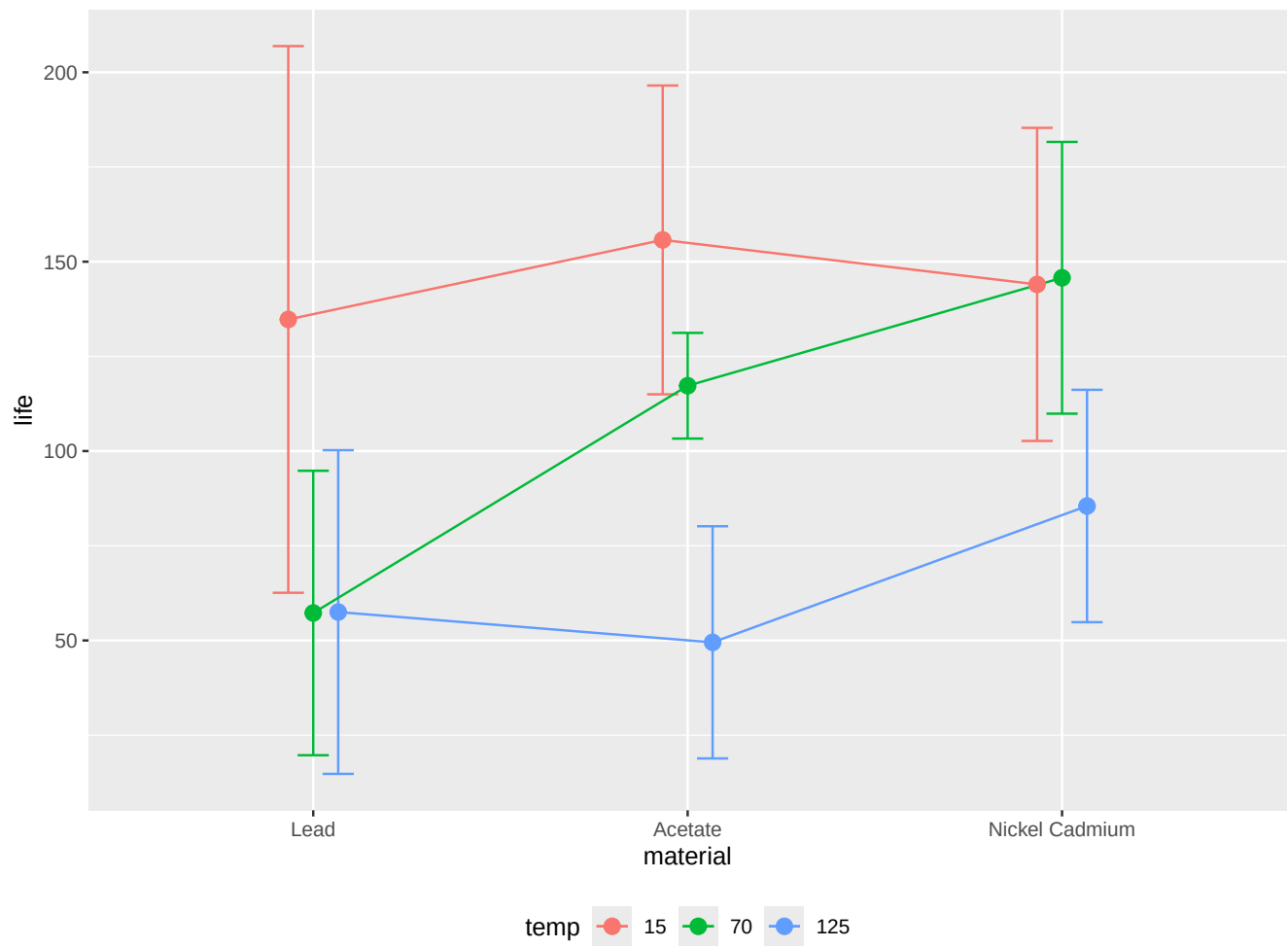


Figure 16: Means Plot for Batteries data

BALANCED COMPLETE FACTORIAL DESIGNS

Consider a **factorial design** (FD) with two factors A and B, with levels $1, \dots, a$ and $1, \dots, b$ respectively, yielding a total of $k = ab$ factor combinations (treatments), and suppose that there are r **replications** in each treatment, giving $n = rab$ observations in total. Let x_{ijt} be the t th replicated observation in the (i, j) th factor-level combination.

- sample mean for Factor A level i

$$\bar{x}_{i.} = \frac{1}{br} \sum_{j=1}^b \sum_{t=1}^r x_{ijt} \quad i = 1, \dots, a$$

- sample mean for Factor B level j

$$\bar{x}_{.j} = \frac{1}{ar} \sum_{i=1}^a \sum_{t=1}^r x_{ijt} \quad j = 1, \dots, b$$

- sample mean for replicates in (i, j) th factor combination

$$\bar{x}_{ij} = \frac{1}{r} \sum_{t=1}^r x_{ijt} \quad i = 1, \dots, a, j = 1, \dots, b$$

- overall sample mean

$$\bar{x}_{..} = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \sum_{t=1}^r x_{ijt}$$

- Sum of Squares for Treatments due to factor A (SST_A)

$$SST_A = \sum_{i=1}^a br(\bar{x}_{i.} - \bar{x}_{..})^2$$

- Sum of Squares for Treatments due to factor B (SST_B)

$$SST_B = \sum_{j=1}^b ar(\bar{x}_{.j} - \bar{x}_{..})^2$$

- Sum of Squares for Interaction (SSI_{AB})

$$SSI_{AB} = \sum_{i=1}^a \sum_{j=1}^b r(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

- Overall Sum of Squares (SS)

$$SS = \sum_{i=1}^a \sum_{j=1}^b \sum_{t=1}^r (x_{ijt} - \bar{x}_{..})^2$$

The following decomposition holds

$$SS = SST_A + SST_B + SSI_{AB} + SSE \quad \therefore \quad SSE = SS - SST_A - SST_B - SSI_{AB}$$

Define

$$MST_A = \frac{SST_A}{a-1} \quad MST_B = \frac{SST_B}{b-1} \quad MSI_{AB} = \frac{SSI_{AB}}{(a-1)(b-1)}$$

and

$$MSE = \frac{SSE}{n-ab}$$

- For testing for a **FACTOR A** effect, use

$$F = \frac{MST_A}{MSE}$$

Under the assumption of **NO FACTOR A EFFECT**, then

$$F \sim F_{a-1, n-ab}$$

which defines the rejection region and p-value in the usual way.

- For testing for a **FACTOR B** effect, use

$$F = \frac{MST_B}{MSE}$$

Under the assumption of **NO FACTOR B EFFECT**, then

$$F \sim F_{b-1, n-ab}$$

- For testing for an **INTERACTION**, use

$$F = \frac{MSI_{AB}}{MSE}$$

Under the assumption of **NO INTERACTION**, then

$$F \sim F_{(a-1)(b-1), n-ab}$$

Note: The only difference between a randomized block design and a factorial design is that in the block design, one of the factors is known or strongly believed to have a significant effect on the response. The method of analysis for interaction and no interaction models are identical.

EXAMPLE Butterfat data (Sokal, R. R. and Rohlf F. J. (1981). *Biometry*, 2nd edition)

The data give the average butterfat content (percentages) for random samples of twenty cows (ten two-year old and ten mature (greater than four years old)) from each of five breeds. The data are from Canadian records of pure-bred dairy cattle. There are 100 observations on two age groups (two year old and mature) and five breeds.

The **response variable** is butterfat level. **Factor A** is the *age* and there are $a = 2$ factor levels:

1. Mature
2. Two years

Factor B is the *breed* and there are $b = 5$ factor levels:

1. Ayrshire
2. Canadian
3. Guernsey
4. Holstein-Fresian
5. Jersey

$r = 10$ replicate measurements were made, yielding $n = 2 \times 5 \times 10 = 100$ data in total.

Results: Levene's test indicates that there may be some differences in the variances

```
> car::leveneTest(butterfat~breed*age, data=Butterfat, center='mean')
+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value Pr(>F)
+ group  9    2.711 0.00769 **
+      90
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> car::leveneTest(butterfat~breed*age, data=Butterfat, center='median')
+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group  9    1.871 0.0664 .
+      90
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now fit the four models

- Model 3: $A + B + A : B$ (main effect plus interaction)
- Model 2: $A + B$ (main effects only)
- Model 1: B (factor B only)
- Model 0: 1 (constant mean)

```
> m3<-aov(butterfat~breed*age, data=Butterfat)
> summary(m3)

+               Df Sum Sq Mean Sq F value Pr(>F)
+ breed          4   34.3     8.58   49.57 <2e-16 ***
+ age            1    0.3     0.27    1.58   0.21
```

```

+ breed:age      4      0.5      0.13      0.74      0.57
+ Residuals     90     15.6      0.17
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m2<-aov(butterfat~breed+age,data=Butterfat)
> summary(m2)

+              Df Sum Sq Mean Sq F value Pr(>F)
+ breed         4   34.3     8.58    50.1 <2e-16 ***
+ age           1    0.3     0.27     1.6   0.21
+ Residuals     94   16.1     0.17
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m1<-aov(butterfat~breed,data=Butterfat)
> summary(m1)

+              Df Sum Sq Mean Sq F value Pr(>F)
+ breed         4   34.3     8.58    49.8 <2e-16 ***
+ Residuals     95   16.4     0.17
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m0<-aov(butterfat~1,data=Butterfat)
> anova(m0,m1,m2,m3)

+ Analysis of Variance Table
+
+ Model 1: butterfat ~ 1
+ Model 2: butterfat ~ breed
+ Model 3: butterfat ~ breed + age
+ Model 4: butterfat ~ breed * age
+   Res.Df    RSS Df Sum of Sq    F Pr(>F)
+ 1         99 50.69
+ 2         95 16.37  4      34.32 49.565 <2e-16 ***
+ 3         94 16.09  1       0.27  1.580  0.212
+ 4         90 15.58  4       0.51  0.742  0.566
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Note that we get the same result if we rearrange terms in the model formula: for example

```

> m3a<-aov(butterfat~age*breed,data=Butterfat)
> print(summary(m3a), digits = 5)

+              Df Sum Sq Mean Sq F value Pr(>F)
+ age           1  0.274   0.2735   1.5801 0.2120
+ breed         4 34.321   8.5803  49.5651 <2e-16 ***
+ age:breed     4  0.514   0.1285   0.7421 0.5658
+ Residuals     90 15.580   0.1731
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The results are simply rearranged in the output.

Taking the results in turn:

1. **Interaction model:** First note that the Levene test **REJECTS** the null hypothesis of equal group variances ($p = 0.008$), so the following ANOVA results are questionable. However, the p-value is not too small, so we proceed but with caution.

There is a **significant difference** due to Factor B (breed, $F_{obs} = 49.565$, p-value < 0.001), but there is no effect of Factor A (age, $F_{obs} = 1.580$, $p = 0.212$), and no significant interaction (age:breed, $F_{obs} = 0.742$, $p = 0.566$).

2. **Factor B only:** If we omit the Factor A and interaction term, and fit model m1, we confirm the strong effect of Factor B ($F_{obs} = 49.8$, $p < 0.000$), and then can estimate the Factor B treatment means. Note how the error degrees of freedom changes when terms in the model are omitted.
3. The anova function can be used to summarize the fits of all the models.

Means plots:

```
> par(mar=c(4, 4, 1, 2))
> buttersum <- Rmisc::summarySE(Butterfat, measurevar="butterfat",
+                               groupvars=c("age", "breed"))
> buttersum

+           age      breed  N butterfat      sd      se      ci
+ 1      Mature Ayreshire 10      4.154 0.246270 0.0778774 0.176171
+ 2      Mature Canadian 10      4.389 0.260532 0.0823873 0.186373
+ 3      Mature Guernsey 10      5.001 0.408315 0.1291205 0.292091
+ 4      Mature Holstein 10      3.676 0.310133 0.0980725 0.221855
+ 5      Mature Jersey 10      5.452 0.506377 0.1601305 0.362240
+ 6 Two year old Ayreshire 10      3.966 0.251802 0.0796269 0.180129
+ 7 Two year old Canadian 10      4.488 0.456869 0.1444745 0.326824
+ 8 Two year old Guernsey 10      4.899 0.566224 0.1790559 0.405052
+ 9 Two year old Holstein 10      3.663 0.212606 0.0672318 0.152089
+ 10 Two year old Jersey 10      5.133 0.667334 0.2110295 0.477382

> pd <- position_dodge(0.2)
> ggplot(buttersum,
+        aes(x=breed, y=butterfat, colour=age, group=age)) +
+   geom_errorbar(aes(ymin=butterfat-ci, ymax=butterfat+ci),
+                 width=0.25, position=pd) +
+   geom_line(position=pd) +
+   geom_point(position=pd, size=3) +
+   theme(legend.position="bottom") + ylim(3, 6.5)
```

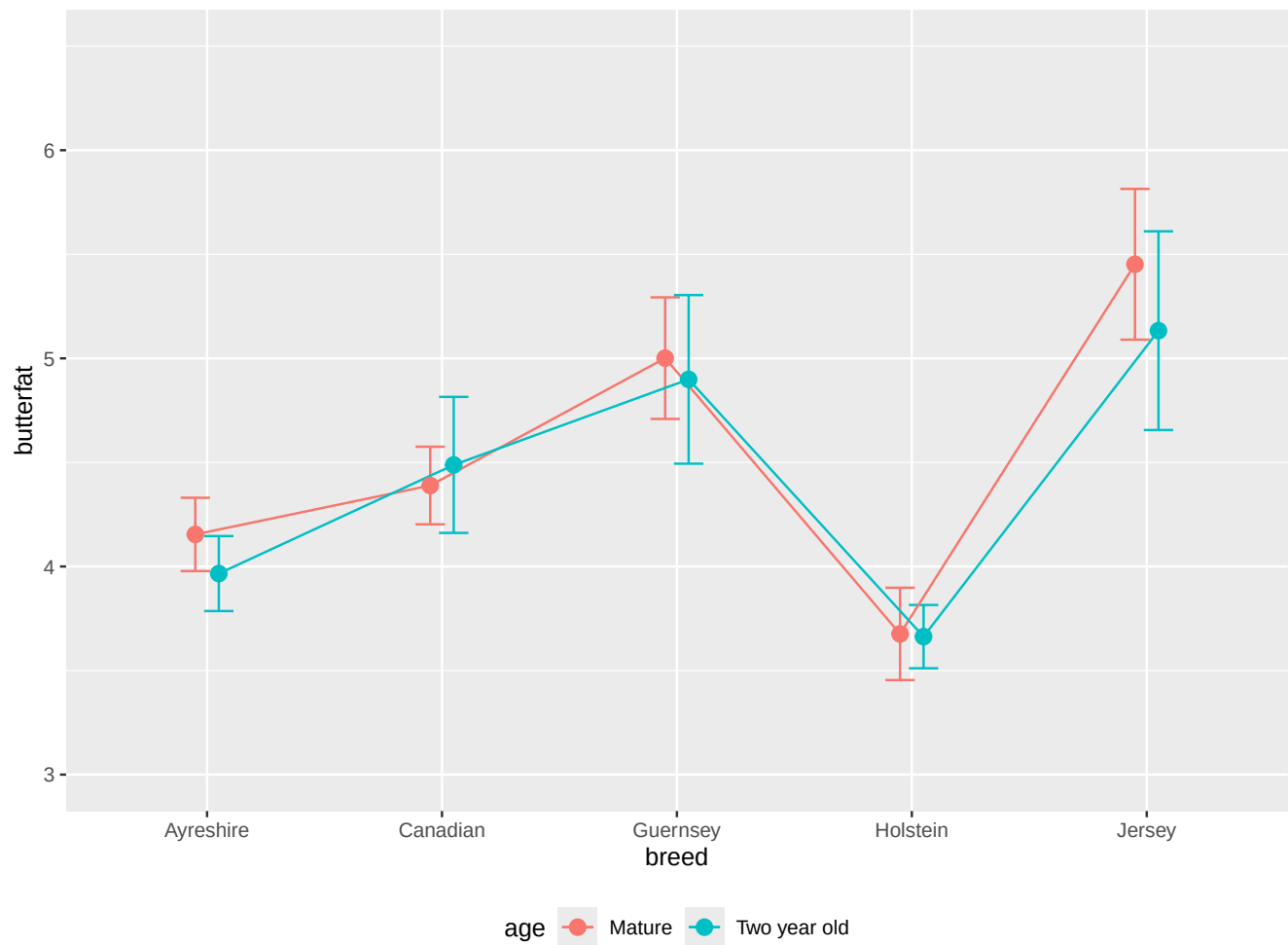


Figure 17: Means Plot by breed

```
> par(mar=c(4, 4, 1, 2))
> pd <- position_dodge(0.2)
> ggplot(buttersum,
+   aes(x=age, y=butterfat, colour=breed, group=breed)) +
+   geom_errorbar(aes(ymin=butterfat-ci, ymax=butterfat+ci),
+     width=0.25, position=pd) +
+   geom_line(position=pd) +
+   geom_point(position=pd, size=3) +
+   theme(legend.position="bottom") + ylim(3, 6.5)
```

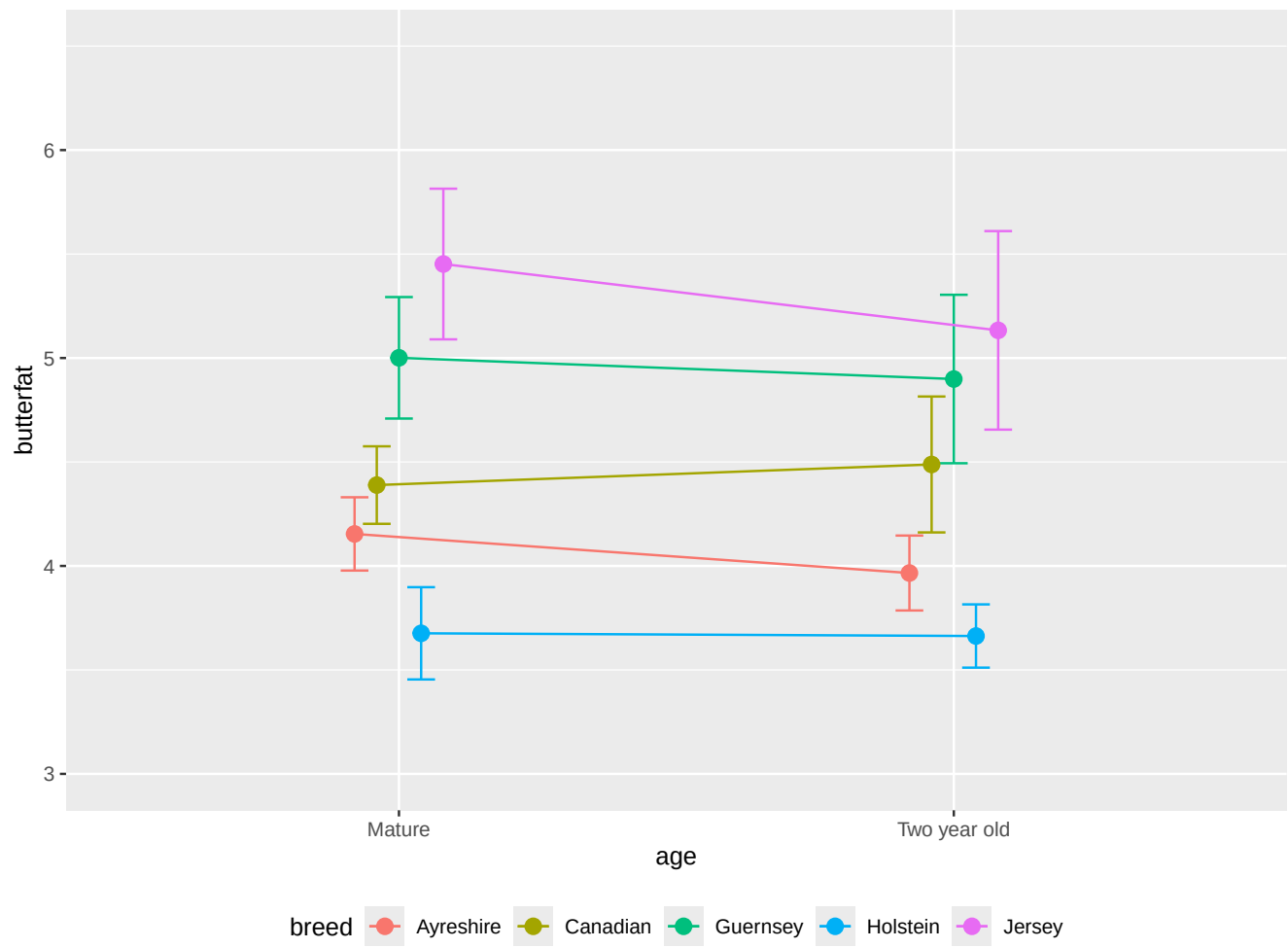


Figure 18: Means Plot by age

EXAMPLE Lyrics data (McClave and Sincich, *Statistics*)

The effect of violent song lyrics on the aggression level of listeners is to be investigated. Two songs (classified as *Violent* and *Non-Violent*) were played to two groups (or “pools”) of students, one volunteer group and one group drawn from a psychology class. The students then rated the songs lyrical content, and from this (by means of a word-association test), the aggression level of the students was computed.

The **response variable** is aggression level. **Factor A** is the *song* and there are $a = 2$ factor levels:

1. Violent
2. Non-violent

Factor B is the *pool* and there are $b = 2$ factor levels:

1. Volunteer
2. Psychology class

$r = 15$ replicate measurements were made, so that $n = 2 \times 2 \times 15 = 60$ data were obtained in total.

```
> car::leveneTest(SCORE~SONG, data=Lyrics, center='mean')
+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value Pr(>F)
+ group 1    0.017  0.897
+      58

> car::leveneTest(SCORE~SONG, data=Lyrics, center='median')
+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group 1    0.008  0.928
+      58
```

```
> car::leveneTest(SCORE~SONG*POOL, data=Lyrics, center='mean')
+ Levene's Test for Homogeneity of Variance (center = "mean")
+      Df F value Pr(>F)
+ group 3    0.329  0.804
+      56

> car::leveneTest(SCORE~SONG*POOL, data=Lyrics, center='median')
+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group 3    0.226  0.878
+      56
```

First note that the Levene test **does not reject** the null hypothesis of equal group variances for the model that fits the SONG factor only, nor for the full factorial model that fits both main effects and interaction.

```

> m3<-aov(SCORE~SONG*POOL,data=Lyrics)
> print(summary(m3), digits = 5)

+           Df  Sum Sq Mean Sq F value    Pr(>F)
+ SONG         1   5.8907   5.8907 26.1144 4.031e-06 ***
+ POOL         1   0.1307   0.1307  0.5793   0.4498
+ SONG:POOL     1   0.3527   0.3527  1.5634   0.2164
+ Residuals    56 12.6320   0.2256
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m2<-aov(SCORE~SONG+POOL,data=Lyrics)
> print(summary(m2), digits = 5)

+           Df  Sum Sq Mean Sq F value    Pr(>F)
+ SONG         1   5.8907   5.8907 25.8588 4.264e-06 ***
+ POOL         1   0.1307   0.1307  0.5736   0.452
+ Residuals    57 12.9847   0.2278
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m1<-aov(SCORE~SONG,data=Lyrics)
> print(summary(m1), digits = 5)

+           Df  Sum Sq Mean Sq F value    Pr(>F)
+ SONG         1   5.8907   5.8907  26.05 3.851e-06 ***
+ Residuals    58 13.1153   0.2261
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m0<-aov(SCORE~1,data=Lyrics)
> anova(m0,m1,m2,m3)

+ Analysis of Variance Table
+
+ Model 1: SCORE ~ 1
+ Model 2: SCORE ~ SONG
+ Model 3: SCORE ~ SONG + POOL
+ Model 4: SCORE ~ SONG * POOL
+   Res.Df  RSS Df Sum of Sq    F    Pr(>F)
+ 1       59 19.01
+ 2       58 13.12  1      5.891 26.114 4.03e-06 ***
+ 3       57 12.98  1      0.131  0.579   0.450
+ 4       56 12.63  1      0.353  1.563   0.216
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Results:

1. **Interaction model:** There is a **significant difference** due to Factor A (song, $F_{obs} = 26.114$, $p\text{-value} < 0.001$), but there is no effect of Factor B (pool, line 5, $F_{obs} = 0.579$, $p = 0.450$), and no significant interaction (song:pool, line 6, $F_{obs} = 1.563$, $p = 0.216$).
2. Fits of the **main-effects** model m2 (Factor A and Factor B but no interaction), and the Factor A only model m1 confirm the results.

EXAMPLE Gravel data

A company produces gravel from a number of quarries and in each quarry there are morning and afternoon shifts of workers. The company wishes to know whether there are differences in the quantity of gravel produced from these quarries and gathers the following data on the amount of gravel produced by each shift in one week (in tonnes). It can be assumed that the week being studied was a typical week, and that there was no systematic differences due to different workers etc.

The **response variable** is amount of gravel produced. **Factor A** is the *shift* and there are $a = 2$ factor levels:

1. AM
2. PM

Factor B is the *quarry* and there are $b = 4$ factor levels:

1. A
2. B
3. C
4. D

$r = 5$ replicate measurements were made, so that $n = 2 \times 4 \times 5 = 40$ data were obtained in total.

```
> car::leveneTest(amount~shift*quarry,data=Gravel,center='median')
+ Levene's Test for Homogeneity of Variance (center = "median")
+      Df F value Pr(>F)
+ group  7    0.168  0.99
+      32
```

```
> m3<-aov(amount~shift*quarry,data=Gravel)
> print(summary(m3), digits = 5)

+              Df Sum Sq Mean Sq F value    Pr(>F)
+ shift          1  135.79   135.792  13.6670 0.0008139 ***
+ quarry         3  596.04   198.679  19.9964 1.742e-07 ***
+ shift:quarry    3   32.75    10.916   1.0986 0.3639819
+ Residuals      32 317.94     9.936
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> m2<-aov(amount~shift+quarry,data=Gravel)
> m1<-aov(amount~quarry,data=Gravel)
> m0<-aov(amount~1,data=Gravel)
```

```
> anova(m0,m1,m2,m3)

+ Analysis of Variance Table
+
+ Model 1: amount ~ 1
+ Model 2: amount ~ quarry
+ Model 3: amount ~ shift + quarry
+ Model 4: amount ~ shift * quarry
+   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
```

```

+ 1      39 1082.5
+ 2      36  486.5  3      596.0 19.996 1.74e-07 ***
+ 3      35  350.7  1      135.8 13.667 0.000814 ***
+ 4      32  317.9  3      32.7  1.099 0.363982
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

1. **Interaction model:** First note that the Levene test **does not reject** the null hypothesis of equal group variances. There is a **significant difference** due to Factor A (shift, $F_{obs} = 13.667$, $p = 0.001$), and due to Factor B (quarry, $F_{obs} = 19.996$, $p < 0.001$), but **no significant interaction** (shift:quarry, $F_{obs} = 1.099$, $p = 0.364$).
2. **Factor A and B only:** If we omit the interaction term, and refit the model, we confirm the strong effect of both factors (shift, $F_{obs} = 13.55$, $p = 0.001$, quarry, $F_{obs} = 19.83$, $p < 0.001$). The conclusion is that there is a difference between the two levels of factor *shift* and the four levels of factor *quarry*, but that there is no interaction, that is, the difference between morning and afternoon shift is the same in each block; this is depicted in the Marginal Means plot.

Note again how the error degrees of freedom changes when terms in the model are omitted.

```

> par(mar=c(4, 4, 1, 2))
> quarrysum <- Rmisc::summarySE(Gravel, measurevar="amount",
+                               groupvars=c("shift", "quarry"))
> quarrysum

+   shift quarry N amount      sd      se      ci
+ 1     AM      A  5  26.96 2.98044 1.33289 3.70070
+ 2     AM      B  5  23.22 2.37318 1.06132 2.94670
+ 3     AM      C  5  22.44 2.86671 1.28203 3.55948
+ 4     AM      D  5  14.64 4.15187 1.85677 5.15522
+ 5     PM      A  5  21.12 3.58148 1.60169 4.44699
+ 6     PM      B  5  21.12 3.65335 1.63383 4.53624
+ 7     PM      C  5  17.34 2.31257 1.03421 2.87144
+ 8     PM      D  5  12.94 2.82719 1.26436 3.51042

> pd <- position_dodge(0.2)
> ggplot(quarrysum,
+        aes(x=shift, y=amount, colour=quarry, group=quarry)) +
+   geom_errorbar(aes(ymin=amount-ci, ymax=amount+ci),
+                 width=0.25, position=pd) +
+   geom_line(position=pd) +
+   geom_point(position=pd, size=3) +
+   theme(legend.position="bottom")

```

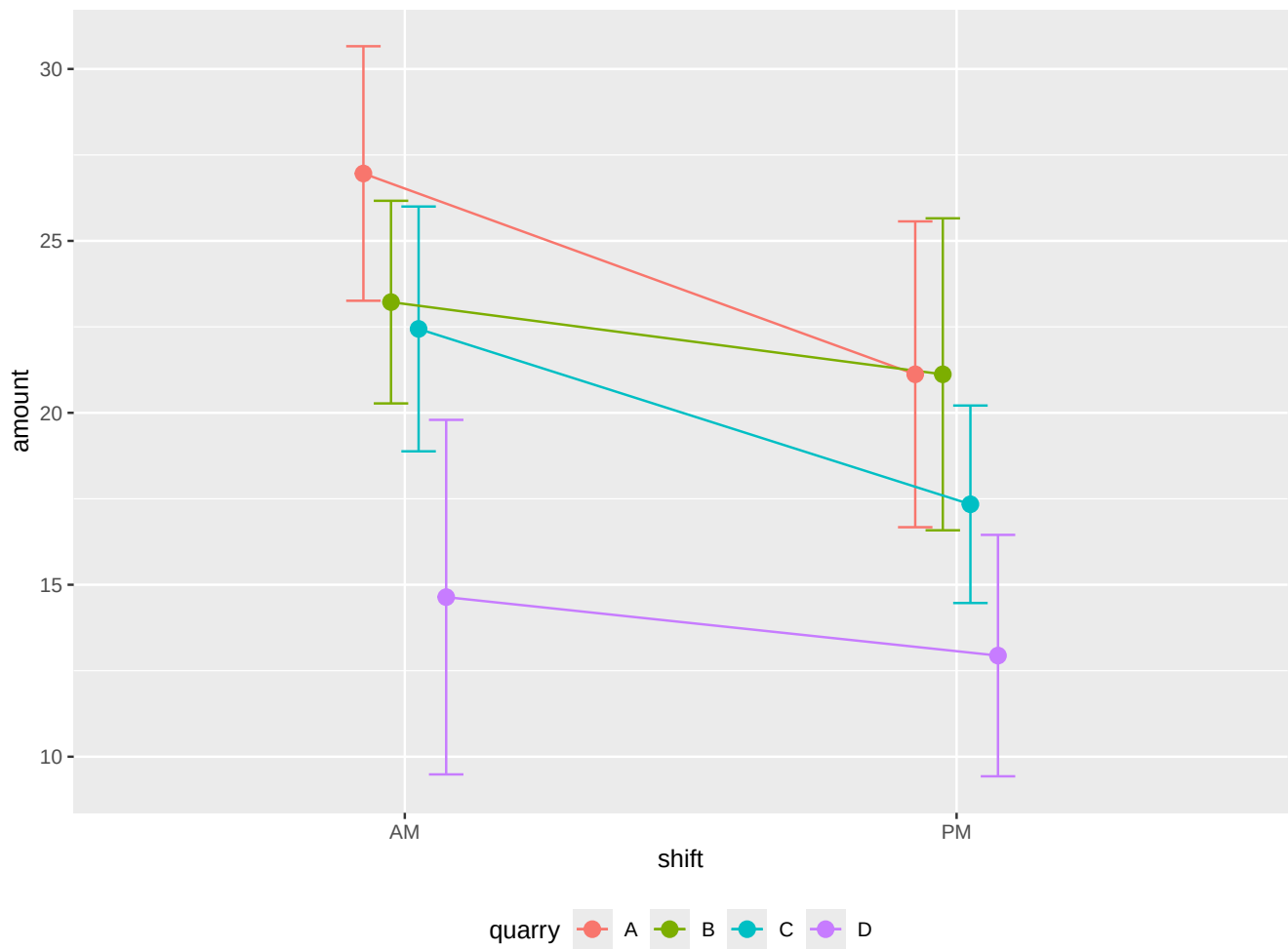


Figure 19: Means plot by shift

```
> par(mar=c(4, 4, 1, 2))
> pd <- position_dodge(0.2)
> ggplot(quarrysum,
+   aes(x=quarry, y=amount, colour=shift, group=shift)) +
+   geom_errorbar(aes(ymin=amount-ci, ymax=amount+ci),
+     width=0.25, position=pd) +
+   geom_line(position=pd) +
+   geom_point(position=pd, size=3) +
+   theme(legend.position="bottom")
```

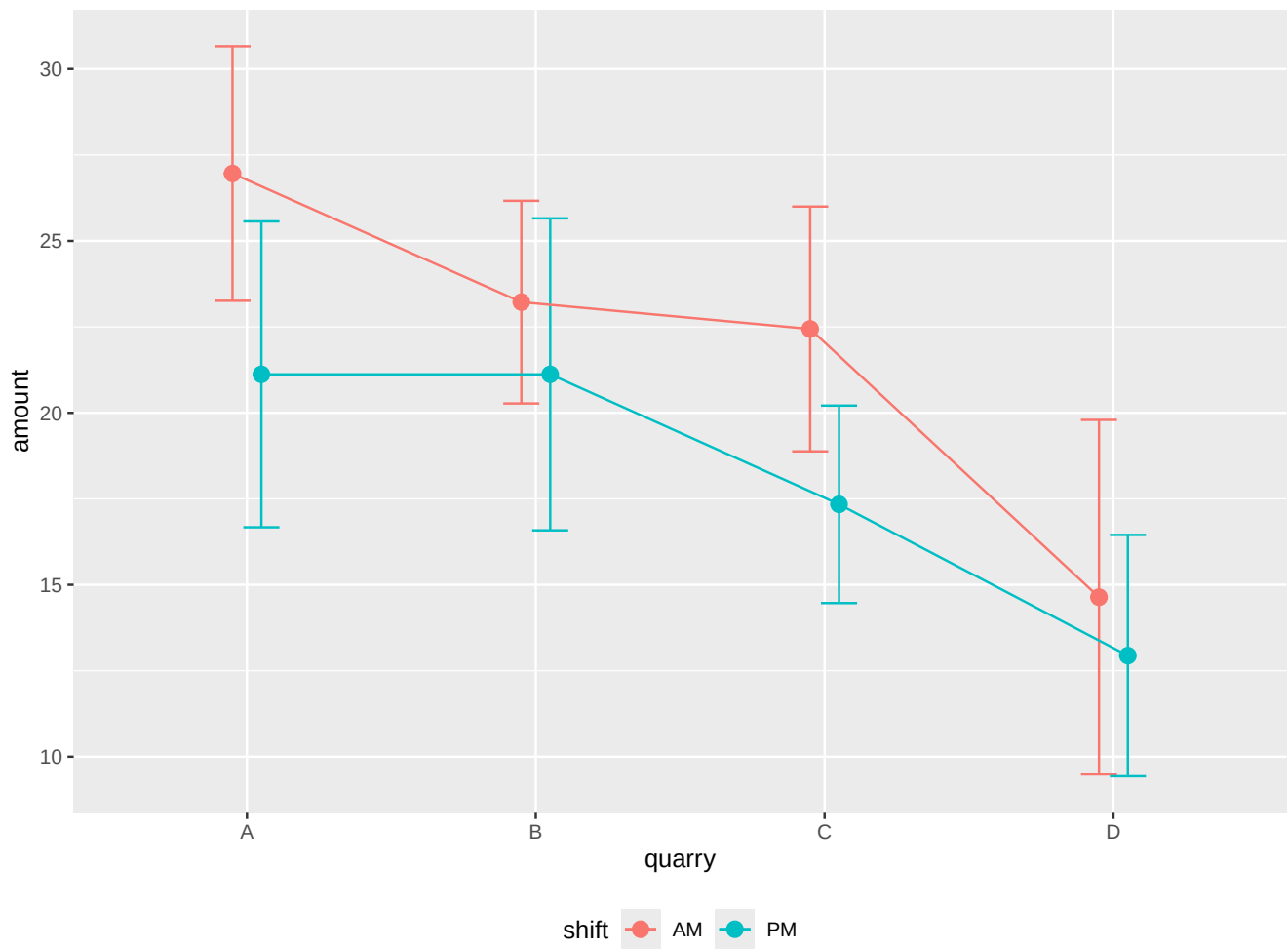


Figure 20: Means plot by quarry

SIMPLE LINEAR REGRESSION

We consider the model for response variable, Y , as a function of the predictor, X , observed to take the value x . Specifically we consider the model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 and β_1 are the **intercept** and **slope** parameters respectively, and ϵ is a random variable with expectation zero and variance σ^2 . In this model

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x.$$

To estimate the parameters β_0 and β_1 from data $(x_i, y_i), i = 1, \dots, n$, we use the **least-squares** criterion, and choose the values $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the **sum of squared errors**

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

It can be shown that the parameter estimates depend on the following sample summary statistics:

- Sample mean of x values:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Sample mean of y values:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- Sum of Squares SS_{xx} :

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

- Sum of Squares SS_{xy} :

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

The **least-squares estimates** are:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

yielding **fitted-values**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and **residual errors** (or **residuals**)

$$\hat{e}_i = y_i - \hat{y}_i.$$

An estimate of the **residual error variance** is given by

$$\hat{\sigma}^2 = \frac{SSE(\hat{\beta}_0, \hat{\beta}_1)}{n - 2}$$

EXAMPLE Blood Viscosity and Packed Cell Volume

The following data are measurements of packed cell volume (PCV) and blood viscosity in samples taken from 32 hospital patients. We wish to model viscosity (y) as a function of PCV (x).

Reference: Begg, C. B. and Hearn, J. B. (1966) Components of Blood Viscosity. The relative contributions of haematocrit, plasma fibrinogen and other proteins, *Clinical Science*, **31**, 87-92.

Unit	x	y	Unit	x	y	Unit	x	y	Unit	x	y
1	40.00	3.71	9	46.75	4.14	17	51.25	4.68	25	49.50	5.12
2	40.00	3.78	10	48.00	4.20	18	50.25	4.73	26	56.00	5.15
3	42.50	3.85	11	46.00	4.20	19	49.00	4.87	27	50.00	5.17
4	42.00	3.88	12	47.00	4.27	20	50.00	4.94	28	47.00	5.18
5	45.00	3.98	13	43.25	4.27	21	50.00	4.95	29	53.25	5.38
6	42.00	4.03	14	45.00	4.37	22	49.00	4.96	30	57.00	5.77
7	42.50	4.05	15	50.00	4.41	23	50.50	5.02	31	54.00	5.90
8	47.00	4.14	16	45.00	4.64	24	51.25	5.02	32	54.00	5.90

- Sample mean of x values: $\bar{x} = 47.938$; sample mean of y values: $\bar{y} = 4.646$
- Sums of Squares

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 615.75 \quad SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 75.386$$

Thus

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 0.122 \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -1.223$$

The estimate of the residual error variance is

$$\hat{\sigma}^2 = \frac{SSE(\hat{\beta}_0, \hat{\beta}_1)}{n - 2} = \frac{2.721}{30} = 0.091$$

```
> m0.visc<-lm(viscosity~pcv,data=Viscosity)
> summary(m0.visc)

+
+ Call:
+ lm(formula = viscosity ~ pcv, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.4881 -0.2215  0.0469  0.1296  0.6492
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -1.2234     0.5842   -2.09   0.045 *
+ pcv           0.1224     0.0121  10.09  3.7e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.301 on 30 degrees of freedom
+ Multiple R-squared:  0.772, Adjusted R-squared:  0.765
+ F-statistic: 102 on 1 and 30 DF, p-value: 3.73e-11
```

```
> par(mar=c(4,4,1,2))
> plot(Viscosity$viscosity~Viscosity$pcv,
+      pch=19,cex=0.8, xlab='pcv', ylab='Viscosity')
> abline(coef(m0.visc),col='red')
```

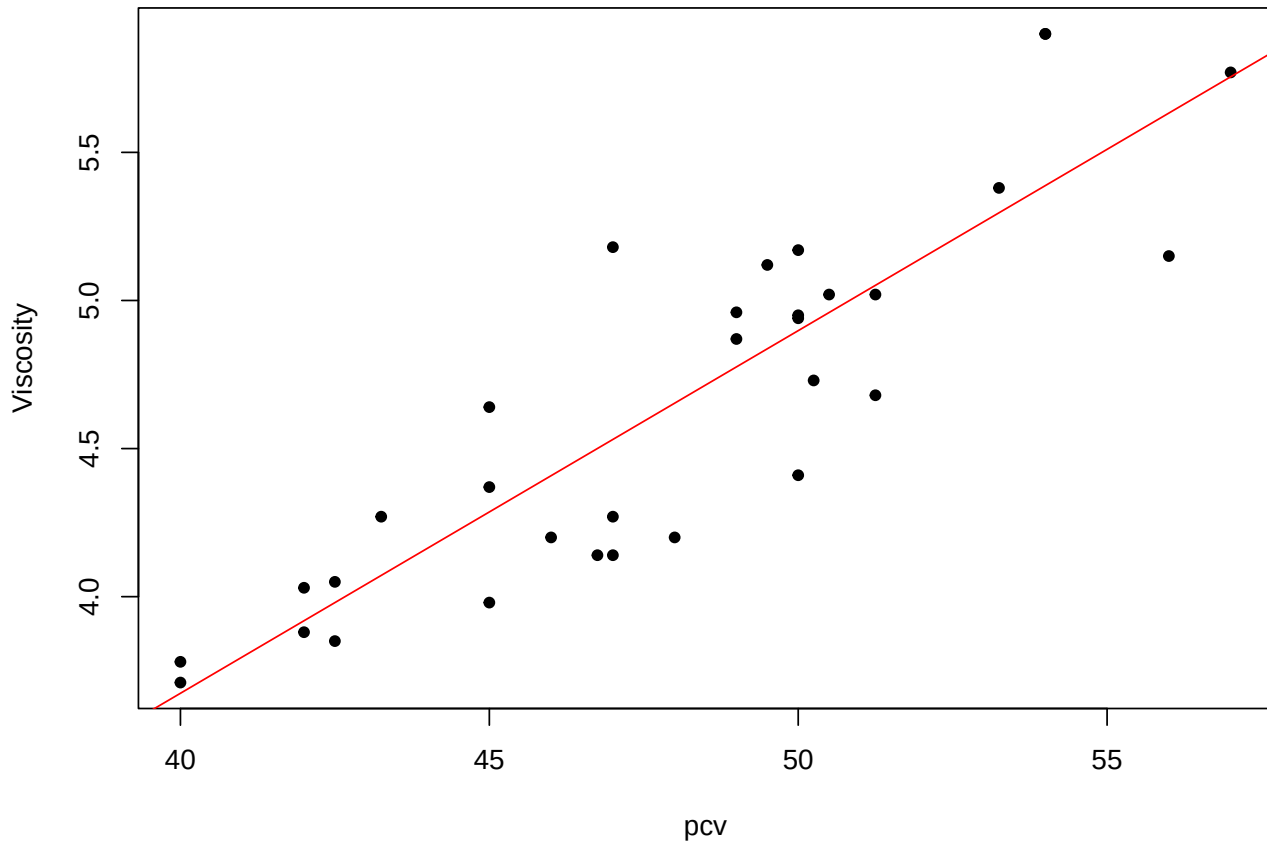


Figure 21: Viscosity data: straight-line fit

```
> sqrt(sum(residuals(m0.visc)^2)/30)
+ [1] 0.301156
```

```
> cor.test(Viscosity$pcv,Viscosity$viscosity)
+
+ Pearson's product-moment correlation
+
+ data: Viscosity$pcv and Viscosity$viscosity
+ t = 10.09, df = 30, p-value = 3.73e-11
+ alternative hypothesis: true correlation is not equal to 0
+ 95 percent confidence interval:
+ 0.764350 0.939586
+ sample estimates:
+ cor
+ 0.878818
```

EXAMPLE Coleman Report Data

Data were collected at 20 US schools, and used to examine the relationship between performance of students in the school in a verbal reasoning test and the socioeconomic status of the catchment area.

Reference: Mosteller and Tukey (1977) *Data Analysis and Regression*

School	Score	Status	School	Score	Status
	y	x		y	x
1	7.20	37.01	11	-12.86	23.30
2	-11.71	26.51	12	0.92	35.20
3	12.32	36.51	13	4.77	34.90
4	14.28	40.70	14	-0.96	33.10
5	6.31	37.10	15	-16.04	22.70
6	6.16	33.90	16	10.62	39.70
7	12.70	41.80	17	2.66	31.80
8	-0.17	33.40	18	-10.99	31.70
9	9.85	41.01	19	15.03	43.10
10	-0.05	37.20	20	12.77	41.01

Results: Here, to test for significant correlation, we use the test statistic

$$T_{obs} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.927}{\sqrt{(1-0.927^2)/(20-2)}} = 10.499$$

which we must compare against the $t_{n-2} \equiv t_{18}$ distribution. For a two-tailed test at the significance level $\alpha = 0.05$, the critical values are $C_R = \pm 2.101$ (from t -tables), so the hypothesis H_0 that the true correlation is zero is **rejected**.

```
> cor.test(Coleman$status, Coleman$testscore)
+
+ Pearson's product-moment correlation
+
+ data: Coleman$status and Coleman$testscore
+ t = 10.5, df = 18, p-value = 4.2e-09
+ alternative hypothesis: true correlation is not equal to 0
+ 95 percent confidence interval:
+ 0.821826 0.971207
+ sample estimates:
+ cor
+ 0.927161
```

The regression analysis is contained below:

```
> m0.cole<-lm(testscore~status, data=Coleman)
> summary(m0.cole)
+
+ Call:
+ lm(formula = testscore ~ status, data = Coleman)
+
+ Residuals:
```

```

+      Min      1Q Median      3Q      Max
+ -8.941 -1.219 -0.055  1.951  6.989
+
+ Coefficients:
+           Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -50.682      5.193   -9.76  1.3e-08 ***
+ status       1.534      0.146   10.50  4.2e-09 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 3.71 on 18 degrees of freedom
+ Multiple R-squared:  0.86, Adjusted R-squared:  0.852
+ F-statistic: 110 on 1 and 18 DF, p-value: 4.2e-09

> print(summary(aov(m0.cole)), digits = 5)

+           Df Sum Sq Mean Sq F value    Pr(>F)
+ status      1 1513.2 1513.21  110.23 4.199e-09 ***
+ Residuals   18  247.1   13.73
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> par(mar=c(4,4,1,2))
> plot(Coleman$testscore~Coleman$status,
+       pch=19,cex=0.8, xlab='status', ylab='testscore')
> abline(coef(m0.cole),col='red')

```

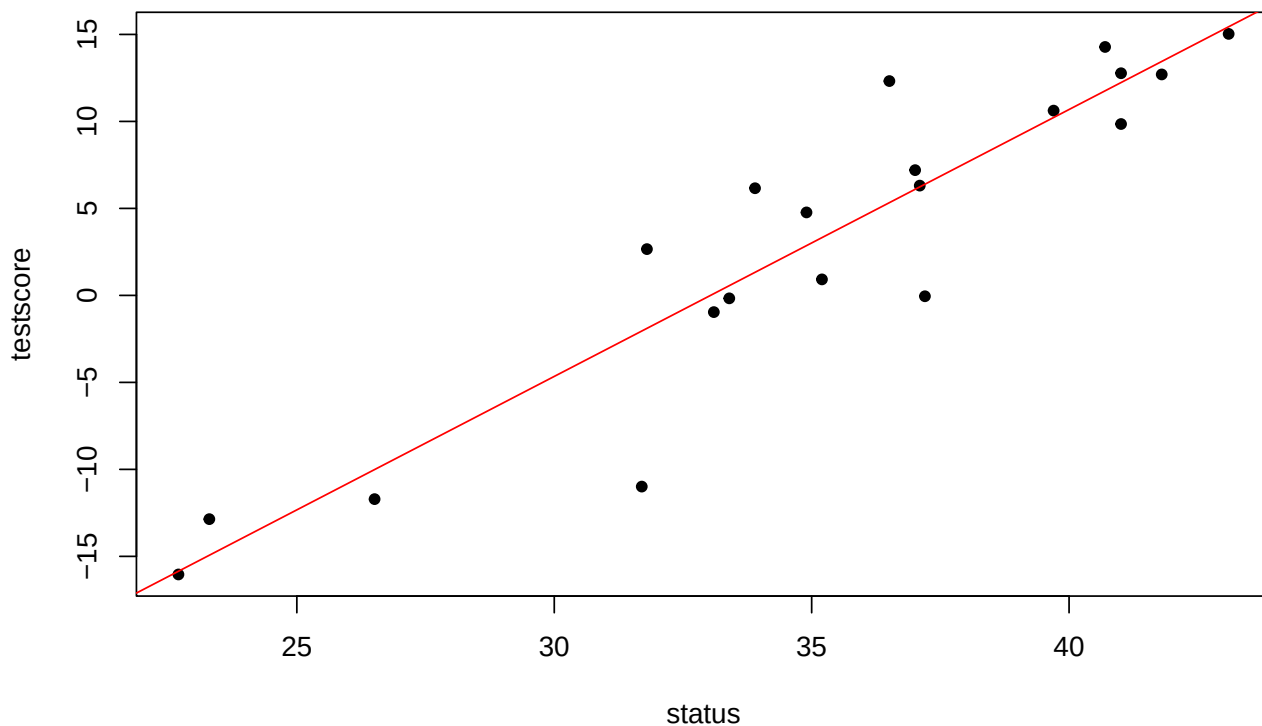


Figure 22: Coleman data: straight-line fit

EXAMPLE Hooker's Temperature and Pressure Data

The following data record the boiling point temperature (in degrees F) of water under different atmospheric pressures. The data were collected in a Himalayan expedition by botanist Joseph Hooker.

Reference: Forbes, J. (1957). Further experiments and remarks on the measurement of heights by boiling point of water. *Transactions of the Royal Society of Edinburgh*, 21, 235-243.

y	x	y	x	y	x	y	x
210.8	29.211	196.4	21.928	189.5	18.869	184.1	16.817
210.2	28.559	196.3	21.654	188.8	18.356	183.2	16.385
208.4	27.972	195.6	21.605	188.5	18.507	182.4	16.235
202.5	24.697	193.4	20.480	185.7	17.267	181.9	16.106
200.6	23.726	193.6	20.212	186.0	17.221	181.9	15.928
200.1	23.369	191.4	19.758	185.6	17.062	181.0	15.919
199.5	23.030	191.1	19.490	184.1	16.959	180.6	15.376
197.0	21.892	190.6	19.386	184.6	16.881		

Here, to test for significant correlation, we use the test statistic

$$T_{obs} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.996}{\sqrt{(1-0.996^2)/(31-2)}} = 59.143$$

which we must compare against the $t_{n-2} \equiv t_{29}$ distribution. For a two-tailed test at the significance level $\alpha = 0.05$, the critical values are $C_R = \pm 2.045$ (McClave and Sincich t -tables), so the hypothesis H_0 that the true correlation is zero is **rejected**.

```
> cor.test(Hooker$pressure, Hooker$temp)
+
+ Pearson's product-moment correlation
+
+ data: Hooker$pressure and Hooker$temp
+ t = 59.14, df = 29, p-value <2e-16
+ alternative hypothesis: true correlation is not equal to 0
+ 95 percent confidence interval:
+ 0.991378 0.998034
+ sample estimates:
+ cor
+ 0.99588

> m0.hooker <- lm(temp ~ pressure, data = Hooker)
> summary(m0.hooker)
+
+ Call:
+ lm(formula = temp ~ pressure, data = Hooker)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -1.67  -0.68   0.22   0.53   1.40
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  197.000     1.000    197.00  <2e-16
+ pressure      0.99600     0.016    62.250  <2e-16
```

```

+ (Intercept) 146.6729      0.7764    188.9   <2e-16 ***
+ pressure      2.2526      0.0381     59.1   <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.806 on 29 degrees of freedom
+ Multiple R-squared:  0.992, Adjusted R-squared:  0.991
+ F-statistic: 3.5e+03 on 1 and 29 DF,  p-value: <2e-16

> print(summary(aov(m0.hooker)), digits = 5)

+              Df  Sum Sq Mean Sq F value    Pr(>F)
+ pressure      1 2272.47 2272.47  3497.9 < 2.2e-16 ***
+ Residuals    29   18.84    0.65
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Quadratic Model:

```

> m0.hooker<-lm(temp~pressure,data=Hooker)
> summary(m0.hooker)

+
+ Call:
+ lm(formula = temp ~ pressure, data = Hooker)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -1.67  -0.68   0.22   0.53   1.40
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept) 146.6729      0.7764    188.9   <2e-16 ***
+ pressure      2.2526      0.0381     59.1   <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.806 on 29 degrees of freedom
+ Multiple R-squared:  0.992, Adjusted R-squared:  0.991
+ F-statistic: 3.5e+03 on 1 and 29 DF,  p-value: <2e-16

> m1.hooker<-lm(temp~pressure+I(pressure^2),data=Hooker)
> summary(m1.hooker)

+
+ Call:
+ lm(formula = temp ~ pressure + I(pressure^2), data = Hooker)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.7991 -0.2631 -0.0158  0.2514  0.7389
+
+ Coefficients:

```

```

+               Estimate Std. Error t value Pr(>|t|)
+ (Intercept)    126.70162     2.11236   59.98 < 2e-16 ***
+ pressure        4.15763     0.19907   20.89 < 2e-16 ***
+ I(pressure^2)  -0.04375     0.00455   -9.61 2.3e-10 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.396 on 28 degrees of freedom
+ Multiple R-squared:  0.998, Adjusted R-squared:  0.998
+ F-statistic: 7.31e+03 on 2 and 28 DF,  p-value: <2e-16

> anova(m0.hooker,m1.hooker)

+ Analysis of Variance Table
+
+ Model 1: temp ~ pressure
+ Model 2: temp ~ pressure + I(pressure^2)
+   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
+ 1       29 18.840
+ 2       28  4.382  1      14.46 92.39 2.29e-10 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

> xv<-seq(15,30,by=0.01)
> fit.lin<-lm(temp~pressure, data=Hooker)
> fit.quad<-lm(temp~pressure+I(pressure^2), data=Hooker)
>
> hooker.fit.lin<-predict(fit.lin, newdata=data.frame(pressure=xv))
> hooker.fit.quad<-predict(fit.quad, newdata=data.frame(pressure=xv))
>
> par(mar=c(4,4,2,2))
> plot(temp~pressure, data=Hooker, pch=19, cex=0.6)
> lines(xv,hooker.fit.lin,col='black')
> lines(xv,hooker.fit.quad,col='black', lty=2)
> legend(16,210, c('Straight-line','Quadratic'), lty=c(1,2))

```

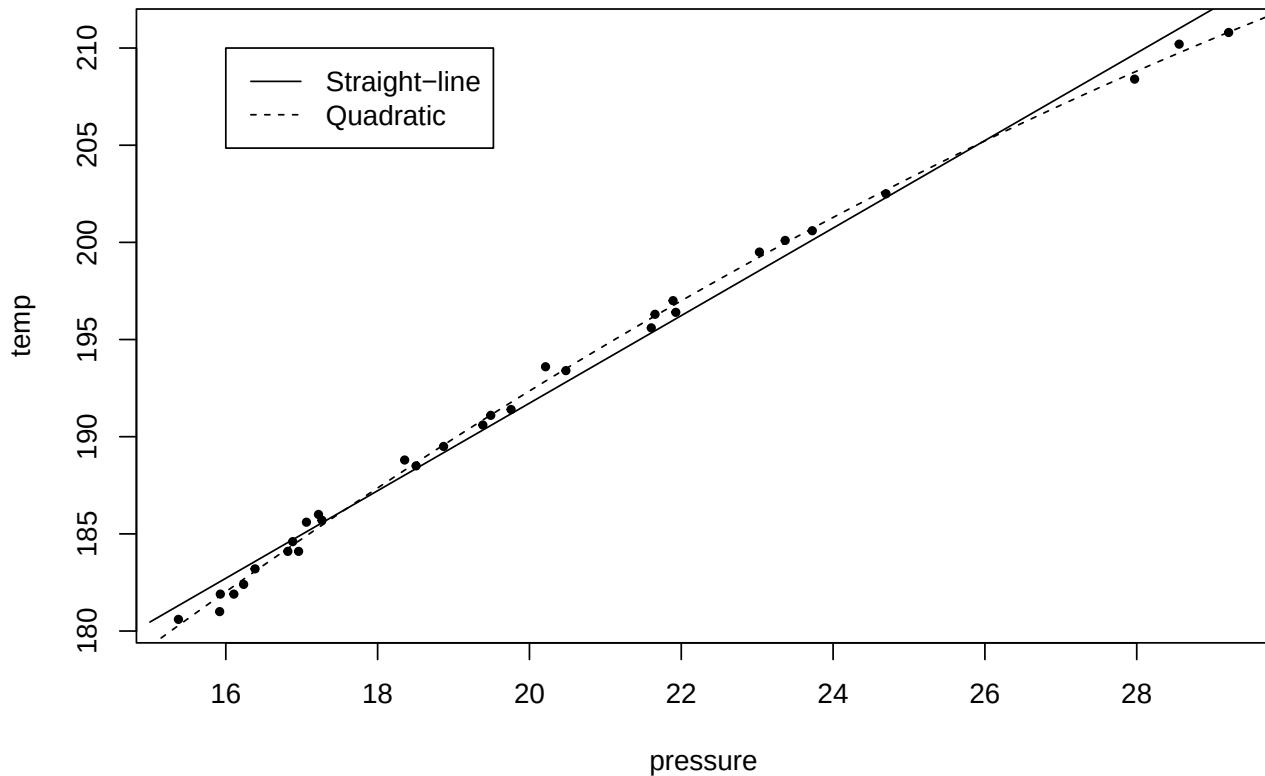


Figure 23: Hooker data: straight-line and quadratic fits

```
> res.lin<-residuals(fit.lin)
> res.quad<-residuals(fit.quad)
> par(mar=c(4,4,1,0), mfrow=c(1,2))
> plot(Hooker$pressure, res.lin, ylim=range(-2,2),
+      pch=19, cex=0.6, xlab='pressure', ylab='Residuals')
> abline(h=0, lty=2)
> text(22,2, 'Straight-line fit')
> plot(Hooker$pressure, res.quad, ylim=range(-2,2),
+      pch=19, cex=0.6, xlab='pressure', ylab='Residuals')
> abline(h=0, lty=2)
> text(22,2, 'Quadratic fit')
```

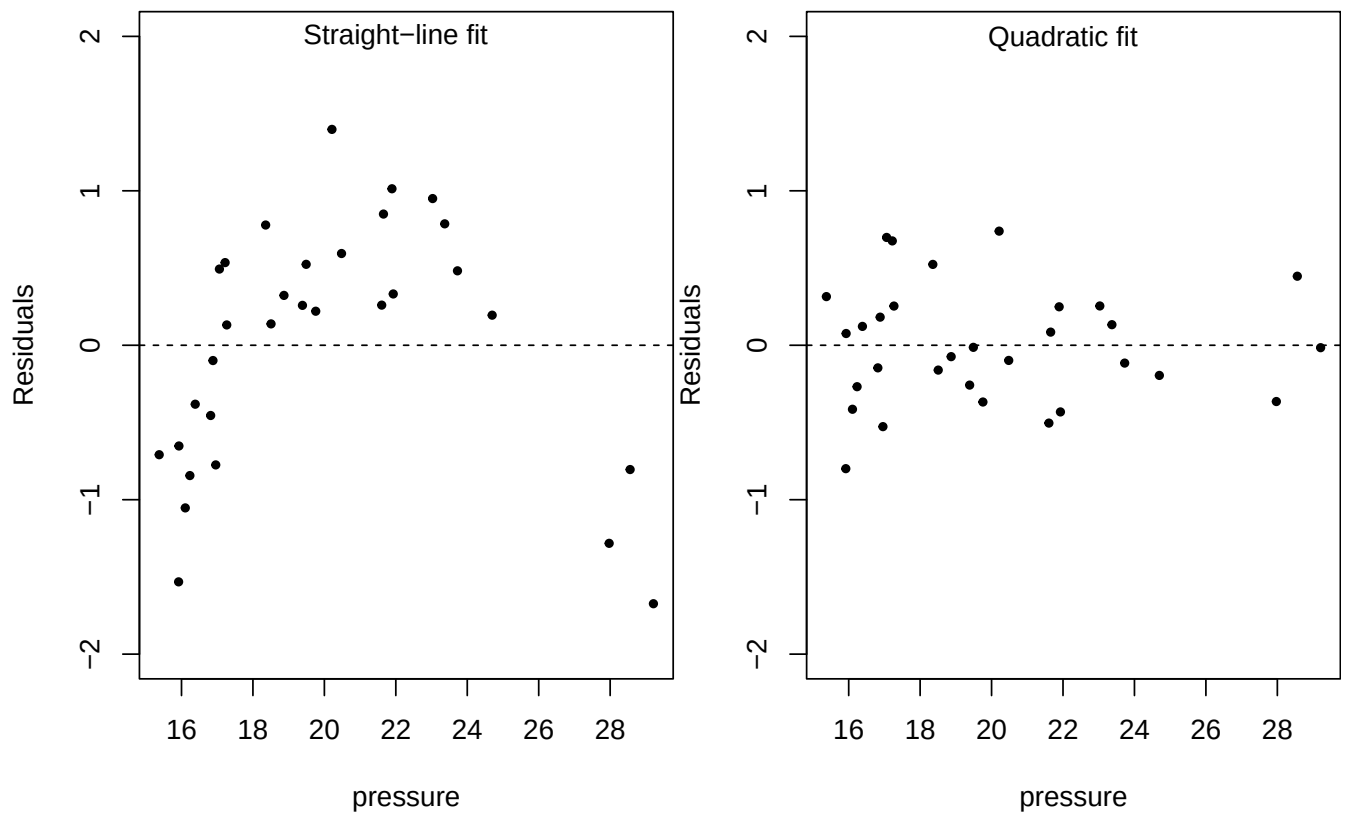


Figure 24: Hooker data: residuals from straight-line and quadratic fits

MATRIX MANIPULATIONS FOR ESTIMATION (MATERIAL NOT EXAMINABLE)

An $r \times c$ **matrix** $\mathbf{A}$ is a rectangular arrangement of numbers with r rows and c columns;

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix}$$

Recall the following rules for manipulating matrices:

- **Transpose:** the transpose operator ^{T} means “flipping” a $r \times c$ matrix into a $c \times r$ matrix. That is

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix} \iff \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{r1} \\ a_{12} & a_{22} & \cdots & a_{r2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1c} & a_{2c} & \cdots & a_{rc} \end{bmatrix}$$

For example, if $r = 2$ and $c = 4$

$$\mathbf{A} = \begin{bmatrix} 5 & -4 & 0 & 1 \\ 3 & 5 & -2 & 0 \end{bmatrix} \iff \mathbf{A}^T = \begin{bmatrix} 5 & 3 \\ -4 & 5 \\ 0 & -2 \\ 1 & 0 \end{bmatrix}$$

A square matrix $\mathbf{A}$ is termed **symmetric** if $\mathbf{A} = \mathbf{A}^T$.

- **Matrix Multiplication:** If $\mathbf{A}$ and $\mathbf{B}$ are two matrices, where $\mathbf{A}$ is a $r_1 \times c$ matrix and $\mathbf{B}$ is a $c \times r_2$ matrix, then the product $\mathbf{A.B}$ (also written $\mathbf{AB}$) is an $r_1 \times r_2$ matrix, with (i, j) th element

$$\sum_{k=1}^c a_{ik} b_{kj} \quad i = 1, \dots, r_1, \quad j = 1, \dots, r_2.$$

For example,

$$\begin{bmatrix} 5 & -4 & 0 & 1 \\ 3 & 5 & -2 & 0 \end{bmatrix} \begin{bmatrix} 3 & 3 & -3 \\ -1 & 2 & -2 \\ 0 & -2 & 0 \\ -5 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 14 & 5 & -6 \\ 4 & 23 & -19 \end{bmatrix}$$

That is, for the first entry in the result matrix, we multiply the **first row** of the first matrix by the **first column** of the second matrix:

$$(5 \times 3) + (-4 \times -1) + (0 \times 0) + (1 \times -5) = 15 + 4 - 5 = 14$$

Note that for matrix multiplication to work, we need the first matrix to have the same number of columns as the number of rows in the second matrix. If this holds, the matrices are termed **conformable**. In general, for rectangular matrices

$$\mathbf{A.B} \neq \mathbf{B.A} \quad \text{and} \quad \mathbf{A.B.C} = \mathbf{A.(B.C)} = (\mathbf{A.B}).\mathbf{C}$$

- **Matrix Identity :** A square $k \times k$ with ones along the main diagonal, and zeros elsewhere, is termed the **identity** matrix, and denoted $\mathbf{I}_k$

$$\mathbf{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad \text{so that} \quad \mathbf{I}_k \cdot \mathbf{A} = \mathbf{A} \quad \text{for any } k \times k \text{ matrix } \mathbf{A}$$

- **Matrix Inversion :** A square $k \times k$ matrix $\mathbf{A}$ has an **inverse**, denoted $\mathbf{A}^{-1}$ if

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}_k$$

MATRICES IN LINEAR REGRESSION

- $n \times 1$ vector $\mathbf{y} = [y_1, \dots, y_n]^\top$
- $n \times 2$ matrix $\mathbf{X}$ given by

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top$$

- 2×1 Parameter estimate vector $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1]^\top$

It can be shown that

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

The other quantities of interest in statistical inference for the simple linear regression are also available in matrix form.

- SSE:

$$SSE = S(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})$$

- Residual error variance estimate, $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{S(\hat{\beta})}{n-2} = \frac{1}{n-2} (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})$$

- Variance/Standard Errors of the Parameter estimates:

$$Var[\hat{\beta}] = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

This is a 2×2 matrix, with diagonal entries equal to the squared estimated standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$, $s_{\hat{\beta}_0}^2$ and $s_{\hat{\beta}_1}^2$ respectively.

- Fitted-values :

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{H}\mathbf{y}$$

say, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$

- Residuals:

$$\hat{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

- Prediction: if $\mathbf{x}_p = [1, x_p]^\top$, then the prediction is at the value x_p is

$$y_p = \mathbf{x}_p^\top \hat{\beta}$$

and the prediction error variances are

$$\begin{aligned} \text{Expected Value} &: \hat{\sigma}^2 \mathbf{x}_p^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_p \\ \text{Individual Value} &: \hat{\sigma}^2 (1 + \mathbf{x}_p^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_p) \end{aligned}$$

MULTIPLE LINEAR REGRESSION

EXAMPLE: BLOOD VISCOSITY AND PACKED CELL VOLUME

The following blood viscosity data studied earlier are a good example of where multiple regression could be used. Recall that the data blood viscosity in samples taken from 32 hospital patients. We wish to model viscosity (y) as a function three covariates

- Packed Cell Volume (PCV), x_1 .
- Plasma Fibrinogen, x_2 .
- Plasma Protein, x_3 .

Unit	Viscosity y	PCV x_1	Plasma Fib. x_2	Plasma Pro. x_3
1	3.71	40.00	344	6.27
2	3.78	40.00	330	4.86
3	3.85	42.50	280	5.09
4	3.88	42.00	418	6.79
5	3.98	45.00	774	6.40
6	4.03	42.00	388	5.48
7	4.05	42.50	336	6.27
8	4.14	47.00	431	6.89
9	4.14	46.75	276	5.18
10	4.20	48.00	422	5.73
11	4.20	46.00	280	5.89
12	4.27	47.00	460	6.58
13	4.27	43.25	412	5.67
14	4.37	45.00	320	6.23
15	4.41	50.00	502	4.99
16	4.64	45.00	550	6.37
17	4.68	51.25	414	6.40
18	4.73	50.25	304	6.00
19	4.87	49.00	472	5.94
20	4.94	50.00	728	5.16
21	4.95	50.00	716	6.29
22	4.96	49.00	400	5.96
23	5.02	50.50	576	5.90
24	5.02	51.25	354	5.81
25	5.12	49.50	392	5.49
26	5.15	56.00	352	5.41
27	5.17	50.00	572	6.24
28	5.18	47.00	634	6.50
29	5.38	53.25	458	6.60
30	5.77	57.00	1070	4.82
31	5.90	54.00	488	5.70
32	5.90	54.00	488	5.70

We consider four analyses:

Multiple regression : $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

Regression on x_1 : $Y = \beta_0 + \beta_1 x_1 + \epsilon$

Regression on x_2 : $Y = \beta_0 + \beta_2 x_2 + \epsilon$

Regression on x_3 : $Y = \beta_0 + \beta_3 x_3 + \epsilon$

```
> mv.visc<-lm(viscosity~pcv+plasfib+plaspro,data=Viscosity)
> summary(mv.visc)

+
+ Call:
+ lm(formula = viscosity ~ pcv + plasfib + plaspro, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.4661 -0.1734  0.0285  0.1693  0.5525
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -1.378238   0.896665  -1.54    0.14
+ pcv          0.116823   0.013609   8.58 2.5e-09 ***
+ plasfib      0.000402   0.000351   1.15    0.26
+ plaspro      0.040036   0.097153   0.41    0.68
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.304 on 28 degrees of freedom
+ Multiple R-squared:  0.784, Adjusted R-squared:  0.761
+ F-statistic: 33.9 on 3 and 28 DF, p-value: 1.88e-09

> m1.visc<-lm(viscosity~pcv,data=Viscosity)
> summary(m1.visc)

+
+ Call:
+ lm(formula = viscosity ~ pcv, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.4881 -0.2215  0.0469  0.1296  0.6492
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -1.2234     0.5842   -2.09   0.045 *
+ pcv          0.1224     0.0121  10.09 3.7e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.301 on 30 degrees of freedom
+ Multiple R-squared:  0.772, Adjusted R-squared:  0.765
+ F-statistic: 102 on 1 and 30 DF, p-value: 3.73e-11
```

```

> m2.visc<-lm(viscosity~plasfib,data=Viscosity)
> summary(m2.visc)

+
+ Call:
+ lm(formula = viscosity ~ plasfib, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -1.175 -0.373 -0.122  0.351  1.219
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  3.870813   0.292450   13.24  4.6e-14 ***
+ plasfib      0.001659   0.000589    2.82  0.0085 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.561 on 30 degrees of freedom
+ Multiple R-squared:  0.209, Adjusted R-squared:  0.183
+ F-statistic: 7.93 on 1 and 30 DF, p-value: 0.0085

> m3.visc<-lm(viscosity~plaspro,data=Viscosity)
> summary(m3.visc)

+
+ Call:
+ lm(formula = viscosity ~ plaspro, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.9797 -0.4863  0.0686  0.3887  1.2330
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)    5.296      1.174    4.51  9.3e-05 ***
+ plaspro       -0.110      0.198   -0.56    0.58
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.628 on 30 degrees of freedom
+ Multiple R-squared:  0.0102, Adjusted R-squared: -0.0228
+ F-statistic: 0.31 on 1 and 30 DF, p-value: 0.582

```

```

> anova(m1.visc,mv.visc)

+ Analysis of Variance Table
+
+ Model 1: viscosity ~ pcv
+ Model 2: viscosity ~ pcv + plasfib + plaspro
+   Res.Df    RSS Df Sum of Sq    F Pr(>F)
+ 1       30 2.721
+ 2       28 2.583  2    0.1384 0.75  0.482

> anova(m2.visc,mv.visc)

+ Analysis of Variance Table
+
+ Model 1: viscosity ~ plasfib
+ Model 2: viscosity ~ pcv + plasfib + plaspro
+   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
+ 1       30 9.451
+ 2       28 2.582  2    6.869 37.24 1.29e-08 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(m3.visc,mv.visc)

+ Analysis of Variance Table
+
+ Model 1: viscosity ~ plaspro
+ Model 2: viscosity ~ pcv + plasfib + plaspro
+   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
+ 1       30 11.828
+ 2       28  2.582  2    9.246 50.12 5.59e-10 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

FACTOR PREDICTOR REGRESSION USING DUMMY VARIABLES

We can fit a factor predictor using **dummy variables**.

Suppose that a **factor predictor**, X , takes L levels, indexed by $l = 1, 2, \dots, L$. We proceed as follows:

1. Define L new "dummy" variables $X_1, \dots, X_L$, where, for $l = 1, \dots, L$,

$$X_l = \begin{cases} 1 & \text{if } X = l \\ 0 & \text{if } X \neq l \end{cases}$$

2. Fit the multiple regression model with $L - 1$ of the dummy variables as continuous co-variates, that is,

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{L-1} x_{L-1} + \epsilon_i$$

Note that we cannot include all of $X_1, X_2, \dots, X_L$ if we have an intercept β_0 in the model; by default, R omits the first level of the factor, so we omit X_1 and regard 1 as the baseline group.

EXAMPLE Diabetes Data Set (more details in upcoming example).

Data set `Diabetes` has three subgroups defined by different patient characteristics, so $L = 3$.

A subset of the data are displayed below, with the new variables X_1 , X_2 and X_3 defined as above.

ID	glutest y	group x	Dummy 1 x_1	Dummy 2 x_2	Dummy 3 x_3
1	356	3	0	0	1
2	289	3	0	0	1
3	319	3	0	0	1
4	356	3	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
87	503	2	0	1	0
88	540	2	0	1	0
89	469	2	0	1	0
90	486	2	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
113	1468	1	1	0	0
114	1487	1	1	0	0
115	714	1	1	0	0
116	1470	1	1	0	0

The three dummy variables will be used in the `lm` function.

Results: In the default parameterization:

```
> summary(lm(glutest~group,data=Diabetes))

+
+ Call:
+ lm(formula = glutest ~ group, data = Diabetes)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -489.4  -44.0   -1.7    39.3   492.6
+
+ Coefficients:
+                Estimate Std. Error t value Pr(>|t|)
+ (Intercept)      1027.4       25.7    39.9  <2e-16 ***
+ groupChemically Diabetic   -533.4       35.4   -15.1  <2e-16 ***
+ groupNormal         -677.4       30.7   -22.1  <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 146 on 141 degrees of freedom
+ Multiple R-squared:  0.777, Adjusted R-squared:  0.774
+ F-statistic: 246 on 2 and 141 DF,  p-value: <2e-16
```

Using the dummy variables:

```
> Y<-Diabetes$glutest
> X1<-as.numeric(Diabetes$group=="Overt Diabetic")
> X2<-as.numeric(Diabetes$group=="Chemically Diabetic")
> X3<-as.numeric(Diabetes$group=="Normal")
> summary(lm(Y~X2+X3))

+
+ Call:
+ lm(formula = Y ~ X2 + X3)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -489.4  -44.0   -1.7    39.3   492.6
+
+ Coefficients:
+                Estimate Std. Error t value Pr(>|t|)
+ (Intercept)      1027.4       25.7    39.9  <2e-16 ***
+ X2              -533.4       35.4   -15.1  <2e-16 ***
+ X3              -677.4       30.7   -22.1  <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 146 on 141 degrees of freedom
+ Multiple R-squared:  0.777, Adjusted R-squared:  0.774
+ F-statistic: 246 on 2 and 141 DF,  p-value: <2e-16
```

Replacing the intercept:

```
> summary(lm(glutest~-1+group,data=Diabetes))

+
+ Call:
+ lm(formula = glutest ~ -1 + group, data = Diabetes)
+
+ Residuals:
+   Min       1Q   Median       3Q      Max
+ -489.4  -44.0   -1.7    39.3   492.6
+
+ Coefficients:
+               Estimate Std. Error t value Pr(>|t|)
+ groupOvert Diabetic    1027.4     25.7   39.9  <2e-16 ***
+ groupChemically Diabetic    493.9     24.3   20.4  <2e-16 ***
+ groupNormal              350.0     16.7   20.9  <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 146 on 141 degrees of freedom
+ Multiple R-squared:  0.946, Adjusted R-squared:  0.944
+ F-statistic: 816 on 3 and 141 DF,  p-value: <2e-16
```

```
> summary(lm(Y~-1+X1+X2+X3))

+
+ Call:
+ lm(formula = Y ~ -1 + X1 + X2 + X3)
+
+ Residuals:
+   Min       1Q   Median       3Q      Max
+ -489.4  -44.0   -1.7    39.3   492.6
+
+ Coefficients:
+      Estimate Std. Error t value Pr(>|t|)
+ X1    1027.4     25.7   39.9  <2e-16 ***
+ X2     493.9     24.3   20.4  <2e-16 ***
+ X3     350.0     16.7   20.9  <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 146 on 141 degrees of freedom
+ Multiple R-squared:  0.946, Adjusted R-squared:  0.944
+ F-statistic: 816 on 3 and 141 DF,  p-value: <2e-16
```

FACTOR PREDICTOR REGRESSION

We need to take some care when combining factor predictors and covariates in the regression model. Suppose that we have only two predictors a **covariate**, x_1 and a **factor predictor**, x_2 , now taking L levels, with the levels being indexed by $l = 1, 2, \dots, L$.

EXAMPLE Binary Factor $L = 2$

Suppose that factor predictor x_2 takes two levels, labelled 0 and 1, that identify two data subgroups. Five models can be considered, that correspond to different straight-line models

- **MODEL 0** : Same intercept, slope zero, in the two subgroups
- **MODEL 1** : Different intercept, slope zero, in the two subgroups
- **MODEL 2** : Same intercept, same non-zero slope, in the two subgroups
- **MODEL 3** : Different intercept, same non-zero slope, in the two subgroups
- **MODEL 4** : Different intercept, different non-zero slopes, in the two subgroups

The general model can be written

$$Y = \begin{cases} \beta_{00} + \beta_{01}x_1 + \epsilon & \text{GROUP 0 } (l = 0) \\ \beta_{10} + \beta_{11}x_1 + \epsilon & \text{GROUP 1 } (l = 1) \end{cases}$$

- **MODEL 0** : $\beta_{00} = \beta_{10} = \beta_0, \beta_{01} = \beta_{11} = 0$
- **MODEL 1** : $\beta_{00} \neq \beta_{10}, \beta_{01} = \beta_{11} = 0$
- **MODEL 2** : $\beta_{00} = \beta_{10} = \beta_0, \beta_{01} = \beta_{11} = \beta_1 \neq 0$
- **MODEL 3** : $\beta_{00} \neq \beta_{10}, \beta_{01} = \beta_{11} = \beta_1 \neq 0$
- **MODEL 4** : $\beta_{00} \neq \beta_{10}, \beta_{01} \neq \beta_{11}$

The numbers of parameters, p , in each model are as follows:

MODEL 0	:	$p = 1$	β_0
MODEL 1	:	$p = 2$	β_{00}, β_{10}
MODEL 2	:	$p = 2$	$\beta_{00} = \beta_{10} = \beta_0, \beta_{01} = \beta_{11} = \beta_1$
MODEL 3	:	$p = 3$	$\beta_{00}, \beta_{10}, \beta_1$
MODEL 4	:	$p = 4$	$\beta_{00}, \beta_{10}, \beta_{01}, \beta_{11}$

The default parameterization used by R is different from the one described above. As in ANOVA, R takes a **baseline** group, and looks for **differences** in the parameters compared to the baseline group. The baseline group is taken to be the **first** listed subgroup for the factor predictor; in the binary example above, the baseline group would be Group 0. The interaction model is therefore

$$Y = [\beta_0 + (1 - x_2)\delta_{10}] + [(\beta_1 + (1 - x_2)\delta_{11})x_1] + \epsilon$$

- δ_{10} is the **change in intercept** from Group 0 to Group 1
- δ_{11} is the **change in slope** from Group 0 to Group 1

EXAMPLE Diabetes Data Set

The data in the data set `Diabetes` contain information on 68 diabetes patients falling into two clinically different categories (overt and chemical diabetics) and 76 normal controls. Measurements of plasma glucose in blood samples when fasting and in a dietary test are recorded.

The objective is to predict the the test glucose levels from the fasting glucose levels in the three subgroups, and to find out if there is any significant difference between the subgroups.

In this analysis, there is a single response variable, one covariate and one factor predictor:

- y : **glutest**, the test glucose level
- x_1 : covariate **glufast**, the fasting glucose level
- x_2 : factor predictor **group**, the diabetes group
 - GROUP 1: Overt Diabetic
 - GROUP 2: Chemical Diabetic
 - GROUP 3: Normal Patients

We recode the baseline group to be 'Normal Patients'.

```
> Dia.copy<-Diabetes
> Dia.copy$group2<-Dia.copy$group
> levels(Dia.copy$group2)<-c('OD','CD','NO')
> Dia.copy$group2<-
+   factor(Dia.copy$group2, levels=c('NO','CD','OD'))
```

After recoding, the data set `Dia.copy` has levels

- GROUP 1: Normal Patients (NO)
- GROUP 2: Chemical Diabetic (CD)
- GROUP 3: Overt Diabetic (OD)

We decide to carry out the analysis on the **log** (ln) scale.

```
> attach(Diabetes)
> par(mar=c(4,4,1,0))
> plot(log(glufast),log(glutest),type="n")
> points(log(glufast[group=="Normal"]),log(glutest[group=="Normal"]),
+   pch=17)
> points(log(glufast[group=="Chemically Diabetic"]),
+   log(glutest[group=="Chemically Diabetic"]),pch=4)
> points(log(glufast[group=="Overt Diabetic"]),
+   log(glutest[group=="Overt Diabetic"]),pch=3)
> legend(4.2,7.3,c("Normal","Chem. Diabetic","Overt Diabetic"),
+   pch=c(17,4,3))
```

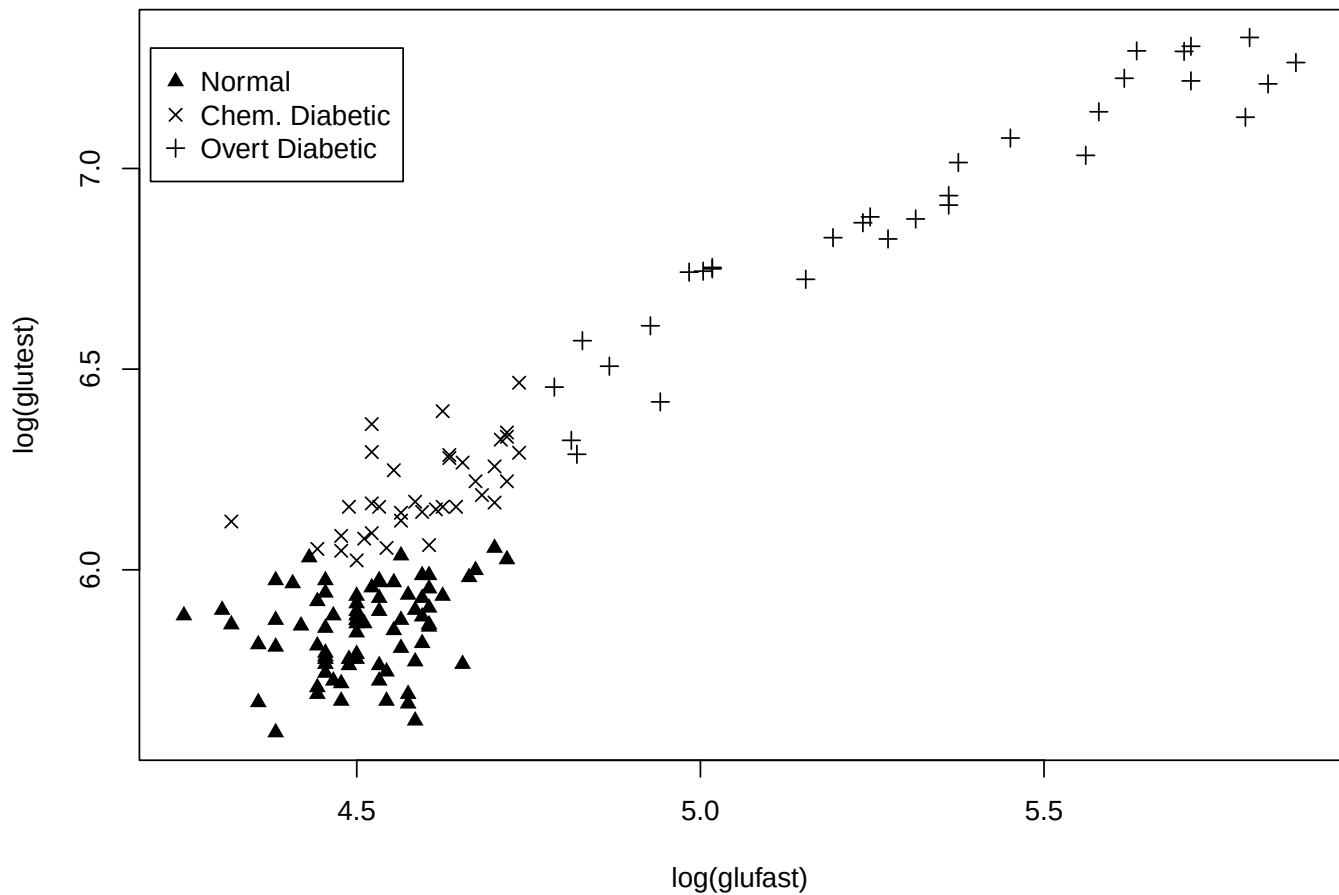


Figure 25: Diabetes data: raw data broken down by group

```
> detach(Diabetes)
```

```
> m.Diabetes<-lm(log(glutest)~log(glufast)*group2,data=Dia.copy)
>
> print(summary(aov(m.Diabetes)), digits = 5)

+               Df  Sum Sq Mean Sq    F value    Pr(>F)
+ log(glufast)      1 24.7664 24.7664 2593.8314 < 2.2e-16 ***
+ group2            2  2.2656  1.1328  118.6414 < 2.2e-16 ***
+ log(glufast):group2 2  0.1547  0.0773   8.0994 0.0004721 ***
+ Residuals       138  1.3177  0.0095
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(m.Diabetes)

+
+ Call:
+ lm(formula = log(glutest) ~ log(glufast) * group2, data = Dia.copy)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.2510 -0.0628  0.0067  0.0690  0.2155
```

```

+
+ Coefficients:
+               Estimate Std. Error t value Pr(>|t|)
+ (Intercept)         4.504      0.559    8.06 3.2e-13 ***
+ log(glufast)         0.299      0.124    2.41 0.0171 *
+ group2CD            -1.436      0.958   -1.50 0.1362
+ group2OD            -2.037      0.619   -3.29 0.0013 **
+ log(glufast):group2CD  0.382      0.210    1.82 0.0709 .
+ log(glufast):group2OD  0.535      0.134    4.00 0.0001 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.0977 on 138 degrees of freedom
+ Multiple R-squared:  0.954, Adjusted R-squared:  0.952
+ F-statistic: 569 on 5 and 138 DF,  p-value: <2e-16

```

The first ANOVA table demonstrates that there **is** a significant interaction between the covariate and the factor predictor ($F_{obs} = 8.099$, p-value 0.000472). This means that there is a **significantly different slope in at least two of the three subgroups**.

The second table gives the slope and intercept parameters in the three groups. The R parameterization is not directly in terms of the slopes and intercepts, but looks at **differences** from baseline subgroup, Group 1. For example, the Group 3 (Overt Diabetic) intercept and slope are

$$\text{INTERCEPT : } 4.504 + (-2.037) = 2.467 \quad \text{SLOPE : } 0.2991 + 0.5349 = 0.8340.$$

```

> attach(Diabetes)
> par(mfrow=c(2,2), mar=c(4,3,1,1))
> plot(log(glufast), log(glutest), type="n")
> points(log(glufast[group=="Normal"]), log(glutest[group=="Normal"]))
+       ,pch=17)
> be0<-coef(m.Diabetes)[1]; be1<-coef(m.Diabetes)[2]
> abline(be0,be1)
> title("Normal")
> legend(4.2,7.3,c("Normal"),pch=c(17))
>
> plot(log(glufast), log(glutest), type="n")
> points(log(glufast[group=="Chemically Diabetic"]),
+       log(glutest[group=="Chemically Diabetic"]),pch=4)
> be0<-coef(m.Diabetes)[1]+coef(m.Diabetes)[3];
> be1<-coef(m.Diabetes)[2]+coef(m.Diabetes)[5]
> abline(be0,be1)
> title("Chemically Diabetic")
> legend(4.2,7.3,c("Chem. Diabetic"),pch=c(4))
>
> plot(log(glufast), log(glutest), type="n")
> points(log(glufast[group=="Overt Diabetic"]),
+       log(glutest[group=="Overt Diabetic"]),pch=3)
> be0<-coef(m.Diabetes)[1]+coef(m.Diabetes)[4]
> be1<-coef(m.Diabetes)[2]+coef(m.Diabetes)[6]
> abline(be0,be1)
> title("Overt Diabetic")

```

```

> legend(4.2, 7.3, c("Overt Diabetic"), pch=c(3))
>
> plot(log(glufast), log(glutest), type="n")
> be0<-coef(m.Diabetes)[1]; be1<-coef(m.Diabetes)[2]
> abline(be0, be1)
> be0<-coef(m.Diabetes)[1]+coef(m.Diabetes)[3]
> be1<-coef(m.Diabetes)[2]+coef(m.Diabetes)[5]
> abline(be0, be1, lty=2)
> be0<-coef(m.Diabetes)[1]+coef(m.Diabetes)[4]
> be1<-coef(m.Diabetes)[2]+coef(m.Diabetes)[6]
> abline(be0, be1, lty=3)
> legend(4.2, 7.3, c("Normal", "Chem. Diabetic", "Overt Diabetic"),
+       lty=c(1, 2, 3))

```

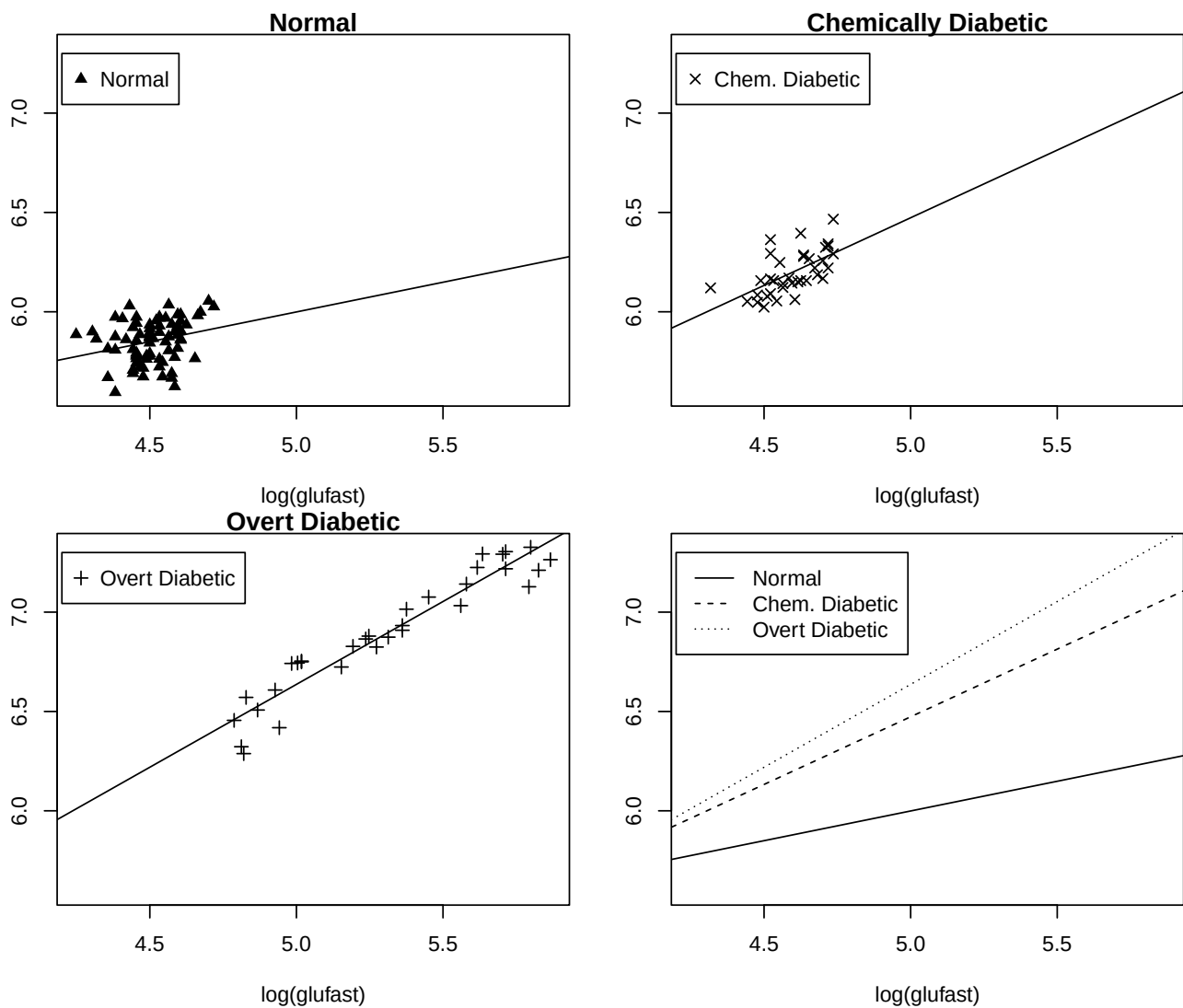


Figure 26: Diabetes data: fit broken down by group

```

> detach(Diabetes)

```

```

> drop1(m.Diabetes, test="F")

+ Single term deletions
+
+ Model:
+ log(glutest) ~ log(glufast) * group2
+
+               Df Sum of Sq   RSS   AIC F value    Pr(>F)
+ <none>                        1.318 -663.9
+ log(glufast):group2  2     0.1547 1.472 -651.9    8.099 0.000472 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The models that can be fitted are as follows:

```

MODEL 0 : Null model
MODEL 1 : Effect of group only
MODEL 2 : Effect of log(glufast) only
MODEL 3 : Additive main effect of group and log(glufast)
MODEL 4 : Main effects plus interaction

```

```

> m0.diabetes<-lm(log(glutest)~1,data=Dia.copy)
> m1.diabetes<-lm(log(glutest)~group2,data=Dia.copy)
> m2.diabetes<-lm(log(glutest)~log(glufast),data=Dia.copy)
> m3.diabetes<-lm(log(glutest)~group2+log(glufast),data=Dia.copy)
> m4.diabetes<-lm(log(glutest)~group2*log(glufast),data=Dia.copy)
>
> print(summary(aov(m0.diabetes)), digits = 5)

+               Df Sum Sq Mean Sq F value Pr(>F)
+ Residuals    143  28.504  0.19933

> print(summary(aov(m1.diabetes)), digits = 5)

+               Df Sum Sq Mean Sq F value    Pr(>F)
+ group2          2 24.344 12.1722  412.57 < 2.2e-16 ***
+ Residuals    141   4.160   0.0295
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> print(summary(aov(m2.diabetes)), digits = 5)

+               Df Sum Sq Mean Sq F value    Pr(>F)
+ log(glufast)    1 24.7664 24.7664  940.85 < 2.2e-16 ***
+ Residuals    142   3.7379   0.0263
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> print(summary(aov(m3.diabetes)), digits = 5)

```

```

+           Df  Sum Sq Mean Sq F value    Pr(>F)
+ group2      2 24.3443 12.1722 1157.43 < 2.2e-16 ***
+ log(glufast) 1  2.6877  2.6877  255.56 < 2.2e-16 ***
+ Residuals   140  1.4723  0.0105
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> print(summary(aov(m4.diabetes)), digits = 5)

+           Df  Sum Sq Mean Sq    F value    Pr(>F)
+ group2      2 24.3443 12.1722 1274.8150 < 2.2e-16 ***
+ log(glufast) 1  2.6877  2.6877  281.4841 < 2.2e-16 ***
+ group2:log(glufast) 2  0.1547  0.0773    8.0994 0.0004721 ***
+ Residuals   138  1.3177  0.0095
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

EXAMPLE Potato damage data

The potato damage data is an unbalanced three-way factorial design:

```
> head(Potato)

+      potato      regime  temp  photo  leak
+ 1 Variety 1 Room Temp  -4 C  4.6386  2.25
+ 2 Variety 1 Room Temp  -4 C  5.7914  4.34
+ 3 Variety 1 Room Temp  -4 C 29.3515  4.25
+ 4 Variety 1 Room Temp  -4 C 18.4173  6.14
+ 5 Variety 1 Room Temp  -4 C  2.2556  2.38
+ 6 Variety 1 Room Temp  -8 C  8.2358 16.30

> Potato.counts<-xtabs(~potato+regime+temp, data=Potato)
> ftable(Potato.counts)

+              temp  -4 C  -8 C
+ potato      regime
+ Variety 1 Room Temp          5    5
+              Cold Room      12   13
+ Variety 2 Room Temp          13   13
+              Cold Room       7    7

> levels(Potato$potato)<-c('V1','V2')
> levels(Potato$regime)<-c('Room','Cold')
> levels(Potato$temp)<-c('-4','-8')
```

```
> m.potato<-lm(leak~potato*regime*temp,data=Potato)
> summary(m.potato)

+
+ Call:
+ lm(formula = leak ~ potato * regime * temp, data = Potato)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -22.151  -3.325  -0.169   3.288  18.589
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)         3.87        3.85   1.01  0.31833
+ potatoV2            18.20        4.53   4.02  0.00015 ***
+ regimeCold          -1.53        4.58  -0.33  0.73913
+ temp-8              2.06        5.45   0.38  0.70673
+ potatoV2:regimeCold -18.12        6.11  -2.97  0.00417 **
+ potatoV2:temp-8      8.19        6.41   1.28  0.20567
+ regimeCold:temp-8    6.59        6.45   1.02  0.31065
+ potatoV2:regimeCold:temp-8 -9.44        8.61  -1.10  0.27670
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 8.61 on 67 degrees of freedom
+ Multiple R-squared:  0.64, Adjusted R-squared:  0.603
+ F-statistic: 17 on 7 and 67 DF, p-value: 1.02e-12
```

```

> print(summary(aov(m.potato)), digits = 5)
+
+               Df Sum Sq Mean Sq F value    Pr(>F)
+ potato          1 3420.2   3420.2  46.1178 3.578e-09 ***
+ regime          1 1951.8   1951.8  26.3173 2.674e-06 ***
+ temp           1 1255.6   1255.6  16.9302 0.0001084 ***
+ potato:regime    1 2089.9   2089.9  28.1805 1.349e-06 ***
+ potato:temp      1   28.9     28.9   0.3894 0.5347194
+ regime:temp      1    6.8      6.8   0.0916 0.7631284
+ potato:regime:temp 1   89.2     89.2   1.2027 0.2767010
+ Residuals       67 4968.9     74.2
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> drop1(m.potato, test="F")
+ Single term deletions
+
+ Model:
+ leak ~ potato * regime * temp
+               Df Sum of Sq  RSS    AIC F value Pr(>F)
+ <none>                                4969 330.5
+ potato:regime:temp 1         89.2 5058 329.8    1.203  0.277

> #Can drop 3-way interaction
>
> m1.potato<-update(m.potato, ~.-potato:regime:temp)
>
> drop1(m1.potato, test="F")
+ Single term deletions
+
+ Model:
+ leak ~ potato + regime + temp + potato:regime + potato:temp +
+       regime:temp
+               Df Sum of Sq  RSS    AIC F value    Pr(>F)
+ <none>                                5058 329.8
+ potato:regime 1         2093.2 7151 353.8   28.140 1.33e-06 ***
+ potato:temp   1           35.4 5094 328.4    0.477   0.492
+ regime:temp   1            6.8 5065 327.9    0.091   0.763
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> #Can drop 2-way interaction regime:temp
>
> m2.potato<-update(m1.potato, ~.-regime:temp)
> drop1(m2.potato, test="F")
+ Single term deletions
+
+ Model:
+ leak ~ potato + regime + temp + potato:regime + potato:temp
+               Df Sum of Sq  RSS    AIC F value    Pr(>F)
+ <none>                                5065 327.9

```

```

+ potato:regime 1      2094.6 7159 351.9 28.535 1.12e-06 ***
+ potato:temp 1      28.9 5094 326.4 0.393 0.533
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> #Can drop 2-way interaction potato:temp
>
> m3.potato<-update(m2.potato,~.-potato:temp)
> drop1(m3.potato,test="F")

+ Single term deletions
+
+ Model:
+ leak ~ potato + regime + temp + potato:regime
+           Df Sum of Sq  RSS    AIC F value    Pr(>F)
+ <none>                5094 326.4
+ temp      1      1226 6319 340.5    16.84 0.000108 ***
+ potato:regime 1      2090 7184 350.2    28.72 1.02e-06 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(m3.potato)

+
+ Call:
+ lm(formula = leak ~ potato + regime + temp + potato:regime, data = Potato)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -21.070  -4.346  -0.008   3.772  19.670
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)      0.857      2.872   0.30  0.76619
+ potatoV2        22.296      3.174   7.02  1.1e-09 ***
+ regimeCold       1.771      3.192   0.55  0.58082
+ temp-8           8.087      1.971   4.10  0.00011 ***
+ potatoV2:regimeCold -22.854      4.264  -5.36  1.0e-06 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 8.53 on 70 degrees of freedom
+ Multiple R-squared:  0.631, Adjusted R-squared:  0.61
+ F-statistic: 29.9 on 4 and 70 DF, p-value: 1.59e-14

> print(summary(aov(m3.potato)), digits = 5)

+           Df Sum Sq Mean Sq F value    Pr(>F)
+ potato      1 3420.2  3420.2  47.002 2.286e-09 ***
+ regime      1 1951.8  1951.8  26.822 2.048e-06 ***
+ temp        1 1255.6  1255.6  17.255 9.096e-05 ***
+ potato:regime 1 2089.9  2089.9  28.721 1.015e-06 ***
+ Residuals   70 5093.7    72.8
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

EXAMPLE Task Distraction data

The task distraction data is a balanced two-way factorial design with a continuous covariate:

```
> head(TaskDistract)

+      Group          Task Distract Errors
+ 1 NonSmoker Pattern Recognition      107      9
+ 2 NonSmoker Pattern Recognition      133      8
+ 3 NonSmoker Pattern Recognition      123     12
+ 4 NonSmoker Pattern Recognition       94     10
+ 5 NonSmoker Pattern Recognition       83      7
+ 6 NonSmoker Pattern Recognition       86     10

> TD<-TaskDistract
> levels(TD$Group)<-c('Non','Del','Act')
> levels(TD$Task)<-c('Pat','Cog','Dri')
> TD.counts<-xtabs(~Group+Task, data=TD)
> ftable(TD.counts)

+      Task Pat Cog Dri
+ Group
+ Non          15  15  15
+ Del          15  15  15
+ Act          15  15  15

> m.Task<-lm(Errors~Group*Task*Distract,data=TD)
> print(summary(aov(m.Task)), digits = 5)

+              Df  Sum Sq Mean Sq  F value    Pr(>F)
+ Group          2    354.5   177.3    3.6645  0.028601 *
+ Task           2 28661.5 14330.8 296.2361 < 2.2e-16 ***
+ Distract       1  5747.0  5747.0 118.7986 < 2.2e-16 ***
+ Group:Task     4   1626.5   406.6    8.4055 5.374e-06 ***
+ Group:Distract 2    496.1   248.1    5.1277 0.007334 **
+ Task:Distract  2   2643.3  1321.6   27.3201 1.835e-10 ***
+ Group:Task:Distract 4    142.9    35.7    0.7386 0.567489
+ Residuals    117  5660.0    48.4
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> drop1(m.Task,test="F")

+ Single term deletions
+
+ Model:
+ Errors ~ Group * Task * Distract
+              Df Sum of Sq  RSS    AIC F value Pr(>F)
+ <none>                5660 540.3
+ Group:Task:Distract  4    142.9 5803 535.7    0.739  0.567

> #Can drop 3-way interaction
>
> m1.Task<-update(m.Task,~.-Group:Task:Distract)
> drop1(m1.Task,test="F")
```

```

+ Single term deletions
+
+ Model:
+ Errors ~ Group + Task + Distract + Group:Task + Group:Distract +
+ Task:Distract
+
+ Df Sum of Sq RSS AIC F value Pr(>F)
+ <none> 5803 535.7
+ Group:Task 4 1824.5 7627 564.6 9.511 1.01e-06 ***
+ Group:Distract 2 414.4 6217 541.0 4.320 0.0154 *
+ Task:Distract 2 2643.3 8446 582.4 27.558 1.37e-10 ***
+ ---
+ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> #Maybe drop Group:Distract
> m2.Task<-update(m1.Task,~.-Group:Distract)
> drop1(m2.Task,test="F")

```

```

+ Single term deletions
+
+ Model:
+ Errors ~ Group + Task + Distract + Group:Task + Task:Distract
+
+ Df Sum of Sq RSS AIC F value Pr(>F)
+ <none> 6217 541.0
+ Group:Task 4 1754 7971 566.6 8.675 3.35e-06 ***
+ Task:Distract 2 2725 8942 586.1 26.955 1.96e-10 ***
+ ---
+ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

> summary(m1.Task)

+
+ Call:
+ lm(formula = Errors ~ Group + Task + Distract + Group:Task +
+ Group:Distract + Task:Distract, data = TD)
+
+ Residuals:
+ Min 1Q Median 3Q Max
+ -16.267 -3.019 -0.467 3.807 23.022
+
+ Coefficients:
+ Estimate Std. Error t value Pr(>|t|)
+ (Intercept) 7.67787 8.62606 0.89 0.37519
+ GroupDel -20.44543 8.96349 -2.28 0.02430 *
+ GroupAct 0.66186 8.75109 0.08 0.93984
+ TaskCog -33.54794 8.45784 -3.97 0.00012 ***
+ TaskDri -2.54922 9.59524 -0.27 0.79094
+ Distract 0.01550 0.07593 0.20 0.83864
+ GroupDel:TaskCog 13.68795 3.59278 3.81 0.00022 ***
+ GroupAct:TaskCog 13.92341 3.64937 3.82 0.00022 ***
+ GroupDel:TaskDri -1.57046 3.65903 -0.43 0.66854
+ GroupAct:TaskDri -7.42416 3.72894 -1.99 0.04874 *
+ GroupDel:Distract 0.18410 0.07711 2.39 0.01852 *

```

```

+ GroupAct:Distract -0.00149 0.07466 -0.02 0.98408
+ TaskCog:Distract 0.45529 0.07133 6.38 3.3e-09 ***
+ TaskDri:Distract 0.02543 0.08105 0.31 0.75425
+ ---
+ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 6.93 on 121 degrees of freedom
+ Multiple R-squared: 0.872, Adjusted R-squared: 0.858
+ F-statistic: 63.4 on 13 and 121 DF, p-value: <2e-16

> print(summary(aov(m1.Task)), digits = 5)

+              Df Sum Sq Mean Sq F value Pr(>F)
+ Group          2  354.5   177.3   3.6964 0.027656 *
+ Task           2 28661.5 14330.8 298.8182 < 2.2e-16 ***
+ Distract       1  5747.0   5747.0 119.8341 < 2.2e-16 ***
+ Group:Task     4  1626.5   406.6   8.4788 4.587e-06 ***
+ Group:Distract 2   496.1   248.1   5.1723 0.006991 **
+ Task:Distract  2  2643.3  1321.6  27.5582 1.373e-10 ***
+ Residuals     121  5802.9    48.0
+ ---
+ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m3.Task<-lm(Errors~Group+Task+Distract,data=TD)
> summary(m3.Task)

+
+ Call:
+ lm(formula = Errors ~ Group + Task + Distract, data = TD)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -25.67  -5.51  -1.07    5.09   27.02
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -28.8305     4.6692  -6.17    8e-09 ***
+ GroupDel      4.4677     1.9197   2.33    0.022 *
+ GroupAct      4.4007     1.9093   2.30    0.023 *
+ TaskCog     27.5102     1.9180  14.34 <2e-16 ***
+ TaskDri     -1.8414     1.9160  -0.96    0.338
+ Distract      0.3162     0.0378   8.38    8e-14 ***
+ ---
+ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 9.05 on 129 degrees of freedom
+ Multiple R-squared: 0.767, Adjusted R-squared: 0.758
+ F-statistic: 84.9 on 5 and 129 DF, p-value: <2e-16

```

PREDICTION FROM GENERAL LINEAR MODELS

Consider the analysis presented includes in a summary of parameter estimates obtained for the Task Distraction data. The model being fitted is

$$\text{Group} + \text{Task} + \text{Distract} + \text{Group.Task} + \text{Task.Distract}$$

where

- `Group` is a factor predictor with three levels
- `Task` is a factor predictor with three levels
- `Distract` is a covariate

We know that `Group` and `Task` are factor predictors as the entries in the parameter estimates table appear in the form displaying a parameter estimate for each level. etc. R takes the first level for each factor to be the baseline level; in this case, each factor has three levels, so the baseline treatment group is the one where

$$\text{Group} = \text{'NonSmoker'} \quad \text{and} \quad \text{Task} = \text{'Pattern Recognition'}$$

To make a prediction from this model, we need to combine the parameter estimates. We do this by considering the model formula, and extracting one numerical value for each term, remembering to also add in the intercept.

Suppose we want to make predictions for

- **Prediction A:** Subject in Group "NonSmoker", doing Task "Driving Simulation", with Distraction level 180.
- **Prediction B:** Subject in Group "Delayed Smoker", doing Task "Cognitive Task", with Distraction level 180.
- **Prediction C:** Subject in Group "Active Smoker", doing Task "Pattern Recognition", with Distraction level 100.

We can go through term by term in the model, as in the following table :

Term	Prediction A	Prediction B	Prediction C
Intercept	7.678	7.678	7.678
Group	0.000	-20.445	0.662
Task	-2.549	-33.548	0.000
Distract	0.015×180	0.015×180	0.015×100
Group.Task	0.000	13.688	0.000
Task.Distract	0.025×180	0.455×180	0.000×100
Total	12.495	85.251	9.740

This can be achieved in R using the `predict` function, and a `newdata` data frame that specifies the values at which the predictions are made. We may compute confidence intervals for the predicted value excluding or including the residual variation.

```

> newTask<-data.frame(Group=c("Non", "Del", "Act"),
+   Task=c("Dri", "Cog", "Pat"), Distract=c(180, 180, 100))
> predict(m1.Task, newdata=newTask, interval="confidence")

+   fit      lwr      upr
+ 1 12.49529  2.71407 22.2765
+ 2 85.25114 76.35698 94.1453
+ 3  9.74008  5.74588 13.7343

> predict(m1.Task, newdata=newTask, interval="prediction")

+   fit      lwr      upr
+ 1 12.49529 -4.34640 29.3370
+ 2 85.25114 68.90868 101.5936
+ 3  9.74008 -4.54011 24.0203

```

Notes:

- when you include covariate terms, remember to multiply the coefficient by the value at which you are predicting; here we multiply the coefficients involving Distract by 180 (in Predictions A and B) and 100 (in Prediction C).
- this approach of going through term by term works for all models. For example, if we fitted the model

Group + Task + Distract

which is model m3.Task, and obtained the parameter estimates

```

> summary(m3.Task)

+
+ Call:
+ lm(formula = Errors ~ Group + Task + Distract, data = TD)
+
+ Residuals:
+   Min       1Q   Median       3Q      Max
+ -25.67  -5.51  -1.07    5.09   27.02
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -28.8305    4.6692   -6.17    8e-09 ***
+ GroupDel      4.4677    1.9197    2.33    0.022 *
+ GroupAct      4.4007    1.9093    2.30    0.023 *
+ TaskCog      27.5102    1.9180   14.34   <2e-16 ***
+ TaskDri     -1.8414    1.9160   -0.96    0.338
+ Distract      0.3162    0.0378    8.38    8e-14 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 9.05 on 129 degrees of freedom
+ Multiple R-squared:  0.767, Adjusted R-squared:  0.758
+ F-statistic: 84.9 on 5 and 129 DF,  p-value: <2e-16

```

then for a prediction in Group 'Delayed Smoker', Task 'Driving Simulation' and

Distraction 140, we would have

$$-28.830 + 4.466 + (-1.841) + (0.316 \times 140) = 18.035$$

For predictions A, B and C, we have

```
> predict(m3.Task, newdata=newTask, interval="confidence")

+      fit      lwr      upr
+ 1 26.2425 19.99076 32.4942
+ 2 60.0619 54.06667 66.0570
+ 3  7.1893  3.60742 10.7712

> predict(m3.Task, newdata=newTask, interval="prediction")

+      fit      lwr      upr
+ 1 26.2425  7.27409 45.2109
+ 2 60.0619 41.17647 78.9472
+ 3  7.1893 -11.07393 25.4525
```

Note that predictions from the full factorial model

Group*Task*Distract

(fitted as m.Task) do not differ greatly from the predictions in the selected model.

```
> predict(m.Task, newdata=newTask, interval="confidence")

+      fit      lwr      upr
+ 1 12.88873 -0.195647 25.9731
+ 2 88.54081 77.868781 99.2128
+ 3  8.46897  4.122008 12.8159

> predict(m.Task, newdata=newTask, interval="prediction")

+      fit      lwr      upr
+ 1 12.88873 -6.10970 31.8872
+ 2 88.54081 71.11578 105.9658
+ 3  8.46897 -5.97525 22.9132
```

- Making predictions near and beyond the end of the Distraction range might lead to inappropriate (for example, negative-valued) predictions. Here the minimum value of Distraction is 64, the maximum is 191.

SUMMARY OF ISSUES IN ANOVA, REGRESSION AND LINEAR MODELLING

1. **Model Assumptions :** The key model assumption is that the residual (measurement) errors are independent and identically distributed Normal random quantities. If this assumption is not met, then none of the hypothesis tests based on the t- and F-distributions are valid.

The validity of this model assumption can be checked by the inspection of the *residuals*, $\hat{e}_i$, or *standardized residuals*, $\hat{z}_i$ where

$$\hat{e}_i = y_i - \hat{y}_i \qquad \hat{z}_i = \frac{\hat{e}_i}{s} = \frac{y_i - \hat{y}_i}{s}$$

for $i = 1, \dots, n$. Plots of the residuals can be used to check for

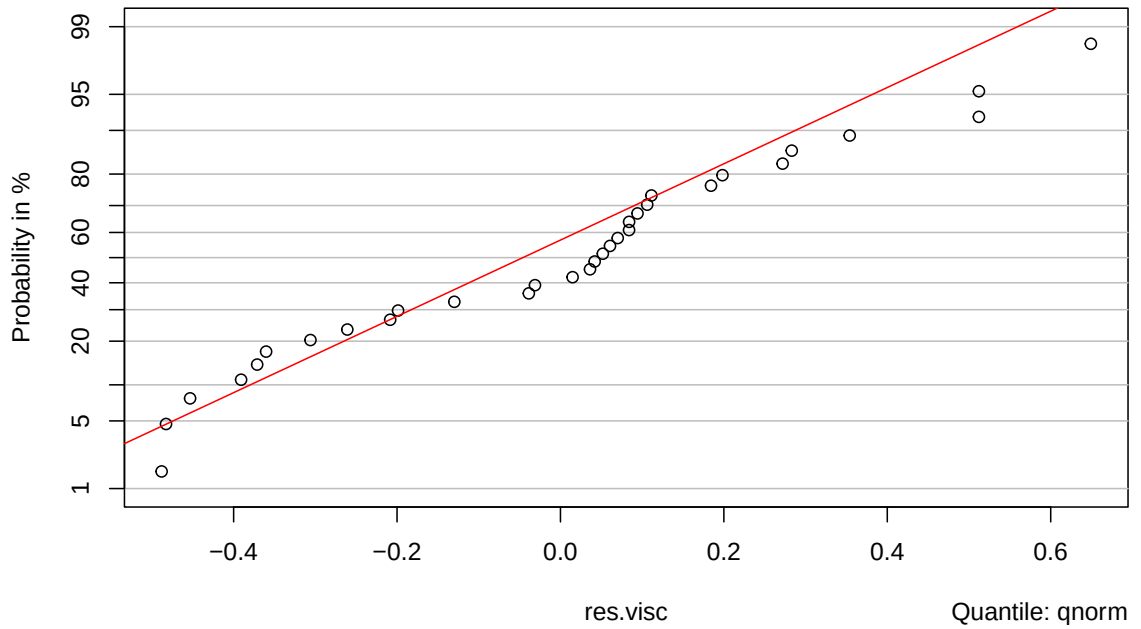
- (i) **Normality** To check the normality of the residuals, a *histogram* or *probability plot* can be used. A probability plot is a plot constructed using the **observed** (standardized) residuals and their **theoretical** counterparts **assuming a normal model**. Points in such a plot should lie on a straight-line with slope one; any deviation from this may indicate deviation from normality.

The difference between P-P plots and Q-Q plots is that the P-P plots use the observed cumulative probability (that is, plots the observed residuals against the **rank** of the observed residuals divided by $n + 1$), whereas Q-Q plots plot theoretical quantiles against observed quantiles from the residual distribution (that is, the $i/(n + 1)$ quantiles for $i = 1, \dots, n$ against the observed residuals sorted into ascending order).

The examples below are from the Viscosity vs PCV data, and the straight-line and quadratic model analysis of the Hooker data. For the Viscosity data, the probability plot indicates that the residual variance is larger in the middle compared to the of the residual range. For the Hooker data straight-line fit, the points in the probability plot do not lie on the straight-line, so again a deviation from normality is indicated; for the quadratic fit, the plot indicates normality of the residuals.

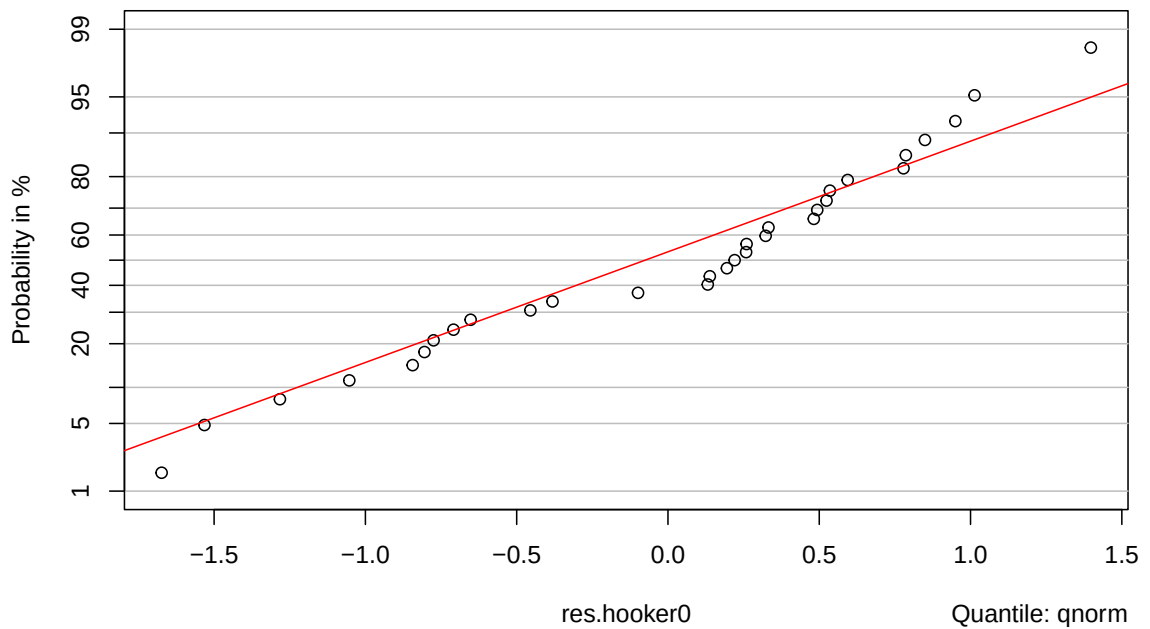
```
> m0.visc<-lm(viscosity~pcv,data=Viscosity)
> m0.hooker<-lm(temp~pressure,data=Hooker)
> m1.hooker<-lm(temp~pressure+I(pressure^2),data=Hooker)
>
> res.visc<-residuals(m0.visc)
> res.hooker0<-residuals(m0.hooker)
> res.hooker1<-residuals(m1.hooker)
>
> #P-P plots
> #library(e1071)
> e1071::probplot(res.visc)
> title('P-P plot for Viscosity residuals')
```

P-P plot for Viscosity residuals



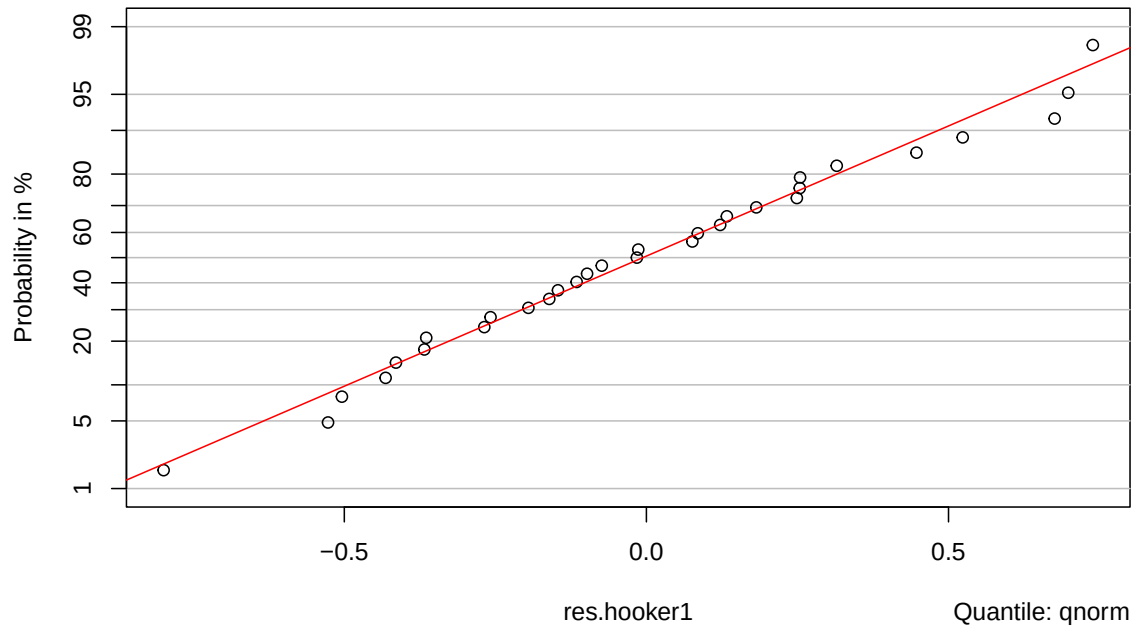
```
> e1071::probplot(res.hooker0)
> title('P-P plot for Linear fit to Hooker data')
```

P-P plot for Linear fit to Hooker data



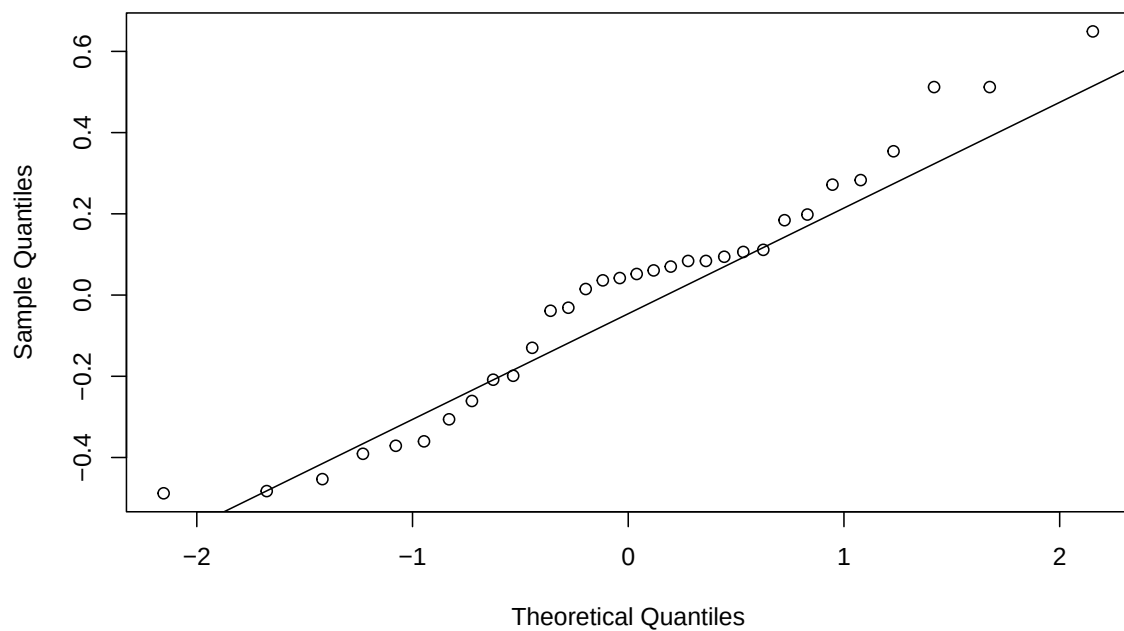
```
> e1071::probplot(res.hooker1)
> title('P-P plot for Quadratic fit to Hooker data')
```

P-P plot for Quadratic fit to Hooker data



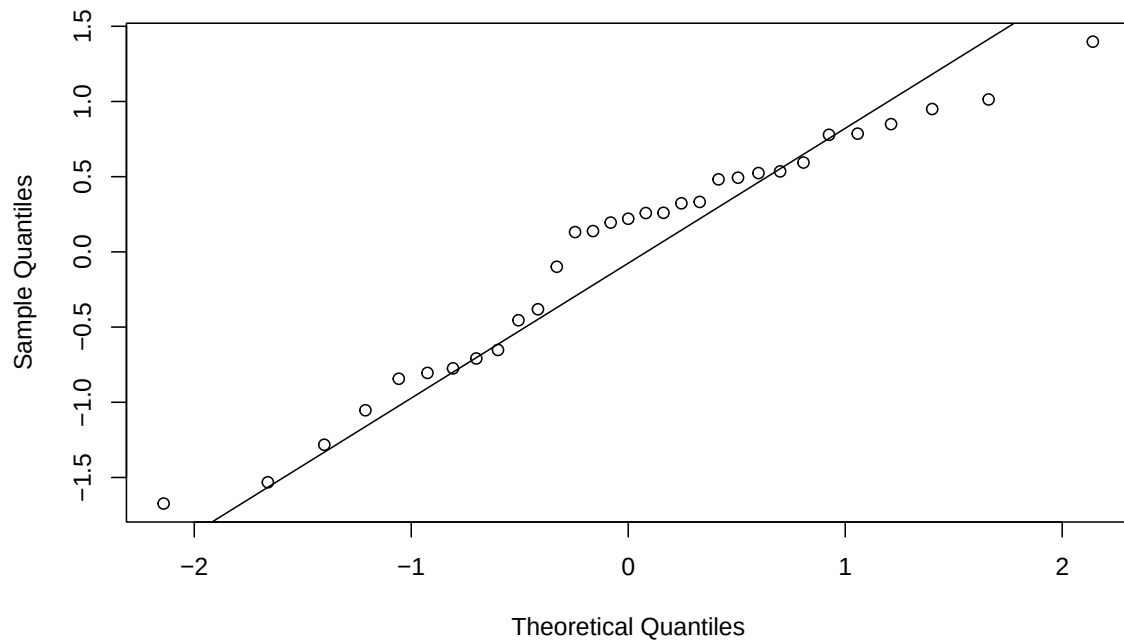
```
> #Q-Q plots
> qqnorm(res.visc,main='Viscosity data: Q-Q plot for residuals')
> qqline(res.visc)
```

Viscosity data: Q-Q plot for residuals



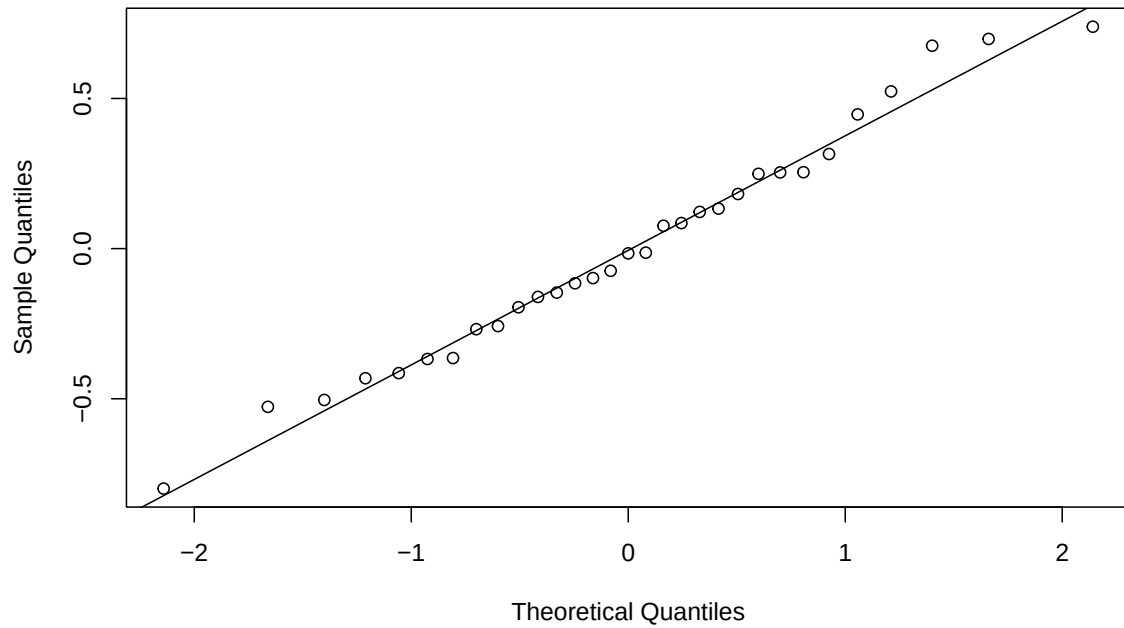
```
> qqnorm(res.hooker0,main='Hooker data: Q-Q plot for Linear fit')
> qqline(res.hooker0)
```

Hooker data: Q-Q plot for Linear fit



```
> qqnorm(res.hooker1, main='Hooker data: Q-Q plot for Quadratic fit ')\n> qqline(res.hooker1)
```

Hooker data: Q-Q plot for Quadratic fit



(ii) *Dependence on the covariates*

(iii) *Constant variance*

(iv) *Outliers*: An *outlier* is an response value that gives rise to a residual which is large in magnitude, indicating that the fit of the model is poor for that data point.

Outliers can significantly alter the fit of a model, and the parameter estimates. If an outlier is suspected, then careful consideration should be given to omitting that data point from the analysis (see below; estimates (standard errors) change in the presence of an outlier).

2. **Data Transformations** : The response variable and continuous covariates can be **transformed** (using log or square-root transformation say) to improve the fit of the model, or to make the model assumptions more appropriate.
3. **Model Selection** : Model selection by means of stepwise selection and sequential ANOVA-F testing can be an effective way of finding the important explanatory variables and interactions. However it must be carried out with care. In general, we aim to select the **simplest model** that provides an adequate fit to the data. The goodness of fit measures R^2 and adjusted R^2 statistics can provide a final assessment of model adequacy.
4. **Multicollinearity** : *Multicollinearity* is the term to describe dependence between the covariates used in a regression model. If the covariates are highly correlated, then the estimated coefficients for those covariates in a multiple regression need careful interpretation.

If two covariates are highly correlated, then if one is a useful predictor of the response, the other will likely appear to be a useful predictor as well, that is if one estimated coefficient is significantly different from zero, then the other will be also. However, in a multiple regression model with both covariates included, it might be that neither coefficient is significantly different from zero.

5. **Predicting outside the Range of the Covariates** : In a regression model, the fitted parameters reflect relationships and dependencies in the *observed* data. The model can be used for prediction, but is only likely to be reliable if the prediction is carried out at x values within the range of the observed x s.

For example, in a simple linear regression, if x takes values on the range $(0, 100)$, predictions at new x values within this range will be reliable, but predictions at, say, $x = 200$ will be much less reliable.

CHI-SQUARED TESTS FOR CATEGORICAL DATA

In a **multinomial** experiment, the independent experimental units are classified to one of k categories determined by the levels of a discrete factor. Let $n_1, n_2, \dots, n_k$ be the counts of the numbers of experimental units in the k categories, where $n_1 + n_2 + \dots + n_k = n$.

The probability that an experimental unit is classified to category i is p_i , for $i = 1, \dots, k$, so that

$$p_1 + p_2 + \dots + p_k = 1.$$

- The **one-way** classification table can be displayed as follows:

Category	1	2	$\dots$	k
Count	n_1	n_2	$\dots$	n_k
Probability	p_1	p_2	$\dots$	p_k

We can test a hypothesis H_0 that fully specifies $p_1, \dots, p_k$, for example

$$H_0 : p_1 = p_1^{(0)}, p_2 = p_2^{(0)}, \dots, p_k = p_k^{(0)}$$

so that, for $k = 3$, we might have

$$H_0 : p_1 = p_2 = p_3 = 1/3 \quad \text{or} \quad H_0 : p_1 = 1/2, p_2 = p_3 = 1/4.$$

We use the test statistic

$$X^2 = \sum_{i=1}^k \frac{(n_i - np_i^{(0)})^2}{np_i^{(0)}} = \sum_{i=1}^k \frac{(\text{Observed Count in Cell } i - \text{Expected Count in Cell } i)^2}{\text{Expected Count in Cell } i}$$

We sometimes write $\hat{n}_i = np_i^{(0)}$. If H_0 is true, $X^2 \approx \chi_{k-1}^2$.

- The **two-way** classification table can also be constructed to represent the cross-classification for two discrete factors A and B with r and c levels respectively.

		Factor B			
		1	2	$\dots$	c
Factor A	1	n_{11}	n_{12}	$\dots$	n_{1c}
	2	n_{21}	n_{22}	$\dots$	n_{2c}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	r	n_{r1}	n_{r2}	$\dots$	n_{rc}

To test the hypothesis

$$H_0 : \text{Factor A and Factor B levels are assigned independently}$$

we use the same test statistic that can be rewritten

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

where

$$\hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n} \quad n_{i.} = \sum_{j=1}^c n_{ij} \quad n_{.j} = \sum_{i=1}^r n_{ij}.$$

The terms $n_{i.}$ and $n_{.j}$ are the row and column totals for row i and column j respectively. If H_0 is true

$$X^2 \approx \chi_{(r-1)(c-1)}^2$$

EXAMPLE DNA Sequence Data

The counts of the numbers of nucleotides (A,C,G,T) in the DNA sequence of the cancer-related gene BRCA 2 are presented in the table below.

Category	1	2	3	4	Total
Nucleotide	A	C	G	T	
Count	38514	24631	25685	38249	127079

so that $k = 4$. To test the hypothesis

$$H_0 : p_1 = p_2 = p_3 = p_4 = 1/4$$

We use the one-way table chi-squared test: here

$$\hat{n}_i = np_i^{(0)} = \frac{127079}{4} = 31769.75$$

so the test statistic is

$$\begin{aligned} X^2 &= \frac{(38514 - 31769.75)^2}{31769.75} + \frac{(24631 - 31769.75)^2}{31769.75} + \frac{(25685 - 31769.75)^2}{31769.75} + \frac{(38249 - 31769.75)^2}{31769.75} \\ &= 5522.597 \end{aligned}$$

We compare this with the $\chi_{k-1}^2 \equiv \chi_3^2$ distribution. From McClave and Sincich tables,

$$\chi_{0.05,3}^2 = 7.815 < X^2$$

so H_0 is **rejected**.

EXAMPLE Eye and Hair Colour Data

The table below contains counts of the number of people in a study with a combination of eye and hair colour.

		Hair				$n_{i.}$
		Black	Brunette	Red	Blonde	
Eyes	Brown	68	119	26	7	220
	Blue	20	84	17	94	215
	Hazel	15	54	14	10	93
	Green	5	29	14	16	64
$n_{.j}$		108	286	71	127	592

so $r = c = 4$. To test the hypothesis

$$H_0 : \text{Eye and Hair colour are assigned independently}$$

we use the X^2 statistic

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Here, for example, for $i = 2$ and $j = 3$

$$\hat{n}_{23} = \frac{n_{2.} \times n_{.3}}{n} = \frac{215 \times 71}{592} = 25.785.$$

In fact, on complete calculation, we find that

$$X^2 = 138.2898.$$

We compare this with the $\chi^2_{(r-1)(c-1)} \equiv \chi^2_9$ distribution. From McClave and Sincich,

$$\chi^2_{0.05,9} = 16.919 < X^2$$

so H_0 is **rejected**

```
> #DNA data Chisq Test
> DNA.data<-c(38514,24631,25685,38249)
> DNA.fitted<-rep(1/4,4)*sum(DNA.data)
> (Xsq<-sum((DNA.data-DNA.fitted)^2/DNA.fitted))

+ [1] 5522.6

> (pval<-1-pchisq(Xsq,df=3))

+ [1] 0

> #DNA data Chisq Test: alternative null
> DNA.data<-c(38514,24631,25685,38249)
> DNA.fitted<-c(0.3,0.2,0.2,0.3)*sum(DNA.data)
>
> (Xsq<-sum((DNA.data-DNA.fitted)^2/DNA.fitted))

+ [1] 31.4923

> (pval<-1-pchisq(Xsq,df=3))

+ [1] 6.69513e-07

> #####
> #Hair & Eye colour
> xtabs(count~eye+hair,data=HairEyes)

+           hair
+ eye      Black Brown Red  Blond
+ Brown      68   119  26     7
+ Blue       20    84  17    94
+ Hazel      15    54  14    10
+ Green       5    29  14    16
```

```

> chisq.test(xtabs(count~eye+hair,data=HairEyes))

+
+   Pearson's Chi-squared test
+
+ data:  xtabs(count ~ eye + hair, data = HairEyes)
+ X-squared = 138.3, df = 9, p-value <2e-16

> #Direct calculation
> HairEyes.table<-xtabs(count~eye+hair,data=HairEyes)
> HE.rowtots<-apply(HairEyes.table,1,sum)
> HE.col tots<-apply(HairEyes.table,2,sum)
>
> #Form the fitted values from the row and column totals
> HE.fitted<-outer(HE.rowtots,HE.col tots,"*")/sum(HairEyes.table)
> Xsq<-sum((HairEyes.table-HE.fitted)^2/HE.fitted)
> (pval<-1-pchisq(Xsq,df=9))

+ [1] 0

```

NON-PARAMETRIC STATISTICAL TESTS

Non-parametric tests are normally based on **ranks** of the data samples, and test hypotheses relating to **quantiles** of the probability distribution representing the population from which the data are drawn. Specifically, tests concern the **population median**, η , where

$$\Pr[\text{Observation} \leq \eta] = \frac{1}{2}$$

The **sample median**, x_{MED} , is the mid-point of the sorted sample; if the data $x_1, \dots, x_n$ are sorted into **ascending** order, then

$$x_{\text{MED}} = \begin{cases} x_m & n \text{ odd}, n = 2m + 1 \\ \frac{x_m + x_{m+1}}{2} & n \text{ even}, n = 2m \end{cases}$$

INDEPENDENT SAMPLES

ONE SAMPLE TEST FOR MEDIAN: THE SIGN TEST

For a single sample of size n , to test the hypothesis $\eta = \eta_0$ for some specified value η_0 we use the **Sign Test**. The test statistic S depends on the alternative hypothesis, H_a .

(a) For **one-sided** tests, to test

$$\begin{aligned} H_0 &: \eta = \eta_0 \\ H_a &: \eta > \eta_0 \end{aligned}$$

we define test statistic S by

$$S = \text{Number of observations } \mathbf{greater\ than} \eta_0$$

whereas to test

$$\begin{aligned} H_0 &: \eta = \eta_0 \\ H_a &: \eta < \eta_0 \end{aligned}$$

we define S by

$$S = \text{Number of observations } \mathbf{less\ than} \eta_0$$

If H_0 is **true**, it follows that

$$S \sim \text{Binomial}\left(n, \frac{1}{2}\right)$$

The p-value is defined by

$$p = \Pr[X \geq S]$$

where $X \sim \text{Binomial}(n, 1/2)$. The rejection region for significance level α is defined implicitly by the rule

Reject H_0 if $\alpha \geq p$.

The Binomial distribution is tabulated in McClave and Sincich.

(b) For a **two-sided** test,

$$H_0 : \eta = \eta_0$$

$$H_a : \eta \neq \eta_0$$

we define the test statistic by

$$S = \max\{S_1, S_2\}$$

where S_1 and S_2 are the counts of the number of observations less than, and greater than, η_0 respectively. The p-value is defined by

$$p = 2 \Pr[X \geq S]$$

where $X \sim \text{Binomial}(n, 1/2)$.

Notes :

- 1 The only assumption behind the test is that the data are drawn independently from a continuous distribution.
- 2 If any data are equal to η_0 , we **discard** them before carrying out the test.
- 3 **Large sample approximation.** If n is large (say $n \geq 30$), and $X \sim \text{Binomial}(n, 1/2)$, then it can be shown that

$$X \approx \text{Normal}(np, np(1-p))$$

Thus for the sign test, where $p = 1/2$, we can use the test statistic

$$Z = \frac{S - \frac{n}{2}}{\sqrt{n \times \frac{1}{2} \times \frac{1}{2}}} = \frac{S - \frac{n}{2}}{\sqrt{n} \times \frac{1}{2}}$$

and note that if H_0 is true,

$$Z \approx \text{Normal}(0, 1).$$

so that the test at $\alpha = 0.05$ uses the following critical values

$$H_a : \eta > \eta_0 \quad \text{then} \quad C_R = 1.645$$

$$H_a : \eta < \eta_0 \quad \text{then} \quad C_R = -1.645$$

$$H_a : \eta \neq \eta_0 \quad \text{then} \quad C_R = \pm 1.960$$

- 4 For the large sample approximation, it is common to make a **continuity correction**, where we replace S by $S - 1/2$ in the definition of Z

$$Z = \frac{\left(S - \frac{1}{2}\right) - \frac{n}{2}}{\sqrt{n} \times \frac{1}{2}}$$

Tables of the standard Normal distribution are given in McClave and Sincich.

TWO SAMPLE TEST: THE MANN-WHITNEY-WILCOXON TEST

For a two **independent** samples of size n_1 and n_2 , to test the hypothesis of **equal population medians**

$$H_0 : \eta_1 = \eta_2$$

we use the **Wilcoxon Rank Sum Test**, or an equivalent test, the **Mann-Whitney U Test**; we refer to this as the

Mann-Whitney-Wilcoxon (MWW) Test

By convention it is usual to formulate the test statistic in terms of the **smaller** sample size. Without loss of generality, we label the samples such that $n_1 > n_2$. The test is based on the **sum of the ranks** for the data from sample 2.

EXAMPLE $n_1 = 4, n_2 = 3$ yields the following ranked data

SAMPLE 1	0.31	0.48	1.02	3.11			
SAMPLE 2	0.16	0.20	1.97				
SAMPLE	2	2	1	1	1	2	1
	0.16	0.20	0.31	0.48	1.02	1.97	3.11
RANK	1	2	3	4	5	6	7

Thus the rank sum for sample 1 is

$$R_1 = 3 + 4 + 5 + 7 = 19$$

and the rank sum for sample 2 is

$$R_2 = 1 + 2 + 6 = 9.$$

Two related test statistics can be used

- **Wilcoxon Rank Sum Statistic**

$$W = R_2$$

- **Mann-Whitney U Statistic**

$$U = R_2 - \frac{n_2(n_2 + 1)}{2}$$

We again consider three alternative hypotheses:

$$H_a : \eta_1 < \eta_2$$

$$H_a : \eta_1 > \eta_2$$

$$H_a : \eta_1 = \eta_2$$

and define the rejection region separately in each case.

- **Large Sample Test** If $n_2 \geq 10$, a large sample test based on the Z statistic

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

can be used. Under the hypothesis $H_0 : \eta_1 = \eta_2$,

$$Z \approx \text{Normal}(0, 1)$$

so that the test at $\alpha = 0.05$ uses the following critical values

$$H_a : \eta_1 > \eta_2 \quad \text{then} \quad C_R = -1.645$$

$$H_a : \eta_1 < \eta_2 \quad \text{then} \quad C_R = 1.645$$

$$H_a : \eta_1 \neq \eta_2 \quad \text{then} \quad C_R = \pm 1.960$$

- **Small Sample Test** If $n_1 < 10$, an **exact** but more complicated test can be used. The test statistic is R_2 (the sum of the ranks for sample 2). The null distribution under the hypothesis $H_0 : \eta_1 = \eta_2$ can be computed, but it is complicated.

The table in McClave and Sincich gives the critical values (T_L and T_U) that determine the rejection region for different n_1 and n_2 values up to 10.

- **One-sided tests:** These are tests at the $\alpha = 0.025$ significance level.

$$H_a : \eta_1 > \eta_2 \quad \text{Rejection Region is} \quad R_2 \leq T_L$$

$$H_a : \eta_1 < \eta_2 \quad \text{Rejection Region is} \quad R_2 \geq T_U$$

- **Two-sided tests:** This is a test at the $\alpha = 0.05$ significance level.

$$H_a : \eta_1 \neq \eta_2 \quad \text{Rejection Region is} \quad R_2 \leq T_L \text{ or } R_2 \geq T_U$$

Notes:

- The only assumption is are needed for the test to be valid is that the samples are independently drawn from two continuous distributions.
- The sum of the ranks across **both** samples is

$$R_1 + R_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

- If there are **ties** (equal values) in the data, then the rank values are replaced by **average** rank values.

DATA VALUE	0.16	0.20	0.31	0.31	0.48	1.97	3.11
ACTUAL RANK	1	2	3	3	5	6	7
AVERAGE RANK	1	2	3.5	3.5	5	6	7

EXAMPLE Sign Test: Water Content Example

The following data are measurements of percentage water content of soil samples collected by two experimenters. We wish to test the hypothesis

$$H_0 : \eta = 9.0$$

for each experiment.

Experimenter 1:	$n = 10$	5.5	6.0	6.5	7.6	7.6	7.7	8.0	8.2	9.1	15.1
Experimenter 2:	$n = 20$	5.6	6.1	6.3	6.3	6.5	6.6	7.0	7.5	7.9	8.0
		8.0	8.1	8.1	8.2	8.4	8.5	8.7	9.4	14.3	26.0

To perform the test, we need tables of the Binomial distribution with $p = 1/2$. The individual probabilities are given by the formula

$$\Pr[X = x] = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \frac{1}{2^n} = \frac{n!}{x!(n-x)!} \frac{1}{2^n} \quad x = 0, 1, \dots, n$$

We test at the $\alpha = 0.05$ level. For the first experiment, with $n = 10$:

- If

$$H_a : \eta > 9.0$$

the test statistic is

$$S = \text{Number of observations **greater than** 9} \quad \therefore \quad S = 2$$

and the p-value is

$$p = \Pr[X \geq 2] = 1 - \Pr[X < 2] = 1 - \Pr[X = 0] - \Pr[X = 1] = 0.9893$$

so we **do not** reject H_0 in favour of this H_a .

- If

$$H_a : \eta < 9.0$$

the test statistic is

$$S = \text{Number of observations **less than** 9} \quad \therefore \quad S = 8$$

and the p-value is

$$p = \Pr[X \geq 8] = \Pr[X = 8] + \Pr[X = 9] + \Pr[X = 10] = 0.0547$$

so we **do not** reject H_0 in favour of this H_a .

- If

$$H_a : \eta \neq 9.0$$

the test statistic is

$$S = \max\{S_1, S_2\} = \max\{2, 8\} = 8$$

and the p-value is

$$p = 2\Pr[X \geq 8] = 2(\Pr[X = 8] + \Pr[X = 9] + \Pr[X = 10]) = 0.1094$$

so we **do not** reject H_0 in favour of this H_a .

For the second experiment, with $n = 20$:

- If $H_a : \eta > 9.0$, the test statistic is $S = 3$. The p-value is

$$p = \Pr[X \geq 3] = 1 - \Pr[X < 3] = 1 - \Pr[X = 0] - \Pr[X = 1] - \Pr[X = 2] = 0.9998.$$

so we **do not** reject H_0 in favour of this H_a .

- If $H_a : \eta < 9.0$, the test statistic $S = 17$. The p-value is

$$p = \Pr[X \geq 17] = \Pr[X = 17] + \Pr[X = 18] + \Pr[X = 19] + \Pr[X = 20] = 0.0013.$$

so we **do** reject H_0 in favour of this H_a .

- If $H_a : \eta \neq 9.0$, the test statistic is $S = \max\{S_1, S_2\} = \max\{3, 17\} = 17$. The p-value is

$$p = 2\Pr[X \geq 17] = 2(\Pr[X = 17] + \Pr[X = 18] + \Pr[X = 19] + \Pr[X = 20]) = 0.0026.$$

so we **do** reject H_0 in favour of this H_a .

In R, the sign test can be carried out using basic functions:

```
> #Sign test: Test statistic
> S<-sum(Sign1$h2o>9)
> #First p-value
> 1-pbinom(S-1,nrow(Sign1),0.5)
+ [1] 0.989258

> #Second p-value
> pbinom(S,nrow(Sign1),0.5)
+ [1] 0.0546875

> #Third p-value
> 2*(1-pbinom(max(S,nrow(Sign1)-S)-1,nrow(Sign1),0.5))
+ [1] 0.109375
```

```
> eta<-9
> binom.test(sum(Sign1$h2o>eta),nrow(Sign1),p=0.5,
+           alternative="greater")
+
+ Exact binomial test
+
+ data: sum(Sign1$h2o > eta) and nrow(Sign1)
+ number of successes = 2, number of trials = 10, p-value = 0.989
+ alternative hypothesis: true probability of success is greater than 0.5
+ 95 percent confidence interval:
+ 0.0367714 1.0000000
+ sample estimates:
+ probability of success
+ 0.2
```

```

> binom.test(sum(Sign1$h2o>eta),nrow(Sign1),p=0.5,
+           alternative="less")
+
+ Exact binomial test
+
+ data:  sum(Sign1$h2o > eta) and nrow(Sign1)
+ number of successes = 2, number of trials = 10, p-value = 0.0547
+ alternative hypothesis: true probability of success is less than 0.5
+ 95 percent confidence interval:
+  0.000000 0.506901
+ sample estimates:
+ probability of success
+                0.2

```

```

> binom.test(sum(Sign1$h2o>eta),nrow(Sign1),p=0.5,
+           alternative="two.sided")
+
+ Exact binomial test
+
+ data:  sum(Sign1$h2o > eta) and nrow(Sign1)
+ number of successes = 2, number of trials = 10, p-value = 0.109
+ alternative hypothesis: true probability of success is not equal to 0.5
+ 95 percent confidence interval:
+  0.0252107 0.5560955
+ sample estimates:
+ probability of success
+                0.2

```

Using the Mann-Whitney-Wilcoxon test:

```

> #Mann-Whitney-Wilcoxon test
> Sample1<-c(0.31,0.48,1.02,3.11)
> Sample2<-c(0.16,0.20,1.97)
> Y<-c(Sample1,Sample2)
> Samp<-rep(c(1,2),c(4,3))
> rY<-rank(Y)
> (R1<-sum(rY[Samp==1]))
+ [1] 19
> (R2<-sum(rY[Samp==2]))
+ [1] 9

```

In R, the function `wilcox.test` carries out the test.

```

> wilcox.test(Y~Samp)
+
+ Wilcoxon rank sum exact test
+
+ data:  Y by Samp

```

```

+ W = 9, p-value = 0.4
+ alternative hypothesis: true location shift is not equal to 0
> wilcox.test(Y~Samp, alternative="less")
+
+ Wilcoxon rank sum exact test
+
+ data: Y by Samp
+ W = 9, p-value = 0.886
+ alternative hypothesis: true location shift is less than 0

```

Using the coin library function wilcox_test

```

> coin::wilcox_test(Y~as.factor(Samp)) #Asymptotic approximation
+
+ Asymptotic Wilcoxon-Mann-Whitney Test
+
+ data: Y by as.factor(Samp) (1, 2)
+ Z = 1.061, p-value = 0.289
+ alternative hypothesis: true mu is not equal to 0
> coin::wilcox_test(Y~as.factor(Samp), distribution="exact") #Exact null
+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: Y by as.factor(Samp) (1, 2)
+ Z = 1.061, p-value = 0.4
+ alternative hypothesis: true mu is not equal to 0

```

We can also use a permutation procedure:

```

> #Approximate null using permutation Monte Carlo
> coin::wilcox_test(Y~as.factor(Samp), distribution="approximate")
+
+ Approximative Wilcoxon-Mann-Whitney Test
+
+ data: Y by as.factor(Samp) (1, 2)
+ Z = 1.061, p-value = 0.397
+ alternative hypothesis: true mu is not equal to 0

```

EXAMPLE Mann-Whitney-Wilcoxon Test: Low Birthweight Example

The birthweights (in grammes) of babies born to two groups of mothers A and B are displayed below: Thus $n_1 = 9$, $n_2 = 8$. From this sample (which has ties, so we need to use average ranks),

Group A:	$n = 9$	2164	2600	2184	2080	1820	2496	2184	2080	2184
Group B:	$n = 8$	2576	3224	2704	2912	2444	3120	2912	3848	

so that the two statistics are $R_1 = 48$, $R_2 = 105$, and

$$\text{Wilcoxon } W = R_2 = 105$$

$$\text{Mann-Whitney } U = R_2 - \frac{n_2(n_2 + 1)}{2} = 105 - 36 = 69$$

- For the **small sample** test, from tables in McClave and Sincich, we find

$$T_L = 51 \quad T_U = 93$$

Thus $W > 93$, so we

Do not reject H_0 against $H_a : \eta_1 > \eta_2$ as $W = R_2 > T_L$

Reject H_0 against $H_a : \eta_1 < \eta_2$ as $W = R_2 > T_U$

Reject H_0 against $H_a : \eta_1 \neq \eta_2$ as $W = R_2 > T_U$

The one-sided tests are carried out at $\alpha = 0.025$, the two-sided test is carried out at $\alpha = 0.05$.

- For the **large sample** test, we find

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = 3.175$$

Thus we

Do not reject H_0 against $H_a : \eta_1 > \eta_2$ as $Z > C_R = -1.645$

Reject H_0 against $H_a : \eta_1 < \eta_2$ as $Z > C_R = 1.645$

Reject H_0 against $H_a : \eta_1 \neq \eta_2$ as $Z > C_{R_2} = 1.960$

All tests are carried out at $\alpha = 0.05$.

```
> R1<-sum(rank(Birthweight$BW)[Birthweight$gp==1])
> R2<-sum(rank(Birthweight$BW)[Birthweight$gp==2])
> n1<-sum(Birthweight$gp==1); n2<-sum(Birthweight$gp==2)
> c(R1, R2, n1)

+ [1] 58 95 10

> W<-R2
> U<-W-0.5*n2*(n2+1)
> Z<-(U-0.5*n1*n2)/sqrt(n1*n2*(n1+n2+1)/12)
```

Using the Mann-Whitney-Wilcoxon tests:

```
> wilcox.test(BW~Group, data=Birthweight)

+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
cannot compute exact p-value with ties
```

```

+
+ Wilcoxon rank sum test with continuity correction
+
+ data: BW by Group
+ W = 3, p-value = 0.00203
+ alternative hypothesis: true location shift is not equal to 0
> wilcox.test(BW~Group,data=Birthweight,alternative="greater")
+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
+ cannot compute exact p-value with ties
+
+ Wilcoxon rank sum test with continuity correction
+
+ data: BW by Group
+ W = 3, p-value = 0.999
+ alternative hypothesis: true location shift is greater than 0
> coin::wilcox_test(BW~Group,data=Birthweight,
+ dist="exact",ties.method="mid-ranks")
+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: BW by Group (A, B)
+ Z = -3.134, p-value = 0.000668
+ alternative hypothesis: true mu is not equal to 0
> coin::wilcox_test(BW~Group,data=Birthweight,
+ dist="exact",alternative="greater")
+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: BW by Group (A, B)
+ Z = -3.134, p-value = 1
+ alternative hypothesis: true mu is greater than 0
> coin::wilcox_test(BW~Group,data=Birthweight,
+ dist="exact",alternative="less")
+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: BW by Group (A, B)
+ Z = -3.134, p-value = 0.000309
+ alternative hypothesis: true mu is less than 0

```

EXAMPLE Mann-Whitney-Wilcoxon Test: Treadmill Test Example

The treadmill stress test times (in seconds) of two groups of patients (disease group and healthy controls) are displayed below:

Disease :	$n = 10$	864	636	638	708	786	600	1320	750	594	750
Healthy :	$n = 8$	1014	684	810	990	840	978	1002	1110		

Thus $n_1 = 10, n_2 = 8$. For this sample (which has ties, so we need to use average ranks)

$$R_1 = 70 \quad R_2 = 101$$

so that the two statistics are

$$\text{Wilcoxon } W = R_2 = 101$$

$$\text{Mann-Whitney } U = R_2 - \frac{n_2(n_2 + 1)}{2} = 101 - 36 = 65$$

- For the **small sample** test, from tables in McClave and Sincich, we find

$$T_L = 54 \quad T_U = 98$$

Thus $W > 98$, so we

Do not reject H_0 against $H_a : \eta_1 > \eta_2$ as $W = R_2 > T_L$

Reject H_0 against $H_a : \eta_1 < \eta_2$ as $W = R_2 > T_U$

Reject H_0 against $H_a : \eta_1 \neq \eta_2$ as $W = R_2 > T_U$

One-sided tests are carried out at $\alpha = 0.025$; two-sided test is carried out at $\alpha = 0.05$.

- For the **large sample** test, we find

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = 2.221$$

Thus we

Do not reject H_0 against $H_a : \eta_1 > \eta_2$ as $Z > C_R = -1.645$

Reject H_0 against $H_a : \eta_1 < \eta_2$ as $Z > C_R = 1.645$

Reject H_0 against $H_a : \eta_1 \neq \eta_2$ as $Z > C_{R_2} = 1.960$

All tests are carried out at $\alpha = 0.05$.

```
> wilcox.test(Time~Group, data=Treadmill)
+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
+ cannot compute exact p-value with ties
+
+ Wilcoxon rank sum test with continuity correction
+
+ data: Time by Group
+ W = 65, p-value = 0.0294
+ alternative hypothesis: true location shift is not equal to 0
```

```

> wilcox.test(Time~Group,data=Treadmill,alternative="greater")

+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
cannot compute exact p-value with ties

+
+ Wilcoxon rank sum test with continuity correction
+
+ data: Time by Group
+ W = 65, p-value = 0.0147
+ alternative hypothesis: true location shift is greater than 0

> wilcox.test(Time~Group,data=Treadmill,alternative="less")

+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
cannot compute exact p-value with ties

+
+ Wilcoxon rank sum test with continuity correction
+
+ data: Time by Group
+ W = 65, p-value = 0.988
+ alternative hypothesis: true location shift is less than 0

> coin::wilcox_test(Time~as.factor(Group),data=Treadmill,
+ dist="exact", ties.method="mid-ranks")

+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: Time by as.factor(Group) (1, 2)
+ Z = 2.222, p-value = 0.0249
+ alternative hypothesis: true mu is not equal to 0

> coin::wilcox_test(Time~as.factor(Group),data=Treadmill,
+ dist="exact", alternative="greater")

+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: Time by as.factor(Group) (1, 2)
+ Z = 2.222, p-value = 0.0125
+ alternative hypothesis: true mu is greater than 0

> coin::wilcox_test(Time~as.factor(Group),data=Treadmill,
+ dist="exact", alternative="less")

+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: Time by as.factor(Group) (1, 2)
+ Z = 2.222, p-value = 0.989
+ alternative hypothesis: true mu is less than 0

```

DEPENDENT SAMPLES

TWO DEPENDENT SAMPLES: WILCOXON SIGNED RANK TEST

Data collected from the same experimental units are in general **dependent**. For example, if data are collected on two occasions (time 1 and time 2, or before and after treatment) from the same n individuals, then the resulting data samples $(y_{11}, \dots, y_{n1})$ and $(y_{12}, \dots, y_{n2})$ are dependent. Such data are often referred to as **paired**. We wish to test whether there is a significant change across the two measurements.

For a **parametric** test, we typically assume that the within-individual differences

$$x_i = y_{i1} - y_{i2} \quad i = 1, \dots, n$$

are **Normally** distributed, and test the hypothesis that the mean difference μ is zero

$$H_0 : \mu = 0$$

using a one-sample Z -test (σ known) or T -test (σ unknown), with statistic

$$z = \frac{\bar{x}}{\sigma/\sqrt{n}} \quad \text{or} \quad t = \frac{\bar{x}}{s/\sqrt{n}}$$

distributed as $\text{Normal}(0, 1)$ or t_{n-1} respectively.

For a **non-parametric** test, we can use the **Wilcoxon Signed Rank** test:

1. Compute the within-individual differences

$$x_i = y_{i1} - y_{i2} \quad i = 1, \dots, n$$

If any $x_i = 0$, then that data point is discarded and the sample size adjusted.

2. Sort the **absolute values** $s_1, \dots, s_n$ of $x_1, x_2, \dots, x_n$ into **ascending** order, and assign ranks 1 up to n . If there are ties, assign **average** ranks.
3. Form the two rank sums T_+ and T_- , where

$$T_+ = \text{Sum of ranks for those } x_i > 0 \quad T_- = \text{Sum of ranks for those } x_i < 0$$

The test statistic is a function of these rank sums. Heuristically, if the statistic T_+ is large and T_- is small, this implies that the experimental units where $y_{i1} > y_{i2}$ have a **larger** (in magnitude) difference than those where $y_{i1} < y_{i2}$. This indicates an overall **decrease** between the first and second measurements. Conversely, if the statistic T_- is large and T_+ is small, this implies that the experimental units where $y_{i2} > y_{i1}$ have a **larger** (in magnitude) difference than those where $y_{i2} < y_{i1}$. This indicates an overall **increase** between the first and second measurements.

We test the null hypothesis

$$H_0 : \text{No change between first and second measurements}$$

against the three alternative hypotheses

- (1) H_a : Significant **decrease** between first and second measurements
- (2) H_a : Significant **increase** between first and second measurements
- (3) H_a : Significant **change** between first and second measurements

To test H_0 vs (1), we perform a one-sided test using the statistic T_- ; the critical value in the test is denoted T_0 , and is determined by the table in McClave and Sincich:

If $T_- \leq T_0$, we **reject** H_0 in favour of H_a (1)

To test H_0 vs (2), we perform a one-sided test using the statistic T_+ ; the critical value is T_0 and

If $T_+ \leq T_0$, we **reject** H_0 in favour of H_a (2)

To test H_0 vs (3), we perform a two-sided test using the statistic $T = \min\{T_-, T_+\}$; the critical value is T_0

If $T \leq T_0$, we **reject** H_0 in favour of H_a (3)

Notes :

- 1 The only assumption behind the test is that the difference data x_i are drawn independently from a continuous distribution.
- 2 **Large Sample Test:** For $n \geq 25$, we can use a large sample version of the test based on T_+ , and the Z statistic

$$Z = \frac{T_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

If H_0 is **true**, then $Z \sim \text{Normal}(0, 1)$, so that the test at $\alpha = 0.05$ uses the following critical values

For H_a (1) use $C_R = 1.645$

For H_a (2) use $C_R = -1.645$

For H_a (3) use $C_R = \pm 1.960$

EXAMPLE Haemodialysis Data

The following data are measurements of the heparin cofactor II (HCII) to plasma protein ratios in a group of patients at baseline and five months after haemodialysis.

Patient	Before	After				
	y_{i1}	y_{i2}	x_i	s_i	Rank	Ave. Rank
1	2.11	2.15	-0.04	0.04	3	3.5
2	1.85	2.11	-0.26	0.26	10	10.0
3	1.82	1.93	-0.11	0.11	8	8.0
4	1.75	1.83	-0.08	0.08	6	6.0
5	1.54	1.90	-0.36	0.36	11	11.0
6	1.52	1.56	-0.04	0.04	3	3.5
7	1.49	1.44	0.05	0.05	5	5.0
8	1.44	1.43	0.01	0.01	1	1.5
9	1.38	1.28	0.10	0.10	7	7.0
10	1.30	1.30	0.00	0.00	-	-
11	1.20	1.21	-0.01	0.01	1	1.5
12	1.19	1.30	-0.11	0.11	9	9.0
						$T_+ = 13.5$
						$T_- = 52.5$

OMIT

From the textbook table, for $n = 12 - 1 = 11$, we find that the $\alpha = 0.025/0.05$ (one/two-sided) significance level critical value is $T_0 = 11$. Thus using T_+ , we **cannot reject** either of the null

hypotheses (2) and (3), as $T_+ > T_0$. Note that $Z = -1.734$, so if the approximation was valid, we would be able to reject (2) at $\alpha = 0.05$.

```
> #Haemodialysis data
> wilcox.test(Haemo$after[-10], Haemo$before[-10],
+   paired=T)

+ Warning in wilcox.test.default(Haemo$after[-10], Haemo$before[-10], paired
= T): cannot compute exact p-value with ties

+
+ Wilcoxon signed rank test with continuity correction
+
+ data: Haemo$after[-10] and Haemo$before[-10]
+ V = 52.5, p-value = 0.0908
+ alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Haemo$after[-10], Haemo$before[-10],
+   paired=T, alternative="greater")

+ Warning in wilcox.test.default(Haemo$after[-10], Haemo$before[-10], paired
= T, : cannot compute exact p-value with ties

+
+ Wilcoxon signed rank test with continuity correction
+
+ data: Haemo$after[-10] and Haemo$before[-10]
+ V = 52.5, p-value = 0.0454
+ alternative hypothesis: true location shift is greater than 0

> #Direct calculation
> xi<-Haemo$after-Haemo$before
> xi<-xi[xi!=0]
> si<-abs(xi)
> n<-length(si)
>
> Tplus<-sum(rank(si)[xi > 0])
> Tminus<-sum(rank(si)[xi < 0])
>
> (Z<-(Tplus-0.25*n*(n+1))/sqrt(n*(n + 1)*(2*n + 1)/24))

+ [1] 1.73376

> pnorm(Z)

+ [1] 0.95852
```

EXAMPLE Asthma data:

```
> PEFRAsthma #PEFRAsthma data
+ Patient before after
+ 1      1      312    300
+ 2      2      342    201
+ 3      3      340    232
+ 4      4      388    312
+ 5      5      296    220
+ 6      6      254    256
+ 7      7      391    328
+ 8      8      402    330
+ 9      9      290    231

> PEFRAsthma$before-PEFRAsthma$after #Differences
+ [1] 12 141 108 76 76 -2 63 72 59

> wilcox.test(PEFRAsthma$after, PEFRAsthma$before, paired=T)
+ Warning in wilcox.test.default(PEFRAsthma$after, PEFRAsthma$before, paired
= T): cannot compute exact p-value with ties
+
+ Wilcoxon signed rank test with continuity correction
+
+ data: PEFRAsthma$after and PEFRAsthma$before
+ V = 1, p-value = 0.0128
+ alternative hypothesis: true location shift is not equal to 0
> wilcox.test(PEFRAsthma$after, PEFRAsthma$before, paired=T,
+ alternative="less")
+ Warning in wilcox.test.default(PEFRAsthma$after, PEFRAsthma$before, paired
= T, : cannot compute exact p-value with ties
+
+ Wilcoxon signed rank test with continuity correction
+
+ data: PEFRAsthma$after and PEFRAsthma$before
+ V = 1, p-value = 0.00639
+ alternative hypothesis: true location shift is less than 0
> #Direct calculation
> xi<-PEFRAsthma$after-PEFRAsthma$before
> xi<-xi[xi!=0]
> si<-abs(xi); n<-length(si)
>
> Tplus<-sum(rank(si)[xi > 0])
> Tminus<-sum(rank(si)[xi < 0])
>
> (Z<-(Tplus-0.25*n*(n+1))/sqrt(n*(n + 1)*(2*n + 1)/24))
+ [1] -2.5471
> pnorm(Z)
+ [1] 0.00543111
```

THREE OR MORE INDEPENDENT SAMPLES

We now seek non-parametric tests that can be used for multiple independent samples, such as those found in the Completely Randomized Design (CRD) and Randomized Block Design (RBD) described in the ANOVA section. The non-parametric equivalents of the F tests for these two designs are

- The **Kruskal-Wallis H test** for a Completely Randomized Design
- **Friedman's test** for a Randomized Block Design

KRUSKAL-WALLIS TEST

In a CRD, we have k independent groups, corresponding to k different treatments, with sample sizes $n_1, \dots, n_k$. Let $n = n_1 + \dots + n_k$. To compute the test statistic, H , we

1. Pool the data, sort them into ascending order, and assign ranks. If there are ties in the data, then average ranks are used.
2. For $j = 1, \dots, k$, compute the rank sum R_j

$$R_j = \text{Sum of ranks for data from sample } j.$$

To test the hypothesis

$$\begin{aligned} H_0 &: \text{No difference between the population distributions of the } k \text{ groups} \\ H_a &: \text{At least two population distributions different} \end{aligned}$$

the test statistic is

$$H = \frac{12}{n(n+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(n+1)$$

If H_0 is **true**, then for large n ,

$$H \approx \chi_{k-1}^2.$$

Notes :

- 1 The test assumes that the k samples are independently drawn from continuous populations.
- 2 For the approximation to be valid, there should be at least **five** observations in each sample, and the number of ties should be small.

EXAMPLE Mucociliary efficiency data

The data are measures of mucociliary efficiency from the rate of removal of dust in normal subjects (Group 1), subjects with obstructive airway disease (Group 2), and subjects with asbestosis (Group 3).

Reference: Myles Hollander, M and Douglas A. Wolfe (1973), *Nonparametric statistical inference*, New York: John Wiley & Sons. pp115-120.

Group	1	1	1	1	1	2	2	2	2	3	3	3	3	3
y	2.9	3.0	2.5	2.6	3.2	3.8	2.7	4.0	2.4	2.8	3.4	3.7	2.2	2.0
Rank	8	9	4	5	10	13	6	14	3	7	11	12	2	1

Hence $R_1 = 36$, $R_2 = 36$ and $R_3 = 33$, and the test statistic $H = 0.7714$. To complete the test, we compare with the $\alpha = 0.05$ quantile of the $\chi^2_{k-1} = \chi^2_2$ distribution. We have

$$\chi^2_{0.05,2} = 5.99 > H \quad \therefore \quad \text{No evidence to reject } H_0$$

and a p-value of $p = 0.680$.

FRIEDMAN TEST

In a RBD, we have k treatment groups, and a blocking factor. For example, we might have k repeated measurements on the same b experimental units, and $n = bk$ observations in total. To compute the test statistic, F_r , we proceed as follows.

1. **Within each block separately**, sort the k data values into ascending order, and assign ranks. If there are ties in the data, then average ranks are used.
2. For $j = 1, \dots, k$, compute the rank sum R_j

$$R_j = \text{Sum of ranks for data from treatment } j.$$

To test the hypothesis

H_0 : No difference between the population distributions of the k treatment groups

H_a : At least two population distributions different

the test statistic is

$$F_r = \frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 - 3b(k+1)$$

If H_0 is **true**, then for large n ,

$$F_r \approx \chi^2_{k-1}$$

Notes :

- 1 The test assumes that the data are drawn independently from continuous populations, with random assignment of treatments within blocks.
- 2 For the approximation to be valid, it is recommended that b or k is at least five, and the number of ties should be small.

EXAMPLE Skin potential under hypnosis

A study was conducted to investigate whether hypnosis has the same effect on skin potential for four different emotions. Eight subjects were asked to display fear, joy, sadness and calmness under hypnosis, and the resulting skin potential (measured in millivolts) was recorded for each emotion. Thus in this experiment, $b = 8$ and $k = 4$.

Subject	Fear		Joy		Sadness		Calmness	
	y	Rank	y	Rank	y	Rank	y	Rank
1	23.1	4	22.7	3	22.5	1	22.6	2
2	57.6	4	53.2	2	53.7	3	53.1	1
3	10.5	3	9.7	2	10.8	4	8.3	1
4	23.6	4	19.6	3	21.1	2	21.6	1
5	11.9	1	13.8	4	13.7	3	13.3	2
6	54.6	4	47.1	3	39.2	2	37.0	1
7	21.0	4	13.6	1	13.7	2	14.8	3
8	20.3	3	23.6	4	16.3	2	14.8	1
Rank Sum	27		20		19		14	

Thus the within-treatment rank sums are $R_1 = 27, R_2 = 20, R_3 = 19$ and $R_4 = 14$ and thus $F_r = 6.45$. To complete the test, we compare with the $\alpha = 0.05$ quantile of the

$$\chi_{k-1}^2 = \chi_3^2$$

distribution. We have

$$\chi_{0.05,3}^2 = 7.81 > F_r \quad \therefore \quad \text{No evidence to reject } H_0$$

and a p-value of $p = 0.092$.

```
> #Mucociliary data
> #Mean ranks
> xtabs(rank(y) ~ group, data=Muco) / xtabs(~group, data=Muco)

+ group
+           Healthy Obstructive airway disease
+           7.2           9.0
+           Asbestosis
+           6.6

> kruskal.test(y ~ group, data=Muco)

+
+  Kruskal-Wallis rank sum test
+
+ data:  y by group
+ Kruskal-Wallis chi-squared = 0.7714, df = 2, p-value = 0.68

> #MemoryTask
> #Mean ranks
> xtabs(rank(Memory) ~ Task, data=Memory) / xtabs(~Task, data=Memory)
```

```

+ Task
+   Counting      Rhyming      Adjective      Imagery Intentional
+      12.95      13.10      31.50      36.60      33.35

> kruskal.test(Memory~Task, data=Memory)

+
+   Kruskal-Wallis rank sum test
+
+ data:   Memory by Task
+ Kruskal-Wallis chi-squared = 25.38, df = 4, p-value = 4.23e-05

```

```

> #Skin potential data (Hypnosis)
>
> friedman.test(SkinResponse~EmoCode | Subject, data=Hypnosis)

+
+   Friedman rank sum test
+
+ data:   SkinResponse and EmoCode and Subject
+ Friedman chi-squared = 6.45, df = 3, p-value = 0.0917

> #Soil
> friedman.test(sulphur~solvent | soil, data=Soil)

+
+   Friedman rank sum test
+
+ data:   sulphur and solvent and soil
+ Friedman chi-squared = 1.08, df = 3, p-value = 0.782

```

RANK CORRELATION

SPEARMAN'S RANK CORRELATION

A measure of association for two samples $x_1, \dots, x_n$ and $y_1, \dots, y_n$ is the **Pearson Product Moment Correlation Coefficient**, r , where

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

where

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

This quantity measures the **linear** association between the X and Y variables. A more general measure of association between the samples $x_1, \dots, x_n$ and $y_1, \dots, y_n$ is the **Spearman Rank Correlation Coefficient**, r_S , which computes the correlation between the **ranks** of the data.

The Spearman Rank Correlation Coefficient is computed as follows:

1. Assign ranks $u_1, \dots, u_n$ and $v_1, \dots, v_n$ to the data $x_1, \dots, x_n$ and $y_1, \dots, y_n$ separately by sorting each sample into ascending order and assigning the ranks in order.
2. Compute r_S as

$$r_S = \frac{SS_{uv}}{\sqrt{SS_{uu} SS_{vv}}}$$

where

$$SS_{uu} = \sum_{i=1}^n (u_i - \bar{u})^2 \quad SS_{vv} = \sum_{i=1}^n (v_i - \bar{v})^2 \quad SS_{uv} = \sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})$$

If there are no ties in the data, then

$$r_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad \text{where} \quad d_i = u_i - v_i \quad i = 1, \dots, n$$

Tests for r_S : If the population correlation is ρ , then we may test the hypothesis

$$H_0 : \rho = 0$$

against different hypotheses using the table of the null distribution in McClave and Sincich. If Spearman_α is the α tail quantile of the null distribution

$$\begin{array}{ll} H_a : \rho > 0 & \text{Reject } H_0 \text{ if } r_S > \text{Spearman}_\alpha \\ H_a : \rho < 0 & \text{Reject } H_0 \text{ if } r_S < -\text{Spearman}_\alpha \\ H_a : \rho \neq 0 & \text{Reject } H_0 \text{ if } |r_S| > \text{Spearman}_{\alpha/2} \end{array}$$

Country	Latitude <i>x</i>	Rank <i>u</i>	DZT Rate <i>y</i>	Rank <i>v</i>
Portugal	40	1.5	6.5	2.0
Greece	40	1.5	8.8	13.0
Spain	41	3.0	5.9	1.0
Bulgaria	42	4.0	7.0	3.0
Italy	44	5.0	8.6	11.5
France	47	6.5	7.1	4.0
Switzerland	47	6.5	8.1	7.5
Austria	48	8.0	7.5	6.0
Belgium	51	9.5	7.3	5.0
FR Germany	51	9.5	8.2	9.0
Holland	52	11.5	8.1	7.5
GDR	52	11.5	9.1	16.0
England & Wales	53	13.5	8.9	14.5
Ireland	53	13.5	11.0	18.0
Scotland	56	15.5	8.9	14.5
Denmark	56	15.5	9.6	17.0
Sweden	60	17.0	8.6	11.5
Norway	61	18.0	8.3	10.0
Finland	62	19.0	12.1	19.0

EXAMPLE Latitude and dizygotic twinning rates

The relationship between the geographical latitude of a country and its dizygotic twinning (DZT) rate is to be investigated. The data are presented and plotted below.

Reference: James, W.H. (1985) Dizygotic twinning, birth weight and latitude, *Annals of Human Biology*, **12**, 5, pp. 441-447.

For these data

$$r_s = \frac{SS_{uv}}{\sqrt{SS_{uu} SS_{vv}}} = \frac{384.5}{\sqrt{567 \times 568.5}} = 0.677 \quad r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} = \frac{118.4}{\sqrt{866.105 \times 38.88}} = 0.645$$

indicating a strong positive association.

```
> cor(DZT$Lat, DZT$DZTRate, method="spearman")
+ [1] 0.677235
> cor(DZT$Lat, DZT$DZTRate, method="pearson")
+ [1] 0.645213
```

```
> par(mar=c(4, 4, 2, 2), pty='s')
> plot(DZT$Lat, DZT$DZT, pch=20, xlab="Latitude (degrees)",
+      ylab="DZT Rate (per 1000, standardized)",
+      xlim=range(37, 65), ylim=range(5, 12.5))
> DZT.ylab<-DZT$DZT+0.25
> DZT.ylab[c(11, 13, 18)]<-DZT.ylab[c(11, 13, 18)]-0.45
> text(DZT$Lat, DZT.ylab, DZT$Country, cex=0.6)
```

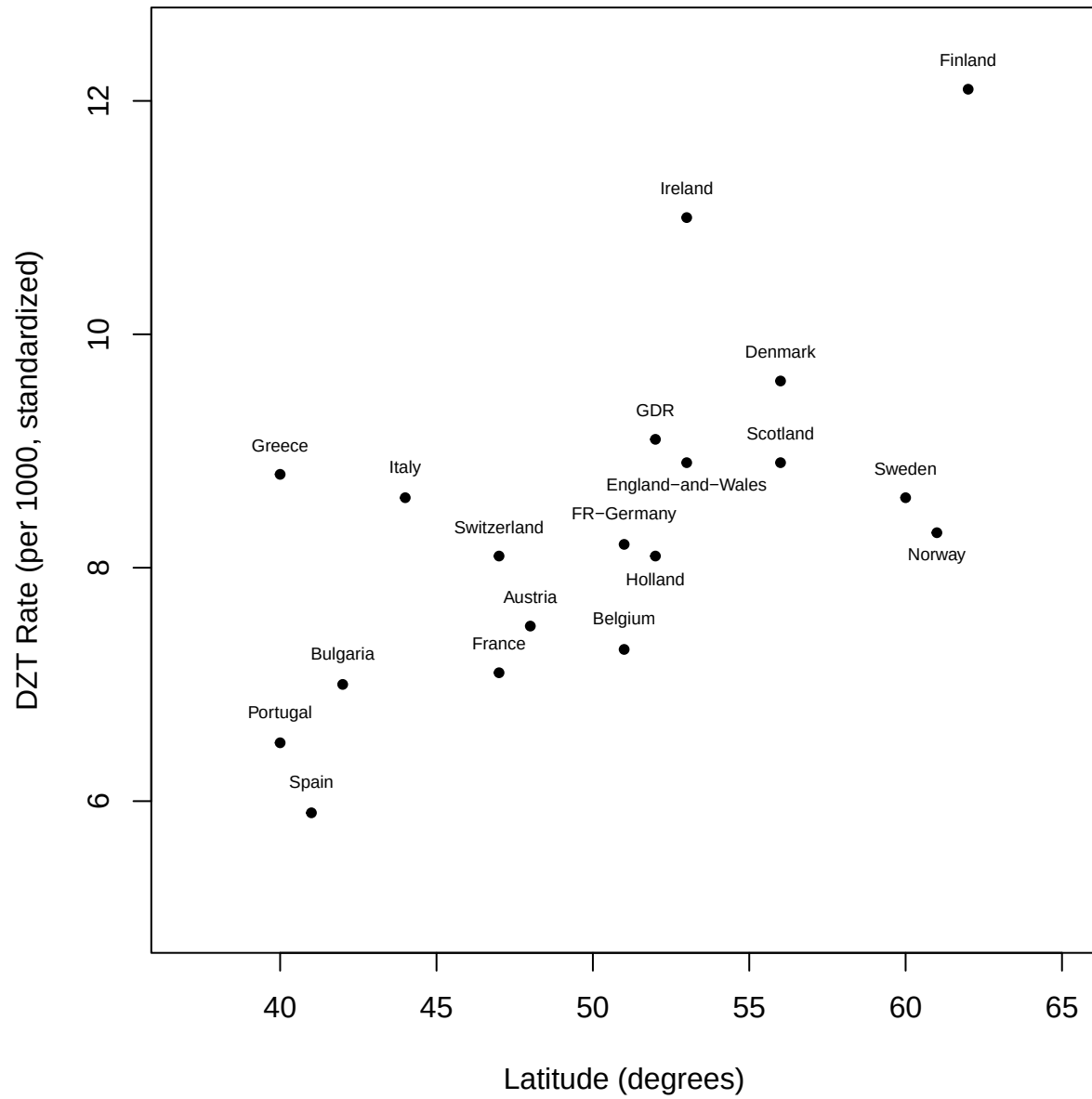


Figure 27: Latitude and dizygotic twinning rates: raw data

RANDOMIZATION AND PERMUTATION PROCEDURES

Randomization or **Permutation** procedures are useful for computing **exact** null distributions for non-parametric test statistics when sample sizes are small.

Suppose that two data samples $x_1 \dots, x_{n_1}$ and $y_1 \dots, y_{n_2}$ (where $n_1 \geq n_2$) have been obtained, and we wish to carry out a comparison of the two populations from which the samples are drawn. The Wilcoxon test statistic, W , is the sum of the ranks for the second sample. The permutation test proceeds as follows:

1. Let $n = n_1 + n_2$. If there are no ties, the pooled and ranked samples will have ranks

$$1 \quad 2 \quad 3 \quad \dots \quad n$$

2. The test statistic is $W = R_2$, the rank sum for sample two items. For the observed data, W will be the sum of n_2 of the ranks given in the list above.
3. If the null hypothesis

$$H_0 : \text{No difference between population 1 and population 2}$$

were **true**, then there should be **no pattern** in the group labels when sorted into ascending order; the sorted data would give rise a **random** assortment of group 1 and group 2 labels.

4. To obtain the exact distribution of W under H_0 (for the assessment of statistical significance), we could compute W for all possible permutations of the group labels, and then form the probability distribution of the values of W . We call this the **permutation null distribution**.
5. But W is a rank sum, so we can compute the permutation null distribution simply by tabulating **all possible subsets** of size n_2 of the set of ranks $\{1, 2, 3, \dots, n\}$.
6. There are

$$\binom{n}{n_2} = \frac{n!}{n_1! n_2!} = N$$

say possible subsets of size n_2 ; for $n = 6$ and $n_2 = 2$, the number of subsets of size n_2 is

$$\binom{6}{2} = \frac{6!}{2! 4!} = 15$$

However, the number of subsets increases dramatically as n increases; for $n_1 = n_2 = 10$, so that $n = 20$, the number of subsets of size n_2 is

$$\binom{20}{10} = \frac{20!}{10! 10!} = 184756$$

7. The exact rejection region and p-value are computed from the permutation null distribution. Let $W_i, i = 1, \dots, N$ denote the value of the Wilcoxon statistic for the N possible subsets of the ranks of size n_2 . The probability that the test statistic, W , is less than or equal to w is

$$\Pr[W \leq w] = \frac{\text{Number of } W_i \leq w}{N}$$

We seek the values of w that give the appropriate rejection region, $\mathcal{R}$, so that

$$\Pr[W \in \mathcal{R}] = \frac{\text{Number of } W_i \in \mathcal{R}}{N} = \alpha$$

It may not be possible to find critical values, and define $\mathcal{R}$, so that this probability is **exactly** α as the distribution of W is **discrete**.

EXAMPLE Simple Example

Suppose $n_1 = 7$ and $n_2 = 3$. There are

$$\binom{10}{3} = \frac{10!}{7! 3!} = 120$$

subsets of the ranks $\{1, 2, 3, \dots, 10\}$ of size 3. The subsets and rank sums are listed below:

Ranks	W	Ranks	W	Ranks	W	Ranks	W
1 2 3	6	1 7 8	16	2 7 10	19	4 6 7	17
1 2 4	7	1 7 9	17	2 8 9	19	4 6 8	18
1 2 5	8	1 7 10	18	2 8 10	20	4 6 9	19
1 2 6	9	1 8 9	18	2 9 10	21	4 6 10	20
1 2 7	10	1 8 10	19	3 4 5	12	4 7 8	19
1 2 8	11	1 9 10	20	3 4 6	13	4 7 9	20
1 2 9	12	2 3 4	9	3 4 7	14	4 7 10	21
1 2 10	13	2 3 5	10	3 4 8	15	4 8 9	21
1 3 4	8	2 3 6	11	3 4 9	16	4 8 10	22
1 3 5	9	2 3 7	12	3 4 10	17	4 9 10	23
1 3 6	10	2 3 8	13	3 5 6	14	5 6 7	18
1 3 7	11	2 3 9	14	3 5 7	15	5 6 8	19
1 3 8	12	2 3 10	15	3 5 8	16	5 6 9	20
1 3 9	13	2 4 5	11	3 5 9	17	5 6 10	21
1 3 10	14	2 4 6	12	3 5 10	18	5 7 8	20
1 4 5	10	2 4 7	13	3 6 7	16	5 7 9	21
1 4 6	11	2 4 8	14	3 6 8	17	5 7 10	22
1 4 7	12	2 4 9	15	3 6 9	18	5 8 9	22
1 4 8	13	2 4 10	16	3 6 10	19	5 8 10	23
1 4 9	14	2 5 6	13	3 7 8	18	5 9 10	24
1 4 10	15	2 5 7	14	3 7 9	19	6 7 8	21
1 5 6	12	2 5 8	15	3 7 10	20	6 7 9	22
1 5 7	13	2 5 9	16	3 8 9	20	6 7 10	23
1 5 8	14	2 5 10	17	3 8 10	21	6 8 9	23
1 5 9	15	2 6 7	15	3 9 10	22	6 8 10	24
1 5 10	16	2 6 8	16	4 5 6	15	6 9 10	25
1 6 7	14	2 6 9	17	4 5 7	16	7 8 9	24
1 6 8	15	2 6 10	18	4 5 8	17	7 8 10	25
1 6 9	16	2 7 8	17	4 5 9	18	7 9 10	26
1 6 10	17	2 7 9	18	4 5 10	19	8 9 10	27

There are 22 possible rank sums, $\{6, 7, 8, \dots, 25, 26, 27\}$; the number of times each is observed is displayed in the table below, with the corresponding probabilities and cumulative probabilities.

W	6	7	8	9	10	11	12	13	14	15	16
Frequency	1	1	2	3	4	5	7	8	9	10	10
Prob.	0.008	0.008	0.017	0.025	0.033	0.042	0.058	0.067	0.075	0.083	0.083
Cumulative Prob.	0.008	0.017	0.033	0.058	0.092	0.133	0.192	0.258	0.333	0.417	0.500
W	17	18	19	20	21	22	23	24	25	26	27
Frequency	10	10	9	8	7	5	4	3	2	1	1
Prob.	0.083	0.083	0.075	0.067	0.058	0.042	0.033	0.025	0.017	0.008	0.008
Cumulative Prob.	0.583	0.667	0.742	0.808	0.867	0.908	0.942	0.967	0.983	0.992	1.000

Thus, for example, the probability that $W = 19$ is 0.075, with a frequency of 9 out of 120. From this table:

$$\Pr[8 \leq W \leq 25] = 0.983 - 0.017 = 0.966$$

implying that the two-sided rejection region for $\alpha = 0.05$ is the set $\mathcal{R} = \{6, 7, 26, 27\}$.

For these sample sizes, the exact calculation can be performed using the `combinat` library

```
> n1<-7;n2<-3;n<-n1+n2
> Rvec<-c(1:n)
>
> #use the combinat library
> Cmat<-combinat::combn(n,n2)
> Scores<-apply(Cmat,2,sum)
> Prob<-as.numeric(table(Scores))/length(Scores)
> CProb<-cumsum(Prob)
> par(mar=c(4,4,1,2))
> plot(unique(Scores),Prob, ylim=range(0,0.15),
+       pch=18,xlab="W",ylab="Prob.")
```

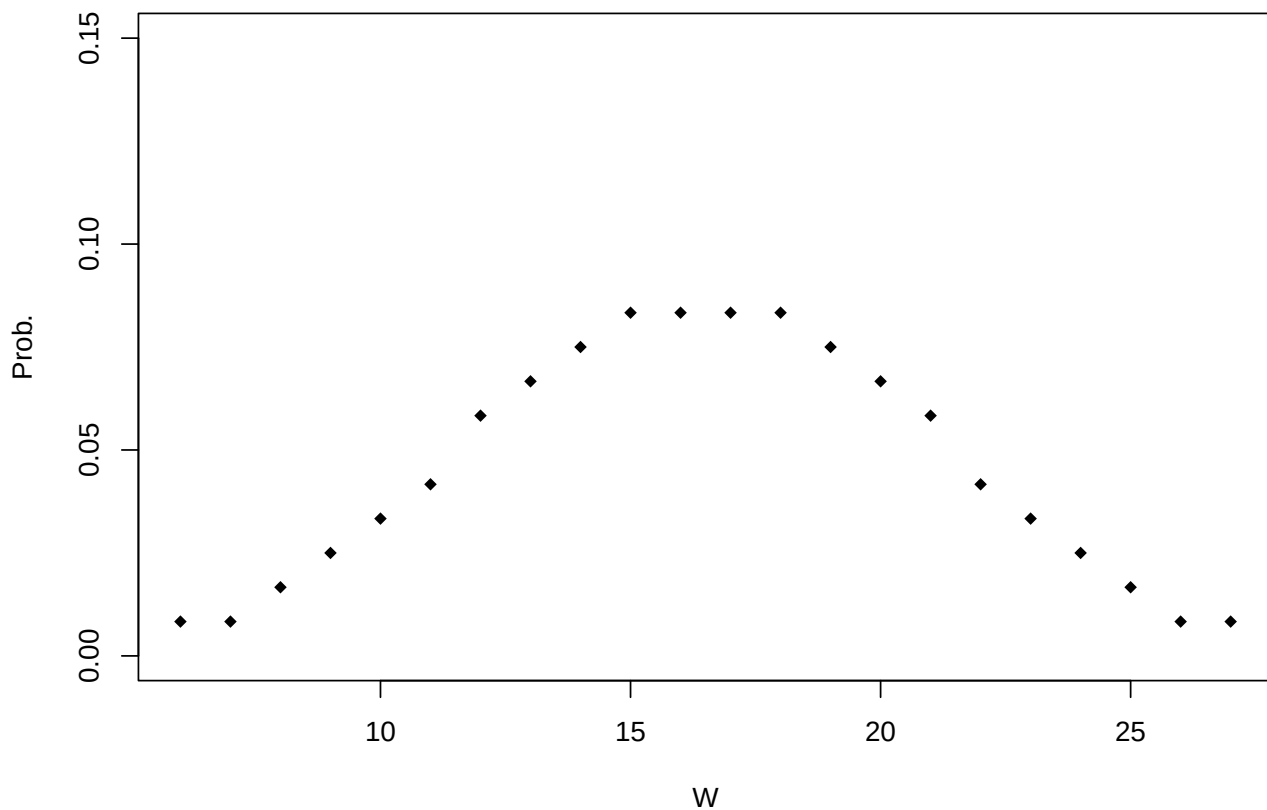


Figure 28: Permutation pmf

```
> par(mar=c(4,4,1,2))
> plot(unique(Scores),CProb,, ylim=range(0,1),
+       pch=18,xlab="W",ylab="Cumulative Prob.")
```

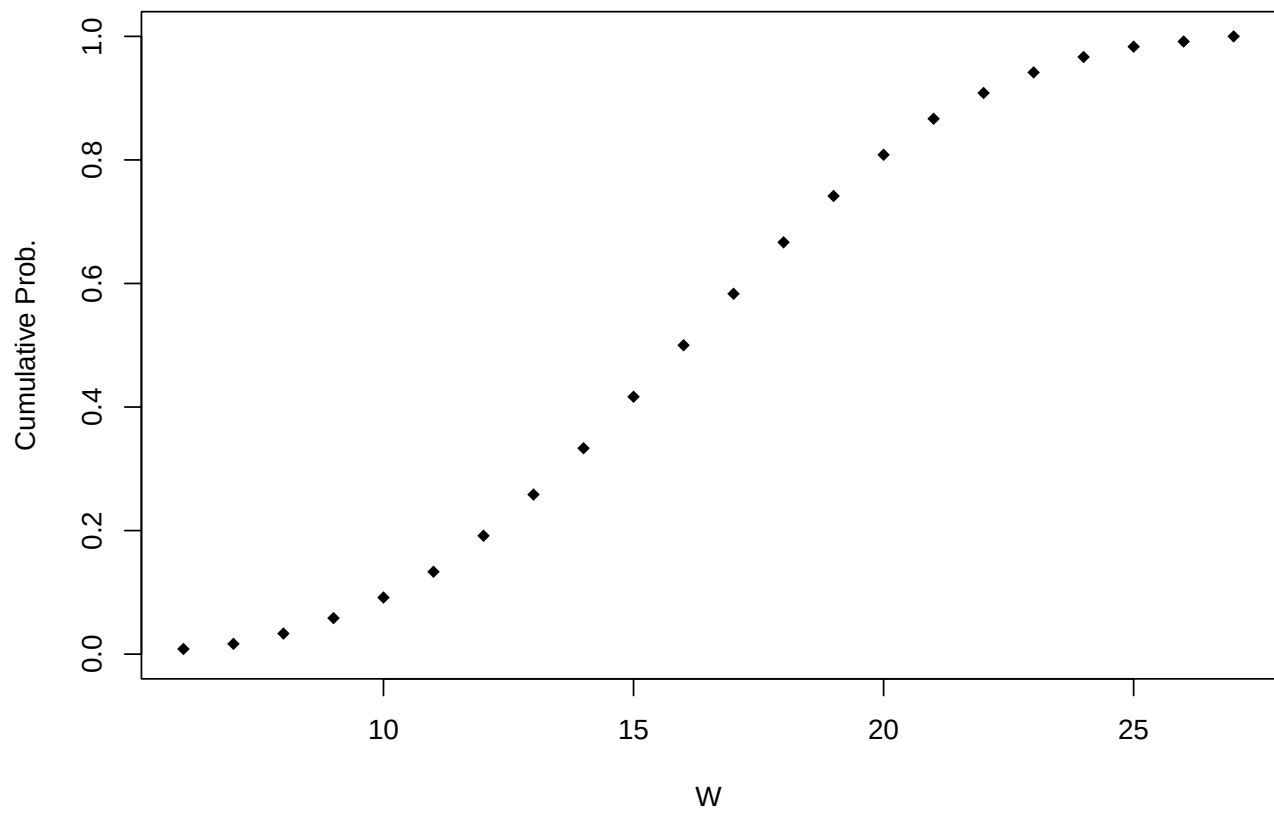


Figure 29: Permutation cdf