Statistical Methods in Medical Research

P. Armitage
MA, PhD
Emeritus Professor of Applied Statistics
University of Oxford

G. Berry
MA, PhD
Professor in Epidemiology and Biostatistics
University of Sydney

J.N.S. Matthews
MA, PhD
Professor of Medical Statistics
University of Newcastle upon Tyne

FOURTH EDITION

Blackwell
Science

Editorial Offices:
Osney Mead, Oxford OX2 0EL
25 John Street, London WC1N 2BS
23 Ainslie Place, Edinburgh EH3 6AJ
350 Main Street, Malden
MA 02148-5018, USA
54 University Street, Carlton
Victoria 3053, Australia
10, rue Casimir Delavigne
75006 Paris, France

Other Editorial Offices:
Blackwell Wissenschafts-Verlag GmbH
Kurfürstendamm 57
10707 Berlin, Germany

Blackwell Science KK
MG Kodenmacho Building
7-10 Kodenmacho Nihombashi
Chuo-ku, Tokyo 104, Japan

Iowa State University Press
A Blackwell Science Company
2121 S. State Avenue
Ames, Iowa 50014-8300, USA

First published 1971
Reprinted 1973, 1974, 1977, 1980, 1983, 1985
Italian editions 1975, 1977, 1979, 1981
Polish edition 1978
Second edition 1987
Reprinted 1988 (twice), 1990, 1991, 1993
Spanish edition 1992
Third edition 1994
Reprinted 1995, 1996
Fourth edition 2002

Set by Kolam Information Services Pvt. Ltd.,
Pondicherry, India
Printed and bound in Great Britain
by MPG Books Ltd, Bodmin, Cornwall

DISTRIBUTORS

Marston Book Services Ltd
PO Box 269
Abingdon, Oxon OX14 4YN
(*Orders*: Tel: 01235 465500
Fax: 01235 465555)

USA
Blackwell Publishing
c/o AIDC
PO Box 20
50 Winter Sport Lane
Williston, VT 05495-0020
(*Orders*: Tel: 800 216 2522
Fax: 802 864 7626)

Canada
Login Brothers Book Company
324 Saulteaux Crescent
Winnipeg, Manitoba R3J 3T2
(*Orders*: Tel: 204 837 2987)

Australia
Blackwell Science Pty Ltd
54 University Street
Carlton, Victoria 3053
(*Orders*: Tel: 3 9347 0300
Fax: 3 9347 5001)

A catalogue record for this title is available from the British Library

ISBN 0-632-05257-0

Library of Congress
Cataloging-in-publication Data

Armitage, P.
Statistical methods in medical research /
P. Armitage, G. Berry,
J.N.S. Matthews.—4th ed.
p. cm.
Includes bibliographical references
and indexes.
ISBN 0-632-05257-0
1. Medicine—Research—Statistical methods.
I. Berry, G. (Geoffrey) II. Matthews, J.N.S.
III. Title.
[DNLM: 1. Biometry. 2. Research—methods.
WA 950 A733s 2001]
R852 .A75 2001
610′.7′27—dc21
00–067992

For further information on Blackwell Science, visit our website: www.blackwell-science.com

	Values of	Values of		
Pair	x_1	x_2	Sum	Difference
1	x_{11}	x_{21}	X_1	Y_1
2	x_{12}	x_{22}	X_2	Y_2
·	·	·	·	·
·	·	·	·	·
·	·	·	·	·
n	x_{1n}	x_{2n}	X_n	Y_n

Denoting by x'_{ij} the deviation of x_{ij} from its expectation,

$$\begin{aligned} \mathrm{cov}(X_i, Y_i) &= \mathrm{E}[(x'_{1i} + x'_{2i})(x'_{1i} - x'_{2i})] \\ &= \mathrm{E}[(x'_{1i})^2 - (x'_{2i})^2] \\ &= \mathrm{var}(x_{1i}) - \mathrm{var}(x_{2i}). \end{aligned} \tag{7.25}$$

A test of the equality of $\mathrm{var}(x_{1i})$ and $\mathrm{var}(x_{2i})$ is, therefore, the same as a test of the hypothesis that the covariance of X_i and Y_i is zero, which means that the correlation coefficient must also be zero. The test of equality of variances may, therefore, be effected by any of the equivalent tests described above for the hypothesis of zero association between X_i and Y_i. This test is due to E.J.G. Pitman. Its adaptation for purposes of estimation is described by Snedecor and Cochran (1989, §10.8).

7.5 Regression to the mean

The term 'regression' was introduced by Sir Francis Galton (1822–1911) to express the fact that for many inherited characteristics, such as height, the measurements on sons are on average closer to the population mean than the corresponding values for their fathers. The regression coefficient of son's height on father's height is less than 1. The regression of father's height on son's height is also less than 1, so that whether one looks forwards or backwards in generations there is a *regression to the mean.*

Similar phenomena are widely observed in various branches of medicine. Suppose, for instance, that a person's systolic blood pressure is determined by the average of five readings to be 162 mmHg. This is somewhat above the likely population mean of, say, 140 mmHg. If a further reading is taken from the same subject, it will *on average* tend to be less than 162 mmHg. Conversely, a subject with a sample mean less than the population mean will tend to show an increase when retested.

The explanation is closely related to the discussion on shrinkage in §6.4, and anticipates a relevant description of *components of variance* in §8.3. Suppose that repeated blood pressure readings on the ith subject follow a distribution with mean μ_i and variance σ^2, and that the values of μ_i vary from one subject to another with a mean μ_0 and variance σ_0^2. (As noted in §8.3, σ^2 and σ_0^2 are 'components of variance' within and between subjects, respectively.) In terms of the discussion of §6.2, the variation between subjects may be thought of as a prior distribution for μ_i, and the distribution of repeated readings on one subject as providing the likelihood. It follows from the discussion in §§6.2 and 6.4 that, for a subject with a mean blood pressure of $\bar{x}_i$ based on n observations, the estimate of the true mean μ_i will tend to be shrunk from $\bar{x}_i$ towards the population mean μ_0, as will the mean value of a future set of replicate readings.

Suppose that for each subject two independent samples of size n are taken, with means $\bar{x}_{i1}$ and $\bar{x}_{i2}$ for the ith subject. The correlation coefficient between $\bar{x}_{i1}$ and $\bar{x}_{i2}$ is

$$\rho = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}. \tag{7.26}$$

If the distributions between and within subjects are normal, the joint distribution of $\bar{x}_1$ and $\bar{x}_2$ will be bivariate normal, as in Fig. 7.6, and (7.26) will also measure the regression coefficient of either of the means on the other one. If there were no within-subject variation ($\sigma^2 = 0$), this regression coefficient would be unity. The regression towards the mean is therefore measured by $1 - \rho$, so the smaller ρ is, the greater is the regression towards the mean. Small values of ρ will occur when σ_0^2 is small in relation to the sampling error of the means, e.g. when there is very little real heterogeneity between subjects. High values of ρ will occur when the sampling error is small relative to variance between individuals, and this may be achieved by increasing n.

In some studies of medical interventions, subjects are selected by preliminary screening tests as having high values of a relevant test measurement. For example, in studies of cholesterol-lowering agents, subjects may be selected as having serum cholesterol levels above some critical value, either on a single or on the mean of repeated determinations. On average, even with no effect of the agent under test, subsequent readings on the same subjects would tend to regress toward the mean. A significant reduction below the pretreatment values thus provides no convincing evidence of a treatment effect.

Nevertheless, if precise estimates of the components of variance, and hence of ρ, are available, the extent of regression to the mean can be calculated and allowed for. Gardner and Heady (1973) studied this effect, in relation to cholesterol, blood pressure and daily calorie intake, and gave formulae for the regression effects for various levels of the initial screening cut-off point. As noted above, the effect is reduced by increasing the number of observations made at the initial screen. Johnson and George (1991) extended this work by distinguishing between two sources of within-subject variation: measurement error and physiological fluctuations. The latter type of variation may result in fluctuations that are not independent, showing perhaps cyclical or other trends, a topic

$$s_i^2 = [S_i - (T_i^2/n_i)]/(n_i - 1),$$

and an approximate or *empirical* weight calculated as

$$w_i' = n_i/s_i^2.$$

We calculate the weighted sum of squares (8.15) as

$$G'' = \sum_i w_i'\bar{y}_i^2 - (\sum_i w_i'\bar{y}_i)^2/\sum_i w_i'.$$

On the null hypothesis of homogeneity of the μ_i, G'' is distributed approximately as $\chi^2_{(k-1)}$, high values indicating excessive disparity between the $\bar{y}_i$. The approximation is increasingly inaccurate for smaller values of the n_i (say, below about 10), and a refinement is given by James (1951) and Welch (1951). For an application of this method in a comparison of sets of pock counts, see Armitage (1957, p. 579).

8.3 **Components of variance**

In some studies which lead to a one-way analysis of variance, the groups may be of no great interest individually, but may nevertheless represent an interesting source of variation. The result of a pipetting operation, for example, may vary from one pipette to another. A comparison between a particular pair of pipettes would be of little interest; furthermore, a test of the null hypothesis that the different pipettes give identical results on average may be pointless because there may quite clearly be a systematic difference between instruments. A more relevant question here will be: how great is the variation between pipettes as compared with that of repeated readings on the same pipette?

A useful framework is to regard the k groups as being randomly selected from a population of such groups. This will not usually be strictly true, but it serves as an indication that the groups are of interest only as representing a certain type of variation. This framework is often called *Model II*, or the *random-effects model*, as distinct from *Model I*, or the *fixed-effects model*, considered in §8.1.

Suppose, in the first instance, that each group contains the same number, n, of observations. (In the notation of §8.1, all the n_i are equal to n.) Let μ_i be the 'true' mean for the ith group and suppose that in the population of groups μ_i is distributed with mean μ and variance σ_B^2. Readings within the ith group have mean μ_i and variance σ^2. The quantities σ^2 and σ_B^2 are called *components of variance* within and between groups respectively. The situation is illustrated in Fig. 8.1. The data at our disposal consist of a random sample of size n from each of k randomly selected groups.

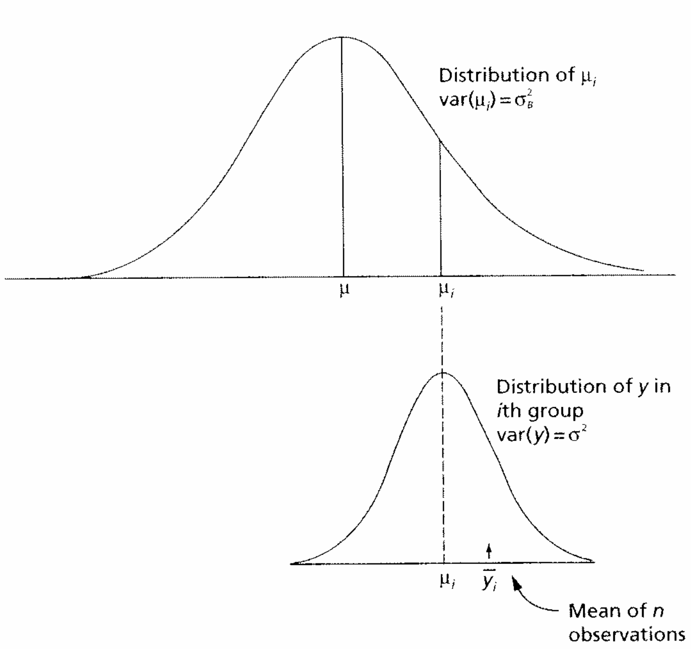

Fig. 8.1 Components of variance between and within groups, with normal distributions for each component of random variation.

Consider first the variance of a single group mean, $\bar{y}_i$. We have

$$\bar{y}_i - \mu = (\mu_i - \mu) + (\bar{y}_i - \mu_i),$$

and the two terms in parentheses represent independent sources of variation—that of μ_i about μ and that of $\bar{y}_i$ about μ_i. Therefore, using (5.10),

$$\begin{aligned} \text{var}(\bar{y}_i) &= \text{var}(\mu_i) + \text{var}(\bar{y}_i, \text{given } \mu_i) \\ &= \sigma_B^2 + (\sigma^2/n). \end{aligned} \tag{8.17}$$

Now, the analysis of variance will have the following structure:

	DF	MSq
Between groups	$k-1$	s_B^2
Within groups	$k(n-1)$	s_W^2
Total	$nk-1$ $(= N-1)$	

The Between-Groups SSq is (with notation for T_i and T as in §8.1)

$$\begin{aligned} &\frac{\sum T_i^2}{n} - \frac{T^2}{N} \\ &= n\left[\sum \bar{y}_i^2 - (\sum \bar{y}_i)^2/k\right], \end{aligned}$$

summations running from $i = 1$ to k. Thus, the Between-Groups MSq, s_B^2,

$$= n\left[\sum \bar{y}_i^2 - (\sum \bar{y}_i)^2/k\right]/(k-1)$$

$=$ n times an unbiased estimate of var$(\bar{y}_i)$
$=$ n times an unbiased estimate of $\sigma_B^2 + (\sigma^2/n)$, from (8.17),
$=$ an unbiased estimate of $n\sigma_B^2 + \sigma^2$.
That is,

$$\mathrm{E}(s_B^2) = \sigma^2 + n\sigma_B^2. \tag{8.18}$$

This important result is similar to (8.9), which is the analogous result for the situation in which the μ_i are fixed.

The Within-Groups MSq is (from §8.1) an unbiased estimate of σ^2. The result (8.18) thus confirms the plausibility of the F test, for the null hypothesis is that $\sigma_B^2 = 0$ and in this case both mean squares are unbiased estimates of σ^2. If $\sigma_B^2 > 0, s_B^2$ will on average be greater than s_W^2, and F will tend to be greater than 1.

To estimate σ_B^2, note that

$$\begin{aligned}\mathrm{E}(s_B^2 - s_W^2) &= \mathrm{E}(s_B^2) - \mathrm{E}(s_W^2)\\ &= (\sigma^2 + n\sigma_B^2) - \sigma^2\\ &= n\sigma_B^2.\end{aligned}$$

Hence, an unbiased estimate of σ_B^2 is given by

$$\hat{\sigma}_B^2 = \frac{s_B^2 - s_W^2}{n}. \tag{8.19}$$

If $s_B^2 < s_W^2$ (as will often be the case if σ_B^2 is zero or near zero), $\hat{\sigma}_B^2$ is negative. There is a case for replacing $\hat{\sigma}_B^2$ by 0 when this happens, but it should be noted that the unbiased property of (8.19) is then lost.

Example 8.2

Bacharach *et al.* (1940) carried out an experiment on 'diffusing factor', a substance which, when present in an inoculation into the skin of rabbits, spreads the blister caused by the inoculation. They gave inoculations of the same dose at six sites on the back of each of six animals. Their experimental design permitted a study of the influence of the particular site and the order of administration, but there was no evidence that these factors had any effect and we shall regard the data as forming a one-way classification: between and within animals. The variable analysed is the area of the blister (square centimetres).

An analysis of variance was as follows:

	SSq	DF	MSq	VR
Between animals	12·8333	5	2·5667	4·39
Within animals	17·5266	30	0·5842	
Total	30·3599	35		

We have

$$\begin{aligned}\hat{\sigma}_B^2 &= (2{\cdot}5667 - 0{\cdot}5842)/6\\ &= 0{\cdot}3304.\end{aligned}$$

The two components of variance are then estimated as follows, where each is expressed as a percentage of the total:

Between animals	$\hat{\sigma}_B^2$	0·3304	36%
Within animals	s_W^2	0·5842	64%
Total	$\hat{\sigma}_B^2 + s_W^2$	0·9146	100%

The sum of the two components is the estimated variance of a single reading from a randomly chosen rabbit, and the analysis shows that of this total variance 36% is estimated to be attributable to systematic differences between rabbits. (For further analysis of these data, see below; see also Example 9.4, p. 260.)

Confidence limits for σ^2 are obtained from the Within-Groups SSq by use of the χ^2 distribution, as in §5.1. Confidence limits for σ_B^2 are rather more troublesome. An approximate solution is recommended by Boardman (1974). For $100(1-\alpha)\%$ confidence limits we need various entries in the F table corresponding to a tabulated one-sided level of $\frac{1}{2}\alpha$. Thus, for 95% confidence limits we need entries corresponding to $P = 0{\cdot}025$. Denoting the entry for degrees of freedom ν_1 and ν_2 as F_{ν_1,ν_2}, and putting $f_1 = k-1$, $f_2 = k(n-1)$, we need

$$\begin{aligned}F_1 &= F_{f_1,f_2}\\ F_2 &= F_{f_1,\infty}\\ F_3 &= F_{f_2,f_1}\\ F_4 &= F_{\infty,f_1}\\ F &= \text{observed value}, s_B^2/s_W^2.\end{aligned}$$

Then the upper limit for σ_B^2 is

$$\hat{\sigma}_{BU}^2 = F_4\left(F - \frac{1}{F_3}\right)\left(\frac{s_W^2}{n}\right) \tag{8.20}$$

and the lower limit is

$$\hat{\sigma}_{BL}^2 = \left(\frac{F - F_1}{F_2}\right)\left(\frac{s_W^2}{n}\right). \tag{8.21}$$

Note that the lower limit is zero if $F = F_1$, i.e. if F is just significant by the usual test. If $F < F_1$, the lower limit will be negative and in some instances the upper limit may also be negative. For a discussion of this apparent anomaly, see Scheffé (1959, §7.2). The validity of these limits will depend rather heavily on the assumption of normality, particularly for the between-groups variation.

Example 8.2, continued

The 95% confidence limits for σ^2 are (from §5.1)

$$\frac{17{\cdot}5266}{46{\cdot}98}$$

and

$$\frac{17{\cdot}5266}{16{\cdot}79},$$

i.e.

$$0{\cdot}373 \text{ and } 1{\cdot}044,$$

the divisors being the appropriate percentiles of the $\chi^2_{(30)}$ distribution.

For confidence limits for σ^2_B we need the following tabulated values of F, writing $f_1 = 5$, $f_2 = 30$:

$$F_1 = 3{\cdot}03, F_2 = 2{\cdot}57, F_3 = 6{\cdot}23, F_4 = 6{\cdot}02,$$

and the observed F is 4·39. Thus, from (8.20) and (8.21),

$$\hat{\sigma}^2_{BU} = (6{\cdot}02)\left(4{\cdot}39 - \frac{1}{6{\cdot}23}\right)(0{\cdot}0974) = 2{\cdot}48$$

and

$$\hat{\sigma}^2_{BL} = \left(\frac{4{\cdot}39 - 3{\cdot}03}{2{\cdot}57}\right)(0{\cdot}0974) = 0{\cdot}052.$$

The wide ranges of error associated with these estimates makes it clear that the percentage contributions of 36% and 64% are very imprecise estimates indeed.

If the numbers of observations from the groups are unequal, with n_i from the ith group, (8.19) must be modified as follows:

$$\hat{\sigma}^2_B = \frac{s^2_B - s^2_W}{n_0}, \qquad (8.22)$$

where

$$n_0 = \frac{1}{(k-1)}\left[N - \frac{\sum n_i^2}{N}\right].$$

A further difficulty is that (8.22) is not necessarily the best way of estimating σ^2_B. The choice of method depends, however, on the unknown ratio of the variance components which are being estimated, and (8.22) will usually be a sensible method if the n_i are not too different. See Searle (1987, §13.3) and Robinson (1998) for a fuller discussion of the issues involved. The situation corresponds to a multilevel model, methods for analysing which are given in §12.5.

8.4 Multiple comparisons

We return now to the fixed-effects model of §8.1. In the analysis of data in this form it will usually be important not to rely solely on the analysis of variance table and its F test, but to examine the differences between groups more closely to see what patterns emerge. It is, in fact, good practice habitually to report the mean values $\bar{y}_i$ and their standard errors, calculated as $s_W/\sqrt{n_i}$, in terms of the Within-Groups MSq s^2_W unless the assumption of constant variance is clearly inappropriate.

The standard error of the difference between two means is given by (8.11), and the t distribution may be used to provide a significance test or to assign confidence limits, as indicated in §8.1. If all the n_i are equal (to n, say), it is sometimes useful to calculate the *least significant difference* (LSD) at a certain significance level. For the 5% level, for instance, this is

$$t_{f_2,\,0{\cdot}05}\, s_W \sqrt{(2/n)},$$

where $f_2 = k(n-1)$, the degrees of freedom within groups. Differences between pairs of means which are significant at this level can then be picked out by eye.

Sometimes interest is focused on comparisons between the group means other than simple differences. These will usually be measurable by a *linear contrast* of the form

$$L = \sum \lambda_i \bar{y}_i, \qquad (8.23)$$

where $\sum \lambda_i = 0$. From (5.10),

$$\text{var}(L) = \sum \lambda_i^2 \,\text{var}(\bar{y}_i),$$

and the standard error of L is thus estimated as

$$\text{SE}(L) = s_W \sqrt{(\sum \lambda_i^2/n)}, \qquad (8.24)$$

and the usual t test or confidence limits may be applied.

Some examples of linear contrasts are as follows.

1 *A contrast of one group with the mean of several other groups.* One group may have a special identity, perhaps as a control group, and there may be some reason for pooling a set of q other groups (e.g. if related treatments have been applied to these groups). The relevant comparison will then be

Example 9.6

Table 9.12 gives counts of particle emission during periods of 1000 s, for 30 aliquots of equal size of certain radioactive material. Each aliquot is placed twice in the counter. There are three sources of random variation, each with its component of variance, as follows.

1 Variation between aliquots, with a variance component σ_2^2. This may be due to slight variations in size or in radioactivity, or to differences in technique between the 30 occasions on which the different aliquots were examined.

2 Systematic variation between replicate counts causing changes in the expected level of the count, with a variance component σ_1^2. This may be due to systematic biases in counting which affect different counts in different ways, or to inconsistency in the apparatus, due perhaps to variation in the way the material is placed in the counter.

3 Random variation from one time period to another, all other conditions remaining constant: variance component σ_0^2. There is no replication of counts under constant conditions, but we know that this form of variation follows the Poisson distribution (§3.7), in which the variance equals the mean. The mean will vary a little over the whole experiment, but to a close approximation we could estimate σ_0^2 by the observed mean for the whole data, 303·6.

Table 9.12 Radioactivity counts during periods of 1000s.

Aliquot	Counts		Aliquot	Counts	
1	281	291	16	325	267
2	309	347	17	284	296
3	316	356	18	255	281
4	289	277	19	347	285
5	322	292	20	326	302
6	287	321	21	347	307
7	338	320	22	292	344
8	333	275	23	322	308
9	319	311	24	294	272
10	258	302	25	307	303
11	338	294	26	281	331
12	319	281	27	284	322
13	307	247	28	287	305
14	279	259	29	318	352
15	326	272	30	307	301

The analysis of variance is that for a simple one-way classification and is as follows:

	SSq	DF	MSq	Expected value of MSq
Between aliquots	19 898	29	686·1	$\sigma_0^2 + \sigma_1^2 + 2\sigma_2^2$
Within aliquots	20 196	30	673·2	$\sigma_0^2 + \sigma_1^2$
Total	40 094	59		
Poisson			303·6	σ_0^2

The expected values of the mean squares follow from §8.3, if we note that the within-aliquots variance component is $\sigma_0^2 + \sigma_1^2$ (since differences between replicate counts are affected by variation of both type **2** and type **3**), and that the between-aliquots component is σ_2^2.

The estimates of the variance components are now obtained:

$$\begin{aligned} \sigma_2^2 &= (686{\cdot}1 - 673{\cdot}2)/2 = 6{\cdot}4 \\ \sigma_1^2 &= 673{\cdot}2 - 303{\cdot}6 \qquad = 369{\cdot}6 \\ \sigma_0^2 &= 303{\cdot}6 \end{aligned}$$

These estimates are, of course, subject to sampling error, but there is clearly no evidence of any large component, σ_2^2, due to aliquot differences. Replicate counts vary, however, by substantially more than can be explained by the Poisson distribution.

9.5. The number of swabs positive for *Pneumococcus* is recorded in families. A simple model might assume that the mean number of swabs is μ and the observation on the jth member of the ith family can be modelled by:

$$y_{ij} = \mu + \varepsilon_{ij}, \tag{12.33}$$

where ε_{ij} is a simple error term, with zero mean and variance σ^2, that is independent from observation to observation. However, this model does not reflect the fact that the observations are grouped within families. Moreover, if the variation between families is larger than that within families, then this cannot be modelled because only one variance has been specified. An obvious extension is to add an extra term, ξ_i, to (12.33) to accommodate variation between families. This term will be a random variable that is independent between families and of the ε_{ij}, has zero mean and variance σ_F^2. Thus, the new model for the jth member of the ith family is:

$$\mu + \xi_i + \varepsilon_{ij}.$$

It should be noted that it is the same realization of the random variable that is applied to each observation within a family; a consequence of this is that observations within a family are correlated. Two observations within a family have covariance σ_F^2 and, as each observation has variance $\sigma_F^2 + \sigma^2$, the correlation is

$$\sigma_F^2/(\sigma_F^2 + \sigma^2). \tag{12.34}$$

This is not surprising; the model is such that families with a propensity to exhibit pneumococcal infection will have a large value for ξ_i and as this is applied to each member of the family, each family member will tend to report a large value—that is, the values are correlated. Clearly, this tendency will be less marked if the within-family variation is substantial relative to that between families; this is reflected in (12.34) because, as σ^2/σ_F^2 becomes larger, (12.34) becomes smaller. It should be noted that correlations generated in this way cannot be negative: they are examples of the *intraclass correlation* discussed in §19.11.

Because they are random variables, the terms ξ and ε are referred to as *random effects* and their effect is measured by a variance or, more accurately, a *component of variance*, such as σ^2 and σ_F^2. More elaborate models can certainly be built. One possibility is to add extra terms that are not random (and so are often referred to as fixed effects) to elaborate on the simple mean μ. In Example 9.5 the families were classified into three categories measuring how crowded their living conditions were. The model could be extended to

$$\mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \xi_i + \varepsilon_{ij}, \tag{12.35}$$

where $x_{1i} = 1$ if the ith family lives in crowded conditions and is 0 otherwise, and $x_{2i} = 1$ if the ith family lives in uncrowded conditions and is 0 otherwise. The and the restriction of response measurements to a random sample of individuals in each cluster, rather than the whole cluster population.

The most important point about cluster randomization is that, although response measurements may be made on individual subjects, the treatment comparisons are less precise than they would be if subjects were individually randomized. The reason is that responses of subjects within the same cluster are likely to be positively correlated. In the extreme case of perfect correlation, the advantage of replication within clusters is completely lost, so that the effective sample size is the number of clusters rather than the number of subjects. Essentially the same point arises with two-stage sampling (§19.2).

Suppose that there are $2K$ clusters, to be assigned randomly to two treatments, with K clusters in each, and that the ith cluster contains n_i individuals, with a total sample size $N = \Sigma n_i$ for each treatment. It can be shown that the variance of the overall mean response in either group is inflated by a factor $1 + \rho[(\Sigma n_i^2/N) - 1]$, called the *design effect*, where ρ is the correlation between responses for individuals in the same cluster (the *intraclass correlation*; see §19.11). If the clusters are all the same size n, this factor becomes $1 + \rho(n - 1)$. If $\rho = 1$, the design effect becomes n, and the variance of the overall mean is proportional to $1/K$ rather than $1/N$, as noted in the previous paragraph. At the design stage, sample sizes determined by the methods described earlier should be multiplied by the design effect (Donner, 1992).

One simple approach to the analysis of a cluster-randomized trial is to summarize the responses in each cluster by their mean and use these cluster means in a standard two-sample analysis—for example, by a two-sample t test or a Mann–Whitney non-parametric test. A theoretical disadvantage in this approach is that the cluster means will have different variances if the n_i differ. This is unlikely to be a serious problem unless the n_i differ grossly.

A more formal approach is to represent the data by a random-effects model (§§8.3, 12.5), with variance components σ_B^2 for the variation between clusters, and σ_W^2 within clusters. These variance components can be estimated from the data (§8.3). If the n_i vary, the modification given by (8.22) may be used. The variance components are related to the intraclass correlation by the equation $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$ (see §19.11), which may be used to estimate the design effect given above.

The more elaborate analysis of §12.5 may be used if the effects of covariates (defined either for the cluster or for the individual subject) are to be taken into account.

A comprehensive account of cluster-randomized trials is given by Murray (1998).

Suppose a new method is available for measuring some variable and it is required to assess the agreement between values obtained using the new method and those obtained with an existing method. Data could be collected in which a number of individuals were measured using each method to give a pair of values for each individual. A high correlation about a line not passing through the origin or with a slope different from 1 would not represent useful agreement between the methods as far as the user was concerned, although if this occurred it might be possible to calibrate the new method to give better agreement. Thus although, unlike the twins, there is no ambiguity on the identification within pairs, we require a measure of correlation with the property that it will only equal unity if the two measurements are identical within each individual.

The intraclass correlation coefficient is a measure of the correlation between the values obtained with any two randomly chosen methods within the same individual (class) and has the above property. The correlation calculated from $2n$ pairs as above is approximately equal to the intraclass correlation coefficient except when n is small. The method is closely related to components of variance (§§8.3 and 9.6), and using this methodology is more convenient and more accurate than forming multiple pairs. In general, there may be more than two methods under test, and we suppose that there are m methods, each assessed on n subjects. Then the design is equivalent to randomized blocks (§9.2), in which the methods are tested within the subjects, i.e. the blocks. There are three sources of variation, each with its component of variance:

1. Variation between subjects, with a variance component σ_s^2.
2. Systematic variation between methods, with a variance component σ_m^2. This variability represents differences between methods.
3. Random variation from one measurement to another, with a variance component σ^2, additional to the sources of variation **1** and **2**.

Then the two-way analysis of variance has the form:

	DF	MSq	Expected MSq
Between subjects (classes)	$n-1$	M_s	$\sigma^2 + m\sigma_s^2$
Between methods	$m-1$	M_m	$\sigma^2 + n\sigma_m^2$
Residual	$(n-1)(m-1)$	M_r	σ^2

The intraclass correlation coefficient is defined as the correlation between any two measurements in the same subject using randomly chosen methods. All three components of variation contribute to the variance of each measurement and, since the two measurements are for the same subject, the variance component representing variation between subjects is common to the two measurements. Therefore the intraclass correlation coefficient is