Patty Solomon, Tyman Stanford
Components of variance
Diagnostic Histopathology, 2016; 22(7):253-262

http://dx.doi.org/10.1016/j.mpdhp.2016.06.006

Components of variance

Patty Solomon, BSc(Hons), DipMathStats(Camb), PhD (Imperial College)[1]
Tyman Stanford, BMa&CompSci(Hons), PhD (Adelaide)[2]

September 5, 2016

[1]Professor of Statistical Bioinformatics
School of Mathematical Sciences
The University of Adelaide
Adelaide SA 5005
Australia
Email: patty.solomon@adelaide.edu.au
Telephone: +61 8 8313 3033
Fax: +61 8 8313 3696
Corresponding Author

[2]Associate Lecturer in Statistics
School of Mathematical Sciences
The University of Adelaide
Adelaide SA 5005
Australia
Email: tyman.stanford@adelaide.edu.au
Telephone: +61 8 8313 9248
Fax: +61 8 8313 3696

No conflicts of interest to declare.

Abstract

Components of variance have a long history and find application in all areas of scientific investigation. This paper introduces components of variance and their importance firstly by examples on blood pressure, proteomic data, breath analysers and oropharyngeal pH monitoring devices. We then present an intuitive geometric representation of analysis of variance and explain how the components of variance can be estimated from the analysis of variance table. We conclude by suggesting practice points for studies which incorporate components of variance, and recommend commonly used statistical software to undertake such analysis.

Keywords: components of variance, analysis of variance, multilevel model, interobserver agreement, Bland-Altman plot.

1 Introduction to the key ideas

To a statistician, the term *components of variance* conjures up a linear statistical model in which all the random variables are assumed to be independent and normally distributed. However, components of variance have a long history dating back several centuries and much of the early work, which was in genetics and sampling, did not require such assumptions. In 1918, R.A. Fisher employed components of variance in his work on Mendelian genetics and introduced the term *Variance*, then popularised the *analysis of variance*[1]. L.C. Tippett later clarified the role of analysis of variance for estimating components of variance using linear models. There was also a great deal of activity in the 1930s by other well-known statisticians including H. Daniels on application to the production of wool yarn and F. Yates on optimal sampling for yield in cereal experiments. We recommend S. Stigler's book[2] to readers who are interested in the early history of statistics.

Components of variance arise when the observed variation is attributable to sources with direct physical meaning. Tomasetti and Vogelstein[3] suggest that "only a third of the variation in cancer risk among tissues is attributable to environmental factors or inherited predispositions. The majority is due to 'bad luck', that is, random mutations arising during DNA replication in normal, noncancerous stem cells." Components of variance find widespread application in the biomedical sciences in which many variables are characterised by their variability. A key question is: what are components of variance used for? A common aim is the estimation of a population mean, such as the true level of a leaf parasite infection in a forest. The researcher has to first come to terms with the variability between leaves on trees and the variability between trees in order to determine how many trees and how many leaves within trees to sample. In this situation, the variance components are of secondary interest to the primary purpose of the study, which is the estimation of the true mean infection level. Alternatively, the study of variability may be of primary interest in its own right. For example, in genetics the variance components may describe total phenotypic variation that may be attributable to genetic and environmental sources, or in industrial statistics the variance components may be due to variability between observers or machines. In such cases, it is of specific interest to know how much variability each source contributes to the overall variability, and statistics provides us with the techniques to determine this.

Statistical models can include two main types of effects, *fixed* or *random*. Fixed effects can be viewed as an attempt to estimate mean values from a population while random effects can be viewed as an attempt to estimate the variance of a population. When both types of effects are present, the model is referred to as *mixed*. Technically of course, all statistical models that include a single error term have a random effect but in these simple cases such a model is not referred to as being mixed. The effects being modelled can also be *nested*, that is, specific to a sub-branch of the hierarchy, as in the leaves on trees hierarchy, or *crossed*, that is, effects apply across branches of the hierarchy, as say, observers would be crossed with machines in the industrial example. Statistical interactions may also be present, as either fixed or random effects, in which

different combinations of effects modify the outcome.

The fitting of variance components models to data was until recently hampered by the complexity of the calculations involved, and the limited availability of computing facilities and appropriate statistical programs. Fortunately, the situation has dramatically improved over the past two decades and we discuss recommendations for current software in Section 5. The relative ease of model-fitting and estimation today has led to a resurgence of interest in variance components models and as their application has become more widespread, so has the terminology. The simplest models are purely nested in which the observed variation is the sum or one or more variance components arising from a hierarchical structure; Example 2 below describes for a one-way nested model for systolic blood pressure. Example 3 analyses data on breath analysers and extends the one-way model to include a fixed effect for the breath analysers as well as a random interaction term; such mixed models are also known as a multilevel, hierarchical or random effects models. We present a more complex mixed effects model for proteomic gastric cancer data in Example 4. An analysis of interobserver agreement for oropharyngeal pH monitoring devices is described in Example 5. We have endeavoured to include examples of interest to the present readership, notwithstanding "The limited number of public datasets for histopathology analysis"[4].

Example 1. Consider a group of adult male patients each of whom has a 'true' but unknown value of systolic blood pressure. Let μ_i denote this value for patient i, where $i = 1, \ldots, I$, and assume one measurement is made on each patient. Then for each μ_i, the corresponding observation Y_i may be written as a model:

$$Y_i = \mu_i + A_i,$$

where on the right hand side of the model equation, A_i is a random variable with mean zero and variance σ_A^2. We write $E(A_j) = 0$ and $\text{var}(A_i) = \sigma_A^2$, and σ_A^2 is called the *component of variance within patients*. In the case of random sampling from a single homogeneous population, σ_A^2 is also referred to as the component of variance for sampling error, or measurement error, for blood pressure.

Example 2. Blood pressure data: Now suppose that the I patients in Example 1 are regarded as a random sample representing the population of adult male patients. Then each μ_i itself may be regarded as an observation on a random variable B_i, with population mean μ and variance σ_B^2, where μ is the unknown true mean systolic blood pressure of the population, and σ_B^2 is the *component of variance between patients*. That is, the model for blood pressure observation Y_i can now be written as

$$Y_i = \mu + B_i + A_i,$$

where $E(B_i) = 0$, $\text{var}(B_i) = \sigma_B^2$, and in the simplest case we assume the between-patient random variables B_i and within-patient random variables A_i are uncorrelated, *i.e.*, $\text{cov}(A_i, B_i) = 0$, where cov stands for covariance. This formulation is an example of the *one-way nested* variance component model, where $E(Y_i) = \mu$ and

$$\text{var}(Y_i) = \sigma_B^2 + \sigma_A^2, \quad i = 1, \ldots, I.$$

We cannot yet separately estimate the two variance components. So now suppose we have repeated blood pressure measurements, $j = 1, \ldots, J$, for each patient. Then the jth replicate observation for the ith patient, Y_{ij} is given by

$$Y_{ij} = \mu + B_i + A_{ij}. \tag{1}$$

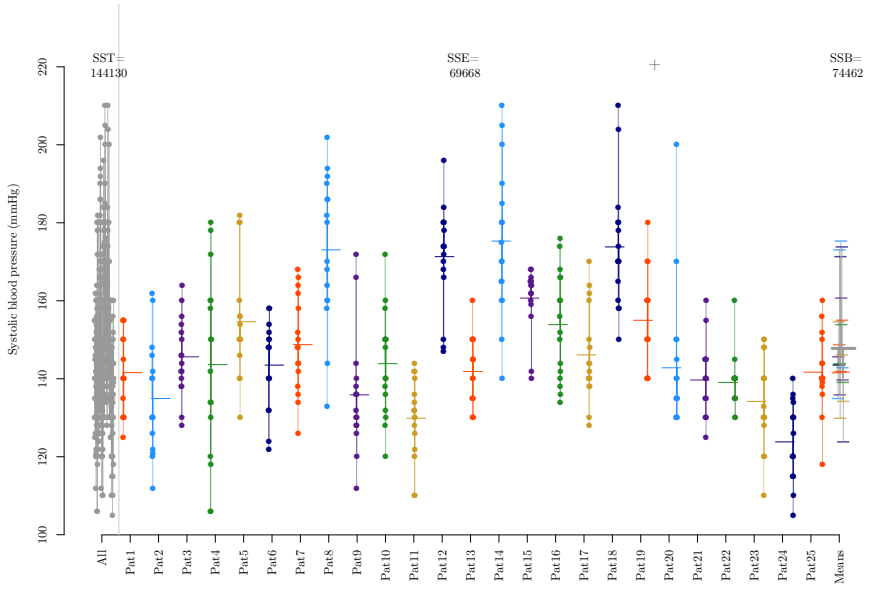
The mean and variance of Y_{ij} are as before, but the responses within patient i are correlated, that is, we can show that $\text{cov}(Y_{ij}, Y_{ij'}) = \sigma_B^2$ for repeated observations j and j'. This then leads to the definition of the *intraclass correlation coefficient* (ICC), which is the correlation between repeated measurements within a patient:

$$\rho_I = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_A^2}. \tag{2}$$

When ρ_I is small (that is, close to zero) the repeated measurements within a patient are only weakly correlated. When ρ_I is close to one, repeated measurements are highly correlated and this feature of the data can be exploited to improve the efficiency of statistical estimators, and to reduce the sample size needed for power calculations at the design stage of a study. In multilevel model terminology, the patient random variables B_i are at level two of the model, and the random variables representing repeated measurements within patients, A_{ij}, are at level one[5].

Cox and Solomon[6] analysed quarterly systolic blood pressure measurements observed on 25 men prescribed oxprenolol in the International Prospective Primary Prevention Study in Hypertension (IPPPSH). The study was one of several large-scale multi-centre randomised clinical trials conducted in the 1980s with the aim of establishing a gold standard target for diastolic blood pressure in the primary prevention of heart disease and stroke. In this trial, patients were examined at three-monthly visits and both systolic and diastolic blood pressure measurements were taken in duplicate, five minutes apart. Figure 1 displays the second recorded (resting) systolic blood pressure measurement for $I = 25$ randomly selected adult males at their $J = 16$ visits over a four-year period. The individual measurements are observed to range from 105 to 240 mmHg and there is considerable variability observed both between patients and between repeated measurements within patients. We ignore here possible time trends or serial correlation in the repeated measurements over time. The overall estimated mean systolic blood pressure is 147.78 mmHg, the estimated component of variance between patients is 182.30 mmHg^2, and the measurement error, that is, the variability between visits within patients, is 185.78 mmHg^2. Thus the repeated measures component of variance is high, and the estimated ratio σ_B^2/σ_A^2 is 0.98. The estimated intraclass correlation coefficient is 0.50, which is modest. The measurement error component is reduced by transforming the raw measurements to the log scale; see[6], but the data nevertheless demonstrate high variability in blood pressure measurements both within and between patients.

- Figure 1 here -



In Section 2, we explain how to obtain the variance component estimates from the analysis of variance table. Throughout the discussion, we assume the data are *balanced* in the sense that each patient has the same number of repeated measurements.

Example 3. Breath analysers: In a recent study, Gullberg[7] compared the performance of six breath analyser machines. 10 replicate blood alcohol measurements were made on three independent adult male subjects on each machine. The comparative performance of the six breath analysers is of primary interest in the study, in which the subjects represent a population of adult males who consume alcohol and the replicate measurements within subjects/machines represent measurement error. Thus, an appropriate model is one which treats the breath analyser machines as fixed effects (α_i) to be estimated and the subject effects (C_j) and measurement error (A_{ijk}) as additive random effects with variance components. We also include a random effect (B_{ij}) for possible interaction between subjects and machines. This leads to the blood alcohol observation on the ith analyser, jth subject and kth replicate, Y_{ijk}, modelled by

$$Y_{ijk} = \mu + \alpha_i + C_j + B_{ij} + A_{ijk}, \tag{3}$$

where μ is the overall mean, the α_i, $i = 1, \ldots, 6$, are the breath analyser fixed effects, the random variables C_j, B_{ij} and A_{ijk} are uncorrelated with zero means, and $\text{var}(Y_{ijk}) = \sigma_C^2 + \sigma_B^2 + \sigma_A^2$.

The results of an analysis of simulated data which approximate Gullberg's results, with estimates of the fixed effects α_i omitted, are shown in Table 1. It should be noted there were significant differences between some of the breath analyser fixed effects. The results showed that 95.7% of the variability is explained by differences between subjects, 1.3% by the interaction between breath analyser machines and subjects and 2.9% is residual variation. It is reassuring that for a given subject, there is relatively little variation between the breath analysers.

- Table 1 here -

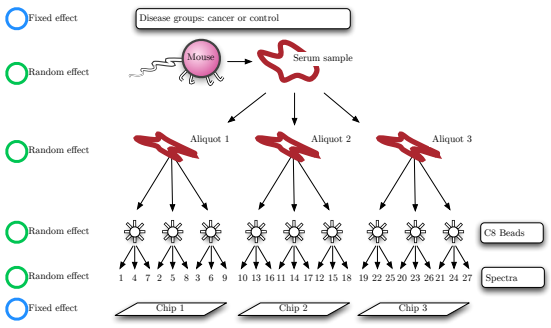
Parameter	Estimate	% Variance
σ^2_C	0.0001820	95.7
σ^2_B	0.0000025	1.3
σ^2_A	0.0000055	2.9
var(Y_{ijk})	0.0001901	100.0

Example 4. Gastric cancer protein expression data:

In a recent collaboration with Megan Penno and colleagues at the Adelaide Proteomics Centre[8] we analysed protein expression levels obtained from a murine gastric cancer experiment. The experiment was conducted using matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry and a primary aim of the study was to detect biomarkers for gastric cancer in serum. Forty mice belonging to five different genotype groups were reared until 12 to 14 weeks of age, at which time blood samples were extracted. The five experimental groups can be grouped into two groups of primary interest: a gastric cancer group and a control group. The gastric cancer genotype group is facilitated using an established murine model where a mutation in glycoprotein 130 protein coding gene in the mice leads to malignant tumours.

Each blood sample was centrifuged to isolate the serum which was then split into three aliquots. Each aliquot was further split into three batches and fractionated with C8 beads which produced nine batches from all aliquots. Finally, each batch was pipetted onto three separate locations on one of the three MALDI-TOF chips used in the experiment, producing 27 samples per mouse. Notably, all samples that originated from a particular aliquot were eventually pipetted onto the same chip, leading to confounding of the chip and aliquot effects. Each chip contained nine samples from the 40 mice, or 360 samples in total. Figure 2 outlines of the process which produced the 27 replicate samples from one mouse. The raw spectra were pre-processed to adjust for experimental and other known systematic effects before downstream statistical analysis.

- Figure 2 here -



A linear mixed model for the pre-processed peak expression data which takes account of the experimental structure is one with four nested levels of variability, for which a single observation $Y_{ijk\ell}$ may written as

$$Y_{ijk\ell} = \mu + \beta_i + \alpha_{2j} + \alpha_{3j} + D_i + C_{ij} + B_{ijk} + A_{ijk\ell}. \tag{4}$$

The fixed effects and random effects in the model are described in Table 2.

- Table 2 here -

				Peak 6821 Da	
Type	Effect	Description	Parameter	Estimate	% Variation
Fixed	μ	Mean expression	μ	9.14	
	β_i	Gastric cancer effect	β_i	0.94	
	α_{2j}	Chip 2 effect	α_{2j}	-0.56	
	α_{3j}	Chip 3 effect	α_{3j}	-0.54	
Random	D_i	Mouse effect	σ^2_D	0.32	55
	C_{ij}	Aliquot effect	σ^2_C	0.07	12
	B_{ijk}	C8 bead effect	σ^2_B	0.09	16
	$A_{ijk\ell}$	Replicate effect	σ^2_A	0.10	17

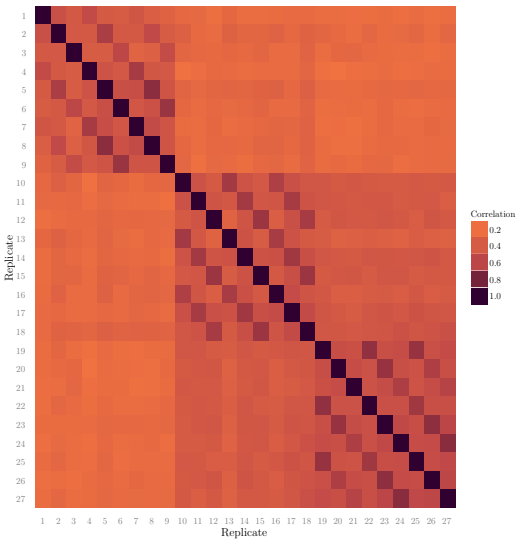
The model was fitted separately to 159 protein peaks with the primary aim of detecting biomarker proteins for gastric cancer. Table 2 gives the results from fitting model (4) to a single peak with low missingness which is a potential biomarker for gastric cancer, specifically, a peptide estimated to weigh 6821 Daltons (Da).

It should be noted that the fixed effects are estimated as differences relative to other fixed effects: α_{2j} and α_{3j} are the effects of chips 2 and 3, respectively, compared to chip 1. Similarly, β_j is the expression effect of gastric cancer mice relative to control mice and is the effect of primary interest. Table 2 shows that for the 6821 Da peptide, the effect of gastric cancer relative to control is an increase in expression of 0.94 units. Chips 2 and 3 are estimated to reduce the expression by 0.56 and 0.54 units, respectively, relative to chip 1. The estimate for μ is interpreted as the mean expression for a control mouse on chip 1.

The estimated variance components reveal some interesting insights into the hierarchal structure of the experimental design. The largest source of variability is the biological variation between mice (55%). The next largest contribution to the total is the residual variance (17%) which is likely to be the result of an under-specified model due to unknown covariates. The smallest contribution to the variance was the third level component of aliquot (12%), suggesting the partitioned serum samples are largely homogeneous. The information about the variance components can be utilised to optimise the experimental design for future studies[6].

Figure 3 is a heat-map of an empirical correlation matrix which demonstrates the different levels of correlation between the spectra derived from each mouse; the correlations are given in descending strength by colour change from dark to light. The heat-map was constructed by calculating all pairwise correlations of peak expressions for the 27 replicates per mouse, then taking the average correlation matrix over the 40 mice. We observe that spectra from the same C8 bead fractionation share more similarity than those from another C8 bead fractionation or spectra derived from a different aliquot/MALDI chip. The observed correlation matrix was very similar to the theoretical matrix hypothesised from the model.

- Figure 3 here -



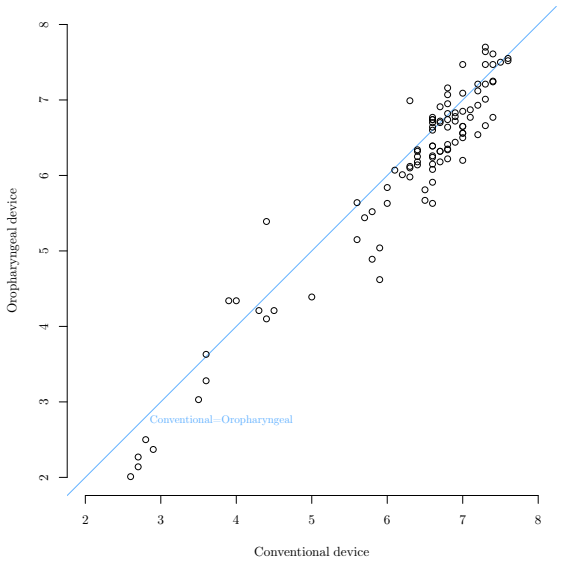
Intraclass correlation coefficients take on a more generalised meaning as 'class' can be interpreted at the different levels of the hierarchical experimental design. For example, the correlation between C8 bead peak expressions within aliquot/mice is

$$\frac{\sigma_D^2 + \sigma_C^2 + \sigma_B^2}{\sigma_D^2 + \sigma_C^2 + \sigma_B^2 + \sigma_A^2},$$

which is estimated to be 0.83 for the single peak related to the 6821 Da peptide, consistent with Figure 3.

Example 5. Comparison of oropharyngeal pH monitoring devices: In 2013, Yuksel *et al*[9] performed in vitro studies to systematically study the performance characteristics of a new oropharyngeal pH monitoring probe with a standard device in patients with chronic laryngitis. 136, 127 observations were made of eosophageal pH using the (new) oropharyngeal device and the conventional device. A random sample of 100 pairs of observations are plotted in Figure 4, which shows a strong positive linear relationship with an estimated correlation coefficient of 0.96. The blue line in Figure 4 is the 'zero line', which represents complete agreement between the two devices. It is apparent that the conventional device produces consistently higher pH readings than the new device. The observed bias can be quantified using the Bland-Atman plot[10], which plots the average pH values versus the differences in pH. We revisit this example in Section 3 on interobserver agreement.

- Figure 4 here -



2 Analysis of variance

The analysis of variance (ANOVA) was introduced by R.A. Fisher as a statistical hypothesis testing approach for comparing two or more population means[1], generalising the two-sample t-test for comparing two population means. ANOVA is based on partitioning the observed variability in the data into parts using *sums of squares* (SS) of the data. A key assumption is that if there are no differences between a set of I population (or group) means, then the corresponding sum of squares between groups behaves like the random error sum of squares. In particular, the ratio of the between groups SS divided by its degrees of freedom to the error SS divided by its degrees of freedom should then be close to one.

The sums of squares result from a linear decomposition of the ith observation in the jth group, Y_{ij}, for which we can write

$$Y_{ij} = Y_{..} + (Y_{i.} - Y_{..}) + (Y_{ij} - Y_{i.}), \tag{5}$$

where the dots are equal to averages over the observed array. For illustration, consider a vector of observations $(Y_{11}, Y_{12}, Y_{13}, Y_{21}, Y_{22}, Y_{23})$. The vector can be split into three uncorrelated parts:

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{pmatrix} = \begin{pmatrix} Y_{..} \\ Y_{..} \\ Y_{..} \\ Y_{..} \\ Y_{..} \\ Y_{..} \end{pmatrix} + \begin{pmatrix} Y_{1.} - Y_{..} \\ Y_{1.} - Y_{..} \\ Y_{1.} - Y_{..} \\ Y_{2.} - Y_{..} \\ Y_{2.} - Y_{..} \\ Y_{2.} - Y_{..} \end{pmatrix} + \begin{pmatrix} Y_{11} - Y_{1.} \\ Y_{12} - Y_{1.} \\ Y_{13} - Y_{1.} \\ Y_{21} - Y_{2.} \\ Y_{22} - Y_{2.} \\ Y_{23} - Y_{2.} \end{pmatrix}$$

The degrees of freedom associated with the three vectors may be explained geometrically as follows. For $2\times3 = 6$ observations, the number of dimensions in which the observations can vary is equal to 6; here $I = 2$ and $J = 3$ hence $IJ = 6$. There is one degree of freedom for the mean $Y_{..}$, because if any one value in the mean vector is known, they are all known. There are $I = 2$ values in the vector of deviations about the overall sample mean, but by knowing one value, the second can be determined. For each I, knowing $3 - 1 = 2$ values determines the last value, leading to $2(3 - 1) = 4$ degrees of freedom for the within-group component. In general, there are $I(J - 1)$ degrees of freedom.

Returning to equation (5) and subtracting $Y_{..}$ from both sides yields,

$$(Y_{ij} - Y_{..}) = (Y_{i.} - Y_{..}) + (Y_{ij} - Y_{i.}). \tag{6}$$

From equation (6) it can be seen that the difference between an observation and the overall mean $(Y_{ij} - Y_{..})$ can be separated into two components. Namely, the difference between the group mean from the overall mean $(Y_{i.} - Y_{..})$ and the difference between the observation from the group mean $(Y_{ij} - Y_{i.})$. Squaring both sides of equation (6) and summing over all groups and observations within groups for the example with $I = 2$ and $J = 3$ leads to the *total adjusted sum of squares* (SST) for this vector of six observations:

$$\sum_{i=1}^{2}\sum_{j=1}^{3}(Y_{ij} - Y_{..})^2,$$

which in turn equals the sum of two terms on the right hand side of equation (6) squared: the first is the between-group sum of squares (SSB) and the second is the within-group sum of squares (SSE):

$$\sum_{i=1}^{2}\sum_{j=1}^{3}(Y_{i.} - Y_{..})^2 + \sum_{i=1}^{2}\sum_{j=1}^{3}(Y_{ij} - Y_{i.})^2.$$

Note that the cross-terms vanish when squaring and summing over i and j on the right hand side of equation (6). The resulting sum of squares identity is simply an application of Pythagoras's Theorem and this is illustrated in Figure 5.

– Figure 5 goes here –

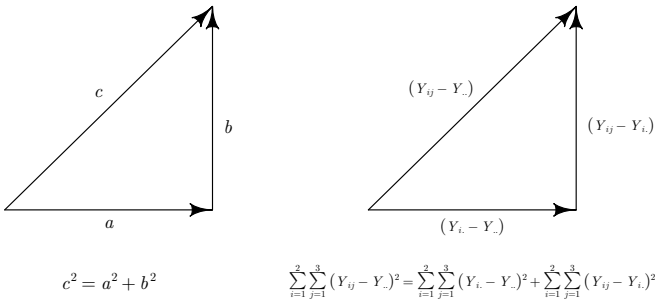
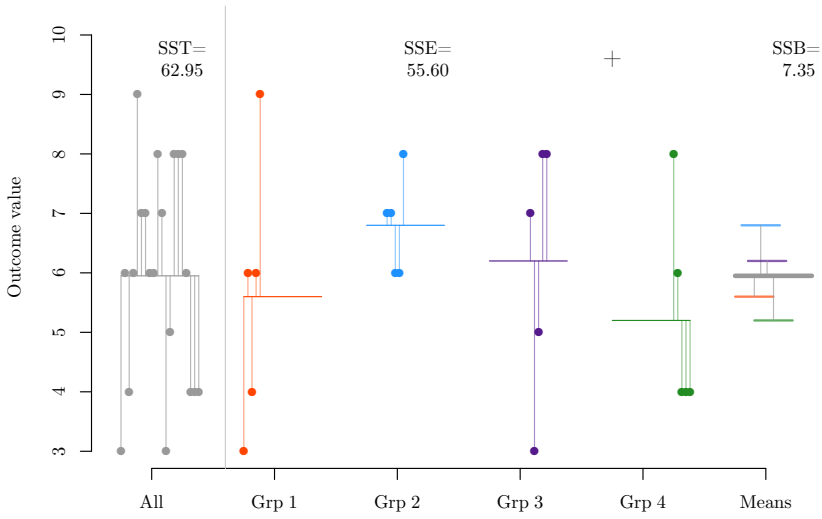


Figure 6 shows how each of the squared differences $(Y_{ij} - Y_{..})^2$, $(Y_{ij} - Y_{i.})^2$, and $(Y_{i.} - Y_{..})^2$, can be visualised on a hypothetical set of data with $I = 4$ and $J = 5$. The data were randomly generated observations from a normal distribution with a standard deviation of two rounded to the nearest integer. Each of the $I = 4$ groups shared a common mean of six so it is expected that the SSB should be small relative to the SST. This is an illustrative scenario similar to the decomposition shown in Figure 1 of Example 2.

- Figure 6 goes here -



In general, for I groups and J repeated measurements within groups, the decomposition of the total sum of squares is given in the analysis of variance table, Table 3.

The mean square (MS) in each case is the sum of squares divided by the appropriate number of degrees of freedom. For example, the mean squares between groups is MSB = SSB/$(I-1)$.

– TABLE 3 goes here –

Estimating the variance components: For the one-way blood pressure model in equation (1), we can show that the expected mean squares between groups is $E(\text{MSB}) = J\sigma_B^2 + \sigma_A^2$ and the error expected mean squares is $E(\text{MSE}) = \sigma_A^2$. This then allows us to estimate σ_B^2 by

$$\frac{\text{MSB} - \text{MSE}}{J},$$

and σ_A^2 by MSE. These are known as the *least squares-based estimates* of the variance components and are unbiased. These are the formulae we used to calculate the variance component estimates for systolic blood pressure given in Example 2. If we additionally assume that all the random variables are normally distributed, formal statistical inference and importantly, significance tests are possible. Specifically, under the normality assumption, the sums of squares are independently distributed proportionally to chi-squared random variables with the appropriate degrees of freedom. Hence, normality gives rise to the familiar F-statistic, MSB/MSE, for testing the null hypothesis of no differences between the group means.

So far, we have assumed that the data are *balanced*, in the sense that there is an equal number of repeated measurements in each group. ANOVA works best when the data are balanced. When the data are *unbalanced*, the statistical principles are the same but the estimation is more complicated. There are numerous computational software solutions available for both ANOVA and the estimation of variance components from balanced and unbalanced data, and we discuss these in Section 5 on software and computational issues.

3 Interobserver agreement

Often the analysis of interest centres on the agreement between two (or more) observers. If the outcome is continuous and each observer measures the outcome, Pearson's correlation coefficient, r, provides an empirical measure of agreement. Further, this is related to the coefficient of determination, r^2, which provides an estimate of the proportional variance shared by two observers. The closer the coefficient of determination is to one, the closer the agreement between the two observers.

To see how correlation and variance are related, consider the following equation for the true correlation, ρ, between two random variables X and Y representing two observers' measurements:

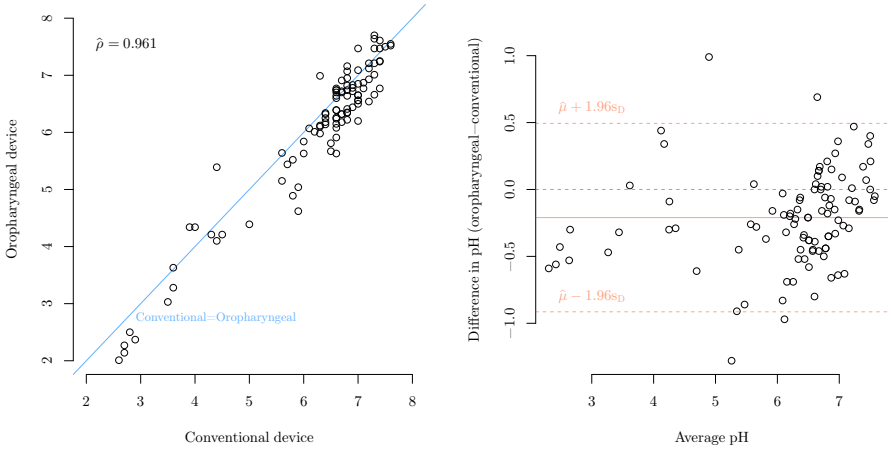
$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.$$

This is the formula which leads to the intraclass correlation coefficient ρ_I given in equation (2) for the one-way model. For intraclass correlation, the random variables X

and Y represent repeated measurements on the same individual, that is, $(X = Y_{ij}, Y = Y_{ij'})$, so that $\text{cov}(X, Y) = \sigma_B^2$, and $\text{var}(X) = \text{var}(Y) = \sigma_B^2 + \sigma_A^2$.

Bland and Altman[10] point out that correlation is not sufficient for measuring inter-observer agreement because, amongst other issues, the correlation coefficient does not provide information about bias. That is, whether one observer consistently reports larger or smaller values than the other observer. Figure 7 shows the data on eosophageal pH from Example 5 again in a scatter plot (left tableau) and a Bland-Altman plot on the right. The Bland-Altman plot gives the average observer values on the x-axis and the difference between the observer values on the y-axis. The plot essentially removes the $y = x$ trend seen in the scatter plot (left) to enable explicit examination of the observer values. The bias can be seen clearly in the right plot with the solid orange line depicting the mean difference in the observer (device) pH measurements. Once again variance is an important element of the agreement analysis. The dotted orange lines give an approximate 95% prediction interval for the differences between the two observers, known as the *limits of agreement*. These are statistical bounds. In practice, such predicted limits of maximum differences should be specified prior to the analysis, and based on clinical knowledge and the variability that would be tolerated between observers.

- Figure 7 here -



Example 5 represents a simple situation in which there are two observers and a continuous outcome variable. Often there are more than two observers, such as in the comparison of breath analysers in Example 3, or the outcomes are not continuous variables, such as diagnostic categories, or the outcomes are not paired. In cases where the interobserver agreement is measured for categorical outcomes, Cohen's Kappa is commonly used.

4 Practice points

Before conducting an experiment or collecting data, the following questions should be answered, ideally in collaboration with a statistician:

- Establish the scientific question to be answered from your data.
- What is the outcome variable of interest? Is it continuous, discrete or ordinal?
- What variables are associated with this outcome? Are they continuous, discrete or ordinal?
- What are the likely systematic influences on the outcome? These determine the fixed effects in the model.
- What are the potential sources of variability? These determine the components of variance in the model.
- Are there repeated measures in the data?
  - If so, are temporal or spatial effects likely in the data?
- Will randomization or random sampling be appropriately employed in the experiment or study?
  - If your data are observational, then batch effects arising from the study design or sampling process can affect the statistical efficiency of clinical effects of interest.

5 Statistical software and computational issues

In most cases, using statistical software to calculate estimates of variance components is recommend since hand calculations can be tedious and prone to errors. Fortunately, there are numerous software packages available that can calculate variance components when provided with data. The following recommendations are not exhaustive and focus on procedures to calculate ANOVA tables and fit linear mixed effects models.

The free software package `R` provides a function called `aov()` to calculate ANOVA tables. When there are two or or more random effects in a model, `lme()` in the `nlme` package is recommended. The function `lmer()` in the `lme4` package can also fit linear mixed models. `SAS` is another popular software package and also has a range of procedures available to estimate variance components. The aptly named `PROC ANOVA` allows users to produce ANOVA tables from data, while `PROC VARCOMP` and `PROC MIXED` will fit linear models with random effects.

Once data are loaded into the statistical software `SPSS`, analysis routines can be selected under 'Analyze' on the toolbar. A simple ANOVA can be selected under `Compare Means > One-Way ANOVA`. `General Linear Model > Univariate` and `General Linear Model > Variance Components` allow fixed and random effects to be specified in the linear model. These routines can also be run using 'Syntax' commands in lieu of the menu driven options mentioned above. The software package `Stata` is also widely used in the health sciences. ANOVA tables can be produced using the `anova` command or by selecting `Statistics > Linear models and related > ANOVA/MANOVA >`

`Analysis of variance and covariance`. Linear models with mixed effects can be fitted using the `mixed` command or by selecting `Statistics > Multilevel mixed-effects models > Linear regression`.

References

[1] Fisher R A. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, 1925.

[2] Stigler S M. The history of statistics: the measurement of uncertainty before 1900. Cambridge, Mass: Belknap Press of Harvard University Press, 1986.

[3] Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science, 2015; 347 (6217): 78-81.

[4] Phouladya H A, Goldgofa D M, Halla L O, and Moutonb P R. Nucleus segmentation in histology images with hierarchical multilevel thresholding. Proc. of SPIE Vol. 9791, 2016. doi: 10.1117/12.2216632

[5] Snidjers T A B and Bosker R J. Multilevel Analysis, Second Edition. London, Sage Publishers, 2012.

[6] Cox D R and Solomon P J. Components of variance. Boca-Raton, Chapman and Hall/CRC, 2003.

[7] Gullberg R G. Employing Components-of-variance to evaluate forensic breath test instruments. Science and Justice 2008; 48: 2-7.

[8] Penno M A S, Klingler-Hoffmann M, Brazzatti J A et al. 2D-DIGE analysis of sera from transgenic mouse models reveals novel candidate protein biomarkers for human gastric cancer. Journal of Proteomics 2012; 77: 40-58.

[9] Yuksel E S, Slaughter J C, Mukhtar N et al. An oropharyngeal pH monitoring device to evaluate patients with chronic laryngitis. Neurogastroenterology and Motility 2013; 25: e315-e323.

[10] Altman D G, and Bland J M. Measurement in Medicine: The Analysis of Method Comparison Studies. Journal of the Royal Statistical Society Series D (The Statistician) 1983; 32(3): 307-317.

Tables

Table 1. Estimated components of variance and percentage of total variance for the breath analyser study; see main text Example 3.

Parameter	Estimate	% Variance
σ_C^2	0.0001820	95.7
σ_B^2	0.0000025	1.3
σ_A^2	0.0000055	2.9
$\text{var}(Y_{ijk})$	0.0001901	100.0

Table 2. Fixed effects and variance component estimates for the linear model (4) fitted to the proteomic expression data; see main text Example 4. The parameter estimates are for the model fitted to the peptide measured at 6821 Da.

				Peak 6821 Da	
Type	Effect	Description	Parameter	Estimate	% Variation
Fixed	μ	Mean expression	μ	9.14	
	β_i	Gastric cancer effect	β_i	0.94	
	α_{2j}	Chip 2 effect	α_{2j}	-0.56	
	α_{3j}	Chip 3 effect	α_{3j}	-0.54	
Random	D_i	Mouse effect	σ_D^2	0.32	55
	C_{ij}	Aliquot effect	σ_C^2	0.07	12
	B_{ijk}	C8 bead effect	σ_B^2	0.09	16
	$A_{ijk\ell}$	Replicate effect	σ_A^2	0.10	17

Table 3. Analysis of variance (ANOVA) table for the one-way model, equation (1)

Source of variation	Sum of squares	df*	Mean squares
Between groups (SSB)	$\sum_{i=1}^{I} J(Y_{i\cdot} - Y_{\cdot\cdot})^2$	$I-1$	$\text{MSB} = \text{SSB}/(I-1)$
Within groups (SSE)	$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\cdot})^2$	$I(J-1)$	$\text{MSE} = \text{SSE}/(I(J-1))$
Total (SST)	$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{\cdot\cdot})^2$	$IJ-1$	

*Degrees of freedom

List of figures

Figure 1. Quarterly systolic blood pressure measurements over four years for 25 adult males from the International Prospective Primary Prevention Study in Hypertension; see main text Example 2. The sums of squares SST, SSB and SSE are defined in Table 3.

Figure 2. Experimental design for the gastric cancer MALDI-TOF mass spectrometry study; see main text Example 4.

Figure 3. Heat-map of the empirical pairwise correlation structure for the peak expressions between the 27 spectra obtained from each mouse in the gastric cancer experiment. The correlations are shown in descending strength in colour change from dark to light; see main text Example 4.

Figure 4. Eosophageal pH for a conventional versus a new monitoring device for a random sample of 100 observations; see main text Example 5.

Figure 5. Pythagoras's Theorem (left tableau). Geometric representation of the decomposition of an observation vector into its uncorrelated parts, resulting in the sums of squares identity for the analysis of variance (right tableau). The decomposition in the balanced case is Pythagoras's Theorem.

Figure 6. Visualisation of the sums of squares leading to the analysis of variance table for a hypothetical one-way model with $I = 4$ groups and $J = 5$ replicate measurements within each group.

Figure 7. Bland-Altman plot for agreement between a conventional device and a new device for monitoring eosophageal pH; see main text Example 5.