Lecture notes on applied statistics

Peter McCullagh
University of Chicago
January 2017

1. Basic terminology

These notes are concerned as much with the logic of inference as they are with computation or statistical methodology. To a certain extent, the statistical models and the associated computations are included to illustrate the logic of statistical inference. At the same time, we do not seek to evade practical or computational considerations.

Mathematically speaking, a statistical model is a fairly complicated construction. It is necessary to develop an understanding of the various parts and how they fit together. Probability is central. A probability model is [a probabilistic description of] a stochastic process, $(Y_u)_{u\in\mathcal{U}}$, which has an index set $u \in \mathcal{U}$, and a state space $Y_u \in \mathcal{S}$. A probability model F is a family of probability distributions, one distribution F_S for each sample $S \subset \mathcal{U}$. In this setting $Y[S]$ is the list of sample values, and $F_S(A) = \text{pr}(Y[S] \in A)$ for each event $A \subset \mathcal{S}^S$. For subsamples $S' \subset S$, these distributions are subject to the standard Kolmogorov consistency condition $F_{S'}(A) = F_S(A \times \mathcal{S}^{S'-S})$ for a stochastic process, which means that $F_{S'}$ is the marginal distribution of F_S after integrating out the unwanted variables. A statistical model is a more complicated object, consisting of a non-empty *family* of stochastic processes $F_{S,\theta}$ indexed by the parameter $\theta \in \Theta$. Each F_θ is a stochastic process, $F_{S,\theta}$ is the finite-dimensional distribution for the sample response $Y[S]$, and $F_{S',\theta}$ is the marginal distribution of $F_{S,\theta}$.

For a randomized experiment, the situation is more complicated, because it is necessary to specify also the probability of treatment assignment $\mathbf{t}$, and the response distribution given each treatment assignment.

We begin with a list of common statistical terms that are commonly used in statistical discussions of observational studies and of experimental investigations—units, population, variable, treatment, randomization, and so on. Subsequent sections adhere closely to the definitions as given, giving particular emphasis to the distinction between sample and population, observational unit versus experimental unit, covariate versus relationship, covariate versus treatment, and so on. All statistical models seek to exploit the *structure of the units* as defined in various ways, mainly by covariates and relationships. Examples, scenarios and vignettes are used freely to explain how to match the various mathematical concepts with physical objects in specific situations.

The reader should be warned that, although most of the terminology is reasonably standard, the definitions differ slightly or substantially from one author to another. In these notes, for example, the treatment assignment vector does not satisfy the conditions for a covariate; that distinction is maintained throughout, and its importance is underlined in the logical structure of probability models for randomized experiments as opposed to observational studies. Likewise, there may be minor or major differences in the definition of experimental unit, in the definition of the state space, the definition of interference, and so on.

1.1. Baseline. Every experiment and every observational study has a temporal component. The baseline is the temporal origin or reference point marking the beginning of the study. Mathematically speaking, the baseline is a point at which the units $u \in \mathcal{U}$ have been assembled, together with all of the information about them that is needed to specify the probability of arbitrary outcomes. All statistical inferences are based on probabilities, and the probability model is said to be *registered at baseline*.

Generally speaking, the units available for study are not homogeneous. The baseline information records sex, age, and, in principle anything else that is available at baseline that could reasonably be deemed to have a bearing on outcome probabilities. In practice, a certain restraint or professional judgement is needed to decide what is likely to be relevant and what is not. In a field experiment, the geometric layout of the plots is ordinarily part of the registered baseline information, and is almost always relevant in that it affects outcome probabilities. Information about crop, treatment and yield in the previous season is sometimes available and might be judged relevant if the new plots were well-aligned with the previous plots. In a clinical trial with human patients, ethnic background might be relevant as a block factor, but the number of letters in the patient's name is unlikely to be considered relevant for clinical outcomes.

For a randomized study, randomization occurs at or immediately after baseline, so the randomization outcome is not a part of the baseline registered information. Model specification begins with randomization probabilities $p(\mathbf{t}) = \mathrm{pr}(T = \mathbf{t})$ for each treatment assignment vector $\mathbf{t} = (t_i)_{i\in S}$, also called the treatment factor.. Even if one assignment list is a permutation of the other, two treatment assignments $\mathbf{t}, \mathbf{t}'$ may have different probabilities depending on baseline information such as covariate or block structure. Most commonly, the randomization is required to be balanced with each treatment level occurring with equal frequency in each block.

Since the probability model is registered at baseline, i.e., pre-randomization, the model specifies the joint distribution for treatment T and response Y. The joint distribution implies a conditional distribution $F(\cdot \,|\, \mathbf{t})$ of the response vector $(Y_i)_{i\in S}$ for *every* treatment assignment $\mathbf{t}$. Randomization subsequently produces a particular treatment configuration, but the conditional distribution associates a response distribution with every configuration having positive probability whether it occurs as the actual randomization outcome or not. In general, the conditional probability $F(A \,|\, \mathbf{t})$ of the event $Y \in A$ may depend on any and all registered baseline information. Every variable measured post-baseline, such as T, is regarded as the outcome of a random process, and, as such, is formally a part of the response.

Baseline need not mean a point in calendar time. In studies of cell development, the baseline would ordinarily be set at a key developmental stage such as fertilization, which is a point in calendar time that may vary from cell to cell. Similar remarks apply to clinical trials where the baseline is usually set at recruitment, which varies from one patient to another on the calendar scale.

1.2. Observational unit. The *observational units* are the objects $u \in \mathcal{U}$ on which variables are defined and measurements may be made. Usually measurements are made only on a small subset of observational units (the sample), so the phrase *measurements may be made* does not imply that measurements are made. The statistical universe almost always includes infinitely many extra-sample units, notional or otherwise, for which probabilistic prediction may be required. Sometimes each unit is a physical object such as a plot, a patient, a rat, a tree, or a M-F pair of fruit flies. Sometimes the units are less tangible, such as time points or time intervals for an economic series, or spatio-temporal points or intervals for a meteorological variable such as temperature or rainfall. Very often, the set of observational units is a Cartesian product set such as

$$\{\text{mice}\} \times \{\text{front, rear}\} \times \{\text{left, right}\} \times \{\text{day0, day1, day2}\}$$

which contains 12 observational units for each mouse. As an index set, time is structured cyclically in a similar way: $\{$*clock times*$\} \times \{$*?? days*$\}$ or $\{$*365 calendar dates*$\} \times \{$*?? years*$\}$. The index set may be stuctured in other ways such as pupils within classrooms within schools, which is a nested or hierachical structure defined by one or more relationships $R(u, u')$ on the units.

1.3. Population. The *population* $\mathcal{U}$ is the complete set of observational units, which is typically infinite; the *sample* is the finite subset of observational units that occurs in the study. By extension, the sample may also include any units for which response predictions are requested. In a meteorological context, the observational units are all points in the plane or sphere, or points in the spatio-temporal product space, so the population is uncountably infinite. For a spatial process, the units may be either points in the plane, or subsets of the plane, or less tangible objects such as signed measures on the plane or planar contrasts. The sample is the finite set of points at which measurements (sample values) are planned or available.

The mathematical population is the *index set* on which the response (yield, health, weather,...) is defined as a stochastic process. As is often the case in mathematics, the mathematical index set is made sufficiently large that it encompasses every conceivable situation that might arise, and many more besides. For a clinical trial in which the experimental units are human patients, the mathematical index set is not finite, and in fact the mathematical subset of units having a specific sex, age and body-mass-index is also infinite. One might object to the fact that the mathematical index set contains more points than there are real physical units. Such objections are not to be taken seriously; they are on a par with rejecting the real number system for engineering or accounting purposes on the grounds that it contains infinitely many 'useless' values that are not needed for billing purposes.

When one talks of a 'Normal population' or a 'Cauchy sample', the reference is not to the population or sample *per se*, but to the population values or sample values or their distribution, usually with independent values for distinct units.

1.4. Sample. The *sample* $S \subset \mathcal{U}$ is the finite subset of observational units on which the response and other variables are recorded. Technically, S is an arbitrary finite ordered list of units, usually but not necessarily distinct, and the recorded response $Y[S]$ is the list of Y-values for $i \in S$ in the same order.

To be clear, the word 'sample' in these notes denotes a finite ordered subset of units. It does not imply a random sample, or a simple random sample.

In settings where prediction or interpolation is involved, it is necessary to consider an extended sample S', which includes S as a sub-sample. Each unit $u \in S' \setminus S$ is called an extra-sample unit. Only $Y[S]$ is actually observed, and prediction refers to the conditional distribution of $Y[S' \setminus S]$ given $Y[S]$, or, in some settings, to the conditional expected value.

1.5. Variable. A *variable* is a function on the observational units, both sample units and extra-sample units. Quantitative examples include 'weight in kg.', 'atmospheric pressure in cm. Hg.', or 'length in cubits; In principle, the variable name includes the physical units of measurement so that the value $x_i \equiv x(i)$ of the variable x for unit i is a real number, not an expression such as '184.5 cm'. Mathematically speaking, weight in kg. and weight in lbs are different variables; in practice, weight is frequently used in everyday speech without specified units. Qualitative examples include *sex* taking values in {*M, F*}; or *occupation* taking values in a suitable set of occupations, one of which may be *'none of the above'*.

If u, v are two variables, the ordered pair (u, v) is also a variable: the value of (u, v) for unit i is $(u, v)(i) = (u_i, v_i)$, which is a point in the Cartesian product space. Each variable is defined on the population and recorded on the sample.

1.6. Feature. Feature is a synonym for variable or attribute—a function on the units. The feature vector takes values in the feature space.

In certain settings, each feature is a component of the response vector, the primary response is a class or characteristic of the unit, and the goal is to classify each unit by computing the conditional distribution over the set of classes given the features.

1.7. Quantitative variable. A *quantitative* variable is a real-valued function on the observational units. More generally, a quantitative variable is a function taking values in a vector space. Dose (of fertilizer or medication in suitable units) is a typical quantitative variable whose values are non-negative. Blood pressure (systolic, diastolic) in mm. Hg. is a quantitative variable taking values in $\mathcal{R}^2$. This means that every realizable value of blood pressure can be found somewhere in $\mathcal{R}^2$; it does not mean that every point in $\mathcal{R}^2$ is realizable as a blood-pressure value for a live human subject. Negative values are in conflict with hydrostatic and hydraulic theories.

If x, z are two quantitative variables taking values in the same vector space, so also is the linear combination $3x + 4z$. If x, z are real-valued variables, so also are the unit-wise products x^2, z^2 and xz.

1.8. Qualitative variable. A *qualitative* variable, also called a *classification factor*, is a function on the observational units taking values in a finite set, called the *factor levels*. Examples include *sex*, *occupation*, *socioeconomic class*, and variables such as *genetic variant* with values 'wild type' and 'mutant'. Often, one level is designated as a reference level. A qualitative variable is sometimes called an *attribute*.

1.9. Response. The *response*, usually denoted by Y, is the variable of primary interest, e.g., yield in kg. per unit area, or time to failure in a reliability study, or stage of disease, or severity of pain, or death in a 5-year period following surgery. There may be secondary or intermediate response variables such as compliance with protocol in a pharmaceutical trial. Synonyms and euphemisms include *yield*, *outcome* and *end point*.

The response is regarded as the realized value of a random variable, or process $u \mapsto Y_u$ taking values in the *state space* $Y_u \in \mathcal{S}$. For an observational study, the distribution is denoted by F; for a randomized study $F(\cdot, \mathbf{t})$ is the joint distribution of Y, T.

1.10. Covariate. A *covariate* x is a function on the observational units that is used in a probability model to permit the outcome distribution for one unit to differ from that of another unit. Ordinarily, the events $Y_i \in A$ and $Y_j \in A$ are presumed to have the same probability if $x_i = x_j$; otherwise, if $x_i \neq x_j$, the probabilities may be different. For this to make operational sense, the covariate must be registered as a variable at baseline. Typical examples include patient age, sex of mouse, type of soil or soil pH. If the set of observational units is a Cartesian product set or a subset thereof, each marginal component is a covariate.

Operationally, a covariate is used in a randomized experiment to reduce 'unexplained' variation and thereby to increase the precision of treatment effect estimates. In an analysis of variance, the total sum of squares for the response is partitioned into various parts, one part associated with registered covariates and block factors, a second part associated with treatment, the remainder being 'unexplained' or residual variation. The part associated with covariates and block factors, the between-blocks variation, is said to be 'eliminated', and the more variation that can be eliminated, the less there is to contaminate the estimates of treatment contrasts. A covariate or block factor is said to be effective for this purpose if the associated mean square is substantially larger than the mean squared residual. This means that the response variation within blocks, the intra-block mean square, should be appreciably smaller than the response variation between blocks, the inter-block mean square.

In practice, it may be acceptable to fudge matters by using as a covariate, a variable measured post-baseline before the effect of treatment has had time to develop, or an external variable whose temporal evolution is known to be independent of treatment assignment for the system under study. At a minimum, it is necessary first to check that the variable in question is indeed unrelated to treatment assignment; otherwise its use as a covariate could be counterproductive. It is well to remember that while measurement pre-baseline is strong positive evidence that no statistical dependence on treatment assignment exists, the most that can be expected of a post-baseline measurement is absence of evidence. For a variable of dubious status, absence of evidence is considerably better than its complement, but it does not provide the same positive assurance as evidence of absence. A concomitant variable of this sort is not counted as a covariate in these notes. It is formally regarded as a component of the response whose dependence on treatment assignment is to be specified as a part of the statistical model. The dependence may be null, but that alone does not give it the status of a covariate. For one example of such a variable, see section ??.

As always, a probability model F allows us to compute whatever conditional distribution is needed for inferential purposes. That includes the conditional distribution given any concomitant or intermediate outcome or the conditional distribution of health values given that the patient is alive, or the conditional distribution of the cholesterol level given that the patient has complied with the protocol, or even the probability of compliance given the cholesterol level. Whether these are the relevant distributions for the purpose at hand is an entirely different matter to be determined by the user.

1.11. Treatment. *Treament* is a function $T\colon \textit{sample units} \to \textit{levels}$ taking values in the set of treatment levels. Treatment is not a covariate because it is not a property of the observational units that is registered at baseline; it is an *intervention* that changes the status quo for the sampled units only. Usually, treatment is a random variable whose value is the outcome of a *randomization scheme*. The components of T for distinct observational units, or even for distinct experimental units, are usually identically distributed, but seldom independent.

In computational work, the observed treatment configuration $(T_u)_{u\in S}$ is called the treatment factor. Although T is defined only for sample units, we must bear in mind that the sample can always be extended indefinitely, at least in principle, so the restriction to S is not a major part of the distinction between a classification factor and a treatment factor. The important distinction is that a pre-baseline variable is a property of the units, whereas treatment level is assigned to units at baseline.

1.12. Randomization. The *randomization scheme* is a probabilistic protocol for the assignment of treatment levels to sample units, often uniformly at random subject to design constraints. For a completely randomized design with 12 sample units and four treatment levels, a balanced randomization scheme is a function $T\colon [12] \to [4]$ (from sample units to treatment levels) chosen [uniformly] at random from the set of $12!/(3!^4) = 369600$ functions having treatment blocks $T^{-1}(1), \ldots, T^{-1}(4)$ of equal size. In the randomized blocks setting, each sample unit is an experimental unit.

Usually, the randomization probabilities depend on the block structure and covariate configuration occurring in the sample units. For a typical randomized blocks design, the joint probability that the pair (u, u') is assigned treatment levels (t, t') depends on whether the units belong to the same block or to different blocks. More generally, the probability $\mathrm{pr}(u \mapsto t; S)$ that treatment level t is assigned to unit u may depend not only on x_u but also on $x_{u'}$ for all other units $u' \in S$. Unless otherwise specified, we assume in these notes that the assignment probabilities $\mathrm{pr}(u \mapsto t; S) > 0$ are strictly positive for every unit and every treatment level. For an exception in which the menu of treatment options may be covariate-restricted, see Example 2.6.

1.13. Experimental unit. The *experimental units* are the objects to which treatment is assigned, i.e., two distinct experimental units may be assigned different treatment levels. Or, to say the same thing in a different way, two distinct experimental units are assigned different treatment levels with strictly positive probability. Each experimental unit consists of one or more observational units, e.g., one mouse consisting of four legs, or one classroom consisting of 20–40 students in the preceding example.

Two observational units u, u' belong to the same experimental unit if the randomization scheme necessarily assigns them to the same treatment level. In mathematical terms, $E(u, u') = 1$ if and only if $T(u) = T(u')$ with probability one. By construction, E is an equivalence relation, which partitions the sample units into disjoint blocks. Each block of E is one experimental unit.

1.14. A/B testing. This phrase, which originates in commercial internet activity, refers to a treatment having two levels A, B, which may be connected with options for on-screen presentation of internet search results. Each search is an observational unit, the response being click/no click. The experimental units may be searches or users or IP addresses, depending on the circumstances.

1.15. Relationship. A *relationship* is a function on *pairs of units* that may be used in the statistical model to distinguish the joint outcome distribution for one pair of units versus another pair. For this to be feasible, the values must be registered at baseline. If the units are points in a metric space, the metric $d(u, u')$ is a non-negative symmetric relationship among them. Experimental unit is a Boolean relationship on S: $E(u, u') = 1$ if u, u' belong to the same experimental unit. Other examples include genetic, familial, neighbour, and adjacency relationships. Ordinarily, the relationship is defined on the population and recorded for the sample.

1.16. Block factor. A *block factor* is a Boolean function on pairs of observational units that is reflexive, symmetric and transitive—an equivalence relation registered at baseline. Each block factor (such as the experimental unit factor) partitions the set of observational units into disjoint non-empty subsets called blocks. The identity function on $\mathcal{U}$ is a block factor whose blocks are all singletons; at the other extreme, the function J such that $J_{u,u'} = 1$ for every pair, has exactly one block.

To each variable or factor x there corresponds a block factor B defined by

$$B_{ij} = 1 \text{ if and only if } x(i) = x(j).$$

Regardless of how the information is stored in an electronic device, the chief mathematical difference between B and x is that the x-blocks are labelled by x-levels, whereas the blocks of B are unlabelled. The x-block $x^{-1}(x(1)) = \{j \in S\colon x(j) = x(1)\}$, i.e., the subset of sample units having the same x-value as unit 1, has the label $x(1)$. Since the blocks of B are unlabelled, a block factor has no reference level or reference block.

At the risk of over-simplification, covariates typically occur in the model for the mean response; block factors and other relationships occur in the model for covariances.

1.17. Statistical effects. In standard probability language, the phrase 'X_i is independent of Y_i' is not so much a statement about the random variables as measurable functions or the pair of outcomes (X_i, Y_i) as numerical values for a particular unit, as it is a statement about probabilities. The joint probability for each product event $(X_i, Y_i) \in A \times B$ is multiplicative. Likewise, when we talk of a statistical effect in a context such as 'the effect of treatment on longevity' or 'the effect of variety on yield', the effect referred to is not a numerical difference of two survival times or two yields, but a difference of

two probabilities or a difference of two probability distributions. For example, if the probability model asserts that the yield in kg/Ha on plot i is distributed as $N(\mu, \sigma^2)$ for variety I and $N(\mu, 2\sigma^2)$ for variety II, the effect of variety (II versus I) is to double the yield variance. The effect of variety on [the probability of] a particular event $Y_i \in A$ is $N(A; \mu, 2\sigma^2) - N(A; \mu, \sigma^2)$, which depends on (μ, σ^2). Similar remarks apply to the effect on linear and non-linear functionals such as means, medians or quartiles of the yield distribution.

Apart from treatment effects, there are other effects of a different nature, such as the effect of aging on mobility or cognitive function. Every treatment effect in these notes is modelled as a group action on probability distributions, which is not necessarily the case for covariate effects.

1.18. Design. The word *design* refers to the arrangement of the sample units by blocks, by covariates, and by restrictions on treatment assignment. Very often, it is helpful to distinguish between two aspects of the design, the *structure of the units*, meaning relationships among them, and the *treatment structure*, which is imposed on them. In a crossover design, where the same physical object occurs as a distinct experimental unit on several successive occasions, the structure of the units includes not only the temporal sequence, but also a block factor whose blocks are the distinct physical objects. In a field experiment, the structure of the units includes the geometric shape of each plot, their physical arrangement in space, and the width of access paths or guard strips separating neighbouring plots.

1.19. Replication. Replication means repeating the experiment independently for different experimental units under essentially identical circumstances in order to gauge the response variation.

1.20. Independence. In the simplest class of statistical models, the responses on distinct *experimental units* are assumed to be distributed independently given the treatment assignment vector, i.e., $Y_u \perp\!\!\!\perp Y_{u'} \,|\, \mathbf{t}$. In other situations such as agricultural field experiments or crossover designs or studies involving infectious diseases, the responses on distinct experimental units cannot reasonably be assumed to be independent given $\mathbf{t}$. For example, there may be geographic or temporal or familial correlations, which may be detectable in the data.

1.21. Interference. If the response Y_u for one experimental unit is statistically independent of the treatment applied to other units, we say there is no interference, or no pairwise interference. Lack of interference is a conditional independence assumption $Y_u \perp\!\!\!\perp \mathbf{t} \,|\, t_u$; it does not imply independence, nor does independence imply lack of interference. Independence and lack of interference are not so much statements of fact or fiction as they are mathematical restrictions on probabilities.

1.22. State space. In a statistical model, the response is regarded as a random variable, a function $u \mapsto Y(u)$ on the observational or experimental units taking values in the *state space* $\mathcal{S}$, (often the real numbers). In certain settings, particularly in observational studies where all variables are regarded as responses on an equal footing, the synonym *feature space* may be used. Usually the feature space is $\mathcal{R}^k$ for some fixed k.

It is important that the state space contain a point for every possible response-related post-baseline event that could possibly be recorded. In a pharmaceutical trial for cholesterol reduction, individual patients give informed consent and agree to abide by the protocol. However, subsequent participation is ultimately voluntary, and not all patients comply by taking their medications on the prescribed schedule. If it is recorded, compliance or the degree of compliance is a response variable, and failure to comply is

one component of the response. The probability model is a probability distribution on the state space, which specifies the compliance probability, the conditional distribution given compliance, and the probability of compliance given the cholesterol levels past and future.

In all cases, the state space is a fixed measurable set, the same set for every unit, either observational unit or experimental unit, regardless of covariates. However, this restriction may lead to mathematical contortions. Consider an animal breeding study where each experimental unit is a family, and the response is measured on individual family members (offspring only) at age six weeks. Suppose that family size x is a covariate recorded at baseline, in which case the response Y_u for a family of size $x(u)$ is a point in $\mathcal{R}^{x(u)}$. The variation of the state space from one experimental unit to another depending on the covariate $x(u)$ appears to violate the definition of state space as a fixed set. But this violation is a mathematical illusion. We can simply re-define the state space to be the disjoint union $\mathcal{S} = \cup_{k\geq 0}\mathcal{R}^k$, and construct the probability distribution on $\mathcal{S}$ in such a way that all of the probability mass for unit u resides in the component $x(u)$ of the state space,

$$\text{pr}(Y_u \in \mathcal{R}^k) = \begin{cases} 1 & x(u) = k \\ 0 & \text{otherwise.} \end{cases}$$

Note that x is not a random variable, so we have not written this as a conditional probability statement.

If the measurements were weights at birth rather than later at six weeks, the baseline would necessarily have to be pre-natal, implying that family size X is a part of the response, not a covariate recorded at baseline. In that setting the response Y is a random variable taking values in $\mathcal{S}$, and the response distribution F determines the distribution of X by $\text{pr}(X = k) = F(\mathcal{R}^k)$ (including $k = 0$). The conditional distribution given X is a function that associates with each integer $k \geq 0$ a probability distribution $F(\cdot \,|\, X = k)$ such that $F(\mathcal{R}^k \,|\, X = k) = 1$.

1.23. Censoring and state-space evolution. In a study of survival times following surgery, each patient is one unit, and the response $Y_u > 0$ is, *prima facie* at least, a point in $\mathcal{R}^+$, the positive real line. Only the most persnickety mathematician would bother to add a point at infinity to cover the remote possibility of immortality, which cannot be ruled out solely on mathematical grounds. However, the response $Y_u^{(s)}$ as it exists today or at the time of analysis, say $s = 1273$ days post-recruitment, is either a failure time in the interval $s^- = (0, s]$, or a not-yet-failure corresponding to the 'point' s^+, which is required to exist as a point in the state space for today. In other words, $\mathcal{S}^{(s)} = s^- \cup \{s^+\}$, the union of a bounded interval and a topologically isolated 'point' exceeding each number in the interval. The limit $\mathcal{S}^{(\infty)} = \mathcal{R}^+ \cup \{\infty\}$ differs from $\mathcal{R}^+$ by one isolated point that exceeds every real number.

To say the same thing in another way, the state space is a filtration, which evolves as an increasing σ-field in calendar time.

Every probability distribution F on $\mathcal{S}^{(\infty)}$ is determined by its hazard measure Λ on $\mathcal{R}^+$ and its survivor function $F(s^+) = \exp\bigl(-\Lambda(s^-)\bigr)$, which is decreasing as a function of s. If the total hazard $\Lambda(\mathcal{R}^+)$ is finite, the atom of immortality $F(\{\infty\}) = \exp(-\Lambda(\mathcal{R}^+))$ is strictly positive; otherwise the atom is zero. With respect to the state of information at time s, the probability density at $y \in \mathcal{S}^{(s)}$ is $\Lambda(dy)\,\exp\bigl(-\Lambda(y^-)\bigr)$ for $0 < y \leq s$, and $\exp\bigl(-\Lambda(y^-)\bigr)$ for $y = s^+$. In particular, if Λ is proportional to Lebesgue measure on $\mathcal{R}^+$, the density is $\lambda e^{-\lambda t}\,dt$ for $0 < t \leq s$ with an atom $e^{-\lambda s}$ at s^+.

Being alive at the time of analysis is one unavoidable form of censoring. In practice, some patients disappear off the radar screen at a certain point $s > 0$, and their subsequent

survival beyond time s cannot be ascertained. These also are typically regarded as censored at the last time they were known to be alive.

1.24. Longitudinal study. In a longitudinal study, also called a panel study, each physical unit is measured at a sequence of time points. Growth studies, of plants or of animals, are of this type, the response $Y(i,t)$ being height or weight. Usually the design calls for measurements to be made at regular intervals, but in practice the intervals tend to be irregular to some degree, particularly for studies involving human subjects.

1.25. Cemetery state. A situation arises in geriatric and other medical studies where, beginning at recruitment, measurements on physical or mental capacity are made annually on patients—but only while they are alive. All patients ultimately die, and the number k_i of measurements on patient i is a major part of the response, which is closely connected with survival time. In this setting, each patient may be regarded as an observational unit, in which case the response $Y_i = (Y_i(0), \ldots, Y_i(k_i - 1))$ is a point in the state space $\cup_{k\geq 0}\mathcal{R}^k$ implying death before time k_i. Alternatively, if each patient-time combination is regarded as one observational unit, it is necessary to add to the real numbers an absorbing state, such that $Y_i(t) = \flat$ implies that patient i is dead at time t. The state space for one observational unit is $\mathcal{R} \cup \{\flat\}$; the state space for one experimental unit (patient) is $\mathcal{S}^{(\infty)} = (\mathcal{R} \cup \{\flat\})^\infty$, each sequence $\flat$-padded on the right where needed.

As always, the state space at calendar time s includes only those events observed or observable up to that time; the state space is censored by the calendar, not by the death of patients.

Exercises 1.

Exercise 1.1: An equivalence relation is a Boolean function $E{:}\mathcal{U} \times \mathcal{U} \to \{0,1\}$ that has three properties (i) $E(u,u) = 1$ for every $u \in \mathcal{U}$ (reflexive); (ii) $E(u,u') = E(u',u)$ for every pair u, u' in $\mathcal{U}$ (symmetric); (iii) $E(u,u') = 1$ and $E(u',u'') = 1$ implies $E(u,u'') = 1$ for every triple (transitive). From condition (ii), $E(u,u') = 1$ implies $E(u',u) = 1$, we may apply (iii) with $u'' = u$ to deduce that $E(u,u) = 1$, and hence that the reflexive property follows from symmetry and transitivity. Give an example to explain the fallacy in this argument.

Exercise 1.2: Consider a horticultural study into the effect of plant foods on the growth of plants in pots. Each pot contains one plant, which is one observational unit. Twelve plants are assigned by randomization to each of four suppliers whose plant food product is to be tested. The product for each supplier is tested at three dose levels—zero, one and two milligrams per pot per day. The assignment of plants to dose levels is done independently by randomization for each supplier subject to the condition that there are four plants for each dose level.

True or false? If ambiguous, clarify; if false, explain.
(i) supplier S is a qualitative factor with four levels.
(ii) this is a randomized blocks design with supplier as the block factor.
(iii) dose D is a quantitative factor with three levels.
(iv) the ordered pair (S, D) is a factor with 12 levels.

Exercise 1.3: For the experiment described in the previous exercise, four possibilities for the effect on growth are as follows.
(i) No product has any effect on growth regardless of dose;
(ii) all four products are equivalent;
(iii) for each supplier, the effect on growth is linear in the dose;

(iv) for each supplier, the effects on growth are non-linear.

Each of these statements corresponds to a certain subspace for a linear model. What are the four subspaces (in standard model notation) and what are their dimensions?

Exercise 1.4: Let $f(t) = \log(1+t^2)/(\pi t^2)$ for $0 < t < \infty$ be the density function of a probability distribution F on $\mathcal{R}^+ \cup \{\infty\}$. Show that $S(t) = 2\,\mathrm{arccot}(t)/\pi + tf(t)$ is positive and satisfies

$$-\frac{d}{dt}\log S(t) = \lambda(t).$$

Hence calculate the hazard density $\Lambda(dt)$ and the atom of immortality $F(\{\infty\})$. What is the median survival time?

Exercise 1.5: Calculate the probability density f of the survival time corresponding to the hazard measure $\Lambda(dt) = \lambda t^{\rho-1}\,dt$ for $\lambda > 0$. This distribution has a name. What is it? Are there any parameter values for which $F(\{\infty\}) > 0$?

Exercise 1.6: True or false: $s^+ = (s, \infty)$. Explain

Exercise 1.7: True or false: For each real number $s > 0$, the interval s^- is the complement of s^+ in $\mathcal{S}^{(\infty)}$. You can be pedantic and answer negatively, or more flexible and answer positively, but you must explain your reasoning.

Exercise 1.8: One of the following sets is finite; two of the other three are [arguably] equal. Identify the sets.

$$(a)\ s^- \cup s^+, \qquad (b)\ \{s^-\} \cup \{s^+\}, \qquad (c)\ \mathcal{S}^{(s)}, \qquad (d)\ \mathcal{S}^{(\infty)}.$$

Exercise 1.9: Suppose that the hazard measure has a density $\lambda(y)$ with respect to Lebesgue measure on $\mathcal{R}^+$. In an earlier paragraph, the survival distribution was said to have a density whose value at $y \in \mathcal{S}^{(s)}$ is $\lambda(y)\exp\bigl(-\Lambda(y^-)\bigr)$ for $0 < y \le s$, and $\exp\bigl(-\Lambda(y^-)\bigr)$ for $y = s^+$. What is the dominating measure on $\mathcal{S}^{(s)}$?

Exercise 1.10: Consider a survival model in which the survival times of patients are independent with hazard measures depending on sex and treatment as follows:

	$T=0$	$T=1$
F	Λ_F	$\Lambda_F \exp(\beta_F)$
M	Λ_M	$\Lambda_M \exp(\beta_M)$

The parameter space consists of a pair of hazard measures Λ_F, Λ_M, plus a treatment effect $\beta = (\beta_F, \beta_M)$ in $\mathcal{R}^2$. Show that the treatment effect is a group action on the space of functions from $\{F, M\}$ into the set of probability distributions on the positive real line. Describe a generic element in the space, a generic element in the group and its action on the space.

Exercise 1.11: Consider the model in the previous exercise with hazard measures restricted to have strictly positive density on the real line. Show that the sex effect is a group action on probability distributions. Describe the group and the action.

Exercise 1.12: Consider the sub-model in which $\Lambda_F(dt) \propto dt$ is proportional to Lebesgue measure, and $\Lambda_M(dt) \propto t\,dt$. What is the density function of survival times for men? Show that the treatment effect is a group action, but the sex effect is not.

2. Examples

Example 2.1 A surgeon decided to investigate the effect of hyperbaric oxygen treatment on the rate at which surgical wounds heal. It is not feasible to experiment on humans, so he used 24 rats, which he partitioned randomly into three groups of eight rats. The procedure called for an incision to be made along the back, from the shoulder to the tail, after which the wound was repaired with surgical staples. The first treatment group was placed for one hour per day in a chamber of pure O_2 *at two atmospheres of pressure; the second for one hour of pure* O_2 *at one atmosphere, and the control group for one hour daily with standard air at standard pressure.*

After two weeks, the animals were sacrificed, i.e., killed. From each animal, five strips of skin perpendicular to the wound were removed. Each strip was one cm. in width, with the surgical scar across the middle. The first strip comes from the shoulders, the last from near the tail, the others uniformly spaced along the back. For each strip, a tensiometer recorded the force needed to break the skin: gruesome but factual.

The goal is to assess the evidence for a beneficial O_2 *effect.*

Observational units: The observational units are rat-site pairs $u = (i, s)$; there are six sites and an endless supply of rats, so the mathematical population is the product set $[N] \times [6]$, where $[N]$ is the set of integers. The sample consists of 24 rats comprising 120 rat-site pairs, $S = [24] \times [6]$, and the response $Y : S \to \mathcal{R}$ is a real-valued function, which is conveniently displayed as a rat-by-site matrix of order 24×6. Even though the response is necessarily positive, there is no loss of generality in taking the state space to be the set of real numbers.

A function $T : S \to [3] \equiv \{1, O_2, O_2^2\}$ assigns a treatment level to each sample unit. Since sites are inseparable from their owner, the design constraint $T(i, 1) = \cdots = T(i, 6)$ for every i implies that each feasible assignment is a function on rats, taking the same value at every site on the same rat. For each treatment level, the subset of rats receiving treatment r,

$$S_r = T^{-1}(r) = \{i : T(i) = r\}$$

consists of eight rats. Balance is a design constraint, not strictly speaking necessary, but desirable for efficient use of experimental resources. There are $24!/(8!)^3$ (roughly 10^{10}) treatment assignment functions consisting of three disjoint treatment-labelled subsets of eight rats, and the randomization scheme chooses T uniformly at random from this set.

Site is a covariate, a function on the units that is recorded at or before baseline. Sites are labelled, and it is entirely possible that the treatment effect, if it is non-null, might have a posterior to anterior trend.

Rat is an equivalence relationship on units such that $R(u, u') = 1$ if the pair $u = (i, s)$ and $u' = (i', s')$ satisfy $i' = i$. When we choose to regard *rat* as a block factor rather than a covariate, we discard the rat labels; we are saying implicitly that the values (T_i, Y_i) on distinct rats are distributed exchangeably.

Each pair of rats has positive probability (16/23) of being assigned a different treatment, and each pair of sites on the same rat is necessarily assigned the same treatment, so each rat is one experimental unit. Experimental unit is an equivalence relation on sample units, which coincides with *rat*, or its restriction to S.

It is possible to proceed with rats as observational units, in which case the state space is $\mathcal{R}^6$, one component for each site.

Exercise 2.1: Why were the control rats put in the hyperbaric chamber each day when they could have had the same air at the same pressure without leaving their cages?

Exercise 2.2: The set of functions $x\colon A \to B$ is denoted by B^A, so the set of functions $[n] \to \mathcal{R}$ is $\mathcal{R}^{[n]}$, or $\mathcal{R}^n$ for brevity. The set $[3]^{[24]}$ of functions $[24] \to [3]$ is finite; how many elements (functions) does it contain?

Exercise 2.3: A function $x\colon [24] \to [3]$ has three disjoint blocks, $x^{-1}(1), x^{-1}(2), x^{-1}(3)$, each block being a subset, possibly empty, of [24]; the block type is the ordered list of block sizes, $(\#x^{-1}(1), \#x^{-1}(2), \#x^{-1}(3))$. How many functions are there of type (7, 8, 9)? How many of type $(8, 8, 8)$?

Exercise 2.4: A partition B of $[n]$ is a set of disjoint non-empty subsets whose union is $[n]$; the block type $1^{m_1}2^{m_2}\cdots n^{m_n}$ is a partition of the integer n, in which m_r is the number of blocks, or parts, of size r. Empty blocks are usually omitted. How many partitions of [24] are there of type $7^1 8^1 9^1$? How many of type 8^3?

Example 2.2 This example is based on an experiment by Solandt, DeLury and Hunter (1943); the data were subsequently used by Cochran and Cox (1957, p. 176–181) to illustrate analysis of covariance for a factorial design.

A number of experiments indicate that electrical stimulation may be helpful in preventing the wasting away of muscles that are denervated. A factorial experiment on albino rats was conducted in order to learn something about the best type of current and the most effective mode of treatment. The factors and their levels are as follows:

A: number of treatment episodes daily: 1, 3, 6;

B: Length of treatment episode (minutes): 1, 2, 3, 5;

C: Type of current: Galvanic; Faradic; AC 60 Hz; AC 25 Hz;

With these levels, there are 48 different treatment combinations, each of which was applied to a different rat. Two replications of the experiment were conducted, using 96 rats in all. Each replicate is one block of 48 rats.

The muscles denervated were the gastrocnemius-solus group on one side of the animal, denervation being accomplished by the removal of a small part of the sciatic nerve in the thigh. Treatment was started on the third day after denervation, and continued for 11 consecutive days. The measure used for judging the effectiveness of treatment was the weight y of the denervated muscle at the end of the experiment. Since the response depends greatly on the size of the animal, the weight x of the corresponding muscle on the other leg was also recorded. Both weights were recorded in units of 0.01 gm at the end of the experiment, i.e., two weeks after denervation.

Clearly A, B, C are treatment factors, two quantitative and one qualitative. Although not stated, it is presumed that treatment levels were assigned at random to the rats. There is an additional unreported treatment factor, the leg (left or right) on which the nerve was cut and to which the stimulation was applied. Solandt *et al* disuss the choice as follows: "*Initially alternate sides were denervated, but ultimately the selection of the limb to be denervated was made at random.*" Presumably the left-right difference is negligible for treatment effects, so although the limb was selected at random, the choice was not recorded. Instead, legs were re-labelled normal and denervated for the purpose of analysis.

One point of view is that each denervated muscle is one experimental unit; these are the objects to which treatment is applied and on which the main measurement $i \mapsto y_i$ is made. What then is the role of x, which is not a physical measurement made on any experimental unit? Despite the notation, x is a post-baseline measurement, not a covariate. Mathematically speaking, it is a function on the units. It is not a physical

property of muscle i, but of the partner of i, here denoted by $i' = \mathrm{pa}(i)$. Thus y is weight at two weeks, $x_i = y_{\mathrm{pa}(i)}$, and the response for unit i is the ordered pair $i \mapsto (y_i, y_{\mathrm{pa}(i)})$, a point in $\mathcal{R}^2$.

An alternative and slightly more straightforward point of view is that each rat is one observational unit and one experimental unit. The treatment levels are those listed above applied to the left or right leg, the partner leg being left as a control. By this count, there are 96 treatment levels in all. The response for rat i is a left-right pair of weights $i \mapsto (z_{il}, z_{ir})$, and the denervation factor tells us which side, left or right, is denervated. The analysis is best done directly on the z-values, but (x_i, y_i) is equal to (z_{il}, z_{ir}) if denervation occurs on the right, and (z_{ir}, z_{il}) if denervation occurs on the left.

Note that if x were total body weight at baseline, the status would be quite different, and the statistical analysis also.

Example 2.3 Consider a meterological time series in which rainfall and temperature are recorded daily at a range of sites throughout the UK. Rainfall is the daily total in mm., and temperature is the daily maximum in °C. The temporal origin or baseline can be chosen arbitrarily, say Jan. 1, 1750. For mathematical purposes, the units are spatio-temporal pairs $u = (s, t)$, where t is an integer, positive or negative, and s is either a point in the plane or a point on the surface of the sphere. Although rainfall cannot be negative, we may take the state space to be $\mathcal{R}^2$. In addition to the marginal variables s and t, there may be site-specific covariates such as elevation and local topography, which are constant in time. Apart from Divine intercession, no attempt is usually made to influence the weather, so there is no randomization and no treatment assignment.

Calendar time has both a cyclic and a linear aspect, both of which are relevant for statistical modelling of meteorological series. For the linear part, each time point is regarded as a real number, and the Euclidean distance $|t - t'|$ is a relationship that is invariant with respect to both temporal and spatial translation. For the cyclic aspect, the calendar date t (mod 365) may be regarded either as a 365-level factor or as a sequence of points equally spaced on the unit circle. Chordal distance, $0 \le \mathrm{Ch}(t, t') \le 2$ is a relationship among units that is periodic and invariant with respect to both temporal and spatial translation. The effect of leap years is not ignorable in a long series, so chordal distance must be defined taking this into account.

Exercise 2.5: With t measured in days beginning at $t = 0$ on Jan 1,1750, give an expression for $\mathrm{Ch}(t, t')$. Bear in mind that 1700, 1800 and 1900 were not leap years, so the mean number of days in one Gregorian year is $365 + 1/4 - 3/400 = 365.2425$.

Example 2.4 The following is a condensed description of a moderately complicated long-term neuro-psychology study concerned with the the physical effect of parental influences on children's outcomes, specifically speech acquisition and the physical effect on cortical thickness as measured by a fMRI brain scan. The example illustrates some of the issues that arise in the interpretation of the various definitions in a specific situation.

Sixty families (parent-child pairs) were recruited for the study. The interaction between parent and child was videotaped for 90-minute periods on 12 occasions every four months from the child's first birthday until the child's fifth birthday. These were not staged encounters in a laboratory setting; they were recorded in the family home at a time when the child was awake, and the intention was to record 'typical' behaviour. On some occasions, the parent, usually the mother, used that time to do the laundry and other household chores. For each parent-child pair, these 18 hours of video were reduced to two numbers: (i) the average number of words uttered by the parent, and (ii) the average number of utterances about 'abstract topics'. The child's sex, parent's IQ, education and household income were recorded at recruitment.

Brain scans were made on the children when they were 10–13 years old. An effort was made to contact each family, and to get parental consent, but in the end 18 children were scanned at age 10 years, and 23 children at age 12 years, 13 of these being repeats. From each fMRI scan, the cortical thickness was measured for six brain regions. All told, therefore, the cortical data consists of six measurements on each of 28 distinct children with 13 repeats, making 41 scans in total. In addition, each scan reports an overall cortical thickness value, an average over the entire brain surface. An added complication is that two different scanners were used, one for the children at age 10, and another for the children at age 12.

The psychologist's main goal is to determine whether the frequency of abstract utterances in childhood has an effect on cortical thickness in any of the six brain regions.

Discussion. *Observational units.* It is natural here to regard the child, or parent-child combination, as the fundamental observational unit, but the observation is recorded for a specific site at a specific age on each child, so the situation is slightly more complicated. Arguably, the observational units are the points (i, t, s) in the product set *children* $\times$ *ages* $\times$ *sites*. To some extent, the choice (i, t) versus (i, t, s) is a matter of taste, but the second form works better for present purposes because each marginal component *child*, *site* and *age at scan* is automatically a function on the units.

Sample. For this study, the sample S is a finite set of 246 triples $u = (i, t, s)$ consisting of 41 distinct (i, t) pairs with six sites for each. As is invariably the case in such studies, the planned sample was much larger, presumably all 60 children at two or more times with six sites for each. We proceed as if the subset S is either fixed at baseline, which is probably not the case here, or S is a random sample independent of the response (cortical thickness). The latter assumption may be close to the truth, but it is not easily tested or verified.

Baseline. For the purposes mentioned in the final sentence of the description, the baseline or recruitment time may notionally be set at age five or ten years, in which case the number of parental utterances and the number of abstract utterances are covariates according to the definition.

Covariates. Additional covariates include sex, which is a property of the child, and IQ, education and household income, which are associated more directly with the parent. Income, IQ and the utterance variables are quantitative; sex is qualitative, and educational attainment is qualitative or semi-quantitative. All four are recorded at age one, so all are covariates according to the definition. Age at scan is also a covariate, as is *site*.

Overall cortical thickness is not a covariate according to the definition; it is technically a component of the response. It is up to the psychologist to decide what the primary response should be, and whether overall thickness is a part of that. For example, the response of interest could be re-defined as the site-specific thickness normalized by overall thickness.

Design variables. fMRI technology did not exist when the children were aged five, so it would have been impossible to plan the study at that time. A baseline of nine or ten years of age is more plausible, in which case *age at scan* and *machine used* may be regarded as covariates or design variables. In the present setting, *age* and *machine* are confounded: they determine the same two-block partition of the units

$$\{u : \mathit{age}(u) = 10\} = \{u : \mathit{machine}(u) = 1\}$$
$$\{u : \mathit{age}(u) = 12\} = \{u : \mathit{machine}(u) = 2\}$$

Treatment. This is an example of an observational study; there is no treatment factor or differential intervention, so there are no experimental units. It may be argued that

video-taping is an intrusion or intervention, but it is not a differential intervention that is applied in one form to a subset of the families and in another form to the complementary subset. If there were two video strategies assigned uniformly at random, each child, or parent-child, would be one experimental unit. In that circumstance, the baseline would have to be moved back to recruitment at age one or earlier, in which case parental utterances are not covariates according to the definition.

Response. Cortical thickness in mm. is the response implied by the final paragraph of the description. With $u = (i, t)$ pairs as observational units, the response $Y_u = (Y_{u1}, \ldots, Y_{u6})$ at age t is a point in $\mathcal{R}^6$, one component for each site. The resulting observation $Y[S]$ is a matrix of order 41×6, one row for each scan, with 13 children occupying two rows each and 15 occupying one row each. This is manageable, but a little awkward. With child-age-site triples $u = (i, t, r)$ as observational units, the response $Y_u = Y_{itr}$ is a real number, and the observation $Y[S]$ is a list of 246 real numbers. The difference is a difference of format, which is largely cosmetic, but the second form is more convenient.

Relationships. Child is a relationship such that $c(u, u') = 1$ if the two units involve the same child; this is a block factor or equivalence relation on the units, which are (i, t, s)-triples. Site is a six-level factor; some pairs of sites are physically closer than others, so there may be a distance function defined on pairs of sites.

Probability model. The preceding discussion says little about how the various covariates and relationships are to be used in a statistical model. There is, however, one guideline. If the relevant parent-child covariate values, *sex, IQ,*... are equal for two children, the responses must have the same joint distribution for all subsequent times and sites. It should be clear also that two observations on the same child at different sites or at different times are likely to be strongly correlated, so *child* as a block factor will certainly occur in the model for covariances. There may also be a temporal trend and a temporal correlation for observations at different times on the same child. There may also be site effects and spatial correlations.

Example 2.5 In a competition experiment such as a chess tournament or the Rugby World Cup, each observational unit is an ordered pair (a, b) of distinct competitors or teams. The sample is a subset consisting of some, but usually not all, ordered pairs. For the preliminary round of the 2015 rugby world cup, the 20 qualifying teams were divided into four pools of five teams each. Within each pool, each pair of teams played once only. The sample S is a collection of 40 pairs, 10 pairs for each pool.

The response may be an ordered pair $Y_{a,b} = (Y_a, Y_b)$ of scores, in which case the state space is $\mathcal{R}^2$, and $Y_{b,a} = (Y_b, Y_a)$ is automatically determined by $Y_{a,b}$. Or the response may be the score difference $Y_{a,b} = Y_a - Y_b$, in which case the state space is $\mathcal{R}$, and Y is a skew-symmetric matrix, regarded as a random variable. Regardless of the nature of the response, the observation $Y[S]$, which is the restriction of Y to the subset S, is not a matrix unless $S \subset [n]^2$ is a product subset. For the 2015 RWC example, $Y[S]$ is a set of four 5×5 matrices, one for each pool. Note that $Y[S]$ is a set of four disjoint sub-matrices, not a block-diagonal matrix which implies zero values off the blocks.

This analysis overlooks one complication, namely that one of the 20 competing teams was the home team. Home team advantage is often substantial, but this advantage was not sufficient in the 2015 RWC to lift the home team out of the also-rans.

In a baseball or football league, where each pair of teams meets more than once, each game is an observational unit, home team and away team are two functions on games, and $\mu_{a,b}$ is the expected response for each game in which a is at home and b is away (assuming that the value is constant throughout the season). Similar remarks apply to

chess tournaments, where the ordered pair (a, b) is a game in which player a is white and player b is black.

Example 2.6 The following is an abbreviated description of an experiment by Sharon et al. (2010) on the effect of diet and commensal bacteria on the breeding behaviour of fruit flies, Drosophila melanogaster. An initial genetically homogeneous population was segregated into two lines for breeding purposes, and maintained separately for 30+ generations. One breeding line was fed with a corn-molasses-yeast diet (C), the other with a starch diet (S). After five breeding generations, a small subset of flies was removed from each breeding line, fed separately for one generation on a neutral diet, and the progeny after 5+1 generations were used for test purposes. Four virgin flies were inserted into each of 30 test wells, one male and one female from the ancestral CMY line, one male and one female from the ancestral S line. All male-female matings occurring during the first hour in each well were recorded by an expert observer (graduate student) as either homogamic, CC or SS, or heterogamic CS or SC.

For background information, one mating, including courtship activities, takes about 10–15 minutes. Following mating, Drosophila females have a refractory period of approximately 12–24 hours during which no mating occurs. Given the opportunity, males may mate several times in one hour.

Exercise 2.6: Each xxxx is one observational unit. Answers: xxxx = (i) fly; (ii) well; (iii) mating; (iv) dietary pair; (iv) M-F pair; (v) M-F pair in the same well. Justify your answer.

Exercise 2.7: What is the treatment, and what are the treatment levels?

Exercise 2.8: What are the experimental units?

Exercise 2.9: How did the student distinguish male flies from female flies?

Exercise 2.10: In the last generation immediately before testing, the C and S flies were housed separately, fed the same neutral diet, and their virgin offspring used for test mating. Why was this step taken?

Exercise 2.11: How did the student distinguish ancestral C flies from ancestral S flies? The information provided above is not sufficient to answer this question, but the answer can be found at `http://www.pnas.org/content/107/46/20051.full`.

Exercise 2.12: Describe briefly the courtship behaviour of fruit flies, and the relevance of this activity to the preceding question.

Example 2.7 Lymphedema, or swelling of the limbs due to poor lymphatic drainage, is a problem that afflicts patients who have had cancer surgery followed by removal of certain lymph nodes. It is also a side-effect of the parasitic disease, Filariasis. The symptoms can be alleviated to some extent by plastic surgery. Depending on the specifics of the case, surgery may involve transplantation of lymph nodes from elsewhere to the affected site, for example from the groin to the armpit. There is also a less invasive bypass procedure in which the lymphatic ducts are attached directly to a nearby vein, bypassing the lymph nodes that were removed in cancer surgery.

Over the past 2.5 years, roughly 225 such surgeries have been performed at the University of Chicago, all on patients who had had cancer surgery in the past, and were

subsequently declared cancer-free. Information from these cases was compiled by going back to the patient records and extracting variables as follows. Pre-operative variables: age, sex, bmi, cancer site, time since cancer surgery, radiation treatment, severity of scarring, lymphedema severity (volume of the affected limb relative to the unaffected limb), and so on. Lymphedema surgery type, with one bypass and four transplant levels, was also recorded.

Outcome measurements included relative limb volume and LLIS (Lymphedema Life Impact Scale), which were recorded at certain times following surgery. All patients had at least one follow-up measurement, but some had multiple follow-ups at irregular times in the year following surgery.

The goal of this study is to compare the effectiveness of the different surgical options, in particular bypass versus transplantation.

Discussion. This is a comparative observational study. It is comparative in the sense that the goal is to compare the outcome distribution for one surgical treatment versus another, (lymph node transplantation versus lymph duct bypass). It is an observational study in the sense that the assignment of patients to one surgical technique or the other was not randomized, so there can be no expectation that the pre-operative prognosis for bypass patients is comparable to that for the transplant patients.

There is a further potential confounding issue in the recording of the response. Follow-up is ultimately voluntary, i.e., dependent either on the good will of the patient or the perceived medical needs. If the well of good-will for patients having successful outcomes is deeper than that for patients whose outcomes are less successful, we should expect successful outcomes to be over-represented in the data. On the other hand, if the appetite for medical attention is higher among the medically needy patients whose surgery was less successful, we should expect unsuccessful outcomes to be over-represented in the data. But we should not necessarily expect that the degree of over-representation should be greater for one surgical group versus the other. The point in all of this is that we cannot be entirely confident that the variation in the frequency of follow-up from one patient to another is statistically independent of the outcome.

Finally, since these data were culled from past records, it is not entirely clear whether any or all bypass patients might have been eligible for transplantation, or transplant patients eligible for bypass. Certainly, no such determination was requested or made at the time of surgery. But the direct comparison of outcome distributions for these groups is likely to be useful as a guideline for future surgeries only if both options are medically feasible and ethically justifiable.

Ordinarily a probability model for a randomized experiment has one response distribution for each treatment level, and the differences among them are causal in the sense that the response distribution for any given experimental unit can be altered by treatment intervention. The situation here is a little different. Each transplant is one of four types, depending in part on the affected limb and the source of the transplanted lymph node. A patient who is eligible for a groin-to-armpit transplant is not eligible for an armpit-to-groin transplant. We can compare success rates for the four types of surgery, but the interpretation of a given difference is causal only if the levels being compared are eligible options for the given patient.

3. Remarks on the definitions

Remark I. It is commonplace in agricultural research to study the effects of several treatments in combination. For example, we might want to compare three grass varieties in combination with four fertilizer mixtures. Then, variety is a treatment factor with three levels, fertilizer is a treatment factor with four levels, and the ordered pair (variety, fertilizer) is a treatment factor with 12 levels. The treatment structure could be more complicated, perhaps with a range of fertilizer doses, or with several dates of planting.

Suppose that the field is rectangular, with six strips of land running N–S, each strip being divided into eight approximately-square plots of equal area. The mechanical constraint is such that it is feasible to plant one variety only in each strip of land, so that each variety occupies two strips. Each fertilizer level occurs twice in each strip.

In relation to variety as the treatment, the experimental units are the strips. All plots in one strip are assigned the same variety, and each pair of strips may be assigned different varieties. In relation to fertilizer, the plots are the experimental units. Each pair of plots may be assigned different fertilizers. In relation to the ordered pair (variety, fertilizer), the plots are once again the experimental units.

Remark II. Statistical inference is ultimately a probabilistic statement about the values that are likely to occur on extra-sample units, whether these be eligible patients, or notional unplanted plots of land, or subsequent time points, or next year's flock. A patient who is not eligible today may become eligible tomorrow, and if the statistical model has any value, it must encompass all events needed to compare this future patient's 5-year survival probability under one treatment arm versus another, i.e., to compare $\mathrm{pr}(Y_i > 5 \,|\, \mathrm{data}, T_i = a)$ with $\mathrm{pr}(Y_i > 5 \,|\, \mathrm{data}, T_i = b)$ for an extra-sample patient who becomes eligible in the future. In many simple situations, this difference is a function of treatment-effect parameters τ_a, τ_b, which may be estimated by least squares or maximum likelihood. The difference of conditional probabilities may, however, depend on the observed data in a more complicated manner, particularly if the values for distinct units are not independent.

Mathematically speaking, the sample is an arbitrary fixed finite subset of the population. In the clinical-trial setting, the sample is the first n eligible patients encountered, who also give informed consent. Selection is used to determine eligibility, but no random sampling is involved. Occasionally, the sample may be generated randomly, but this is rare; provided that the sample (as a random subset) is independent of the response, the response distribution for the random sample is the same as the response distribution for a comparable fixed sample (having the same size and covariate structure).

Remark III. Randomization does not mean haphazard assignment of treatment levels, but rather an objective assignment by an overtly random device such as a random number generator. The purpose of randomization is primarily to avoid bias arising from preferential assignment of selected units, or healthier patients, to a particular treatment. It is immaterial whether the preferential assignment is caused by conscious or unconscious bias, by systematic bias, by accidental bias, or by cheating; randomization is the best device available to guard against all bias. A secondary purpose is to obtain correct standard errors for treatment effects.

If the units are human patients, what they are told is part of the treatment. A patient who is given the control drug A and told that he is in the control group, is not comparable to a patient on the same drug who is not told. If the units are patients, it is advisable to maintain blindness, (so that the patient is unaware which treatment has been used), and randomization is helpful in achieving blindness. Where it is feasible, it is advisable that the doctor should also be blinded. For a good example of what can go wrong, and how the errors might have been avoided by better design, see the 1931 paper *The Lanarkshire milk experiment* by Student (W.S. Gosset) *Biometrika 23, 398–406.*

Remark IV. Blocking means the construction of a block factor, i.e., dividing the set of units into disjoint subsets. The purpose of blocking is to improve the precision of estimation by identifying and removing a part of the variation that would otherwise contaminate the estimate of treatment effects. Blocks should be chosen so that the units within each block are similar and homogeneous. Thus, blocking is effective if the within-blocks variation in the response is appreciably smaller than the between-blocks variation. Ideally, each block should contain enough units so that each treatment level occurs at least once in each block, but this should not be done at the expense of introducing excess within-blocks heterogeneity. If the blocks are selected by randomization, the between-blocks mean square and the within-blocks mean square are approximately equal and exactly equal in the limit, in which case blocking is ineffective.

When experimental material (chemical reagents) is supplied in batches over time, or the processing is done in batches, there are usually substantial 'batch effects', which are not easily controlled. In situations of this sort, it is best to regard each batch as a block.

Remark V. Observational units versus experimental units. In the surgical trial of hyperbaric oxygen, the observational units are mouse-site pairs, while the experimental units are mice. There are 24 experimental units, five sites, and 120 observational units. Eight mice assigned to each treatment level, so the estimated treatment effects relative to control are differences of sample averages, each average involving eight experimental units and 40 observational units.

The distinction between observational and experimental unit is important because, as a rule of thumb, there cannot be more degrees of freedom for the estimation of treatment-effect variances than there are experimental units. In the absence of site-by-treatment interaction, we could equally well work with the 24 mouse averages, in which case there are two degrees of freedom for additive treatment effects and 21 degrees of freedom for the estimation of residual variance. In most settings of this type, it would be regarded as a gross statistical error to pool the between-mouse sum of squares on 21 degrees of freedom and the within-mouse sum of squares on 92 degrees of freedom, and to use the pooled mean square to estimate the variance of treatment contrasts. In this case, the correct calculation yields a standard error of 0.44 for treatment contrasts; the incorrect calculation gives 0.27, which gives a misleading impression of the accuracy achieved.

Rules of thumb are very helpful most of the time, but the issues are sometimes not so clear-cut in practice as they appear to be in the mathematics. The *Drosophila* mating experiment by Sharon et al. (2010) is a case in point. Treatment is diet, which has two levels, C and S. Diet is an intervention assigned to flies, but the intervention occurs only to flies in generation zero. Treatment assignment by randomization would involve segregation of individuals, labelling the flies 1–N, generating complementary random subsets C and S of size $N/2$, and finally re-housing each breeding line communally. The authors do not claim to have engaged in this activity, which is difficult to imagine and almost certainly unnecessary. All *Drosophila* flies look more or less alike, so a haphazard partition of a large batch into two subsets of approximately equal size is unlikely to result in appreciable genetic differences between the two lines. However, if the flies in generation zero were delivered in two batches, it would be better to eliminate the possibility of batch differences by splitting each batch in two, or to mix before splitting.

Since there is no treatment assignment or intervention for subsequent generations, what was a treatment randomly assigned in generation zero becomes more like a classification factor for later generations. Each test fly in generation 5+1 is of ancestral type C or S, and there is no possibility of switching heritage.

Each mating well is one observational unit. That much is clear. Each well contains

four flies designated MC, FC, MS, and FS. Each mating is one of four types

$$\mathcal{M} = \{CC,\ CS,\ SC,\ SS\}.$$

The state space of observable ordered M-F matings is $\cup_{k\geq 0}\mathcal{M}^k$, including the one-point subset $\mathcal{M}^0 = \{\star\}$ for no matings. On account of the female refractory period of 12+ hours, a female can mate only once, so each $\mathcal{M}^k$ for $k \geq 3$ has zero probability. Eight of the 16 points in $\mathcal{M}^2$, including the four diagonal points, also have zero probability. Ignoring order, the points having positive probability are

$$\mathcal{M}^0 \cup \mathcal{M} \cup \{\{CC, CS\},\ \{CC, SS\},\ \{CS, SC\},\ \{SC, SS\}\}.$$

Any set that includes these nine events is adequate for the state space.

Events in the state space may be partitioned into activity classes according to the number of sexually active males and the number of sexually active females. Up to order $(2, 2)$, these classes are as follows:

$$\begin{array}{ll}
(0,0) & \star \\
(1,1) & \mathcal{M} \\
(1,2) & \{CC, CS\},\ \{SC, SS\} \\
(2,1) & \{CC, SC\},\ \{CS, SS\} \\
(2,2) & \{CC, SS\},\ \{CS, SC\},
\end{array}$$

with class sizes 1, 4, 2, 2, 2 respectively. If diet has no effect on behaviour, the probability is necessarily constant in each activity class. For reasons mentioned above, class $(2, 1)$ has zero probability, and there is no reason to think that diet might alter this. Note that $\mathrm{diag}(\mathcal{M}^2)$ is regarded as a different activity class than $\mathcal{M}$.

Remark VI. Classification factor, block factor, treatment factor: The chief distinctions roughly in order of importance are: (i) The first two are recorded pre-baseline, the third at or post-baseline. (ii) one is a random variable generated by randomization, implying a particular intervention. (iii) one is an equivalence relation, the others are functions on the units. (iv) as a mathematical function, treatment level may be restricted to sample units only. (The last two are mathematical distinctions, which are less important than the first two. First, every function determines an equivalence relation, so every classsification and treatment factor determines a block factor. Second, the sample can always be extended, at least in a mathematical or conceptual sense, for whatever purpose such as prediction.

The set of levels of a classification or treatment factor usually refers to the levels occurring in the sample, but sometimes refers to the levels in the population, an ambiguity that may need to be clarified. A treatment factor such as dose that is tested with two settings in the design usually has infinitely many unused settings. A quantitative factor usually has an infinite number of levels in the population, but '*age at last birthday*' is quantitative and (arguably) has a finite set of levels. A quantitative factor has the potential to be used linearly or quadratically in a model; a qualitative variable does not; an ordinal variable is somewhere in between.

4. Vector space synopsis.

The word *space* in mathematics means a set with additional structure. Thus, we talk of a Euclidean space, a metric space, a topological space, a Hilbert space, a measurable space, a Polish space, a probability space, and many other types. A *vector space* is a non-empty set $\mathcal{V}$ whose elements v are called vectors. This set is a group, which closed under addition: $u, v \in \mathcal{V}$ implies $u + v = v + u \in \mathcal{V}$. It is also closed under scalar multiplication: $v \in \mathcal{V}$ implies $\alpha v \in \mathcal{V}$ for all scalars α. The group property implies the existence of a vector $0 \in \mathcal{V}$ such that $v + 0 = v$ for every v, and also the inverse $-v$ such that $(-v) + v = 0$. Additional conditions are needed to ensure that scalar multiplication and vector addition are compatible, for example that $1v = v$, $0v = 0$, $(-1)v = -v$, $v + v = 2v$, and so on.

The word *scalar* means a 'point in the field of coefficients', which is either the field of real numbers or the field of complex numbers, as the context requires. (Finite fields are not considered in these notes.) The real plane, $\mathcal{R}^2$, is a two-dimensional real vector space; the complex plane is a one-dimensional complex vector space (same set of points, different fields, different vector spaces). These notes deal with finite-dimensional real vector spaces.

4.1. Subspace. Let $\mathcal{V}$ be a vector space. A subset $\mathcal{U} \subset \mathcal{V}$ that is closed under vector space operations is called a subspace. For example, to each vector $v \in \mathcal{V}$ there corresponds a subset $\{\alpha v \colon \alpha \in \mathcal{R}\}$ consisting of all scalar multiples of v. For each $v \neq 0$, this ray is a subspace of dimension one.

Let $\mathcal{V}$ be either a vector space in its own right, or a subspace of a vector space $\mathcal{W}$. The statement that the vectors $\{v_1, \ldots, v_k\}$ *span* $\mathcal{V}$ means that each of the vectors v_r belongs to $\mathcal{V}$, and each vector $x \in \mathcal{V}$ is expressible as a linear combination $x = \alpha_1 v_1 + \cdots + \alpha_k v_k$ with scalar coefficients $\alpha_1, \ldots, \alpha_k$. The statement that $\{v_1, \ldots, v_k\}$ is *a basis* for $\mathcal{V}$ means also that $\{v_1, \ldots, v_k\}$ are linearly independent vectors spanning $\mathcal{V}$. In the latter case, the coefficients in the linear combination are unique. Every basis contains the same number of vectors, and k is called the dimension of the space or subspace.

In the real plane $\mathcal{R}^2$, the real line consisting of points $(x, 0)$ is closed under addition and scalar multiplication. It is a subspace. In the complex plane, the real line is a subset that is closed under addition but not under scalar multiplication (meaning complex multiplication).

4.2. Span and intersection of subspaces. Let $\mathcal{U}, \mathcal{V}$ be two subspaces in a vector space $\mathcal{W}$. The intersection is a subset that is closed under vector-space operations, which implies that $\mathcal{U} \cap \mathcal{V}$ is a subspace of $\mathcal{W}$. The two subspaces cannot be disjoint because $0 \in \mathcal{W}$ belongs to both. However, if $\mathcal{U} \cap \mathcal{V} = \mathbf{0}$ is the zero subspace, we say that $\mathcal{U}, \mathcal{V}$ are *non-overlapping* subspaces. The symbol $\mathcal{U} + \mathcal{V}$ denotes the span of the two subspaces, which is the set of linear combinations $u + v$ with $u \in \mathcal{U}$ and $v \in \mathcal{V}$. The span is a subspace of $\mathcal{W}$, the smallest subspace containing both $\mathcal{U}$ and $\mathcal{V}$. (The union $\mathcal{U} \cup \mathcal{V}$ is a subset containing both $\mathcal{U}$ and $\mathcal{V}$, but it is not a subspace unless $\mathcal{U} \subset \mathcal{V}$ or $\mathcal{V} \subset \mathcal{U}$.) If $\mathcal{U} \cap \mathcal{V} = \mathbf{0}$ and $\mathcal{U} + \mathcal{V} = \mathcal{W}$, the subspaces are said to be *complementary* in $\mathcal{W}$. The condition $\mathcal{U} + \mathcal{V} = \mathcal{W}$ implies that each vector $w \in \mathcal{W}$ is expressible as $w = u + v$ with $u \in \mathcal{U}$ and $v \in \mathcal{V}$. The additional condition $\mathcal{U} \cap \mathcal{V} = \mathbf{0}$ implies that this decomposition is unique. The direct-sum symbol $\mathcal{W} = \mathcal{U} \oplus \mathcal{V}$ implies that the subspaces are complementary in $\mathcal{W}$, and that $\dim(\mathcal{W}) = \dim(\mathcal{U}) + \dim(\mathcal{V})$. Consider, for example, the following three subspaces of the plane: $\mathbf{1} = \{(x, x) \colon x \in \mathcal{R}\}$, $\mathcal{U} = \{(x, 0) \colon x \in \mathcal{R}\}$, $\mathcal{V} = \{(0, x) \colon x \in \mathcal{R}\}$. Each pair of subspaces is non-overlapping and complementary.

4.3. Linear transformation. A transformation $A \colon \mathcal{V} \to \mathcal{U}$ from one vector space to itself or another vector space is called *linear* if $A(\alpha u + \beta v) = \alpha A u + \beta A v$ for all vectors

u, v and scalar coefficients α, β. Linearity implies $A0 = 0$. Every linear transformation has an *image*

$$\mathrm{Im}(A) = \{Av \colon v \in \mathcal{V}\},$$

which is a subspace of $\mathcal{U}$, and a *kernel*

$$\ker(A) = \{v \colon Av = 0\},$$

which is a subspace of $\mathcal{V}$, also called the *null space* of A. If $\mathcal{U} = \mathcal{V}$, the image and kernel may overlap, and one may be a subspace of the other. A linear transformation $A \colon \mathcal{V} \to \mathcal{V}$ is invertible with a linear inverse if and only if $\ker(A) = \mathbf{0}$. The identity $I \colon \mathcal{V} \to \mathcal{V}$ defined by $Ix = x$ for every x, is a linear transformation whose kernel is $\mathbf{0}$.

A transformation $A \colon X \to X$ on an arbitrary set X is called a *projection* if $A^2 = A \circ A = A$, i.e., $A(Ax) = Ax$ for every $x \in X$. A transformation $P \colon \mathcal{W} \to \mathcal{W}$ on a vector space to itself is called a *linear projection* if it is linear and it is a projection. The image and kernel of a linear projection are complementary subspaces in $\mathcal{W}$. To every linear projection P there corresponds a *complementary projection* $Q = I - P$, such that $\mathrm{Im}(P) = \ker(Q)$ and $\mathrm{Im}(Q) = \ker(P)$. Conversely, to every pair of subspaces $\mathcal{U}, \mathcal{V}$ that are complementary in $\mathcal{W}$, i.e., $\mathcal{W} = \mathcal{U} \oplus \mathcal{V}$, there corresponds a unique projection P such that $\mathrm{Im}(P) = \mathcal{U}$ and $\ker(P) = \mathcal{V}$. For a concrete example, let α be a real number. The matrix

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ \alpha & 1-\alpha \end{pmatrix}$$

is a projection $\mathcal{R}^2 \to \mathcal{R}^2$ with image $\mathbf{1}$ and $\ker(P) = \mathrm{span}((\alpha - 1, \alpha))$, which is complementary to $\mathbf{1}$.

For a linear transformation, the condition $A^2 = A$ implies that the eigenvalues satisfy $\lambda^2 = \lambda$, i.e., $\lambda = 0$ or $\lambda = 1$ only. Accordingly, the trace of a projection, which is the sum of the eigenvalues, is an integer equal to the rank: $\mathrm{tr}(A) = \dim(\mathrm{Im}(A))$.

4.4. Inner product. An *inner product* $\langle u, v \rangle$ in a real vector space is a strictly positive-definite symmetric bi-linear function (real-valued). Symmetry means $\langle u, v \rangle = \langle v, u \rangle$ for every pair; bi-linear means linear in each argument; strict positive definiteness means $\langle u, u \rangle \geq 0$ for every u and $\langle u, u \rangle = 0$ implies $u = 0$. The inner product defines a length or norm for each vector: $\|u\| = \langle u, u \rangle^{1/2}$. (For complex vector spaces, replace bi-linear with conjugate-linear.) Some authors use the alternative term *scalar product*, meaning a scalar-valued product of two vectors, not the product of a scalar with a vector.

A vector space $\mathcal{W}$ with an inner product is called an *inner-product space*, or sometimes a Euclidean space or a Hilbert space. Two vectors u, v are called *orthogonal* if $\langle u, v \rangle = 0$. Two subspaces $\mathcal{U}, \mathcal{V}$ in an inner-product space are orthogonal if every $u \in \mathcal{U}$ is orthogonal to every $v \in \mathcal{V}$. To each subspace $\mathcal{U} \subset \mathcal{W}$ of an inner-product space there corresponds a specific complementary subspace, called the *orthogonal complement of* $\mathcal{U}$ and denoted by $\mathcal{U}^\perp$, which is the set of vectors $v \in \mathcal{W}$ that are orthogonal to every vector $u \in \mathcal{U}$. Evidently, $\mathbf{0}^\perp = \mathcal{W}$, $\mathcal{W}^\perp = \mathbf{0}$ and $(\mathcal{U}^\perp)^\perp = \mathcal{U}$.

In statistical work, a vector space is often understood to have an associated inner product. The standard inner product in $\mathcal{R}^n$ is $\langle u, v \rangle = \sum u_i v_i$, which is invariant with respect to component-wise permutation; other possibilities are of the form $\sum_{i,j} W_{ij} u_i v_j$ for positive definite matrices W, not necessarily diagonal.

In an inner product space, a linear transformation $\mathcal{W} \to \mathcal{W}$ is called *self-adjoint* (or symmetric) if $\langle Au, v \rangle = \langle u, Av \rangle$ for every pair u, v. The image and kernel of a self-adjoint linear transformation are orthogonal as subspaces; they are also complementary, so $\mathrm{Im}(A) = \ker(A)^\perp$. A linear projection that is self-adjoint is called an *orthogonal projection.*

A linear transformation $A\colon \mathcal{V} \to \mathcal{U}$ from one inner-product space into another such that $\langle Au,\, Av\rangle = \langle u,\, v\rangle$ for every pair of vectors $u, v \in \mathcal{V}$ is an *isometry* preserving linear combinations and inner products. For $\mathcal{U} = \mathcal{V}$, such transformations are called *orthogonal.*

4.5. Matrix notation. Let X be a matrix of order $n \times p$ whose columns are a basis for the subspace $\mathcal{X} \subset \mathcal{R}^n$. This statement means that the columns of X are linearly independent, so X has full rank $p \le n$. Let L be the matrix of a linear transformation $\mathcal{R}^n \to \mathcal{R}^{n-p}$ such that $\ker(L) = \mathrm{span}(X) = \mathcal{X}$. This statement means that L is of order $(n-p) \times n$ of rank $n-p$, and $Lv = 0$ if and only if $v \in \mathcal{X}$, i.e., if $v = X\beta$ for some coefficient vector β. In particular, $LX = 0$.

Consider now the following $n \times n$ matrices:

$$A = X(X'WX)^{-1}X'W, \quad B = \Sigma L'(L\Sigma L')^{-1}L,$$

in which W and Σ are symmetric positive-definite of order n, but otherwise arbitrary matrices. Observe that $A^2 = A$ and $B^2 = B$, so both are projections. Since the trace of a projection is the sum of its eigenvalues, which are zero or one only,

$$\begin{aligned}\mathrm{tr}(A) &= \mathrm{tr}(X(X'WX)^{-1}X'W) = \mathrm{tr}((X'WX)^{-1}X'WX) = \mathrm{tr}(I_p) = p\\ \mathrm{tr}(B) &= \mathrm{tr}(\Sigma L'(L\Sigma L')^{-1}L) = \mathrm{tr}(L\Sigma L'(L\Sigma L')^{-1}) = \mathrm{tr}(I_{n-p}) = n-p.\end{aligned}$$

Also $AX = X$ implies $\mathcal{X} \subset \mathrm{Im}(A)$, and $BX = 0$ implies $\mathcal{X} \subset \ker(B)$. Together with the rank conditions, this implies $\mathrm{Im}(A) = \ker(B) = \mathcal{X}$. However, $\mathrm{Im}(B) \ne \ker(A)$ in general.

In the displayed equation immediately above, we have exploited the property

$$\mathrm{tr}(DEF) = \mathrm{tr}(FDE) = \mathrm{tr}(EFD),$$

i.e., that the matrices can be permuted *cyclically* within the trace.

Now regard $\mathcal{R}^n$ as an inner-product space with inner-product matrix W, so that $\langle u,\, v\rangle = u'Wv$, and let $\Sigma = W^{-1}$. Then

$$\begin{aligned}\langle u,\, Av\rangle &= u'WAv = (Au)'Wv = \langle Au,\, v\rangle\\ \langle u,\, Bv\rangle &= u'W\Sigma L'(L'\Sigma L)^{-1}Lv = \langle Bu,\, v\rangle\end{aligned}$$

so that both projections are self-adjoint. The reciprocal relation $\Sigma = W^{-1}$ implies that A and B are complementary projections, i.e., $\mathrm{Im}(B) = \ker(A)$ or

$$\Sigma = W^{-1} \Longrightarrow B = I - A.$$

In effect, W is the matrix of the inner product in one vector space, and W^{-1} is the matrix of the inner product in the dual space of linear functionals.

4.6. Vector multiplication. The inner product is a scalar-valued function $\mathcal{V}\times\mathcal{V} \to \mathcal{R}$: no concept of vector multiplication as a function $\mathcal{V}\times\mathcal{V} \to \mathcal{V}$ is defined for general vector spaces. However, nearly every variable arising in statistical work is a real-valued function $v\colon S \to \mathcal{R}$ on the sample units, so $v \in \mathcal{R}^S$. If S is a set of plots, u is the plot width and v is the plot length in suitable units, then the component-wise product is a function that assigns to each plot i its area $(uv)_i = u_iv_i$, which is a point $uv \in \mathcal{R}^S$. Functional, or component-wise, multiplication is a commutative binary operation, which is distributive with respect to addition: $(u+v)x = ux + vx$. This gives $\mathcal{R}^S$ additional structure that is sometimes relevant in statistical work; it is a *commutative ring.* Every subset generated by a factorial model formula is a subspace of $\mathcal{R}^S$; certain subspaces such as `~1`, `~A` and `~A:B` (but not `~A+B`) are also sub-rings, closed under vector multiplication.

4.7. Quotient space. Let $\mathcal{U}$ be a subspace of the vector space $\mathcal{W}$. To each vector $x \in \mathcal{W}$ there corresponds a subset $x + \mathcal{U} = \{x + u : u \in \mathcal{U}\}$, called a *coset*. Each coset is an orbit generated by the action of $\mathcal{U}$ by addition on $\mathcal{V}$, as a group acts on a set or a subgroup acts on a group. The symbol $\mathcal{W}/\mathcal{U}$ denotes the set of cosets, so each point in $\mathcal{W}/\mathcal{U}$ is an orbit $x + \mathcal{U}$. The points $x + \mathcal{U}$ and $x' + \mathcal{U}$ are equal as orbits if the difference $x - x'$ belongs to $\mathcal{U}$. The set of cosets is a vector space with addition and linear combinations defined in the obvious way; the zero element in $\mathcal{W}/\mathcal{U}$ is the unique coset $0 + \mathcal{U} = \mathcal{U}$ that contains $0 \in \mathcal{W}$.

Let $\mathcal{W} = \mathcal{U} \oplus \mathcal{V}$, so that $\mathcal{V}$ is complementary to $\mathcal{U}$ in $\mathcal{W}$. Complementary means that to each $x \in \mathcal{W}$ there correspond unique points $u \in \mathcal{U}$ and $v \in \mathcal{V}$ such that $x = u + v$. This implies that to each coset $x + \mathcal{U}$ there corresponds a unique representative element $v \in \mathcal{V}$ such that $(x + \mathcal{U}) \cap \mathcal{V} = \{v\}$. Furthermore, the mapping $\mathcal{W}/\mathcal{U} \to \mathcal{V}$ that sends the coset $x + \mathcal{U}$ to $v \in \mathcal{V}$ by coset intersection is linear and invertible. For a given subspace $\mathcal{U} \subset \mathcal{W}$ there are multiple complementary subspaces, but only one quotient space $\mathcal{W}/\mathcal{U}$. If $\mathcal{W}$ is also an inner-product space, there is a natural correspondence between $\mathcal{W}/\mathcal{U}$ and $\mathcal{V} = \mathcal{U}^\perp$ such that the mapping $\mathcal{W}/\mathcal{U} \to \mathcal{U}^\perp$ is an isometry preserving inner products.

In $\mathcal{R}^4$, the set $\mathcal{V}$ of vectors $(x_1, x_2, 0, 0)$ for real x_1, x_2 is a subspace. Likewise the set $\mathcal{U}$ of vectors $(0, 0, x_3, x_4)$ for real x_3, x_4 is also a subspace. Moreover the two subspaces are complementary. Without specifying an inner product, we cannot say whether the subspaces are orthogonal. For statistical purposes, each point y in the quotient space $\mathcal{R}^4/\mathcal{U}$ can be envisaged as a vector of the form $(y_1, y_2, \star, \star)$, where y_1, y_2 are real numbers, and $\star$ is interpreted as 'any real number', or 'some value not yet recorded'. This interpretation is helpful only if the cosets are aligned with the coordinate axes, as they often are in statistical work. A point $y \in \mathcal{V}$ implies $y_3 = y_4 = 0$, whereas a point $y \in \mathcal{R}^4/\mathcal{U}$ has no such implication. Thus, the algebraic statement that $\mathcal{R}^4/\mathcal{U}$ and $\mathcal{V}$, or $\mathcal{R}^4/\mathcal{U}$ and $\mathcal{U}^\perp$, are in a natural one-to-one correspondence must not be interpreted as a statement about the equivalence of points as observations.

4.8. Tensor product space. (Workman's definition.) Consider a finite set $[m]$ together with the vector space $\mathcal{R}^{[m]}$ of real-valued functions $f : [m] \to \mathcal{R}$. Consider now a second finite set $[n]$ together with the vector space $\mathcal{R}^{[n]}$. Each $f \in \mathcal{R}^{[m]}$ is a function $i \mapsto f(i)$ or a vector $f = (f_1, \ldots, f_m)$. Each $g \in \mathcal{R}^{[n]}$ is a function $r \mapsto g(r)$ or a vector $g = (g_1, \ldots, g_n)$. The tensor product $f \otimes g$ is a function $[m] \times [n] \to \mathcal{R}$ defined by $(f \otimes g)(i, r) = f(i)g(r)$, which is also the Kronecker product or outer product of the two vectors. The tensor product space $\mathcal{R}^{[m]} \otimes \mathcal{R}^{[n]} \cong \mathcal{R}^{[m]\times[n]}$ is the span of all vectors $f \otimes g$ with $f \in \mathcal{R}^{[m]}$ and $g \in \mathcal{R}^{[n]}$.

If $f^{(1)}, \ldots, f^{(m)}$ is a basis in $\mathcal{R}^{[m]}$ and $g^{(1)}, \ldots, g^{(n)}$ is a basis in $\mathcal{R}^{[n]}$, the set of mn vectors $f^{(i)} \otimes g^{(r)}$ is a basis in the tensor product space. The tensor product of two indicator vectors is the indicator vector for the ordered pair in $[m] \times [n]$. Note that every element in this basis is a rank-one matrix, but a linear combination of rank-1 matrices usuallu y has rank greater than one, so not every vector in the tensor product-space is a rank-one matrix.

If the component spaces are also inner-product spaces, the inner product is defined multiplicatively for the basis vectors by

$$\langle f^{(i)} \otimes g^{(r)},\, f^{(j)} \otimes g^{(s)} \rangle = \langle f^{(i)},\, f^{(j)} \rangle \times \langle g^{(r)},\, g^{(s)} \rangle$$

and thence by linearity for arbitrary vectors in $\mathcal{R}^{[m]} \otimes \mathcal{R}^{[n]}$.

In statistical work, it is usual to choose the first basis vector in each component space to be the constant function $f^{(1)} = \mathbf{1}_m$; $g^{(1)} = \mathbf{1}_n$. In that case we may identify the component space $A = \mathcal{R}^{[m]}$ with the subspace $\mathcal{R}^{[m]} \otimes 1_n$, of functions in the tensor product space that are constant on $[n]$. Similarly $B \cong \mathbf{1}_m \otimes \mathcal{R}^{[n]}$.

4.9. The space of arrays. Let $[m]$ be a finite set. The vector space $\mathcal{R}^{[m]}$ of real-valued functions $y\colon [m] \to \mathcal{R}$ inherits a natural basis of indicator functions $e_1, \ldots, e_m$, one indicator function e_r for each element $r \in [m]$. The indicator function is $e_r(i) = 1$ for $i = r$ and zero for $i \neq r$, so that $y = (y_1, \ldots, y_m)$ can be expressed as

$$y = \sum_{r\in[m]} y_r e_r.$$

We say that the value $y(r) \equiv y_r$ is the rth component of y with respect to the indicator basis.

The subset $\mathbf{1}_m \subset \mathcal{R}^{[m]}$ of constant functions $y(1) = \cdots = y(m)$ is a subspace of dimension one; it consists of all scalar multiples of the one-function $1 = (1, \ldots, 1)$. The subset

$$\mathcal{V}_m = \big\{ y \in \mathcal{R}^{[m]} : \sum_{r\in[m]} y(r) = 0 \big\}$$

is also a subspace, closed under addition and scalar multiplication. The intersection consists of constant functions adding to zero, i.e., $\mathbf{1}_m \cap \mathcal{V}_m = \mathbf{0}$ is the zero subspace, so the two subspaces are non-overlapping. Every vector $y \in \mathcal{R}^{[m]}$ can be expressed as the sum

$$y = (y_1, \ldots, y_m) = \bar{y}(1, \ldots, 1) + (y_1 - \bar{y}, \ldots, y_m - \bar{y}) = P_1 y + Q_1 y$$

where $\bar{y} = (y_1 + \cdots + y_m)/m$ is the equally-weighted average, $P_1 y = \bar{y}1$ is a projection onto $\mathbf{1}_m$, and $Q_1 = I - P_1$ is the complementary projection. Evidently, $\mathcal{R}^{[m]} = \mathbf{1}_m \oplus \mathcal{V}_m$, so the subspaces are complementary and $\mathcal{V}_m$ has dimension $m - 1$.

It is sometimes convenient to make $\mathcal{R}^{[m]}$ into an inner-product space with the standard inner product. In that case $\mathcal{V}_m = \mathbf{1}_m^\perp$, so that P_1 and Q_1 are orthogonal projections. The pythagorean sum-of-squares decomposition is

$$\sum_{r\in[m]} y_r^2 = \|y\|^2 = \|P_1 y\|^2 + \|Q_1 y\|^2 = m\bar{y}^2 + \sum_{r\in[m]} (y_r - \bar{y})^2.$$

The subspaces $\mathbf{1}_m, \mathbf{1}_m^\perp$ are closed under coordinate permutation, the standard inner product is invariant, and the pythagorean decomposition is also invariant.

The situation is much the same for a Cartesian-product set $[m] \times [n]$, and an associated array of real numbers $y\colon [m] \times [n] \to \mathcal{R}$, which is a point in the tensor product space $\mathcal{R}^{[m]} \otimes \mathcal{R}^{[n]}$. The standard subspace decomposition consists of four mutually orthogonal subspaces

$$\begin{aligned} \mathcal{R}^{[m]} \otimes \mathcal{R}^{[n]} &= \big(\mathbf{1}_m \oplus \mathbf{1}_m^\perp\big) \otimes \big(\mathbf{1}_n \oplus \mathbf{1}_n^\perp\big) \\ &= (\mathbf{1}_m \otimes \mathbf{1}_n) \oplus (\mathbf{1}_m \otimes \mathbf{1}_n^\perp) \oplus (\mathbf{1}_m^\perp \otimes \mathbf{1}_n) \oplus (\mathbf{1}_m^\perp \otimes \mathbf{1}_n^\perp). \end{aligned}$$

of dimensions 1, $n-1$, $m-1$ and $(m-1)(n-1)$ respectively. The subspace decomposition corresponds to the arithmetic identity

$$y_{ij} = \bar{y}_{\bullet\bullet} + (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}) + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})$$

where $y_{\bullet j}$ is the jth column total, $\bar{y}_{\bullet j} = y_{\bullet j}/m$ is the column average, and so on. For example, a vector v in $\mathbf{1}_m \otimes \mathbf{1}_n^\perp$ is an array of real numbers, constant over rows within each column and adding to zero over columns: the first row of v is an n-component vector in $\mathbf{1}_n^\perp$ whose components add to zero, and each row of v is the same vector. The associated sum of squares is

$$\sum_{ij} y_{ij}^2 = mn\bar{y}_{\bullet\bullet}^2 + m\sum_{j=1}^{n} (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2 + n\sum_{i=1}^{m} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_{ij} (y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})^2.$$

4.10. Group action: symmetric group. The symmetric group $\mathcal{S}_m$ of permutations $\rho\colon [m] \to [m]$ acting on the domain $[m]$ of a function y has an induced action by composition as indicated by the diagram

$$[m] \xrightarrow{\rho} [m] \xrightarrow{y} \mathcal{R}.$$

Composition by ρ changes the function y to $y \circ \rho$, so that the vector $y = (y_1, \ldots, y_m)$ in $\mathcal{R}^{[m]}$ is sent to

$$\rho^* y = (y_{\rho(1)}, \ldots, y_{\rho(m)}),$$

which is a vector in $\mathcal{R}^{[m]}$ having the same set of components as y, but the original order $1 : n$ is switched to $\rho(1), \ldots, \rho(n)$. The induced transformation $y \mapsto \rho^* y$ is a linear transformation $\rho^*\colon \mathcal{R}^{[m]} \to \mathcal{R}^{[m]}$, which preserves group composition, albeit with reversal of compositional order: $(\rho\sigma)^* = \sigma^*\rho^*$. Equivalently, the mapping $\rho \mapsto \rho^*$ is a representation of the symmetric group by linear transformations.

Let $\mathcal{U} \subset \mathcal{R}^{[m]}$ be a subspace. We say that the group *acts on* $\mathcal{U}$ if, for every permutation $\rho \in \mathcal{S}_m$, the induced linear transformation $\rho^*\colon \mathcal{R}^{[m]} \to \mathcal{R}^{[m]}$ satisfies $\rho^*\mathcal{U} \subset \mathcal{U}$, (which implies $\rho^*\mathcal{U} = \mathcal{U}$ since ρ^* is, by definition, invertible.) In other words, $y \in \mathcal{U}$ implies $\rho^* y \in \mathcal{U}$ for every permutation ρ. A subspace that is closed under the action of the group on $\mathcal{R}^{[m]}$ is called a sub-representation.

Since y and $\rho^* y$ are two vectors having the same *set* of components, the sum of the components is invariant, the same for every ρ. It follows that the symmetric group acts on each of the subspaces $\mathbf{1}_m$, $\mathbf{1}_m^\perp$, so each of these subspaces is a sub-representation. These subspaces are in fact irreducible: if $\mathcal{U} \subset \mathcal{R}^{[m]}$ is closed under the group action, then $\mathcal{U}$ is one of the subspaces $\mathbf{0}$, $\mathbf{1}_m$, $\mathbf{1}_m^\perp$ or $\mathcal{R}^{[m]}$.

The situation is much the same for the product group $\mathcal{S}_n \times \mathcal{S}_n$ acting on arrays by independent permutation of rows and columns. The tensor product decomposition

$$\begin{aligned}\mathcal{R}^{[m]} \otimes \mathcal{R}^{[n]} &= \big(\mathbf{1}_m \oplus \mathbf{1}_m^\perp\big) \otimes \big(\mathbf{1}_n \oplus \mathbf{1}_n^\perp\big) \\ &= (\mathbf{1}_m \otimes \mathbf{1}_n) \oplus (\mathbf{1}_m \otimes \mathbf{1}_n^\perp) \oplus (\mathbf{1}_m^\perp \otimes \mathbf{1}_n) \oplus (\mathbf{1}_m^\perp \otimes \mathbf{1}_n^\perp).\end{aligned}$$

generates four irreducible non-isomorphic sub-representations of dimensions 1, $n-1$, $m-1$ and $(m-1)(n-1)$ respectively. The associated decomposition of the total sum of squares is invariant under permutation of rows and under permutation of columns.

4.11. Group action: cyclic group. It is possible to decompose the row or column sum of squares further if the context suggests it or requires it. For example, if each row is a day of the week or a month of the year, it may be more reasonable to ask for a decomposition that is invariant with respect to cyclic permutations than one that is invariant with respect to all permutations. Considering the cyclic group on seven letters, we observe that the vectors

$$\begin{aligned}C &= \big(1, \cos(2\pi/7), \cos(4\pi/7), \cos(6\pi/7), \cos(8\pi/7), \cos(10\pi/7), \cos(12\pi/7)\big) \\ S &= \big(0, \sin(2\pi/7), \sin(4\pi/7), \sin(6\pi/7), \sin(8\pi/7), \sin(10\pi/7), \sin(12\pi/7)\big)\end{aligned}$$

in $\mathcal{R}^7$ have components that add to zero, so they belong to $\mathbf{1}^\perp$. This fact becomes apparent when we examine the components of $Z = C + iS$ as seven points equally spaced on the unit circle in the complex plane. The effect of a one-unit cyclic shift is $Z \mapsto e^{2\pi i/7} Z$, so the real and imaginary parts of $\rho^* Z = e^{2\pi i/7}(C + iS)$ are both linear combinations of C and S. In other words, the vector subspace $H_1 = \operatorname{span}(C, S) \subset \mathbf{1}_7^\perp$ of first-order harmonics is closed under the linear transformation ρ^* acting on $\mathcal{R}^{[7]}$. It is a two-dimensional sub-representation in $\mathcal{R}^{[7]}$ of the cyclic group acting on [7]. It is

in fact an irreducible representation. By the same argument, the real and imaginary parts of $Z^2 = C_2 + iS_2$ define a subspace $H_2 = \text{span}(C_2, S_2)$ of dimension two, which is also closed under cyclic shifts. These harmonics are the Fourier subspaces, which are mutually orthogonal, and satisfy $H_r = H_{7-r}$ with $H_0 = \mathbf{1}_7$. Thus,

$$\mathcal{R}^{[7]} \cong H_0 \oplus H_1 \oplus H_2 \oplus H_3$$

is a decomposition of $\mathcal{R}^{[7]}$ by cyclic irreducibles.

For a function $y\colon [m] \to \mathcal{R}$, the projection into H_r is

$$P_{H_r} y = m^{-1} \langle Z^r,\, y \rangle Z^r$$

[Fix this up later] The *power spectrum* associates with each Fourier frequency r the value $\|P_{H_r} y\|^2$, the squared norm of the projection onto H_r.

4.12. Gaussian space. Let P be a probability distribution on $\mathcal{V}$ and let $\theta\colon \mathcal{V} \to \mathcal{R}$ be a linear functional, i.e., $\theta \in \mathcal{V}'$ is a point in the dual space of linear functionals. The linear functional sends the distribution P to a one-dimensional marginal distribution θP on the real line, all distributions defined on Borel sets. In terms of random variables, X is a random point distributed as P on $\mathcal{V}$, and θX is a random point on the real line distributed as θP, which is defined by $(\theta P)(A) = P(\theta^{-1} A)$. The probability that θX belongs to $A \subset \mathcal{R}$ is the probability that X belongs to the inverse image subset $\theta^{-1} A \subset \mathcal{V}$. For this section, we suppose that $\mathcal{V}'$ is an inner-product space such that θP is zero-mean Gaussian with variance $\|\theta\|^2$. In other words,

$$\int_{x \in \mathcal{V}} e^{\theta x}\, P(dx) = \exp(\|\theta\|^2).$$

For $\mathcal{V} = \mathcal{R}^n$ and $P = N_n(0, \Sigma)$, the marginal distribution θP is $N_1(0, \theta' \Sigma \theta)$, so $\|\theta\|^2 = \theta' \Sigma \theta$ defines the inner product in the dual space. Here Σ is strictly positive definite with inverse matrix W, and $\langle u,\, v \rangle = u' W v$ is the natural inner product in $\mathcal{V}$, which defines the log density of P.

Exercise 4.1: Let D, E, F be three matrices, not necessarily square. Show that, if both products are defined, $\text{tr}(DE) = \text{tr}(ED)$. Show that $\text{tr}(DEF) = \text{tr}(EFD) = \text{tr}(FDE)$ and that $\text{tr}(DFE) = \text{tr}(FED) = \text{tr}(EDF)$, but $\text{tr}(DEF) \neq \text{tr}(EDF)$ in general.

Exercise 4.2: Let v be an arbitrary vector in $\mathcal{R}^n$ as an inner-product space with the standard inner product, Show that $A = vv'/\|v\|^2$ is a rank-one projection, and that the projection is orthogonal.

Exercise 4.3: If P is a linear projection, and ξ is an eigenvector with $P\xi = \lambda\xi$, show that λ is either zero or one. Hence deduce that $\text{tr}(P)$ is an integer equal to the number of non-zero eigenvalues.

Exercise 4.4: Let $v_1, \ldots, v_n$ be a basis consisting of mutually orthogonal vectors in $\mathcal{V}$, and let

$$y = \alpha_1 v_1 + \cdots + \alpha_n v_n$$

be the representation of the vector y as a linear combination of the basis vectors. Show that the coefficients are $\alpha_r = \langle y,\, v_r \rangle / \|v_r\|^2$ and that

$$\|y\|^2 = \sum_{r=1}^{n} \frac{|\langle y,\, v_r \rangle|^2}{\|v_r\|^2} = \sum_{r=1}^{n} |\alpha_r|^2\, \|v_r\|^2.$$

Exercise 4.5: Let $\mathcal{X} \subset \mathcal{V}$ be a subspace, and let v be a vector not in $\mathcal{X}$, and let $\mathcal{X}^+ = \mathcal{X} + v$ be the span of $\mathcal{X}, v$. Let the orthogonal projection of y onto $\mathcal{X}^+$ be $x + \alpha v$ for some $x \in \mathcal{X}$. Show that

$$\alpha = \frac{\langle y,\, Qv \rangle}{\langle v,\, Qv \rangle}$$

and that the reduction in residual sum of squares is

$$\|P^+ y\|^2 - \|Py\|^2 = |\alpha|^2 \|Qv\|^2,$$

where P is the orthogonal projection onto $\mathcal{X}$, $Q = I - P$ is the complementary projection, and P^+ is the orthogonal projection onto $\mathcal{X}^+$.

5. Completely randomized design

We begin with the simplest sort of design with no block structure or spatial structure or covariate structure. Consider a completely randomized design with 12 observational units and four treatment levels, using the randomization scheme described above. How is the scheme implemented in practice, and what analysis follows? First generate a coded list of 12 units 1:12, and an initial systematic design as shown below:

$$T' = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ a & a & a & b & b & b & c & c & c & d & d & d \end{matrix}$$

where the teatment levels are coded a, b, c, d. This means $T'(1) = a$ and so on up to $T'(12) = d$. Then generate a uniform random permutation $\rho\colon [12] \to [12]$, for example using `order(runif(12))` or `rank(runif(12))` and replace the top row with the permuted sequence. This gives something like

$$T = \begin{matrix} 11 & 3 & 6 & 4 & 9 & 8 & 2 & 10 & 5 & 12 & 7 & 1 \\ a & a & a & b & b & b & c & c & c & d & d & d \end{matrix}$$

or, equivalently,

$$T = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ d & c & a & b & c & a & d & b & b & c & a & d \end{matrix}$$

meaning $T(1) = d$, $T(2) = c$, and so on. Equivalently, variety a is planted in plots 3, 6, 11; variety b in plots 4, 8, 9, and so on. In this case, the observational units and the experimental units coincide.

Alternatively, given the systematic design T as a list of levels, one can compute the randomized list `T[order(runif(12))]`, which is equivalent to the composition $T\rho$

$$[n] \xrightarrow{\rho} [n] \xrightarrow{T'} \{a, b, c, d\}$$

with $\rho =$ `order(runif(12))`. The composite function $T = T'\rho$ is the treatment assignment factor.

Note that $T'\rho$ has the same distribution as $T'\rho^{-1}$, but the realizations are different. Both are acceptable for randomization.

In the vector space $\mathcal{R}^{[12]}$ of real-valued functions on the observational units, the *treatment subspace* is spanned by the treatment-level indicator functions as follows. For the randomization outcome shown above, these vectors are

$$\begin{pmatrix} \text{unit} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ T \equiv a & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ T \equiv b & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ T \equiv c & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ T \equiv d & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

In settings such as this, the symbol T does double duty, denoting both the randomization outcome as the list of 12 levels shown above, and the four-dimensional treatment subspace $T \subset \mathcal{R}^{12}$ containing the one-dimensional subspace $\mathbf{1}$:

$$\mathbf{1} \subset T \subset \mathcal{R}^{12}.$$

Not only is T a subspace of $\mathcal{R}^{12}$, but it is also a subspace with a particular basis, the indicator basis for treatment levels. It is a tendency of linear algebra to avoid emphasizing one basis over another, but the indicator basis and the subspace $\mathbf{1}$ are of particular importance in statistical work.

Exercise 5.1: With `rho <- order(runif(12))`, explain what the subsequent operations `order(rho)` and `rank(rho)` produce.

Exercise 5.2: Denote the indicator basis vectors for the subspace T by $\{T_a, T_b, T_c, T_d\}$. Show that $\{\mathbf{1}, T_b, T_c, T_d\}$ is also a basis, where $\mathbf{1}$ is the constant vector with all components equal to one. Each vector $v \in T$ is expressible as a linear combination $v = v_a T_a + v_b T_b + v_c T_c + v_d T_d$ with scalar coefficients v_a, v_b, v_c, v_d. Each vector is also expressible as $v = \beta_1 \mathbf{1} + \beta_2 T_b + \beta_3 T_c + \beta_4 T_d$. Show that $\beta_1 = v_a$. Express the other coefficients β in terms of (v_a, v_b, v_c, v_d).

Exercise 5.3: Suppose that the levels of T are ordered and equally spaced, taking values $1, 2, 3, 4$ on some physical scale (dose). The polynomial basis vectors spanning the treatment subspace are $T_0 = \mathbf{1}$

$$
\begin{aligned}
T_1 &= -3T_a - T_b + T_c + 3T_d \\
T_2 &= T_a - T_b - T_c + T_d \\
T_3 &= -T_a + 3T_b - 3T_c + T_a
\end{aligned}
$$

With $x = T_a + 2T_b + 3T_c + 4T_d$ equal to the dose level, show that $T_1 = 2x - 5$; Express T_2 and T_3 as polynomials in x.

6. Pythagorean arithmetic for a CRD

Suppose that the response vector is

$$y = (14.1, 9.1, 14.5, 15.7, 10.6, 6.7, 12.4, 22.0, 12.9, 12.2, 11.6, 2.2)$$

in plot order. The orthogonal projection of y onto the subspace $\mathbf{1}$ is

$$P_{\mathbf{1}} y = \langle y, \mathbf{1}\rangle \mathbf{1} / \langle \mathbf{1}, \mathbf{1}\rangle$$

which is simply $\bar{y}\mathbf{1}$, i.e., the sample mean $\bar{y} = 12.0$ repeated $n = 12$ times. Here, we use the standard Euclidean inner product whose justification comes from considerations of exchangeability, or invariance (of the response distribution) with respect to permutation of units. Then $Q_{\mathbf{1}} y = y - \bar{y}\mathbf{1}$ has components $y_i - \bar{y}$, so that

$$\|P_{\mathbf{1}} y\|^2 = n\bar{y}^2 = 1728.0; \quad \|Q_{\mathbf{1}} y\|^2 = \sum (y_i - \bar{y})^2 = 260.02$$

is the Pythagorean decomposition of the total sum of squares $\sum y_i^2 = \sum (y_i - \bar{y})^2 + n\bar{y}^2$ associated with the subspace $\mathbf{1} \subset \mathcal{R}^n$.

The orthogonal projections of y onto the subspaces T and $T^\perp$ are

$$P_T y = (9.57, 10.63, 10.93, 16.87, 10.63, 10.93, 9.57, 16.87, 16.87, 10.63, 10.93, 9.57)$$
$$Q_T y = (4.53, -1.53, 3.57, -1.17, -0.03, -4.23, 2.83, 5.13, -3.97, 1.57, 0.67, -7.37)$$

Thus $P_T y$ assigns to each plot the relevant treatment average, and $Q_T y = y - P_T y$ is the residual, the orthogonal projection with kernel T. Putting all of this together, the sequence $\mathbf{1} \subset T \subset \mathcal{R}^n$ gives rise to a decomposition of the total sum of squares into three parts as follows:

subspace	dim	SS	SS	MS
$\mathbf{1}$	1	$\|P_{\mathbf{1}} y\|^2 = n\bar{y}^2$	1728.0	1728.0
$T/\mathbf{1}$	3	$\|P_T y\|^2 - \|P_{\mathbf{1}} y\|^2$	97.833	32.611
$\mathcal{R}^n/T$	8	$\|y\|^2 - \|P_T y\|^2$	162.187	20.273

The quotient symbol $T/\mathbf{1}$ refers to the subspace $\mathbf{1}^\perp \cap T$ of dimension 3, and $\mathcal{R}^n/T$ to the subspace $T^\perp$ of dimension 8. Each sum of squares (SS) is associated with a subspace whose dimension is called the *degrees of freedom*; for example $T/\mathbf{1}$ has dimension $4-1 = 3$. The mean square (MS) is the sum of squares divided by the degrees of freedom. This is called an *analysis of variance table*, or an ANOVA table. Each row in an ANOVA table is indexed by a *pair* of subspaces U/V with $V \subset U \subset \mathcal{R}^n$, for which the associated sum of squares is $\|P_U y\|^2 - \|P_V y\|^2$. The statistical interpretation of the various terms depends will be discussed below.

7. Randomized blocks design I

Suppose that the observational material consists of $n = 24$ units arranged in three blocks of eight units each. In practice, the blocks are constructed or defined in such a way that the response values for two units in the same block are expected to be more similar than the responses for two units in different blocks. For example, nearby plants tend to have more similar yields than more distant plants, so the blocks in a field experiment tend to consist of square or nearly square arrangements of plots. In a veterinary experiment, each block could be a herd or a farm, or a housing unit, or perhaps a breed.

Technically, a block factor is a partition with unlabelled blocks. For this section, however, the blocks are labelled I–III as in a classification factor, and the units are numbered 1–24 in three consecutive blocks of eight. For a balanced design with two treatment replicates in each block we begin with a systematic design of the type desired, i.e.,

unit	1	2	3	4	5	6	7	8	9	10	$\cdots$	23	24
T	a	b	c	d	a	b	c	d	a	b	$\cdots$	c	d

Randomization consists of a permutation $\rho\colon [n] \to [n]$, which is usually chosen uniformly at random from the set of permutations that preserve the block structure, and thus the balance of the systematic design. In symbols, this means the set of permutations such that $B(\rho_i, \rho_j) = B(i, j)$, where B is the block factor. These permutations form a subgroup of size $(8!)^3 \times 3!$: permute elements independently within each block, and then permute the block labels. After that is done, the second row is replaced by the permuted list $T\rho$ with components $T(\rho_1), \ldots, T(\rho_{24})$ giving something like

$$T\rho = (b, a, b, c, d, c, d, a;\ c, d, d, b, a, b, a, c;\ a, d, a, c, b, d, b, c)$$

each block containing two replicates of each treatment.

The treatment subspace $T \subset \mathcal{R}^n$ is spanned by the four indicator functions

$$\begin{aligned}
T_a &= (0,1,0,0,0,0,0,1,0,0,0,0,1,0,1,0,1,0,1,0,0,0,0,0)\\
T_b &= (1,0,1,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,1,0,1,0)\\
T_c &= (0,0,0,1,0,1,0,0,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,1)\\
T_d &= (0,0,0,0,1,0,1,0,0,1,1,0,0,0,0,0,0,1,0,0,0,1,0,0)
\end{aligned}.$$

and the block subspace is spanned by three indicator functions

$$\begin{aligned}
B_1 &= (1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)\\
B_2 &= (0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0)\\
B_3 &= (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1)
\end{aligned}$$

The identity $T_a + T_b + T_c + T_d = B_1 + B_2 + B_3 = \mathbf{1}$ is a consequence of the fact that each unit has one level, and only one level, of each factor.

The ordered pair (B, T) is a variable or factor $(B, T)(i) = (B(i), T(i))$ taking values in the product set $\{I, II, III\} \times \{a, b, c, d\}$. For this design, each of the 12 possible values $(I, a), (II, a), \ldots, (III, d)$ occurs twice, so there are 12 non-zero indicator functions spanning the subspace $BT \subset \mathcal{R}^n$. The indicator functions are in fact the functional products $B_1T_a, B_1T_b, \ldots, B_3T_d$ of the indicator functions for B and T. Although (B, T) is not the same variable as (T, B), they generate the same subspace $BT = TB \subset \mathcal{R}^n$ of dimension 12. For this design, $BT \cong B \otimes T$ is the *tensor product space* of dimension $\dim(B) \times \dim(T)$. More generally, some of the ordered pairs of levels may not occur in the design, in which case BT is the *restriction of the tensor product space* to the sample units. Then $\dim(BT) \le \dim(B) \times \dim(T)$ is the number of ordered pairs that happen to occur in the design. Regardless of the design, B and T are both subspaces of $BT \equiv TB$.

In statistical work, the subspace is denoted by $B.T$, $B*T$ or $B : T$. Although all three symbols denote the same subspace, the implied basis vectors may be different.

In addition, $B + T = \text{span}(B, T)$ is a subspace of dimension

$$\dim(B + T) = \dim(B) + \dim(T) - \dim(B \cap T) = 3 + 4 - 1$$

since $B \cap T = \mathbf{1}$.

Putting all of this together, the simplest randomized-blocks design gives rise to five non-trivial subspaces

$$\mathbf{1} \subset \{B, T\} \subset B + T \subset BT \subset \mathcal{R}^n.$$

To each of the irreducible pairs

$$\mathcal{R}^n/BT, \quad BT/(B + T), \quad (B + T)/B, \quad (B + T)/T, \quad B/\mathbf{1}, \quad T/\mathbf{1}, \quad \mathbf{1}/0.$$

there corresponds a sum of squares in a complete ANOVA table. The first three are

Source	*Space*	*Sum of squares*
Replicates	$\mathcal{R}^n/BT$	$\|y\|^2 - \|P_{BT}y\|^2$
Interaction	$BT/(B + T)$	$\|P_{BT}y\|^2 - \|P_{B+T}y\|^2$
Treatments	$(B + T)/B$	$\|P_{B+T}y\|^2 - \|P_B y\|^2$

The treatment sum of squares is sometimes called the treatment sum of squares eliminating additive block effects, or adjusted for additive block effects.

The other sums of squares are usually less important for statistical analyses, and might not be listed in an ANOVA table. In the order listed above, they are *the block sum of squares eliminating additive treatment effects*, *the block sum of squares ignoring treatment effects*, *the treatment sum of squares ignoring block effects*, and the sum of squares for the mean, which is $n\bar{y}^2$.

8. Geometric orthogonality

The design described above is balanced, with equal block sizes, and each treatment occurring an equal number of times in each block. In such a setting, the subspaces B and T are *geometrically orthogonal*, which means that the non-overlapping subspaces

$$(B \cap T)^\perp \cap B, \quad \text{and} \quad (B \cap T)^\perp \cap T$$

are orthogonal in the ordinary sense. In this case, $B \cap T = \mathbf{1}$, so geometric orthogonality means that

$$B/\mathbf{1} \cong \mathbf{1}^\perp \cap B, \quad \text{and} \quad T/\mathbf{1} \cong \mathbf{1}^\perp \cap T$$

are in fact orthogonal, so that

$$B + T = \mathbf{1} + (\mathbf{1}^\perp \cap B) + (\mathbf{1}^\perp \cap T)$$

is a decomposition of $B + T$ into orthogonal subspaces. In that special case, the sum of squares for treatment ignoring block effects ($T/\mathbf{1}$) coincides with the sum of squares for treatments eliminating additive block effects, ($(B + T)/B$). In the general unbalanced case, the latter is usually of more interest than the former.

The subspace $\mathbf{1}^\perp \cap B$ is spanned by the vectors $B_1 - B_2$, $B_1 - B_3$, $B_2 - B_3$ where B_r is the indicator vector for B at level r. The subspace has dimension two. Likewise $\mathbf{1}^\perp \cap T$ is spanned by the vectors $T_r - T_s$, and there are three linearly independent vectors of this type. It is straightforward to check that $T_a - T_b$ is orthogonal to each of the vectors $B_r - B_s$ if and only if each treatment occurs the same number of times in each block.

9. Factors, variables and model formulae

For the discussion that follows, S is a sample of $n = 24$ units labelled $u_1, \ldots, u_n$, or simply `1:n`, A is a four-level treatment factor, C is a three-level classification factor, and x is a quantitative covariate. This terminology means that C and x are recorded pre-baseline, and A is to be generated at baseline by randomization. Suppose that the three classes are labelled c_1, c_2, c_3 and that the units are listed by classes

```
units <- 1:n; C <- gl(3,8,n); levels(C) <- paste("c", 1:3, sep="")
```

For an initial non-randomized treatment assignment, it is natural to set

```
A0 <- gl(4, 1, n)    or    A0 <- gl(4, 2, n)
```

so that each C-class contains two replicates of each treatment. The non-randomized design including x is

```
D0 <- rbind(A0, C, x); colnames(D0) <- units; D0
```

Randomization means random assignment of units to treatments. Let $\sigma\colon [24] \to [24]$ be a permutation of the units. The permuted design

```
D <- rbind(A0[sigma], C, x); colnames(D) <- units; D
```

is guaranteed to have all four treatment levels each occurring six times (as in `A0`). But it is not guaranteed to have the more desirable balance property that two replicates of each treatment occur in each C-class. The initial treatment assignment `A0` is balanced with respect to C. It is natural to ask for a permutation that preserves each class: $C_u = C_{\sigma(u)}$, so that the composition $A(u) = A0(\sigma(u))$, i.e., `A <- A0[sigma]`, is also balanced. The subset of C-preserving permutations is a sub-group $\mathcal{S}_8^3 \subset \mathcal{S}_{24}$ of size $(8!)^3$.

The random permutation

```
sigma <- c(order(runif(8)), order(runif(8))+8, order(runif(8))+16)
```

is uniformly distributed on the C-preserving sub-group. It does not preserve x: in general $x_u \neq x_{\sigma(u)}$. The randomized design consists of the units $1\colon n$, with C and x as given, and treatment assignment `A=A0[sigma]` by composition with the random permutation.

In a statistical computer package such as R, each of the symbols A, C, x is a list, or a spreadsheet column, with one row or component for each sample unit. Each list also has attributes such as factor, factor levels, and so on, which govern how it is to be used and interpreted in specific contexts. A *model formula* is an expression beginning with `~...`, which identifies a subspace of $\mathcal{R}^n$, sometimes called the model subspace or, if the model is linear, the mean-value subspace. Every term in a model formula such as `A` or `x`, and every compound term such as `A:B` or `A*B` or `A:x` is a subspace. An expression or sub-expression such as `~A+x` or `~x1+x2` is not a vector sum, but the span of two subspaces, so the meaning of the operators '`+`' and '`*`' and '`:`' changes according to the context.

To understand some of the differences (as implemented in R), execute the following code and observe the outputs.

```
A <- 3; B <- 7; A:B; A*B
n <- 6; A <- gl(2,1,n); B <- gl(2,2,n); A:B; A*B
cbind(model.matrix(~A:B), model.matrix(~A*B))
cbind(model.matrix(~A:B-1), model.matrix(~A*B-1))
y <- rnorm(n); lm(y~A:B-1)$coef; lm(y~A*B)$coef
```

Not only is every term in a model formula a subspace, but it is a subspace with a particular basis. Two terms such as `~A:B` and `~B:A` represent the same subspace with the same basis vectors, but the basis vectors are not listed in the same order. The same holds for expressions such as `~A+B` and `~B+A`, or `~A*B` and `~B*A`. The order in which the

basis vectors appear is immaterial for many purposes such as linear projection, but it may have an effect on certain other operations.

Two terms such as `~A:B` and `~A*B`, which represent the *same subspace*, may have different implied bases. The same is true for expressions such as `~A` and `~A-1` (same subspace, different bases), but `~x` and `~x-1` are different as subspaces.

An expression such as `A*B` that has a meaning in one context may be meaningless in another.

Exercise 9.1: Run the code

```
n <- 24; set.seed(57721)
units <- 1:n; C <- gl(3, 8, n); levels(C) <- paste("c", 1:3, sep="");
A0 <- gl(4, 2, 24); A <- A0[1:n]; x <- round(runif(24), 2);
D <- rbind(A, C, x); colnames(D) <- units; D;
sigma <- c(order(runif(8)), order(runif(8))+8, order(runif(8))+16); sigma;
A <- A0[sigma]; D <- rbind(A, C, x); colnames(D) <- units; D;
```

In the initial treatment assignment A_0, units 21 and 22 are assigned the same treatment. What is the probability that they are assigned the same treatment after randomization (for a seed other than 57721)? What is the probability that units 1 and 15 are assigned the same treatment?

Exercise 9.2: Let A and B be two factors having three levels each. The expressions

```
lm(y~A:B-1)$coef;    lm(y~B*A)$coef
```

give the coefficient vectors for the implied basis vectors. The subspaces are the same. Assuming that the design is complete, express the second coefficient vector linearly in terms of the first.

Exercise 9.3: Two possibilities were suggested above for the initial non-randomized design:

`A0 <- gl(4, 1, n)` and `A1 <- gl(4, 2, n).`

The choice matters in the sense that for generic σ, the permuted assignments $A_0 \circ \sigma$ and $A_1 \circ \sigma$ are different. Does the initial choice matter statistically? Is the orbit of A_0 the same as the orbit of A_1 under the action of $\mathcal{S}_8^3$? If the orbits are equal, is the distribution of $A_0 \circ \sigma$ the same as the distribution of $A_1 \circ \sigma$? Explain.

Exercise 9.4: Suppose that the group $\mathcal{S}_{24}$ of all permutations of $[n]$ is used for randomization purposes starting off from A_0 as given above. What is the probability that the treatment assignment $A = A_0 \circ \sigma$ is in fact balanced with two replicates of each treatment in each C-class?

10. Randomized blocks design II

For the second RBD, we suppose that there are 24 units in three blocks of eight, but only three treatment levels. For the initial systematic design, we choose

unit	1	2	3	4	5	6	7	8	9	10	⋯	23	24
T	a	b	c	a	b	c	a	b	c	a	⋯	b	c

In block I, levels a, b occur three times, and c occurs twice; in block II, levels a, c occur three times, and b occurs twice; in block III, levels b, c occur three times, and a occurs twice. After randomization, the treatment assignment is $T\rho$ and the response is y.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$T\rho$	b	c	a	c	c	b	a	a	a	c	b	a	b	c	c	b	b	b	a	a	c	a	b	c
y	16.3	7.8	11.5	10.1	15.6	16.2	16.0	9.0	6.7	0.5	7.4	4.6	9.7	10.5	8.4	10.2	6.9	11.5	0.5	4.5	13.9	3.4	12.4	9.7

The orthogonal projection onto any subspace $\mathcal{X} = \text{span}(X) \subset \mathcal{R}^n$ is a linear transformation, the matrix version of which is $P_X = X(X'X)^{-1}X'$, where X is any matrix whose columns form a basis for $\mathcal{X}$. The squared length of the projection of y onto $\mathcal{X}$ is $\|P_X y\|^2 = y'P_X y$, and $\dim(\mathcal{X}) = \text{tr}(P_X)$ is the dimension of $\mathcal{X}$.

Associated with each consecutive pair in the nested sequence of subspaces

$$1 \subset B \subset B + T \subset BT \subset \mathcal{R}^n$$

there is a sum of squares, which is irreducible in a certain sense. For example the projection $P_{BT}y$ is a list of nine averages, one average for each BT combination, each repeated as many times as the combination occurs in the design. Thus $\|P_{BT}y\|^2 = \sum n_r \bar{y}_r^2 = 2402.23$ is less than $\|y\|^2 = 2556.57$, and the difference 154.34 is the sum of squares for replicates on 15 degrees of freedom. Likewise, $\|P_{B+T}y\|^2 = 2328.23$ is the squared length onto the additive subspace, so $2402.23 - 2328.23 = 74.00$ is the interaction sum of squares, or non-additivity sum of squares, on four degrees of freedom, which is $\dim(BT) - \dim(B + T)$.

Exercise 10.1: The R code shown below generates a random permutation $[24] \to [24]$ preserving block structure in successive blocks of eight:

```
set.seed(57721)
b <- sample(c(0,8,16), 3)
rho <- c(b[1]+sample(1:8, 8), b[2]+sample(1:8, 8), b[3]+sample(1:8, 8))
T <- rep(c("a","b","c","d"), 6)
T[rho]
```

With the seed as shown, b happens to be $(8, 16, 0)$. What is the effect of b versus $b = (0, 8, 16)$ on the outcome ρ? What is the effect of b versus $b = (0, 8, 16)$ on the outcome $T\rho$? For a randomly selected seed, what is the effect of b versus $b = (0, 8, 16)$ on the distribution of $T\rho$?

Exercise 10.2: The R code shown below generates a random permutation $[24] \to [24]$ preserving block structure in successive blocks of eight:

```
set.seed(57721)
trt <- rep(c("a","b","c"), 8)
b <- sample(c(0,8,16), 3)
rperm <- c(b[1]+sample(1:8, 8), b[2]+sample(1:8, 8), b[3]+sample(1:8, 8))
X <- model.matrix( blk+trt - 1)
mu <- X %*% c(11, 10, 9, 2, -1)
y <- as.vector(mu) + rnorm(n)*5;  y <- round(y, 1)
blk <- gl(3, 8, 24);  T <- as.factor(trt[rperm])
anova(lm(y ~ blk*T))
```

Check that this code generates the unbalanced randomized blocks design described above.

Exercise 10.3: Compute the projection matrices P_{BT} and P_{B+T} directly, and compute their traces. Check that the difference $D = P_{BT} - P_{B+T}$ is a projection matrix by computing $D^2 - D$. Use the two projection matrices to compute the squared norms $\|P_{BT}y\|^2$ and $\|P_{B+T}y\|^2$, and hence the interaction, or non-additivity, sum of squares.

Exercise 10.4: Compute the sums of squares associated with $(B+T)/B$ and $T/1$, verifying that these are different.

Exercise 10.5: Explain the difference between `anova(lm(y~T*blk))` and the corresponding `anova()` statement shown above.

Exercise 10.6:

The data for the surgical experiment on rats is given below, one row for each rat, one column for each site in order from the shoulder to near the tail. The rats have been relabelled so that 1–8 are the controls; 9–16 is the first treatment group, and 17–24 the second.

```
 9.1   9.7  10.1   8.4   7.1
10.5   9.8   9.7   9.8   8.8
 8.0   9.3   8.1   6.4   6.3
 9.4   7.9   7.4   7.0   8.8
10.4   9.5  10.4  10.9  12.4
 8.9   9.1   7.6   9.2   8.4
 9.4   9.8  11.2   8.3  10.3
 8.6  10.0   9.6  10.2  10.7
10.5   8.2   9.3   8.1   8.2
 6.7   9.4   9.9   8.6   6.8
10.1   9.7   9.0   9.3   7.8
11.1   9.4   9.5  10.5   9.0
 7.9   8.3   9.8  10.3   7.9
 9.8   9.3   9.1  10.6   8.5
12.7  10.7  10.1  11.2  10.6
11.6  10.4   9.6  10.7   9.7
11.7  10.3  10.7   8.6  10.2
 9.2   9.6  10.5  10.4  10.6
 9.5   8.2  10.1   9.2   8.8
 8.6   7.9   8.2  10.1   8.3
10.7  10.6   9.1  10.3   8.5
 8.8  11.0   9.6   8.3   9.0
11.5   9.2  10.9  10.2  10.8
10.3  10.4  11.2  10.2   9.6
```

Explain why *trt* as a vector subspace is a subspace of *rat*. What are the dimensions?

For each individual site, compute the treatment effect, and the sums of squares associated with the vector space sequence $1 \subset trt \subset rat$.

For the vector space sequence

$$\mathbf{1} \subset trt \subset rat \subset site + rat \subset rat + site * trt \subset \mathcal{R}^{120},$$

compute the five 'irreducible' sums of squares.

Exercise 10.7: The main effect of treatment is the mean response for treated rats minus the mean response for untreated rats. Compute the treatment effect estimate for these data. How would you compute a standard error for this statistic? Report the numerical value.

Exercise 10.8: Let $\bar{y}_{i\bullet}$ be the average for rat i, and $\bar{y}_{\bullet s}$ the average for site s. Verify that the SS for $rats/\mathbf{1}$ is $5\sum_i \bar{y}_{i\bullet}^2 - 120\bar{y}_{\bullet\bullet}^2$ and the SS for $sites/\mathbf{1}$ is $24\sum_s \bar{y}_{\bullet s}^2 - 120\bar{y}_{\bullet\bullet}^2$. Obtain a similar expression for the SS for $rats/trt$.

Exercise 10.9: If the available rats consist of 15 males and 9 females, what adjustments would you make to the design? Explain. With these adjustments, how many treatment asignments are possible?

11. Complete factorial designs

A complete factorial design with k factors $A, B, \ldots$ is one in which each combination of the levels of each factor occurs in the design. The design is said to be balanced if each combination occurs an equal number of times: in the discussion that follows, we assume that each combination occurs once. Most of the decompositions discussed in this section rely on orthogonality of certain subspaces with respect to the standard inner product, and orthogonality relies on balance in the design. Both the arithmetic of the linear algebra and the distribution of the various coefficients are more complicated for unbalanced designs.

Each factor may be either a treatment or a classification factor with ordered or unordered levels. In the decomposition that follows, a block factor is handled as if it were a classification factor with labelled levels.

Consider first a design with two factors A, B having two levels each, labelled zero and one. Each factor combination is an ordered pair (A-level, B-level) of levels, shown below in spreadsheet format with one row for each unit.

unit	$A(u)$	$B(u)$		y
u_1	0	0	(1)	3.4
u_2	1	0	a	5.6
u_3	0	1	b	2.9
u_4	1	1	ab	4.9

The units are listed in standard AB treatment order, which means that the sequence of ordered pairs (B-level, A-level) is the sequence of numbers 0–3 in binary representation. The fourth column shows an alternative representation in a more algebraic style, $a^{A-level}b^{B-level}$, and the last column shows the numerical value used later for illustration.

11.1. Additive effects and interactions. The 'effect of A' is the difference between the mean response with A at the high level and the mean response with A at the low level. Where necessary, we distinguish between the mean difference in the sample, which is the numerical value of a random variable, and the mean difference in the population, which is a model parameter. In either case, the effect of A may depend on the level at which B is set, so the two sample effects are

$$\begin{aligned} \{u\colon B(u) = 0\}: &\quad a - 1 = 5.6 - 3.4 = 2.2 \\ \{u\colon B(u) = 1\}: &\quad ab - b = 4.9 - 2.9 = 2.0 \end{aligned}$$

The sample average effect of A, averaged over levels of B, is $(ab-b+a-1)/2 = 2.1$. The expression $ab-b+a-1$ denotes a linear combination of sample values, which is a linear functional $\mathcal{R}^4 \to \mathcal{R}$ from the sample space into the real numbers. A *linear contrast* is a linear functional whose components add to zero, so the average effect of A is one half the linear contrast $(a-1)(b+1)$.

If the effect of A at the high level of B is the same as the effect of A at the low level of B, then the linear contrast $a-1$ is equal to the contrast $b(a-1)$, i.e., the difference $(a-1)b-(a-1) = (a-1)(b-1)$ is zero. The interaction contrast, usually $(a-1)(b-1)/2$, is a linear functional $\mathcal{R}^4 \to \mathcal{R}$ applied either of the sample means or of the population means. When we say that the interaction is null or zero, we mean that the difference of population means is zero, or that the sample statistic as a random variable has zero mean. The sample contrast is then expected to be small in absolute value, but not zero.

The effect of B is defined in the same way, one contrast $b-1$ for A at the low level, and another $a(b-1)$ for A at the high level. The average effect of B is the contrast

$(a+1)(b-1)/2$. The interaction of B with A is the contrast $(a-1)(b-1)/2$, which is the same as the interaction of A with B.

For three factors A, B, C, each having two levels, there are eight factor combinations, either $\{0,1\}^3$ in standard Cartesian product notation or $a^{\{0,1\}}b^{\{0,1\}}c^{\{0,1\}}$ in the alternative algebraic notation introduced above. The sum of the eight components is the linear combination $(a+1)(b+1)(c+1)$ in which each factor combination has unit coefficient. The average or main effect of A is the average response for $A=1$ minus the average for $A=0$, which is the contrast $(a-1)(b+1)(c+1)/4$. Likewise for the main effects of B and C. On the subset for which $C=1$, the interaction of A and B is $(a-1)(b-1)c/2$; on the complementary subset, the interaction is $(a-1)(b-1)/2$, so the average AB interaction is $(a-1)(b-1)(c+1)/4$. The three-factor ABC interaction is $(a-1)(b-1)(c-1)/4$.

11.2. Yates's 2^k algorithm. Yates's algorithm for the computation of factorial contrasts is a linear transformation from one basis to another—from the unit-indicator basis $\{e_u\}$ to the basis of factorial contrasts, which is $\{(a\pm1)(b\pm1)\cdots\}$ for a 2^k design. The algorithm exploits the tensor product structure of the space and the factorial basis. We illustrate the calculations for the 2^2 design shown above, first in algebraic notation, then with numerical values substituted.

Level	A-step	B-step	divisor	SS
(1)	$a+1$	$(a+1)(b+1)$	2^k	$\lvert(a+1)(b+1)\rvert^2/2^k$
a	$(a+1)b$	$(a-1)(b+1)$	2^k	$\lvert(a-1)(b+1)\rvert^2/2^k$
b	$a-1$	$(a+1)(b-1)$	2^k	$\lvert(a+1)(b-1)\rvert^2/2^k$
ab	$(a-1)b$	$(a-1)(b-1)$	2^k	$\lvert(a-1)(b-1)\rvert^2/2^k$

Level	y	A-step	B-step	divisor	SS
(1)	3.4	9.0	16.8	4	70.56
a	5.6	7.8	4.2	4	4.41
b	2.9	2.2	−1.2	4	0.36
ab	4.9	2.0	−0.2	4	0.01

The first table shows algebraically how the linear functionals are built up stage by stage, and the second table shows the value of each functional for the vector y. The basis vectors $(a\pm1)(b\pm1)\cdots$ are mutually orthogonal, each with squared norm 2^k. In the expression for y as a linear combination of these vectors, the coefficients are

$$\langle(a\pm1)(b\pm1)\cdots,\, y\rangle/2^k$$

which is the value of the linear functional obtained at the kth Yates step divided by 2^k. Note that $\sum y_u = 16.8$ is the total, and

$$\sum_u (y_u-\bar y)^2 = 4.78 = 4.41+0.36+0.01$$

is the sum of the standardized squared factorial contrasts in the final column omitting the mean.

11.3. Yates's 3^k algorithm. For a 3^2 design with two factors having three levels each, it is conventional to label the levels $0,1,2$ and to denote the nine combinations by $a^{\{0,1,2\}}b^{\{0,1,2\}}$, i.e.,

$$(1),\ a,\ a^2,\ b,\ ab,\ a^2b,\ b^2,\ ab^2,\ a^2b^2$$

in AB standard order. We may choose any orthogonal bases for the component spaces A and B. If the levels are ordered, it is natural to choose the orthogonal polynomial basis

$$(1,1,1),\ (-1,0,1),\ (1,-2,1)$$

denoted by $1, A_L, A_Q$, whose squared norms are $3, 2, 6$ respectively. With a similar basis in B, the factorial tensor product basis in $A \otimes B \cong \mathcal{R}^9$ consists of the nine orthogonal vectors $\{1, A_L, A_Q\} \otimes \{1, B_L, B_Q\}$. The norm of $A_Q \otimes B_L$ is the product of the norms of the components $\|A_Q \otimes B_L\| = \|A_Q\| \times \|B_L\|$.

Yates's algorithm begins by arranging the observations in standard order with A-levels varying fastest. It proceeds by computing the dot product in blocks of three with each basis vector for A, three dot products for 1, followed by three for $(-1,0,1)$ and three for $(1,-2,1)$. If the same basis vectors are used for B, the second step is identical to the first.

Level	A-step	B-step	divisor	Factorial contrast
(1)	$1+a+a^2$	$(1+a+a^2)(1+b+b^2)$	9	I
a	$(1+a+a^2)b$	$(a^2-1)(1+b+b^2)$	6	A_L
a^2	$(1+a+a^2)b^2$	$(a-1)^2(1+b+b^2)$	18	A_Q
b	(a^2-1)	$(1+a+a^2)(b^2-1)$	6	B_L
ab	$(a^2-1)b$	$(a^2-1)(b^2-1)$	4	A_LB_L
a^2b	$(a^2-1)b^2$	$(a-1)^2(b^2-1)$	12	A_QB_L
b	$(a-1)^2$	$(1+a+a^2)(b-1)^2$	18	B_Q
a^2b	$(a-1)^2b^2$	$(a^2-1)(b-1)^2$	12	A_LB_Q
a^2b^2	$(a-1)^2b^2$	$(a-1)^2(b-1)^2$	36	A_QB_Q

Note that A_QB_L is the tensor product or Kronecker product

$$A_Q \otimes B_L = \begin{pmatrix} -1 & 0 & 1 \\ 2 & 0 & -2 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 0 & 1 \end{pmatrix},$$

in essence a 3×3 array read as a vector in AB-order, $(-1,2,-1,0,0,0,1,-2,1)$, i.e., the linear functional $(a-1)^2(b^2-1)$. In the expression for y as a linear combination of the tensor product basis vectors, the coefficient of A_QB_L is

$$\langle (a-1)^2(b^2-1),\ y \rangle / 12,$$

the ratio of the linear functional divided by the squared norm of the basis vector. The divisor is the sum of squared coefficients in the linear functional.

11.4. Sparsity and half-normal plots. It is a common presumption in factorial experiments that all but a few of the coefficients in the factorial basis are negligible. Here, we mean that most of the factorial effect *parameters* are negligible; the sample values are subject to stochastic variation and may not be negligible. The goal is to identify the small subset of non-negligible effects in the large set of effects estimated with noise.

Usually a few of the factorial effects are either of little interest or are known not to be zero—or at least they cannot be presumed to be zero. These include the overall mean, the main effects of block factors which are of little interest, and perhaps the main effects of classification factors. Sometimes a few of the main effects of treatment factors are

so clearly non-zero that it is unnecessary to include them in the ensemble of potentially non-zero effects for more detailed examination.

We assume that these identifiable and uninteresting contrasts have been removed. For the remaining m contrasts, we compute the standardized value, which is

$$X_v = \frac{\text{raw contrast}_v}{\sqrt{d_v}} = \frac{\langle v,\, y\rangle}{\|v\|}$$

where v is the contrast vector and $d_v = \|v\|^2$ is the divisor. There is one X-value for each contrast apart from the few eliminated in the preceding paragraph. On the assumption that the responses Y_u for each unit are independent with the same variance σ^2, the raw contrasts are independent with variance $\sigma^2\|v\|^2$, so each X_v has the same variance σ^2. The goal is to identify the contrasts whose mean is non-zero, and to do so in a way that limits the rate of false positives. In the absence of replication, the difficulty is that no independent estimate of σ^2 is available.

Daniel's half-normal plot is an informal graphical procedure for judging whether or not the largest observed effects can reasonably be attributed to chance or whether chance alone is implausible as an explanation. First order the absolute standardized factorial contrasts in increasing order

$$|X_{(1)}| \le |X_{(2)}| \le \cdots \le |X_{(m)}|,$$

and plot these against the expected half-normal order statistic as a scatter plot with points

$$\left(\Phi^{-1}\Big(\frac{m+i}{2m+1}\Big),\ |X_{(i)}|\right).$$

Both vectors are monotone increasing in i. If all but a few of the effects are null, we expect to see roughly a straight line through the origin with slope σ, with possibly a few of the largest values deviating from the straight line.

In practice, one would be very reluctant to identify an interaction as significantly non-zero if the marginal effects were not similarly identified. Interaction between treatment and a classification factor means that the effect of treatment among one class of units is not the same as the effect in another class, so there cannot be an interaction without a treatment effect in one or other class. The average effect of treatment, averaged over the two classes, could be zero. For example, treatment that is beneficial for males might have a detrimental effect on women. In cases of this sort, it is better to split the units into the various classes, and to report qualitatively different conclusions separately for the two classes. Other sorts of interactions do arise, but in general, a strong interaction with small main effects is a red flag—a phenomenon that calls out for closer investigation.

11.5. Analysis of covariance. In this section, z is a quantitative covariate recorded pre-baseline, and y is the response in a complete factorial design. The factors $A, B, C, \ldots$ may be treatment or classification or block factors. If they are of the latter type, they are treated as classification factors and used in computations as additive adjustments for the mean vector.

To accommodate the covariate effect, it is necessary to establish a maximal subspace of factorial effects, a proper subspace $\mathcal{X} \subset A \otimes B \otimes \cdots$, usually a subspace containing all additive main effects $A + B + \cdots$, but possibly containing some or all of the two-factor treatment interactions. Block by treatment subspaces are candidates for exclusion.

Given the subspace $\mathcal{X}$, the standard linear least-squares regression coefficient of y on z is

$$\hat{\beta} = \frac{\langle y, Qz\rangle}{\langle z, Qz\rangle},$$

where Q is the orthogonal projection with kernel $\mathcal{X}$. Least-squares computations for covariate-adjusted factorial effects proceed in the standard manner using Yates's algorithm with y replaced by $y - \hat{\beta}z$. In this analysis, only factorial effects included in $\mathcal{X}$ are to be considered.

On the assumption that $\mathcal{X}$ is the span of some subset of the factorial contrast vectors, all of the computations needed are available from the output of Yates's algorithm applied twice, once to y and once to z. Each factorial contrast is a linear functional associated with a vector in $\mathcal{X}$ or with a vector in $\mathcal{X}^\perp$. The numerator of the least-squares regression coefficient is a sum of products over the finite set of factorial contrast vectors in $\mathcal{X}^\perp$

$$\langle y, Qz\rangle = \sum_{v\in\mathcal{X}^\perp} \frac{\langle v, y\rangle \times \langle v, z\rangle}{\|v\|^2},$$

where $\|v\|^2$ is the divisor. For the denominator, replace y with z.

The covariate-adjusted residual sum of squares is $\mathrm{RSS} = \|Qy\|^2 - \hat{\beta}^2\|Qz\|^2$, and the residual mean square is $s^2 = \mathrm{RSS}/(\dim \mathcal{X}^\perp - 1)$. The raw covariate-adjusted factorial contrasts for $v \in \mathcal{X}$ are obtained by subtraction

$$\langle y - \hat{\beta}z, v\rangle = \langle y, v\rangle - \hat{\beta}\langle z, v\rangle$$

and the covariate-adjusted SS on one degree of freedom is

$$\mathrm{SS}_v = \big|\langle y - \hat{\beta}z, v\rangle\big|^2 / \|v\|^2.$$

If some or most of the parameters associated with factorial contrasts in $\mathcal{X}$ are null, the half-normal plot of the ordered absolute contrasts $\mathrm{SS}_v^{1/2}$ should have a slope approximately equal to s, with positive deviations for the non-null effects.

11.6. Fractional factorials 2^{k-1}. Consider a 2^4 factorial design with $n = 2^4$ units or observations labelled by factor levels in standard algebraic order.

$$(1),\ a,\ b,\ ab,\ c,\ ac,\ bc,\ abc,\ d,\ ad,\ bd,\ abd,\ cd,\ acd,\ bcd, abcd.$$

Every contrast such as $(a-1)(c-1)$ splits this set into two subsets of equal size, a positive half on which the contrast takes the value $+1$, and a negative half on which the value is -1. The contrast $(a-1)(b-1)(c-1)(d-1)$, which takes the value $+1$ if there is an even number of factors at the high level, and -1 if there is an odd number, splits the set of factor levels as follows

$$(1),\ ab,\ ac,\ bc,\ ad,\ bd,\ cd,\ abcd; \qquad a,\ b,\ c,\ abc,\ d,\ abd,\ acd,\ bcd.$$

These are the complementary half fractions of a 2^4 factorial design split by the defining contrast, here $ABCD$, or equivalently, $(a-1)(b-1)(c-1)(d-1)$. Each half is a 2^{4-1} factorial design with four factors but only eight factor combinations observed.

Suppose that the first half is observed in the order shown. There are four combinations with A at the high level and four at the low level, which means that $(a-1)$ splits

the units into a positive half and a negative half. Similarly for B, C and D as shown below:

	(1)	ab	ac	bc	ad	bd	cd	$abcd$
$(a-1)$	$-$	$+$	$+$	$-$	$+$	$-$	$-$	$+$
$(b-1)$	$-$	$+$	$-$	$+$	$-$	$+$	$-$	$+$
$(c-1)$	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$
$(d-1)$	$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$

Evidently these four contrasts are mutually orthogonal, not only on the full factorial 2^4 design, but also on this half fraction. The six products are not mutually orthogonal because they satisfy the identities $(a-1)(b-1) = (c-1)(d-1)$ and $(a-1)(c-1) = (b-1)(d-1)$ and $(a-1)(d-1) = (b-1)(c-1)$. Each pairwise product is equal to the complementary pairwise product. However, the three pairs $(a-1)(b-1)$, $(a-1)(c-1)$ and $(a-1)(d-1)$ are orthogonal to each other and to the first-order contrasts. These seven vectors plus the constant vector constitute an orthogonal basis for the observation space $\mathcal{R}^8$.

For the complementary half fraction of the 2^4 design, the first-order contrasts are

	a	b	c	abc	d	abd	acd	bcd
$(a-1)$	$+$	$-$	$-$	$+$	$-$	$+$	$+$	$-$
$(b-1)$	$-$	$+$	$-$	$+$	$-$	$+$	$-$	$+$
$(c-1)$	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$
$(d-1)$	$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$

The reader may verify that the four first-order contrasts and the same three second-order contrasts are also mutually orthogonal on the complementary half fraction. For each of the complementary pairs, however, we have a sign inversion, for example, $(a-1)(b-1) = -(c-1)(d-1)$. Complementary pairs of two-factor interaction contrasts are said to be *aliased*, or indistinguishable from one another. In addition, $(a-1)(b-1)(c-1) = -(d-1)$, so each main effect is aliased with the complementary three-factor contrast.

In order to understand the alias structure of a fractional factorial design, it is helpful to introduce a little elementary group theory. For a 2^k design, the group is generated by k elements $\mathrm{A}, \mathrm{B}, \mathrm{C}, \ldots$ with commutative multiplication subject to the condition

$$\mathrm{A}^2 = \mathrm{B}^2 = \mathrm{C}^2 = \mathrm{D}^2 = 1.$$

For $k = 4$, the group elements are

1, A, B, AB, C, AC, BC, ABC, D, AD, BD, ABD, CD, ACD, BCD, ABCD.

The pair $\{1, \mathrm{ABCD}\}$ is a sub-group, and each of the eight complementary pairs $\{\mathrm{A}, \mathrm{BCD}\}$, $\{\mathrm{B}, \mathrm{ACD}\}, \ldots$, $\{\mathrm{AB}, \mathrm{CD}\}, \ldots$ is a coset.

Each group element may be identified with a factorial contrast or with the associated subspace of $\mathcal{R}^{2^k}$ in the obvious way. There are other ways to make the group elements and the group operation concrete, some of which generalize more naturally than others to 3^k designs and mixed-level designs. But they are all isomorphic for 2^k designs.

If the full 2^4 design is split in half by the defining contrast ABCD, we have, on that half, either $\mathrm{ABCD} = 1$ or $\mathrm{ABCD} = -1$. Group multiplication by A gives $\mathrm{BCD} = \mathrm{A}$ on one half and $\mathrm{BCD} = -\mathrm{A}$ on the other. Similarly for the other aliased pairs in each coset.

If the full 2^4 design is split in half by the level of ABC, we have, on that half, either $\mathrm{ABC} = 1$ or $\mathrm{ABC} = -1$. Group multiplication by D gives $\mathrm{ABCD} = \mathrm{D}$ on one half and $\mathrm{ABCD} = -\mathrm{D}$ on the other. In either case, $\mathrm{A} = \mathrm{BCD}$ as subspaces. Similarly for the other aliased pairs. In this case, however, we also have $\mathrm{A} = \mathrm{BC}$, $\mathrm{B} = \mathrm{AC}$ and $\mathrm{C} = \mathrm{AB}$, so three

of the four main effects are aliased with two-factor interactions. On the full design, A and BC are mutually orthogonal subspaces; restricted to this half fraction generated by ABC, they are coincident subspaces, and the additive effects coming from either source cannot be distinguished.

11.7. Yates's decomposition for a half fraction. Suppose that the half fraction

$$(1),\ ab,\ ac,\ bc,\ ad,\ bd,\ cd,\ abcd$$

generate by ABCD = 1 is observed. In order to decompose the sum of squares using Yates's algorithm, we must first reduce the number of letters from four to three, and then list the values in standard order for the letters that remain. Suppose we delete b, giving new labels

$$(1),\ a,\ ac,\ c,\ ad,\ d,\ cd,\ acd,$$

without altering the order. Now rearrange the labels and values in standard order, say DCA, giving

$$(1),\ d,\ c,\ cd,\ a,\ ad,\ ac,\ acd.$$

Yates's algorithm with factor labels declared in the order DCA produces eight contrasts labelled

$$I,\ D,\ C,\ DC,\ A,\ DA,\ DC,\ DCA,$$

each with a divisor $n = 2^3 = 8$. The total sum of squares of y is thus decomposed into $n\bar{y}^2$ associated with $I \equiv ABCD$ plus seven other terms, one for each alias pair of factorial contrasts

$$D \equiv ABC;\ C \equiv ABD;\ DC \equiv AB;\ A \equiv BCD;\ DA \equiv BC;\ DC \equiv AB;\ DCA \equiv B.$$

Note that the effect associated with the deleted factor label, B in this illustration, emerges from the algorithm as the three-factor interaction of the factors retained: $B \equiv ACD$.

If interactions can safely be assumed to be null, three degrees of freedom are available for the estimation of the residual variance.

The analysis for the negative half fraction follows the same lines except that the aliased pairs of effects have opposite signs. This has no effect on sums of squares, but it does affect the interpretation of coefficients, in particular, the direction of the B-effect.

11.8. Confounding. In settings where each factor is a treatment whose level can be assigned by randomization, it may be possible to partition the units into two blocks of eight units each, and assign treatments to units in such a way that the positive half fraction of the defining contrast is assigned to one block and the negative half fraction to the other. All treatment combinations are observed, but the block contrast is indistinguishable from, or confounded with, the defining treatment contrast, usually $blk = ABCD$. The factorial decomposition is the same as the decomposition for a complete factorial except that the defining contrast is associated with the block effect, which is typically of little interest, and is ignored in subsequent analyses. On the assumption that there is no block-by-treatment interaction, four three-factor interactions are available for the estimation of residual variance and for testing two-factor interactions.

Exercise 11.1: Consider the nested pair of linear models in which the null model or subspace is $E(Y) = X\beta$, i.e., $E(Y) \in \mathcal{X}$, and the alternative is $E(Y) = X\beta + z\gamma$, where z is a vector in $\mathcal{R}^n$ not in $\mathcal{X}$. Let Q be the orthogonal projection $\mathcal{R}^n \to \mathcal{R}^n$ whose kernel is $\mathcal{X}$, and let $P = I - Q$. By writing $z = Pz + Qz$, show that the least squares coefficient of z is

$$\hat{\gamma} = \frac{\langle Y,\, Qz\rangle}{\langle z,\, Qz\rangle}.$$

Obtain an expression for the reduction in residual sum of squares due to z, i.e., the SS on one degree of freedom associated with $(\mathcal{X} + z)/\mathcal{X}$.

Exercise 11.2: (Tukey 1DOFNA) In the setting of the previous exercise, let $\hat{\mu} = PY$ be the null fitted vector, and let $z_i = \hat{\mu}_i^2$ be the vector of squared fitted values. The residual sum of squares associated with z is called the Tukey's one degree of freedom for non-additivity. Explain why the standardized coefficient ratio $\hat{\gamma}/\operatorname{se}(\hat{\gamma})$ as produced by `lm(y~X+z)` is distributed *exactly* as Student's t under the null Gaussian model. What are the degrees of freedom?

Exercise 11.3: Let x be a quantitative covariate, and suppose that $\mathcal{X} = \operatorname{span}(1, x)$ is of dimension two. Show that Tukey's one degree of freedom is equivalent to setting $z = x^2$, and checking for linearity by testing whether the quadratic coefficient is zero.

Exercise 11.4: Under what conditions on $\mathcal{X}$ can we be sure, for a generic vector $\hat{\mu} \in \mathcal{X}$, that z does not belong to $\mathcal{X}$. Give examples of model formulae such that z belongs to $\mathcal{X}$.

12. Gaussian distribution theory

All of the distribution-theory needed for regression and analysis of variance is based on the Gaussian distribution and distributions derived from the Gaussian. We begin with a summary of Gaussian and chi-squared distributions, with density functions $f(x)$ and Laplace transforms $\int e^{-tx} f(x)\,dx$.

Table 1: *Gaussian, chi-squared, Student's t and Fisher's F distributions*

Random variable	Distribution	Density at x	Laplace transform
X	$N(0,1)$	$\dfrac{e^{-x^2/2}}{\sqrt{2\pi}}$	$e^{t^2/2}$
X^2	χ^2_1	$\dfrac{x^{-1/2}e^{-x/2}}{\sqrt{2\pi}}$	$(1+2t)^{-1/2}$
$SS = X_1^2 + \overset{i.i.d.}{\cdots} + X_k^2$	χ^2_k	$\dfrac{x^{k/2-1}e^{-x/2}}{2^{k/2}\,\Gamma(k/2)}$	$(1+2t)^{-k/2}$
$\dfrac{X}{\sqrt{MS}}$	t_k	$\dfrac{C_k}{(1+x^2/k)^{(k+1)/2}}$	
$\dfrac{MS_1}{MS_2} = \dfrac{\chi^2_f/f}{\chi^2_k/k}$	$F_{f,k}$	$\dfrac{C_{f,k}\,x^{f/2-1}}{(1+fx/k)^{(f+k)/2}}$	

The symbol SS stands for a sum of squares; the distribution theory shown above is for a sum of squares of k independent zero-mean standard Gaussian variables $X_1, \ldots, X_k$. By construction $E(SS) = k$, which is called the degrees of freedom, and $\mathrm{var}(SS) = 2k$. The symbol $MS = SS/k$ stands for a mean square; the distributon-theory in the last two lines of the table is for ratios in which the numerator and denominator are independent. As defined above, each mean square has expected value one and variance $2/k$, inversely proportional to the degrees of freedom. In applications where $X \sim N(0, \sigma^2)$, each sum of squares is distributed as $\sigma^2\chi^2_k$, so each mean square has expectation σ^2 and variance $2\sigma^4/k$. The Student-t and Fisher-F-ratios are unaffected by the scalar multiple.

It is evident from the definitions that $R \sim t_k$ implies $R^2 \sim F_{1,k}$, i.e., if R is distributed according to Student's t distribution on k degrees of freedom, then R^2 is distributed according to Fisher's F distribution on $1, k$ degrees of freedom. Also, $1/F_{f,k} = F_{k,f}$ in the distributional sense that the reciprocal transformation on the positive real line sends one distribution to the other. In particular, $1/F_{k,k} = F_{k,k}$ is self-inverse for every k.

Exercise 12.1: Let X be a standard normal random variable with density ϕ. Differentiate the cumulative distribution function

$$F_{\chi^2}(t) = \mathrm{pr}(X^2 \le t) = 2\int_0^{\sqrt{t}} \phi(x)\,dx$$

to obtain the density function of χ^2_1.

Exercise 12.2: For $X \sim N(0,1)$ show that $E(e^{-tX^2}) = (1+2t)^{-1/2}$. Hence deduce that $E(e^{-t(X_1^2+\cdots+X_k^2)}) = (1+2t)^{-k/2}$ for the sum of squares of independent $N(0,1)$ random variables.

Exercise 12.3: Under the assumptions used in Table 1, show that $X_1^2 + X_2^2$ is distributed according to the exponential distribution with mean 2.

Exercise 12.4: Without typing them into the computer, what values are given by the R expressions

$$\texttt{pchisq}(\texttt{1.96\^{}2}, \texttt{df} = \texttt{1}), \qquad \texttt{pchisq}(\texttt{2} * \texttt{log}(\texttt{20}), \texttt{df} = \texttt{2})$$

Explain.

Exercise 12.5: Let X, Y be independent unit exponential random variables. Show that the sum $S = X + Y$ and the ratio $R = X/(X + Y)$ are also independent. Find their distributions, i.e., find the density functions.

Exercise 12.6: Let $X \sim \chi^2_f$ and $Y \sim \chi^2_k$ be independent random variables. Show that the sum $S = X + Y$ and the ratio $R = X/Y$ are also independent. Find their distributions, hence obtain the constant $C_{f,k}$ in the density of Fisher's F distribution.

13. Distributions for ANOVA

Let $\mathcal{R}^n$ be an inner product space with inner product matrix W, i.e., $\langle u, v\rangle = u'Wv$. Let

$$\mathcal{U}_0 \subset \mathcal{U}_1 \subset \cdots \subset \mathcal{U}_k = \mathcal{R}^n$$

be any nested sequence of subspaces. By definition,

$$\mathcal{U}_0, \quad \mathcal{U}_0^\perp \cap \mathcal{U}_1, \quad \mathcal{U}_1^\perp \cap \mathcal{U}_2, \quad \mathcal{U}_2^\perp \cap \mathcal{U}_3, \ldots, \mathcal{U}_{k-1}^\perp \cap \mathcal{U}_k$$

is a set of complementary orthogonal subspaces. If P_r is the W-orthogonal projection onto $\mathcal{U}_r$, the orthogonal projections onto the mutually orthogonal subspaces are

$$P_0, \quad P_1 - P_0, \quad P_2 - P_1, \ldots, P_k - P_{k-1}$$

where $P_k = I_n$. The dimension of the space is the trace of the projection. Explain why.

Suppose now that $Y \sim N_n(0, \Sigma)$, where Σ is strictly positive definite with inverse matrix W. This means that the probability density at y is proportional to $\exp(-\|y\|^2/2)$. Since

$$\|y\|^2 = \|P_0 y\|^2 + \|(P_1 - P_0)y\|^2 + \cdots + \|(P_k - P_{k-1})y\|^2,$$

is additive, the joint density factors as a product

$$\exp(-\|y\|^2) = \exp(-\|P_0 y\|^2) \times \cdots \times \exp(-\|(P_k - P_{k-1})y\|^2).$$

It follows that random variables

$$P_0 Y, \quad (P_1 - P_0)Y, \quad (P_2 - P_1)Y, \ldots, (P_k - P_{k-1})Y$$

are independently distributed zero-mean Gaussian variables, their squared norms

$$\|P_0 Y\|^2, \quad \|(P_1 - P_0)Y\|^2, \cdots, \|(P_k - P_{k-1})Y\|^2,$$

are independent chi-squared random variables, and the ratios are distributed according to Fisher's F distribution on the appropriate degrees of freedom.

By a similar argument, if $\Sigma = \sigma^2 W^{-1}$ is proportional to W^{-1}, the random variables

$$P_0 Y, \quad (P_1 - P_0)Y, \quad (P_2 - P_1)Y, \ldots, (P_k - P_{k-1})Y$$

are independently distributed zero-mean Gaussian variables, and their squared norms

$$\|P_0 Y\|^2, \quad \|(P_1 - P_0)Y\|^2, \cdots, \|(P_k - P_{k-1})Y\|^2,$$

are independent $\sigma^2 \chi^2_{f_r}$ random variables with degrees of freedom

$$f_r = \operatorname{rank}(P_r - P_{r-1}) = \operatorname{tr}(P_r) - \operatorname{tr}(P_{r-1}) = \dim(\mathcal{U}_r) - \dim(\mathcal{U}_{r-1}).$$

If $Y \sim N_n(\mu, \Sigma)$ has a non-zero mean with the constraint $\mu \in \mathcal{U}_r$, the sequence of projections is independent Gaussian, but the first r means are not zero. The subsequent sequence beyond $\mathcal{U}_r$, i.e.,

$$(P_{r+1} - P_r)Y, \ldots, (P_k - P_{k-1})Y$$

is independent zero-mean Gaussian, and the squared norms are independent χ^2 or $\sigma^2\chi^2$ as before.

In this setting, if the columns of X constitute a basis for $\mathcal{U}$, $P_r = X(X'WX)^{-1}X'W$ is the orthogonal projection onto $\mathcal{U}$, and WP_r is symmetric. For $r \leq s$, the products satisfy $P_r P_s = P_s P_r = P_r$, so that $(P_s - P_r)P_r = 0$.

In the simplest application of this result, the inner product is invariant with respect to permutation, which means that W is a positive linear combination of the identity and the matrix $J = \mathbf{1}\mathbf{1}'$. Suppose first that $W = I_n$, and that $\Sigma = \sigma^2 I_n$.

14. Polynomial subspaces

Suppose that the data available for $n = 12$ units is as follows

x	1	2	3	3	4	5	6	7	7	8	9	9
y	7.5	5.9	7.2	6.0	3.0	9.3	13.2	10.6	10.7	11.4	13.8	11.6

where x is a quantitative covariate and y is the response. The space $\mathcal{R}^{12}$ has a natural decomposition by polynomials

$$\mathcal{U}_0 \subset \mathcal{U}_1 \subset \mathcal{U}_2 \subset \cdots$$

associated with the given covariate x, with $\mathcal{U}_r = \text{span}(1, x, \ldots, x^r)$. The first two terms in this decomposition are

$$\|P_0 y\|^2 = n\bar{y}^2 = 1012.00, \quad \|P_1 y\|^2 - \|P_0 y\|^2 = \Big(\sum (x_i - \bar{x}) y_i\Big)^2 \Big/ \sum (x_i - \bar{x})^2 = 73.16$$

The subsequent values $\|P_r y\|^2 - \|P_{r-1} y\|^2$ for $r = 2, \ldots, 8$ are

$$2.455, 10.301, 1.035, 7.756, 0.002, 19.610, 0.976,$$

and finally

$$\|y\|^2 - \|P_8 y\|^2 = 3.145 = (7.2 - 6.0)^2/2 + (10.6 - 10.7)^2/2 + (13.8 - 11.6)^2/2$$

is the sum of squares for the three replicate pairs. These data were generated with $\mu \in \mathcal{U}_1$, in such a way that all other sums of squares are distributed as $\sigma^2 \chi^2_f$ with $\sigma^2 = 4$. The largest value (19.6) is at the 97% point of χ^2_1, which is not partcularly extreme for the largest of seven iid values. The probability that the maximum value is 19.6 or larger is $1 - 0.973^7 = 0.174$, or about 1 in 6.

14.1. Orthogonal polynomials. Suppose that the covariate is quantitative with a small number of levels, 4–5, say, and that there is an equal number of responses at each factor level, as illustrated below

x	0	0	0	1	1	1	2	2	2	3	3	3
y	7.5	5.9	7.2	6.0	3.0	9.3	13.2	10.6	10.7	11.4	13.8	11.6

At the start, it is best to compress the replicates into one, so that $\{0, 1, 2, 3\}$ is the index set, and the identity function $x(i) = i$ is the four-component vector $x = (0, 1, 2, 3)$ in $\mathcal{R}^4$. The four polynomials x^0, x^1, x^2, x^3 are linearly independent over the index set, but they are not orthogonal [with respect to the standard inner product]. However, we may use the Gram-Schmidt procedure to construct an alternative basis $\{v_0, v_1, v_2, v_3\}$ that is orthogonal and also polynomial.

$$\begin{aligned}
v_0 &= x^0 = (1, 1, 1, 1) \\
v_1 &= x - \frac{\langle x,\, v_0\rangle v_0}{\|v_0\|^2} = x - \frac{6 \times (1, 1, 1, 1)}{4} = \frac{(-3, -1, 1, 3)}{2} \\
v_2 &= x^2 - \frac{\langle x^2,\, v_0\rangle v_0}{\|v_0\|^2} - \frac{\langle x^2,\, v_1\rangle v_1}{\|v_1\|^2} = \frac{(1, -1, -1, 1)}{1} \\
v_3 &= x^3 - \frac{\langle x^3,\, v_0\rangle v_0}{\|v_0\|^2} - \frac{\langle x^3,\, v_1\rangle v_1}{\|v_1\|^2} - \frac{\langle x^3,\, v_2\rangle v_2}{\|v_2\|^2} = \frac{3(-1, 3, -3, 1)}{10}
\end{aligned}$$

Evidently v_r is a linear combination of $x^0, \ldots, x^r$, i.e., a polynomial of degree r. These are the polynomials that are mutually orthogonal over four equally spaced points. They can be normalized to have unit length; for hand computations, it is more convenient to normalize so that each component is an integer.

For the illustration shown above, the space of polynomials in x is embedded as a subspace of dimension four in $\mathcal{R}^{12}$ by repeating each component as many times as there are replicates of that level. For the example, the replicate numbers are all the same (three), so the polynomials v_r are also orthogonal with respect to the standard inner product in $\mathcal{R}^{12}$. The y totals for the four levels of x are $20.6, 18.3, 34.5, 36.8$ with mean 27.55. The sum of squares on three degrees of freedom for x as a factor is

$$\frac{1}{3}(20.6^2 + 18.3^2 + 34.5^2 + 36.8^2 - 4 \times 27.55^2) = 89.243.$$

The fraction $1/3$ comes from the number of replicates in each total, not from the degrees of freedom. This SS can be partitioned into a sum of three squares, one for each basis vector orthogonal to v_0. The numerical decomposition by polynomial subspaces

$$89.243 = 69.984 + 1.763 + 17.496$$

shows, as expected, that the linear contribution is dominant. The ratio of the square for cubics to the residual mean square is $17.496/3.649 = 4.79$, which is moderately large, falling at the 94th percentile of the $F_{1,8}$ distribution. Given that we have selected the larger of the two non-linear polynomial contributions for this comparison, the tail p-value corrected for selection is $2 * 0.06 = 12\%$, which is not particularly extreme.

The fact that the non-linear SS splits so unevenly, the cubic being nearly ten times as large as the quadratic, might seem surprising. But a ratio bigger than 10 or less than $1/10$ occurs with probability nearly 40% in the $F_{1,1}$ distribution. Nothing there to write home about!

For five equally-spaced levels, the orthogonal polynomials other than v_0 are

$$\begin{aligned}
v_1 &\propto (-2, -1, 0, 1, 2) \\
v_2 &\propto (2, -1, -2, -1, 2) = v_1^2 - 2 \\
v_3 &\propto (-1, 2, 0, -2, 1) \\
v_4 &\propto (1, -4, 6, -4, 1)
\end{aligned}$$

15. Trigonometric subspaces

Consider the points $[n]$ equally spaced from $j = 1$ to $j = n$ on the real line, and the corresponding points Ω

$$\omega_j = \exp(2\pi ij/n) = \cos(2\pi j/n) + i\sin(2\pi j/n)$$

equally spaced on the unit circle in the complex plane. Note that $i^2 = -1$ is the imaginary number, $\omega = \omega_1 = \exp(2\pi i/n)$ is a primitive nth root of unity, so the points in Ω are the powers $\omega, \omega^2, \ldots, \omega^n = 1$, and $\bar{\omega}_j = \omega_{n-j}$ are complex conjugate pairs.

Consider now the Euclidean space of complex-valued functions on $[n]$, or the equivalent vector space of complex-valued functions on Ω with Hermitian inner product

$$\langle u,\, v\rangle = \sum_{z\in\Omega} u(z)\bar{v}(z).$$

Since

$$\langle z^r,\, z^s\rangle = \sum_{j=1}^{n} \omega^{j(r-s)} = \begin{cases} n & r = s \pmod n \\ 0 & \text{otherwise,} \end{cases}$$

the monomial powers $\{z^r\}$ are mutually orthogonal, and therefore linearly independent for $r = 0, \ldots, n-1$. Thus, any complex-valued function w defined on Ω may be expressed as a linear combination of monomials

$$w(z) = c_0 z^0 + \cdots + c_{n-1} z^{n-1},$$

with coefficients $c_r = \langle w,\, z^r\rangle / \langle z^r,\, z^r\rangle = \langle w,\, z^r\rangle / n$. The list of coefficients $\{c_0, \ldots, c_{n-1}\}$ is called the *discrete Fourier transform* of z; it can be computed using the *fast Fourier transform* `fft(z)`.

Monomial orthogonality implies, $\|w\|^2 = \sum |c_r|^2 \, \|z^r\|^2 = n \sum |c_r|^2$, so the discrete Fourier coefficients provide a decomposition of the total sum of squares into orthogonal components, one component for each monomial z^r. The list of moduli or squared moduli $|c_r|^2$, ignoring phases, is closely related to the *power spectrum*. Since $z^{-r} = \bar{z}^r$, the monomial basis functions z^r, z^{n-r} span the subspace associated with the same Fourier frequency r. It is more natural to arrange the coefficients in pairs $\{c_r, c_{n-r}\}$ by frequency rather than by monomial. The *power spectrum* is a frequency-based decomposition of the total sum of squares, which is invariant with respect to rotation and reflection acting on Ω (or cyclic permutation and reflection acting on $[n]$). For the power spectrum, we would ordinarily plot $n(|c_r|^2 + |c_{n-r}|^2)$ against r for $r = 1, \ldots, \lfloor n/2 \rfloor$, ignoring c_0.

In certain physical systems, $n(|c_r|^2 + |c_{n-r}|^2)$ is called the *energy at frequency* r. For example, thermal radiation comes in a range of frequencies in the electromagnetic spectrum. Planck's law for black-body radiation energy at frequency ν is proportional to $\nu^3/(e^{k\nu/T} - 1)$ at absolute temperature T.

If w happens to be a real-valued function, as is frequently the case in statistical work, then $c_{n-r} = \bar{c}_r$, so $c_{n-r} z^{n-r}$ is the complex conjugate of $c_r z^r$. For $r = 0, \ldots, \lfloor n/2 \rfloor$, the subspace associated with Fourier frequency r is spanned by the functions $\cos(2\pi jr/n), \sin(2\pi jr/n)$. If $r = 0$, the subspace is $\mathbf{1}_n$ of real dimension one; for each $r \geq 1$, the real dimension is two unless n is even and $r = n/2$, in which case the space is spanned by the single function $\cos(\pi j) = (-1)^j$.

Exercise 15.1: Compute and plot the power spectrum for the Islay monthly rainfall data (`~islay.rain`). Segregate the values at frequencies that are integer multiples of yr^{-1}, from the others and plot them separately, (omitting the annual and 6-month values).

Comment briefly on the pattern observed, and re-scale the plot if necessary. The Wilson-Hilferty cube-root transformation may help.

Explain why the squared modulus of each Fourier coefficient in a stationary series should be approximately exponentially distributed and independent for each Fourier frequency. How is the expected value related to the spectrum of the process?

Apart from the annual and semi-annual cycles, how does this series differ from white noise? Use an exponential model (GLM) to fit a suitable model to the spectrum.

Exercise 15.2: Compute and plot the power spectrum for the Central England daily temperatures 1772–2016, segregating the annual from non-annual frequencies. How does the non-annual energy vary with frequency?

Compute and plot the annual average temperature as a function of year in the range 1772–2016. Does the process appear to be stationary?

Compute the average daily temperature as a function of calendar date, and plot the values. Superimpose on the plot the fitted curve spanned by the first-order harmonics only, and the fitted curve spanned by first and second-order harmonics. On what calendar date does the minimum mean temperature occur?

The data for this exercise are available in the file `~cetdl1772on.dat` in a format that requires some manipulation for analysis; see the file `~cetdl1772on.R` for details.

16. Linear and quadratic forms

Let Y be a random vector with mean vector $E(Y) = \mu$ and covariance matrix $\Sigma = \mathrm{cov}(Y)$. The components of Y are $Y_1, \ldots, Y_n$, and the components of Σ are $\Sigma_{ij} = \mathrm{cov}(Y_i, Y_j)$.

A linear form is a linear combination $a_1 Y_1 + \cdots + a_n Y_n$ with coefficients $a_1, \ldots, a_n$. Note the difference in terminology: the vector Y has *components* $Y_1, \ldots, Y_n$; the linear functional a, which is a vector in the dual space of linear functionals over $\mathcal{R}^n$, has *coefficients* $a_1, \ldots, a_n$, which are the components with respect to the dual basis. To make this distinction explicit in the notation, some authors write $a^1, \ldots, a^n$ for the coefficients of the linear functional. The value of the linear functional a at Y is

$$a(Y) = a'Y = a^i Y_i = \sum_{i=1}^{n} a^i Y_i.$$

The expression $a^i Y_i$ employs the Einstein implicit summation convention for each repeated index, which should occur exactly once as a subscript and once as a superscript.

The Einstein convention for a quadratic form is

$$Y'AY = a^{ij} Y_i Y_j$$

where $a^{ij} = a^{ji}$ by convention. The expected values are

$$\begin{aligned} E(a'Y) &= a^i E(Y_i) = a^i \mu_i = a'\mu, \\ E(Y'AY) &= a^{ij} E(Y_i Y_j) = a^{ij}(\mu_i \mu_j + \Sigma_{ij}) \\ &= \mu' A \mu + \mathrm{tr}(A\Sigma) \end{aligned}$$

where $\mathrm{tr}(A)$ is the sum of the diagonal components of a square matrix.

The covariance of two linear forms is

$$\mathrm{cov}(a^i Y_i, b^j Y_j) = a^i b^j \,\mathrm{cov}(Y_i, Y_j) = a^i b^j \Sigma_{ij} = a'\Sigma b.$$

The covariance of two quadratic forms is

$$\mathrm{cov}(a^{ij} Y_i Y_j, b^{kl} Y_k Y_l) = a^{ij} b^{kl} \,\mathrm{cov}(Y_i Y_j, Y_k, Y_l).$$

If these are zero-mean variables, or if $A\mu = B\mu = 0$ as occurs frequently in linear-model calculations, the expected value may be simplified to

$$a^{ij} b^{kl} (\Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk} + \kappa_{ijkl}) = 2\,\mathrm{tr}(A\Sigma B\Sigma) + a^{ij} b^{kl} \kappa_{ijkl}$$

where the fourth cumulant κ_{ijkl} is zero for Gaussian variables.

17. Additive effects and interactions

Let A be a two-level factor, which is coded as 0/1 or low/high. For any quantitative response Y, the *effect of A on the distribution of Y* is defined conventionally to be

$$\text{effect of } A = \big(\text{mean of } Y \text{ for units at } A = 1\big) - \big(\text{mean of } Y \text{ for units at } A = 0\big).$$

This definition presupposes that the difference between the response distribution for $A = 1$ versus the distribution for $A = 0$ is primarily in the mean rather than the variance or the shape. Likewise, the effect of B on Y is

$$\text{effect of } B = \big(\text{mean of } Y \text{ for units at } B = 1\big) - \big(\text{mean of } Y \text{ for units at } B = 0\big).$$

The word 'mean' can be interpreted either as the mean of the distribution, or as the arithmetic mean of sample values.

When we examine the effect of two factors A, B jointly on Y, there are four mean values $\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11}$ to be considered. Here μ_{rs} is the mean response for each unit having $A = r$ and $B = s$. For those units having $B = 1$, the effect of A is $\mu_{11} - \mu_{01}$, the mean response with A at the high level minus the mean response for A at the low level. For those units having $B = 0$ the effect is $\mu_{10} - \mu_{00}$. If these effects are equal, we say that there is no interaction between A and B.

The *interaction of A, B on Y* is defined to be

$$\begin{aligned}\text{interaction of } A, B &= \big(\text{effect of } A \text{ given } B = 1\big) - \big(\text{effect of } A \text{ given } B = 0\big) \\ &= \mu_{11} - \mu_{01} - \mu_{10} + \mu_{00}\end{aligned}$$

which is also the interaction of B, A. Thus, additive interaction is symmetric.

The interaction is zero if and only if the effect of A is the same for each level of B; in this case, the effect of B is the same for each level of A. To say the same thing in another way, the interaction is zero if and only if there exist coefficients $\alpha_0, \alpha_1, \beta_0, \beta_1$ such that

$$\mu_{rs} = \alpha_r + \beta_s$$

is an additive function of the two effects. In that case $\alpha_1 - \alpha_0$ is the effect of A and $\beta_1 - \beta_0$ is the effect of B. Thus *additivity of effects* is the same as *no interaction* between the factors.

18. Parameterization of effects and interactions

This section is concerned with the parameterization of effects and interactions in typical computer packages that use linear-model formulae for the specification of subspaces, and use linear or iterative linear projection for computing coefficients and fitted values. Examples include `lm(y~A+B+x,...)`, `glm(y~A+B+x,...)` and `regress(y~..., V~...)` in R. For a variety of reasons, the conventions are not the same as those used in the decomposition of sums of squares in balanced factorial designs.

Suppose that the design and response for 10 units are as follows

unit	1	2	3	4	5	6	7	8	9	10
A	0	1	1	0	0	0	1	1	1	1
B	0	0	0	1	1	1	1	1	1	1
y	5.9	6.5	5.8	4.6	6.0	5.3	5.7	6.0	5.6	5.4
$P_{AB}y$	5.90	6.15	6.15	5.30	5.30	5.30	5.675	5.675	5.675	5.675
$P_{A+B}y$	5.84	6.18	6.18	5.32	5.32	5.32	5.660	5.660	5.660	5.660

The penultimate line shows the response average for each of the four AB combinations, and the final line shows the projection of y onto the additive subspace.

In a linear model formula, the symbol A refers to the subspace spanned by the two indicator vectors

$A \equiv 0$	1	0	0	1	1	1	0	0	0	0
$A \equiv 1$	0	1	1	0	0	0	1	1	1	1

and likewise for B. Ordinarily, the constant vector $\mathbf{1}$ is understood to be included in the model formula, so $1 + A$ refers to the vector space spanned by $\mathbf{1}$ and the two indicator functions A_0, A_1. These three vectors are linearly dependent, and each pair spans the same space. The projection of y onto the subspace A is

$$\begin{aligned} P_A y &= \bar{y}_0 A_0 + \bar{y}_1 A_1 = \bar{y}_1 \mathbf{1} + (\bar{y}_0 - \bar{y}_1) A_0 = \bar{y}_0 \mathbf{1} + (\bar{y}_1 - \bar{y}_0) A_1 \\ &= 5.45 A_0 + 5.83 A_1 = 5.83 \mathbf{1} - 0.38 A_0 = 5.45 \mathbf{1} + 0.38 A_1 \end{aligned}$$

Notice how the coefficients depend on the choice of basis. For the indicator basis A_0, A_1, the coefficients are the average y-values for the four observations at $A = 0$ and the six at $A = 1$. Since $\mathbf{1} = A_0 + A_1$, the coefficients for the other pairs are obtained by substitution. The default in most statistical packages is to declare a *reference level* for each factor, usually the first level, and to set the reference coefficient to zero by omitting the reference indicator vector from the basis. In that case, the coefficient of A_r is the difference $\bar{y}_r - \bar{y}_0$ between the response average at $A = r$ and the response average at the reference level. Notice that the coefficient of $\mathbf{1}$ is the reference-level average, not the overall average.

Similar remarks apply to models that include interactions. The symbol AB or $A{*}B$ or $A{:}B$ in a model formula refers to the subspace spanned by the indicator vectors

$$A_0 B_0, \quad A_0 B_1, \quad A_1 B_0, \quad A_1 B_1$$

for each ordered pair of levels. The orthogonal projection of y onto AB is

$$\begin{aligned} P_{AB} y &= \bar{y}_{00} A_0 B_0 + \bar{y}_{01} A_0 B_1 + \bar{y}_{10} A_1 B_0 + \bar{y}_{11} A_1 B_1 \\ &= 5.90 A_0 B_0 + 5.3 A_0 B_1 + 6.15 A_1 B_0 + 5.675 A_1 B_1. \end{aligned}$$

However, the basis implied by $A{*}B$ in a model formula is

$$\mathbf{1}, \quad A_1, \quad B_1, \quad A_1 B_1,$$

omitting all reference levels or setting the reference-level coefficients to zero. Using the identities $A_0 = \mathbf{1} - A_1$ and $B_0 = \mathbf{1} - B_1$, we obtain the coefficients

$$\begin{aligned} P_{AB}y &= \bar{y}_{00}\mathbf{1} + (\bar{y}_{10} - \bar{y}_{00})A_1 + (\bar{y}_{01} - \bar{y}_{00})B_1 + (\bar{y}_{11} - \bar{y}_{01} - \bar{y}_{10} + \bar{y}_{00})A_1B_1 \\ &= 5.90\mathbf{1} + 0.25A_1 - 0.60B_1 + 0.125A_1B_1 \end{aligned}$$

by substitution into the formula for $P_{AB}y$. Notice that the coefficient of A_1B_1 is the sample estimate of the non-additivity or interaction. Notice also that the coefficient of A_1 is not the effect of A or the average effect of A, but the effect of A at the reference level of B. Conversely, the coefficient of B_1 is the effect of B at the reference level of A, and the coefficient of $\mathbf{1}$ is the mean response at the joint reference level. *This mis-interpretation of apparent main-effect coefficients in a model with interaction is a very common error for novice statisticians!*

If we were to switch the reference level of B, the interaction coefficient remains the same except for sign, but the coefficient of A_1 becomes the effect of A at the other level of B, which is an entirely different number.

When the model includes an interaction between two factors such as treatment and sex, the question being addressed is whether treatment has the same effect for males as it does for females. The answer to the question lies in the interaction coefficient, its estimate and its standard error. If no interaction is found, we may then talk of the effect of treatment (regardless of sex). If interaction is present, the 'average treatment effect', however it may be defined, is seldom of great interest, particularly if the interaction is substantial. In such cases, it is more natural to split the population by the classification factor sex, and to discuss separately the effect of treatment in the two sub-populations.

19. Distribution of treatment contrasts

Assume that treatment has k levels and that

(i) the responses for distinct units are independent random variables with finite variances;
(ii) two units assigned the same treatment have the same response distribution;
(iii) the effect of treatment is additive on the response.

These assumptions imply that, whatever the assignment of treatments to units $T = (T_1, \ldots, T_n)$ may be, the response distribution is such that

$$Y_i = \tau_{T(i)} + \epsilon_i,$$

where $\tau_1, \ldots, \tau_k$ are the means for each treatment level, and $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed random variables with finite variance $\sigma^2 > 0$, the same for each unit. In the Gaussian case, the joint distribution is $Y \sim N_n(X\tau, \sigma^2 I_n)$, where the columns of X are the indicator basis vectors for the assigned treatment levels.

Let $n_r = \#\{i{:}\, T(i) = r\}$ be the number of plots assigned to treatment r. Provided that $n_r > 0$, the least squares estimate is $\hat{\tau} = (X'X)^{-1}X'y$. This means that $\hat{\tau}_r$ is the sample average

$$\hat{\tau}_r = \sum_{i:T(i)=r} Y_i/n_r$$

which has variance σ^2/n_r.

An elementary treatment contrast is a pairwise difference $\tau_r - \tau_s$, or $\hat{\tau}_r - \hat{\tau}_s$ for the estimator, which has variance

$$\operatorname{var}(\hat{\tau}_r - \hat{\tau}_s) = \sigma^2\Big(\frac{1}{n_r} + \frac{1}{n_s}\Big)$$

in which σ^2 is estimated by the residual mean square, or replicate mean square, from the $\mathcal{R}^n/T$-line in the ANOVA table. The average variance of all pairwise contrasts is

$$\frac{\sigma^2}{k(k-1)} \sum_{r \neq s} \frac{1}{n_r} + \frac{1}{n_s} = \frac{\sigma^2}{k(k-1)} 2(k-1) \sum_r \frac{1}{n_r} = \frac{2\sigma^2}{k} \sum_r \frac{1}{n_r}.$$

If n is an integer multiple of k, this is minimized by setting $n_1 = \cdots = n_k = n/k$, which means that each treatment is replicated an equal number of times. Otherwise, if n/k is not an integer, the design that minimizes the average variance of treatment contrasts has replicate numbers differing by at most one.

20. Treatment effects in a block design

The analysis for a block design is a little more complicated. We assume that the block and treatment effects are additive, so that the response is a sum of block, treatment and unit effects:

$$Y_i = \tau_{T(i)} + \beta_{B(i)} + \epsilon_i.$$

The unit effects are iid with finite variance σ^2.

Model I treats the block effects $\beta_1, \ldots, \beta_b$ as arbitrary constants to be estimated: In essence, B is a classification factor with b levels, so that the mean vector $\mu = E(Y)$ lies in the subspace $B + T$, and $\mathrm{cov}(Y) = \sigma^2 I_n$.

Model II treats the block effects as random variables distributed independently of ϵ. This model has several versions depending on the assumptions made about the joint distribution of $\beta_1, \ldots, \beta_b$. The most common assumption (Model IIa) is that the block or batch effects are independent and identically distributed with variance σ_b^2. For this model, the batch effects contribute only to the covariances, so the mean $\mu = E(Y)$ lies in the subspace T, and $\mathrm{cov}(Y) = \sigma^2 I_n + \sigma_b^2 B$. In other words

$$\mathrm{cov}(Y_i, Y_j) = \sigma_0^2 \delta_{ij} + \sigma_b^2 B_{ij},$$

where δ_{ij} is Kronecker's delta, and $B_{ij} = 1$ if and only if i, j belong to the same block. Note that $\mathrm{var}(Y_i) = \sigma_0^2 + \sigma_b^2$ is the sum of the two variances, whereas $\mathrm{cov}(Y_i, Y_j) = \sigma_b^2$ for $i \neq j$ in the same block, or zero if they are in different blocks.

The assumption in Model IIa that the block effects are identically distributed is often reasonable, but independence is much less plausible in many areas of application. For example, if the blocks represent sequential batches of a reagent or ingredient, or if they have a physical arrangement in space, serial or spatial dependence might be expected. A satisfactory model for the block effects is important in the estimation of treatment effects, particularly the assessment of their standard errors.

The reader should note that the symbol B has been used in this section for at least three purposes, distinct but related. First B is a list of block labels, in essence the classification factor in Model I. Second B is the block subspace $B \subset \mathcal{R}^n$. Third B is a block factor or Boolean matrix with components B_{ij} in Model IIa. Regardless of how it is used, B is almost always stored as a list of levels. Note that the subspace $B \subset \mathcal{R}^n$ does not occur in Model II; in Model IIa, the subspace has been replaced by a symmetric matrix B whose non-zero eigenvectors are the indicator functions for the blocks. The eigenvalues are the block sizes.

Let P be the orthogonal projection onto the subspace $B \subset \mathcal{R}^n$, and let $Q = I - P$ be the complementary projection. If all the blocks are of equal size, P is proportional to the matrix B; otherwise $P_{ij} = B_{ij}/\mathrm{blocksize}_i$. For all versions of the model, $E(QY) = Q\mu$ lies in the subspace QT, and $\mathrm{cov}(QY) = \sigma^2 Q$, independent of block effects. Provided that we agree to work only with intra-block contrasts QY, the choice of model is immaterial. The differences among the models lie in how they use the information from block averages PY.

Various packages are available in R for fitting the variance-components model IIa. For flexibility, we use the function `regress()`, which uses two model formulae, one for the mean and one for the covariances. Thus, if B and T are declared as factors, the two models can be fitted by

```
fit1 <- regress(y~B+T),    fit2 <- regress(y~T, ~B)
```

and `summary(fit)` gives a summary of the fitted parameters. Note that the intercept is included by default in the first formula, and the identity matrix is included by default in the second.

21. A block model applied to the rat surgery data

The following commands

```
fit1 <- regress(y~trt+site)
fit2 <- regress(y~trt+site, ~rat)
fit3 <- regress(y~trt, ~rat+site)
```

applied to the rat data give exactly the same point estimates for treatment effects $\hat{\tau} = (0.000, 0.335, 0.610)$. But the standard errors reported for the pairwise treatment contrasts are 0.269 by the first model, and a more realistic 0.442 by the second and third. The variance-component estimates are $\hat{\sigma}^2 = 1.443$ for the first model, $(\hat{\sigma}_0^2, \hat{\sigma}_b^2) = (0.880, 0.606)$ for the second, and $(0.880, 0.606, 0.039)$ for the third.

The second model, which treats the rat effects as independent and identically distributed random variables, is in line with the randomization scheme with rats as experimental units. It is not unreasonable to treat the site effects as identically distributed random variables, but independence of site effects is more difficult to justify because an anterior-posterior trend is always a possibility that needs to be checked in circumstances of this sort. This is most easily done by examination of the site means

$$\bar{y} = (9.792, 9.488, 9.613, 9.450, 9.046)$$

for which model 2 implies $\text{cov}(\bar{Y}_r, \bar{Y}_s) = \sigma_0^2/24\delta_{rs} + \sigma_b^2/24$. Alternatively, the fitted site coefficients are

$$(0.000, -0.304, -0.179, -0.342, -0.746),$$

which are the site averages minus the average for site 1. The variances and covariances of site contrasts depends only on σ_0^2.

There is a suggestion of a linear trend whose significance is not immediately apparent from the pairwise contrasts whose variances are $\sigma_0^2(1/24 + 1/24)$. By projection, we can extract from these site means or site coefficients a linear trend and a residual; the sums of squares are 0.234 on one degree of freedom for the linear trend, and 0.070 for the residuals on three degrees of freedom, so 77% of the variation among the site means is accounted for by the linear anterior-posterior trend. To construct a formal test, we could use the F-ratio $F = 0.234/(0.070/3) = 10.0$ on $1, 3$ degrees of freedom, which tells us that the variability is not evenly distributed across the polynomial subspaces. In this setting, it is better to compare both sums of squares with the estimated variance $\hat{\sigma}_0^2/24 = 0.037$ on 92 df. This gives $F = 6.32$ on 1,92 df for the linear trend, and $F = 0.63$ on 3,92 df for the remainder. The first F-value is close to the 99th percentile of the reference distribution, and the second is at the 40th percentile. So it appears from the first F-value that a linear trend is needed and, from the second, that the linear trend suffices.

22. Likelihoods for Gaussian models

All of the models considered in this section are multivariate Gaussian with mean and covariance

$$E(Y) = \mu = X\beta, \qquad \operatorname{cov}(Y) = \Sigma(\theta).$$

The mean model specifies a subspace $\mathcal{X} \subset \mathcal{R}^n$ spanned by the columns of X, which is assumed to have full rank $p \leq n$. In the second part, $\theta \mapsto \Sigma(\theta)$ is a function taking values in the space of positive definite matrices. The inverse matrix $W_\theta = \Sigma_\theta^{-1}$ gives the observation space a parameter-dependent inner product $\langle x, y\rangle_\theta = x'W_\theta y$, with θ-orthogonal projection matrices

$$P_\theta = X(X'W_\theta X)^{-1}X'W_\theta, \qquad Q_\theta = I_n - P_\theta.$$

For an observation point Y, the ordinary log likelihood function is

$$\begin{aligned} l(\beta, \theta) &= -\tfrac{1}{2}\log\det\Sigma_\theta - \tfrac{1}{2}(Y - X\beta)'W_\theta(Y - X\beta) \\ &= \tfrac{1}{2}\log\det W_\theta - \tfrac{1}{2}\|Y - \mu\|_\theta^2. \end{aligned}$$

For each fixed θ, the maximum-likelihood estimate of β satisfies

$$X\hat{\beta}_\theta = P_\theta Y$$

and the profile log likelihood for θ is

$$l(\hat{\beta}_\theta, \theta) = \tfrac{1}{2}\log\det W_\theta - \tfrac{1}{2}\langle Y, Q_\theta Y\rangle_\theta = \tfrac{1}{2}\log\det W_\theta - \tfrac{1}{2}\|Q_\theta Y\|_\theta^2.$$

The `regress()` function with the `kernel=0` option aims to maximize this expression over $\theta \in \Theta$. It reports the parameter values $\hat{\theta}$, $\hat{W} = \Sigma^{-1}(\hat{\theta})$ and

$$\hat{\beta} = (X'\hat{W}X)^{-1}X'\hat{W}Y, \qquad \operatorname{cov}(\hat{\beta}) = (X'\hat{W}X)^{-1},$$

together with the maximum value achieved by the log likelihood.

22.1. Residual likelihood. Residual likelihood means the marginal likelihood based on the residual. For present purposes, the residual is any linear transformation $Y \mapsto LY$ whose kernel is the mean-value subspace: $\ker(L) = \mathcal{X}$. Another way of saying the same thing is to identify the residual space with the quotient space $\mathcal{R}^n/\mathcal{X}$, observe that $W_\theta Q_\theta$ is the matrix of the induced inner product, so the residual log likelihood is

$$\tfrac{1}{2}\log\operatorname{Det}(W_\theta Q_\theta) - \tfrac{1}{2}Y'W_\theta Q_\theta Y,$$

where $\operatorname{Det}(A)$ denotes the product of the non-zero eigenvalues.

We now show directly that the preceding argument gives the correct likelihood. The condition $\ker(L) = \mathcal{X}$ means $Lx = 0$ if and only if $x \in \mathcal{X}$. It is helpful in what follows to pick a linear transformation $\mathcal{R}^n \to \mathcal{R}^{n-p}$, so that the matrix L of order $n - p \times n$ has full rank $n - p$. The distribution is $LY \sim N(0, L\Sigma_\theta L')$, and the residual log likelihood is

$$-\tfrac{1}{2}\log\det(L\Sigma_\theta L') - \tfrac{1}{2}Y'L'(L\Sigma_\theta L')^{-1}LY + \operatorname{const}(L, Y).$$

For a variety of reasons, it is desirable to avoid the explicit construction of L, and to express the residual log likelihood solely in terms of X. We observe first that $\Sigma_\theta L'(L\Sigma_\theta L')^{-1}L$ is a W_θ-orthogonal projection matrix whose kernel is $\mathcal{X}$. In other

words, $\Sigma_\theta L'(L\Sigma_\theta L')^{-1}L = Q_\theta$, so that the matrix of the quadratic form in the residual likelihood is

$$L'(L\Sigma_\theta L')^{-1}L = W_\theta Q_\theta,$$

which is the quotient-space inner-product matrix. Hence the residual log likelihood

$$-\tfrac{1}{2}\log\det(L\Sigma_\theta L') - \tfrac{1}{2}Y'W_\theta Q_\theta Y + \text{const}(L, Y)$$

is the same as the profile log likelihood except for a modification of the determinant. To ensure that every invertible function of the residuals gives exactly the same value, the constant is chosen to be $\frac{1}{2}\log\det(LL')$.

To simplify the determinant, write

$$H = \begin{pmatrix} L \\ X' \end{pmatrix}, \qquad H\Sigma H' = \begin{pmatrix} L\Sigma L' & L\Sigma X \\ X'\Sigma L' & X'\Sigma X \end{pmatrix}$$

in the form of a partitioned matrix in which H is invertible. Then

$$\det(H\Sigma H') = \det(\Sigma)\det(HH') = \det(\Sigma)\det(LL')\det(X'X).$$

In addition, using the formula for the determinant of a partitioned matrix,

$$\begin{aligned}\det(H\Sigma H') &= \det(L\Sigma L')\det\big(X'\Sigma X - X'\Sigma L'(L\Sigma L')^{-1}L\Sigma X\big) \\ &= \det(L\Sigma L')\ \det(X'(I-Q)\Sigma X) \\ &= \det(L\Sigma L')\ \det\big(X'X(X'WX)^{-1}X'X\big) \\ &= \det(L\Sigma L')\det{}^2(X'X)/\det(X'WX).\end{aligned}$$

It follows that

$$\frac{\det(L\Sigma L')}{\det(LL')} = \frac{\det(\Sigma)\ \det(X'WX)}{\det(X'X)},$$

so that the residual log likelihood reduces to

$$-\tfrac{1}{2}\log\det(\Sigma_\theta) - \tfrac{1}{2}\log\det(X'W_\theta X) + \tfrac{1}{2}\log\det(X'X) - \tfrac{1}{2}Y'W_\theta Q_\theta Y.$$

For numerical work, it is best to compute the determinantal factor

$$\det(W_\theta)\det(X'X)/\det(X'W_\theta X)$$

directly as the product of the non-zero eigenvalues of $W_\theta Q_\theta$.

The default option in most computer packages including `lm(y~...)` is first to compute the variance or covariance parameter $\hat\theta$ that maximizes the residual log likelihood. This value determines the fitted covariance matrix $\Sigma(\hat\theta)$ and its inverse $\hat W$. The regression parameter estimates are computed by weighted least squares $\hat\beta = (X'\hat WX)^{-1}X'\hat WY$, and their estimated covariance matrix is reported as $(X'\hat WX)^{-1}$.

22.2. Kernel subspaces. The program `regress()` has an option for specifying the kernel $\mathcal{K} \subset \mathcal{R}^n$ as the span of the columns of a matrix K, which may be different from X. Usually, $\mathcal{K} \subseteq \mathcal{X}$, but this is not strictly necessary mathematically and is not demanded. The chief purpose of this option is the correct computation of likelihood ratio statistics for hypotheses concerning treatment effects, and in such settings $\mathcal{K}$ is usually chosen to be the smaller of the two mean-value subspaces.

For the Gaussian model $Y \sim N(\mu, \Sigma)$, the log likelihood function with kernel $\mathcal{K}$ is

$$\tfrac{1}{2}\log\text{Det}(W_\theta Q_\theta^{\mathcal{K}}) - \tfrac{1}{2}(Y-\mu)'W_\theta Q_\theta^{\mathcal{K}}(Y-\mu),$$

where $Q_\theta^{\mathcal{K}}$ is the W_θ-orthogonal projection with kernel $\mathcal{K}$, $W_\theta Q_\theta^{\mathcal{K}}$ is the matrix of the inner product in $\mathcal{R}^n/\mathcal{K}$, and $\mathrm{Det}(\cdot)$ is the product of the non-zero eigenvalues. In particular, $\mathcal{K} = 0$ implies the ordinary likelihood, and $\mathcal{K} = 1$ implies the likelihood based on simple contrasts. The default option is $\mathcal{K} = \mathcal{X}$.

If $\mathcal{K}$ is a proper subspace of $\mathcal{X}$, $Q_\theta \mu$ is not zero, but the profile log likelihood for θ is

$$\tfrac{1}{2} \log \mathrm{Det}(W_\theta Q_\theta^{\mathcal{K}}) - \tfrac{1}{2} Y' W_\theta Q_\theta^{\mathcal{X}+\mathcal{K}} Y,$$

where $Q_\theta^{\mathcal{X}+\mathcal{K}}$ is the W_θ-orthogonal projection with kernel $\mathcal{X} + \mathcal{K}$.

As always, the reported regression coefficients are obtained by weighted least squares

$$\hat{\beta} = (X'\hat{W}X)^{-1} X'\hat{W}Y, \qquad \mathrm{cov}(\hat{\beta}) = (X'\hat{W}X)^{-1},$$

which is often reasonable for $\mathcal{K} \subset \mathcal{X}$, but not otherwise. If $\mathcal{K}$ is not a subspace of $\mathcal{X}$, the variance-component estimates and the log likelihood may be correct, but the reported regression coefficients are not kernel-adjusted, and are best ignored.

22.3. Likelihood and log likelihood ratios. For the rat data in section?? it is possible formally to test for the adequacy of a linear trend by comparing the log likelihood values achieved by two block-factor models using `regress()`. The larger model uses `site` as a five-level factor; the sub-model replaces it with the quantitative covariate `xsite=as.numeric(site)`. For technical reasons associated with the nature of the REML likelihood, this comparison is not entirely straightforward, and the default must be overridden. The option `kernel=0` implies ordinary maximum likelihood both for the estimation of variance components, and the reported log-likelihood value

```
fit0 <- regress(y~trt, ~rat, kernel=0)
fit1 <- regress(y~trt+xsite, ~rat, kernel=0)
fit2 <- regress(y~trt+site, ~rat, kernel=0)
2*(c(fit0$llik, fit1$llik,fit2$llik) - fit0$llik)
# 0.000 6.306 8.285
```

The likelihood-ratio statistic of 6.31 on one df points to a significant linear trend, and the subsequent increment of 1.98 on three df shows no evidence of departures from linearity of the anterior-posterior trend. The technical point here is that the log likelihood values reported by `regress()` for two or more models are not comparable unless the likelihoods are computed using the same σ-field, or the same kernel subspace. The REML default takes the model subspace $\mathcal{X}$ for the kernel, and the three models shown above have three different model subspaces, so the default REML likelihoods are not comparable.

For comparison, the reported REML log likelihood increments are $(0.000, 6.162, 7.689)$, which are not comparable. If the response is changed from y to $2y$, the REML increments are different.

It is also necessary to check for additivity of site and treatment effects. Because of the default method of estimation used by `regress()`, care is needed to avoid making inappropriate likelihood comparisons. The following choices for $\mathcal{K}$ are both mathematically valid options for testing whether site and treatment effects are additive:

```
K <- model.matrix(~site+trt)
fit0 <- regress(y~site+trt, ~rat, kernel=K)
fit1 <- regress(y~site*trt, ~rat, kernel=K)
2*(fit1$llik - fit0$llik) ### 8.893 on 8 df
fit0 <- regress(y~site+trt, ~rat, kernel=0)
fit1 <- regress(y~site*trt, ~rat, kernel=0)
2*(fit1$llik - fit0$llik) ### 9.280 on 8 df
```

The log likelihood reported by `regress()` with kernel set to $\mathcal{K}$ is the maximized marginal likelihood based on an arbitrary linear transformation TY such that $\ker(T) = \mathcal{K}$. The two choices shown above give different likelihood ratios because they are based on different data. For the comparison of two nested linear models, the recommended default for $\mathcal{K}$ is the smaller of the two model subspaces.

If we take $\mathcal{K}$ to be the larger subspace *site*trt*, the comparison is mathematically correct in the minimal sense that it is the log likelihood ratio based on TY. But all of the information about non-additivity resides in the kernel subspace, which is ignored in the marginal likelihood, so the likelihood ratio statistic is exactly zero.

```
K <- model.matrix(~site*trt)
fit0 <- regress(y~site+trt, ~rat, kernel=K)
fit1 <- regress(y~site*trt, ~rat, kernel=K)
2*(fit1$llik - fit0$llik) # 0.000
```

This code returns the likelihood ratio statistic of zero, which is mathematically correct because the distribution of any linear transformation TY such that $\ker(T) = \mathcal{K}$ is independent of site and treatment effects, whether additive or otherwise. This choice is not recommended for testing additivity!

22.4. Exercises.

The questions that follow refer to the following incomplete block design

blk	1	1	1	2	2	2	3	3	3	4	4	4
trt	1	2	3	4	1	2	3	4	1	2	3	4
y	6.22	3.97	5.09	6.81	5.4	4.51	4.18	4.06	4.65	3.72	2.97	2.85

Exercise 22.1: Calculate the log likelihood ratio statistic on one degree of freedom for testing the significance of the iid random block effect. This means fitting two nested models, both including a treatment effect

```
fit1 <- regress(y~trt);    fit2 <- regress(y~trt, ~blk)
```

and comparing twice the log likelihood increment with χ^2_1.

Exercise 22.2: When we say that the block effects are random, we do not mean that they are independent or Gaussian or that they are identically distributed. The matrix

```
V <- exp(-abs(outer(blk, blk, "-"))/10)
```

is positive semi-definite and defines a Gaussian process that is constant on blocks with identically distributed but positively correlated components. Repeat exercise 1 with `blk` replaced by `V`.

Exercise 22.3: Calculate the log likelihood ratio statistic on three degrees of freedom for testing the significance of the treatment effect, making allowance for additive block effects. This means fitting two nested models, both including an additive block effect

```
fit0 <- regress(y~blk)
fit3 <- regress(y~blk+trt, kernel=model.matrix(~blk))
```

The F-ratio is a monotone function of the log likelihood ratio. To compute the upper tail significance probability exactly, report the observed F-ratio as a percentile relative to the $F_{3,5}$ distribution. For an approximate upper tail value, you may compare twice the log likelihood increment with χ^2_3.

Exercise 22.4: Calculate the log likelihood ratio statistic on three degrees of freedom for testing the significance of the treatment effect, making allowance for additive iid Gaussian block effects. This means fitting two nested models, both including an iid block effect

```
fit0 <- regress(y~1, ~blk);    fit3 <- regress(y~trt, ~blk, kernel=1)
```

For significance testing, you may compare twice the log likelihood increment with χ^2_3 or report the observed F-ratio as a percentile relative to the $F_{3,5}$ distribution.

Exercise 22.5: The ratio of the residual variance in `fit3` in exercise 4 to that in `fit2` should be 7:11. Explain where this ratio comes from.

Exercise 22.6: Let e be the indicator vector for one unit, the first, say, and let K be the matrix generated by the model formula `~trt+e`. Explain why the commands

```
fit2 <- regress(y[-1]~trt[-1], ~blk[-1]);  c(fit0$llik, fit0$sigma)
fit1 <- regress(y~trt, ~blk, kernel=K);    c(fit1$llik, fit1$sigma)
```

produce exactly the same output `-4.3817 0.7866 0.7364`.

22.5. Further exercises. The data for the following exercises are taken from Bailey (2004, p. 155). An experiment to test three alfalfa varieties used six blocks in the field. Each block was split into three sub-blocks, or whole plots; each variety was assigned to one of the whole plots in each block. Each whole plot was divided into four subplots, one subplot for each of four cutting schemes. In the summer, the alfalfa on each subplot was cut twice, the second cutting taking place on 27 July. The cutting schemes were (A) no further cut; or a third cut on Sept 1, Sept 20, or Oct 7. The table below gives the yields in the following year in tons/acre. The values have been rearranged from the randomized field order to clarify the pattern.

	3rd cutting	Block					
Variety	cutdate	I	II	III	IV	V	VI
Ladak	A	2.17	1.88	1.62	2.34	1.58	1.66
	Sept1	1.58	1.26	1.22	1.59	1.25	0.94
	Sept20	2.29	1.60	1.67	1.91	1.39	1.12
	Oct7	2.23	2.01	1.82	2.10	1.66	1.10
Cossack	A	2.33	2.01	1.70	1.78	1.42	1.35
	Sept1	1.38	1.30	1.85	1.09	1.13	1.06
	Sept20	1.86	1.70	1.81	1.54	1.67	0.88
	Oct7	2.27	1.81	2.01	1.40	1.31	1.06
Ranger	A	1.75	1.95	2.13	1.78	1.31	1.30
	Sept1	1.52	1.47	1.80	1.37	1.01	1.31
	Sept20	1.55	1.61	1.82	1.56	1.23	1.13
	Oct7	1.56	1.72	1.99	1.55	1.51	1.33

Unless otherwise indicated, you may assume in what follows that all effects are additive.

The output of the R function

```
regress(yield~variety+cutdate, ~block+wholeplot)
```

includes variety and treatment effects plus three estimated variance components

```
Id         0.0288
block      0.0578
wholeplot  0.0269
```

Exercise 22.7: The variety means are 1.666, 1.572, 1.553 for Ladak, Cossack and Ranger. On the basis of the output shown above, compute a standard error for the mean difference of yields for two varieties.

Exercise 22.8: The mean yields by cutting date are 1.781, 1.341, 1.574, 1.6911. On the basis of the output shown above, compute a standard error for the difference in yields between two cutting dates.

Exercise 22.9: The researcher, who had learned about interaction in her graduate classes, decided it would be best to check whether the effect of the cut date might depend on the variety. Deciding to use a likelihood ratio test, she typed the following commands

```
fit0 <- regress(yield~variety+cutdate, ~block+wholeplot)
fit1 <- regress(yield~variety*cutdate, ~block+wholeplot)
2*(fit1$llik - fit0$llik)
```

and she was surprised by the value -13.95 returned at the last line. Why was she surprised? What do you conclude from this statistic?

Exercise 22.10: The researcher's colleague typed the same commands, but he had the yields measured in tonnes/Ha. What 'log-likelihood-ratio' value did the colleague's computer return at the last line?

Exercise 22.11: As a partition of the observational units, *wholeplot* is a sub-partition of *block*, so the vector subspace $wholeplot \subset \mathcal{R}^{72}$ contains the subspace *block*. Consequently, $wholeplot + block = wholeplot$. The researcher, who had been a math major in college, reckoned that if u is a random block effect, and v is a random wholeplot effect, the vectors $u \in block$ and $v \in wholeplot$ imply $u + v \in wholeplot$ whether the effects are random or not. However, the R commands with *block* and *wholeplot* as random effects

```
fit  <- regress(yield~variety+cutdate, ~block+wholeplot)
fitb <- regress(yield~variety+cutdate, ~block)
fitw <- regress(yield~variety+cutdate, ~wholeplot)
2*c(fit$llik, fitb$llik, fitw$llik)
```

produce the values `35.77 22.21 29.92` at the last line.

How do you respond to the researcher's subspace-inclusion statement? Is the reasoning correct? If it is incorrect, where is the error? If it is correct, why are the first and third fitted models different?

Exercise 22.12: What do you conclude from the computer output in the previous part?

Exercise 22.13: The R commands

```
K <- model.matrix(~variety+cutdate)
fit0 <- regress(yield~variety+cutdate, ~block+wholeplot, kernel=K)
fit1 <- regress(yield~variety*cutdate, ~block+wholeplot, kernel=K)
2*(fit1$llik - fit0$llik)
```

produce the value 7.89 at the last line. What do you conclude from this?

Exercise 22.14: The researcher's colleague typed the same commands, but he had the yields measured in tonnes/Ha. What value did the colleague's computer return at the last line?

23. Recovery of inter-block information

Consider an incomplete-block design with four blocks, four treatments, and three plots per block. The treatments 1–4 or $t_1, \ldots, t_4$, are assigned at random to the units subject to the conditions for a balanced incomplete-blocks design, the outcome of which is as follows:

Treatment assignment				Response values			
Blk I	t_3	t_1	t_4	Blk I	4.23	5.77	5.85
Blk II	t_1	t_3	t_2	Blk II	8.05	3.41	4.51
Blk III	t_2	t_4	t_1	Blk III	4.37	4.06	4.65
Blk IV	t_3	t_2	t_4	Blk IV	3.53	3.16	2.85

We may estimate the treatment effects and their standard errors adjusted for additive block effects in one of two ways. The simplest is to accommodate block effects as arbitrary additive constants in the mean subspace, $trt + blk$ of dimension seven, with block and treatment parameters estimated in the usual way by ordinary least squares. This is an eight-parameter model with seven regression coefficients and one variance. The alternative investigated here is to impose the constraint that the block effects are zero-mean Gaussian random variables, so that their effect is entirely on response variances and covariances. Specifically, if the four block effects are iid $N(0, \sigma_1^2)$, the response distribution is Gaussian with moments

$$E(Y_u) = \tau_{t(u)}, \quad \operatorname{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 B_{u,u'}.$$

This is six-parameter model with four regression coefficients and two variance components. Although the dimension is reduced by the imposition of a distributional constraint, the second model is not a sub-model of the first.

The fitted treatment effects as contrasts with level 1, and their standard errors as reported by

$$\texttt{regress}(\texttt{y}\tilde{}\,\texttt{trt} + \texttt{blk}) \quad \text{and} \quad \texttt{regress}(\texttt{y}\tilde{}\,\texttt{trt}, \tilde{}\,\texttt{blk})$$

are as follows:

Parameter	Est_0	SE_0	Est_1	SE_1
$\tau_2 - \tau_1$	−1.622	0.980	−1.977	0.942
$\tau_3 - \tau_1$	−2.295	0.980	−2.389	0.942
$\tau_4 - \tau_1$	−1.338	0.980	−1.723	0.942
σ_0^2	1.281		1.281	
σ_1^2			0.226	

The block coefficients have been suppressed in the output for first model, and the intercept has been suppressed in both.

To understand where these treatment-effect estimates come from, and why they differ from one version of the block model to the other, we label the units 1–12 by rows in the diagram of treatment assignments. Observe that, in either version of the model, the three linear contrasts

$$\begin{aligned} Y_6 - Y_4 &\sim N(\tau_2 - \tau_1, 2\sigma_0^2) \\ Y_7 - Y_9 &\sim N(\tau_2 - \tau_1, 2\sigma_0^2) \\ Y_{11} - Y_2 - (Y_{10} + Y_{12} - Y_1 - Y_3)/2 &\sim N(\tau_2 - \tau_1, 3\sigma_0^2) \end{aligned}$$

are unbiased for the treatment effect $\tau_2 - \tau_1$. They are also mutually independent. The optimum unbiased linear combination has coefficients $(3, 3, 2)/8$ adding to one and inversely proportional to the variances

$$\begin{aligned} 8T_{21} &= 3(Y_6 - Y_4) + 3(Y_7 - Y_9) + 2(Y_{11} - Y_2) - (Y_{10} + Y_{12} - Y_1 - Y_3) \\ &= Y_1 - 2Y_2 + Y_3 - 3Y_4 + 3Y_6 + 3Y_7 - 3Y_9 - Y_{10} + 2Y_{11} - Y_{12} \\ \operatorname{var}(8T_{21}) &= 48\sigma_0^2. \end{aligned}$$

For the data shown above, the numerical value is $T_{21} = -1.6225$ with variance $3\sigma_0^2/4$, or estimated standard error 0.980 as reported in the output for the fixed-effects model.

The parameter estimates and standard errors for the random-effects model are different because the exchangeability assumption for the block effects affords access to inter-block contrasts. Specifically, the difference between block totals (IV versus I) is distributed as

$$D_{21} = Y_{10} + Y_{11} + Y_{12} - Y_1 - Y_2 - Y_3 \sim N(\tau_2 - \tau_1,\, 6\sigma_0^2 + 18\sigma_1^2).$$

Moreover, this combination is independent of all intra-block contrasts such as T_{21}. The numerical value is -6.310.

From the numerical estimates of the two variance components, $\mathrm{var}(T_{21}) = 0.961$ and $\mathrm{var}(D_{21}) = 11.751$, it is apparent that the additional information coming from inter-block contrasts is not large. The optimum linear combination is

$$(T_{21}/0.961 + D_{21}/11.751)/(1/0.961 + 1/11.751) = -1.977$$

with variance 0.942^2, as reported in the output shown above.

The analysis thus far focuses on treatment contrasts, taking the variance components as given. To understand where the variance-component estimates come from, consider the residual sum of squares $\|QY\|^2$ for the subspace *trt* of dimension four and the subspace *blk+trt* of dimension seven. Their expected values under the random effects model are

$$E(Y'QY) = \sigma_0^2\,\mathrm{tr}(Q) + \sigma_1^2\,\mathrm{tr}(QB),$$

which reduces to $8\sigma_0^2 + 8\sigma_1^2$ for the subspace *trt*, and $5\sigma_0^2$ for the subspace *trt+blk*. The numerical values of the two residual sums of squares are 12.055 and 6.405, giving estimates

$$\hat\sigma_0^2 = 6.405/5 = 1.281, \qquad \hat\sigma_1^2 = (12.055 - 8\hat\sigma_0^2)/8 = 0.226.$$

For this balanced setting, these are exactly the REML estimates reported above.

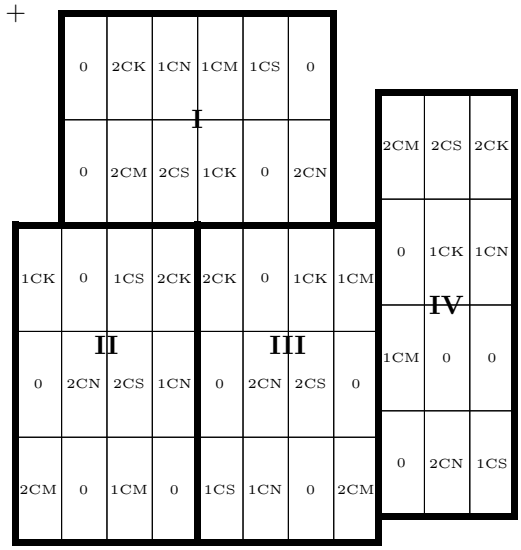
If the block mean square is smaller than the residual mean square the estimate of the block variance will be negative (or zero if a positivity condition is in force). Provided that the overall combination remains strictly positive definite, most computer packages tolerate negative coefficients but allow positivity constraints.

If each treatment appears an equal number of times in every block, the block totals are uninformative for treatment contrasts. For a complete-block design, no information is gained from the exchangeability assumption for block effects.

24. Spatial effects versus block effects

One of the objections to the use of random constants on contiguous rectangular blocks as a device for accommodating spatial correlation is that the implied value, considered as a function $\mathcal{R}^2 \to \mathcal{R}$, has a discontinuity along each block boundary. If the boundaries are based on topography or land management practices the discontinuities in the model may also be present in the field, which is good. But more often than not the blocks are defined in a slightly arbitrary manner, and the discontinuities are an unintended but unavoidable artifact of blocking.

The plan of the field layout for the experiment described in the next section provides an opportunity to compare the effectiveness of the iid random block model with a range of spatial alternatives, each of which is a random process whose realizations are spatially continuous. Each of the four blocks is a compact rectangular region consisting of twelve plots, and each block contains the same set of eight active treatments and four controls. Each plot has an aspect ratio of 2.8:1, the longer direction running north-south as shown in the plan below.



Actual field layout of 48 plots in four blocks. Experiment using fumigants to control eelworms in an oat field. (Bailey, 2008, p. 73).

In the exercises that follow, we consider only the simplest isotropic Gaussian process with exponential covariance function proportional to $e^{-\|x-x'\|/\rho}$ with range $\rho > 0$. Provided that ρ is at least 5–10 plot widths, we find that the spatial model gives an appreciably better fit than the iid block model. Part of the improvement is due to the accommodation of spatial variation within blocks, but an even larger part is due to the elimination from the block model of discontinuities that are not present at block boundaries in the field. The conclusions are not sensitive to the particular value of ρ, which means that any moderately large value suffices. However, the infinite limit is well defined and there is compelling evidence that this limit works reasonably well for a wide range of agricultural processes, as it does here.

The infinite limit with generalized covariance function $-\|x - x'\|^p$ is an instance of an incomplete Gaussian process defined on planar contrasts. Here, $p = 1$ for the exponential limit, $0 < p < 2$ in general, and smoother processes have larger index. Such processes defined on contrasts are sometimes called 'intrinsic'. This aspect of things causes no difficulty in likelihood calculations provided that the likelihood is based on

spatial contrasts, i.e., on differences $Y_u - Y_{u'}$ between points or plot totals. In other words, the kernel subspace must contain $\mathbf{1}$, so `kernel=0` is not permitted.

The power index controls the smoothness of the spatial process. A process observed at discrete points or averaged over plots seldom contains much information about local properties, so there is rarely much information in the data about p. Generally speaking, however, smaller values $p \leq 1$ tend to work slightly better for naturally-occurring processes.

For $p = 2$, the planar process with generalized covariance function

$$K(x, x') = \|x - x'\|^2 \log \|x - x'\|$$

is also an intrinsic process, but its kernel is the three-dimensional vector subspace of linear functions $\mathcal{R}^2 \to \mathcal{R}$, (meaning linear in the polynomial sense that the degree is one or less). This process is smooth with continuous first derivatives, which means that it is not well-suited for most naturally-occurring processes. Nonetheless, there is no difficulty handling such processes provided that the kernel argument declared in `regress(...)` contains the subspace of linear functions.

24.1. A block design for testing eelworm fumigants. Bailey (2008, p. 73) describes a field experiment conducted at Rothamsted in 1935 which was designed to test the effectiveness of four chemical fumigants for reducing the eelworm count in the soil where oats were grown. The same experiment is described in Cochran and Cox (p. 46) as a randomized-blocks design without spatial information. Each fumigant was tested with a single and a double dose. Treatments were assigned randomly to the plots subject to standard constraints for a randomized blocks design with eight 'active' treatment levels and four controls per block. After the first eelworm count was made in spring, the assigned fumigants were ploughed in and oats were sown. After harvest, a sample of 400 grams of soil was taken from each plot, and the number of eelworm cysts was counted. The log ratio of counts is the variable of primary interest.

Seven variables are provided in the file `eelworm.dat` as follows:

```
x1, x2, block, dose, fumigant, eelcount0, eelcount1
```

Relative to the north-west reference point marked + in the plan, the northwest corner of plot i is at $(\texttt{x2}(i), 2.8\texttt{x1}(i))$ in plotwidth units. The four fumigants `CN`, `CM`, `CS` and `CK` are coded 1–4 in the file, and the plot assignment is also shown in the plan. The randomization probabilities were not based on spatial proximity, but on block membership, so adjacent plots may be assigned the same treatment if they occur in different blocks, for example `2CK` in blocks II and III.

For questions 1–7, take the log ratio of eelworm counts as the response.

Exercise 24.1: Analyze the log ratio data as a randomized blocks design with additive iid block effects, making no assumptions about linearity or additivity of dose and fumigant effects, but assuming that treatment and block effects are additive. Comment on the magnitude of the estimated block variance relative to the residual variance. Report the REML log likelihood ratio statistic for testing the hypothesis that the block variance is zero.

Exercise 24.2: The R command `anova(lm(y~block+fumigant:fdose))` (with three qualitative factors) produces an anova table. Under the model of additive random block effects, show that the expected value of the between-blocks mean square is $\sigma_0^2 + k\sigma_1^2$ for some particular number k. Hence obtain the natural unbiased estimate of the between-blocks variance. Explain how these calculations are related to the REML estimates in part 1.

Exercise 24.3: For every range $\rho > 0$, the exponential function $K(x, x') = \rho\, e^{-\|x-x'\|/\rho}$ is positive definite on Euclidean spaces. The planar Gaussian process $Z(x)$ with covariance function K is everywhere continuous, but nowhere differentiable. In particular, there is no discontinuity at the boundary between two blocks. Repeat the analysis in the first part with the iid block factor replaced by this continuous process with range $\rho = 10$ and variance $\sigma_1^2 \geq 0$ to be estimated.

For purposes of this exercise, the covariance of response values for two plots may be taken as

$$\operatorname{cov}\big(Y(u), Y(u')\big) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 K(x(u), x(u'))$$

where $x(u)$ is the reference point or central point in plot u. (Ordinarily, if Y_u is a plot total, one should integrate over pairs (x, x') in $u \times u'$.)

Comment on the parameter estimates and their standard errors as reported by the block model and the spatial model.

Exercise 24.4: Discuss briefly how you might test whether there is (i) additional spatial variation after taking block variation into account; or (ii) additional inter-block variation after taking spatial variation into account. For this setting, all variance components are necessarily non-negative. Compute a likelihood-ratio statistic and report the significance level for each part.

Exercise 24.5: This exercise concerns the long-range limit $\lim_{\rho\to\infty} K(x, x')$ of the exponential covariance function. For large ρ, we have

$$K(x, x') = \rho - \|x - x'\| + O(\rho^{-1}) \simeq \text{const} - \|x - x'\|.$$

However large it may be, the additive constant has no effect on the variance of contrasts $Z(x) - Z(x')$, so the limiting covariance of contrasts is well-defined:

$$\operatorname{cov}\big(Z(x_1) - Z(x_2),\, Z(x_3) - Z(x_4)\big) = -\|x_1 - x_3\| - \|x_2 - x_4\| + \|x_1 - x_4\| + \|x_2 - x_3\|.$$

If covariances can be restricted to spatial contrasts, the long-range limit of $K(x, x')$ may be taken as $-\|x - x'\|$, ignoring the constant. This limit is said to be [conditionally] positive definite on contrasts. Repeat the analysis in part 3 with K replaced by its long-range limit. Comment briefly on the standard errors that are reported for the fitted regression coefficients, particularly that reported for the intercept.

Exercise 24.6: In the fitted block model, all of the active treatment contrasts have the same variance. In the spatial model, some pairwise contrasts have smaller estimated variance than others. Explain this phenomenon.

Exercise 24.7: Consider now the sub-model in which the mean response for each fumigant is linear in the dose. Using the long-range limit model for spatial correlations, test the hypothesis of linearity. Report an appropriate likelihood ratio statistic, explaining exactly which version of the likelihood ratio you have used, and why. Compute the tail p-value and explain whether the evidence points towards linearity or non-linearity.

Exercise 24.8: Using the long-range limit model for spatial correlations, fit the linear dose model with log(initial worm count) as a covariate. Report the estimated coefficient and its standard error.

Exercise 24.9: The planar Gaussian process $Z(x)$ with covariance function

$$\mathrm{cov}(Z(x), Z(x')) \propto \big(1 + \|x - x'\|/\rho\big) \exp(-\|x - x'\|/\rho)$$

is everywhere continuous with two continuous spatial derivatives. This is sometimes called the $\mathrm{Bessel}_{3/2}$ process as opposed to the rougher $\mathrm{Bessel}_{1/2}$ process which has the exponential covariance function and no spatial derivatives. Compare the effectiveness of the smoother with the rougher process for accommodating spatial variation in eelworm counts. Use the most favourable values of ρ, which are roughly three plotwidths for the smooth process and infinity for the rough process.

Exercise 24.10: Summarize your conclusions in one page.

25. Probability model

Statistical inferences are based on a probability model F, which is not given as an an accompaniment with the data, but is invented as a mathematical framework within which statistical inferences may be made—in essence by computing the conditional probability given the data. We defer to later the strategic problem of constructing a suitable *family* of probability distributions, parameter estimation within the family, model testing, predictive distributions, and so on. For the moment the focus is on the nature of the probability distribution F, what its properties are, how it incorporates inhomogeneities among the units, how it incorporates treatment effects, and so on.

We begin with the simplest setting of an observational study in which there is no intervention, no treatment effect, $Y = (Y_u)_{u\in\mathcal{U}}$ is a process indexed by the units, and F is the probability distribution of Y.

25.1. Congruent samples. Recall that a covariate is a function $\mathcal{U} \to \mathcal{X}$ on the units, and a relationship is a function on pairs of units, whose purpose is to summarize the allowable inhomogeneities in outcome probabilities. Let x be the set of all covariates and relationships, which are, by definition, registered at baseline. Two samples S, S' (finite ordered subsets of $\mathcal{U}$) are said to be congruent if the sample configurations are equal: $x[S] = x[S']$. Since the identity function on $\mathcal{U}$ is automatically a registered relationship, congruence implies $\#S = \#S'$, so congruent samples consisting of distinct units are of equal size. Congruent samples also have the same list of covariate values, the same list of pairwise relationships and so on. In short, congruent samples are indistinguishable by registered baseline information.

Congruence is an equivalence relation, which partitions the set of samples into congruence classes.

25.2. Exchangeable distributions. A probability model F determines the distribution F_S for the response $Y[S] = (Y_u)_{u\in S}$ on the sample S: $F_S(A) = \mathrm{pr}(Y[S] \in A)$. It is a fundamental statistical principle, but an assumption nonetheless, that congruent samples must have the same response distributions

$$x[S] = x[S'] \Longrightarrow F_S = F_{S'}.$$

This statement means that the finite-dimensional random variables $Y[S]$ and $Y[S']$ have the same joint distribution whenever $x[S] = x[S']$; there is one distribution for each congruence class.

It is obvious that inference in the minimal sense of extension from one sample to another is impossible without some notion of exchangeability—the same distribution for congruent samples. In that sense, the assumption is obvious and unavoidable. But it is not innocuous, and it forces us to consider carefully what is and what is not registered at baseline. Consider a spatial process where the set of units is $\mathcal{U} = \mathcal{R}^2$, and the only registered inhomogeneity is the inter-unit Euclidean distance $x(u, u') = \|u - u'\|$. Since $x(u, u) = 0$ for every u, all samples of size one are congruent. Exchangeability modulo x implies that the one-dimensional distributions $Y(u)$, and $Y(u')$ are equal. In addition, two samples $S, S' \subset \mathcal{R}^2$ differing by planar translation and orthogonal transformation $\mathcal{R}^2 \to \mathcal{R}^2$ are congruent, so exchangeability in this setting is equivalent to stationarity and isotropy. If this is not the intention, it is essential to register further covariates or relationships, perhaps of higher order, that are needed to discriminate between the distributions. Statistical modelling is a balance between countervailing forces: it is counterproductive to exclude functions that are needed, and it is counterproductive to include functions that are not needed.

25.3. Consistent distributions. For notational simplicity, we suppose that $Y{:}\mathcal{U} \to \mathcal{R}$ is a process indexed by $u \in \mathcal{U}$ taking values in the state space $\mathcal{R}$. Suppose that $S \subset S'$ is a sub-sample of S' of size $m \leq n$. Kolmogorov consistency requires that F_S be the marginal distribution of $F_{S'}$ after integrating out the $n - m$ variables $(Y_u)_{u \in S' \setminus S}$. In other words,

$$F_S(A) = F_{S'}(A \times \mathcal{R}^{n-m})$$

for every event $A \subset \mathcal{R}^S$ generated by the variables $Y[S]$. However, $F_S(A)$ is a function of $x[S]$, while the right side is, in general, a function of $x[S']$, including values $x(u)$ for $u \notin S$. The Kolmogorov condition for a stochastic process requires that the marginal distribution $F_{S'}(A \times \mathcal{R}^{n-m})$ on the sub-sample $S \subset S'$ should coincide with $F_S(A)$.

This condition is easy to overlook because it is easy to satisfy in settings where the components of Y are independent. But it is not so easily satisfied in other settings, and it does sometimes fail, even in settings where the components are independent. One instance of failure is described in the example that follows.

Mutual consistency of the finite-dimensional distributions is a logical condition without which it is impossible to determine the conditional distribution of $Y[S']$ given $Y[S]$, which is the essential ingredient, the *sine qua non* of statistical inference. Such finite-dimensional conditional distributions are sometimes called predictive distributions. The view adopted here is somewhat broader than prediction narrowly construed. It also includes inferences for 'population parameters', such as sub-population averages, by taking $S' = \mathcal{U}$ and considering $F(A \,|\, Y[S])$ for selected events A in the tail σ-field.

Example 25.1 *For an example of an observational study where the Kolmogorov condition fails, suppose that each observational unit $u \in \mathcal{U}$ is a newborn lamb, the only covariates being sex (M/F) and sib count. A sib count of zero implies a single lamb, and one implies a twin. Ewe is an equivalence relation on $\mathcal{U}$ taking the value one if u and u' have the same mother. The response $Y(u)$ is weight in kg. at six weeks. The marginal distribution is $N(16.0, 1.2)$ for single males and $N(15.0, 1.2)$ for single females. The values for twin pairs are also Gaussian, but correlated ($\rho = 0.15$) with variance 1.0. The means are $MM : (13.6, 13.6)$, $MF : (14.4, 11.2)$ and $FF : (12.0, 12.0)$, depending on the sexes, with FM the transpose of MF.*

The point here is not the pair correlation or the dependence of the variance on sib count. Kolmogorov's consistency condition fails because the response distribution for lamb u depends not only on the declared covariates sex and sib count of u, but also, if the sib count is one, on the sex of the twin sibling $u' \neq u$. Given the declared covariates, every sample of size one belongs to one of four types or congruence classes, single male, single female, twin male or twin female. The distributions for the first two classes are $N(16.0, 1.2)$ and $N(15, 1.2)$ respectively. From the information given, the distribution for a twin male is either $N(13.6, 1.0)$ if the sib is male, or $N(14.4, 1.0)$ if the sib is female; the distribution for a twin female is either $N(11.2, 1.0)$ if the sib is male, or $N(12.0, 1.0)$ if the sib is female. Males whose twin sibling is female are, on average, 0.8 kg heavier than males whose twin sibling is male; females whose twin sibling is female are, on average, 0.8 kg heavier than females whose twin sibling is male. Provided that there are no triplets, the difficulty may be evaded by extending the set of covariates to include sex(sib(u)) as a covariate taking values in $\{M, F, \star\}$, where $\star$ means no sib. This extension makes six singleton congruence classes as needed. Alternatively, each ewe may be treated as one observational unit, in which case the state space varies from one unit to another, either $\mathcal{R}$ for ewes with one lamb, or $\mathcal{R}^2$ for ewes with twins.

26. Statistical model

26.1. Null model. In the absence of treatment intervention, a statistical model is a non-empty family

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

in which each element F_θ is a probability distribution on the same space, satisfying the exchangeability and consistency conditions listed in the preceding section. In particular, $x[S] = x[S']$ implies $F_{\theta,S} = F_{\theta,S'}$; congruent samples have the same response distribution for every $\theta \in \Theta$.

A statistical model is a parametric family of probability distributions on the space of outcomes. The model is said to be finite, finite-dimensional or infinite-dimensional if the parameter space is finite, finite-dimensional or infinite-dimensional. An infinite-dimensional parametric model is frequently called 'non-parametric'. For an example of the latter, let Θ be the set of Borel probability distributions on the real line, and let $F_{\theta,S} = \theta^S$ be the product distribution on $\mathcal{R}^S$. This is equivalent to saying that the components $(Y_u)_{u\in\mathcal{U}}$ are independent and identically distributed $Y_u \sim \theta$ for some distribution θ on the real line. Non-parametric does not imply the absence of parameters or freedom from distributional assumptions such as exchangeability or independence of components.

26.2. Group action. A group $\mathcal{G}$ is said to act on a space $\mathcal{X}$ if, to each $g \in \mathcal{G}$ there corresponds a transformation $g^\dagger\colon \mathcal{X} \to \mathcal{X}$ such that the mapping $g \mapsto g^\dagger$ is a homomorphism preserving group properties. A homomorphism has two implications: (i) if g is the group identity, $g^\dagger$ is the identity transformation $\mathcal{X} \to \mathcal{X}$; (ii) composition: $(g_1 g_2)^\dagger = g_1^\dagger g_2^\dagger$. These conditions imply that $(g^{-1})^\dagger = (g^\dagger)^{-1}$; the image of the group inverse is the inverse of the image, so each $g^\dagger$ is invertible $\mathcal{X} \to \mathcal{X}$. In every instance of group action in these notes, $\mathcal{G}, \mathcal{X}$ are topological spaces, and the group action $(g, x) \mapsto g^\dagger x$ is also continuous (in both g and x). We say that the group acts continuously on $\mathcal{X}$.

The group action is called faithful if the homomorphism $g \mapsto g^\dagger$ is one-to-one. Equivalently, there is exactly one $g \in \mathcal{G}$ such that $g^\dagger$ is the identity $\mathcal{X} \to \mathcal{X}$.

For most of the examples that follow, $\mathcal{G}$ is an additive group, either $\mathcal{R}$ or $\mathcal{R}^k$.

26.3. Treatment effects. Treatment is a baseline intervention, which gives rise to a modification of the response distribution; mathematically speaking, that modification is a group action $\mathcal{F} \to \mathcal{F}$ on the statistical model. Consider first the homogeneous case in which T is the randomization outcome, $Y[S]$ is independent of T, and $F_0 \in \mathcal{F}$ is the joint null distribution. The non-null distribution with treatment effect $g \in \mathcal{G}$ is obtained by group action $F_g = g^\dagger F_0$, which leaves the randomization distribution invariant. In effect, $\mathcal{G}$ acts on the space of conditional distributions given T.

To understand better what this means, consider the effect on the response distribution for one experimental unit u, comprising multiple observational units, say one plant or animal over multiple days in a study of growth patterns. If F_0 is the null distribution of Y_u, i.e., the conditional distribution given $T_u = 0$, and $g \in \mathcal{G}$ is the treatment effect (active versus control), then $F_g = g^\dagger F_0$ is the conditional distribution given $T_u = 1$. Conversely, if the roles are reversed, the treatment effect (control versus active) is the group element that sends F_g to F_0, which is g^{-1}, and is unique if the action is faithful. The important point here is that both distributions belong to the same family $\mathcal{F}$, and the set of distributions $(F_g)_{g\in\mathcal{G}}$ constitutes one group orbit in $\mathcal{F}$.

If treatment is qualitative with k levels, and there is one independent parameter for each level, $\mathcal{G}$ is usually replaced with $\mathcal{G}^k/\mathcal{G} \cong \mathcal{G}^{k-1}$, so that $g \in \mathcal{G}^k/\mathcal{G}$ has components

$(g_0, g_1, \ldots, g_{k-1})$ with $g_0 = I$, the group identity. If F_0 is the response distribution for the control level, $F_{g_r} = g_r^\dagger F_0$ is the response distribution for treatment level r, and the k distributions $F_0, \ldots, F_{g_k}$ lie ons the same $\mathcal{G}$-orbit.

26.4. Examples of group action. Every group action on the state space or outcome space has an induced action on probability distributions. Suppose that $\mathcal{G} = \mathcal{R}$, that the state space is the real line and that $\mathcal{F} = N(\star, \star)$ is the family of Gaussian distributions on the real line. Translation is a group action on the state space $y \mapsto y+g$ which also acts on distributions, sending $N(\mu, \sigma^2)$ to $N(\mu + g, \sigma^2)$; scalar multiplication is also a group action on the state space $y \mapsto ye^g$ which also acts on distributions, sending $N(\mu, \sigma^2)$ to $N(\mu e^g, (\sigma e^g)^2)$. Multiplication on the positive real line $y \mapsto e^g y$ has an induced action on gamma distributions, sending $\mathrm{Ga}(\mu, \nu)$ to $\mathrm{Ga}(e^g\mu, \nu)$ keeping the index fixed.

The transformation that sends $N(\mu, \sigma^2)$ to $N(\mu, \sigma^2 e^g)$ is a continuous group action on the family of Gaussian distributions; it associates with each $g \in \mathcal{G} = \mathcal{R}$ a transformation $g^\dagger\colon \mathcal{F} \to \mathcal{F}$ on Gaussian distributions by variance multiplication. This group action is different in a fundamental way from those described in the preceding paragraph, each of which is an action on probability distributions that is induced by an action $y \mapsto g^\dagger y$ on the state space. Every group action on the state space has an induced action on probability distributions, but not every group action on probability distributions is associated with an action on values in the state space.

The great majority of examples of group action in applied statistics are not induced through an action on the state space. Exponential tilting by $g \in \mathcal{R}$ is a group action which sends Poisson distributions to Poisson distributions $\mathrm{Po}(\mu) \mapsto \mathrm{Po}(\mu e^g)$, and binomial distributions $B(m, \pi)$ to $B(m, g^\dagger\pi)$ by odds multiplication $g^\dagger\pi/(1 - g^\dagger\pi) = e^g\pi/(1 - \pi)$. Likewise, hazard multiplication is a group action on survival distributions. Each distribution F on $\mathcal{R}^+$ has a survivor function $F((t, \infty))$, and the group action sends F to $g^\dagger F$

$$-\log(g^\dagger F)((t,\infty)) = -\log F((t,\infty))\, e^g.$$

by hazard multiplication. (The hazard function is the negative derivative of the log survivor function.)

In all settings, the parameter space for treatment effects is a group $\mathcal{G}$, not necessarily Abelian, which acts on probability distributions: $g^\dagger\colon \mathcal{F} \to \mathcal{F}$. For the simplest setting where the treatment effect is a scalar, the group is the real numbers with addition; the group action is a part of the model specification. Translation, exponential tilting, and hazard multiplication are three instances of group action commonly encountered in applied work. In more complicated settings where treatment has two qualitatively different effects, one on the mean and one on the variance, the group may be $\mathcal{R}^2$ with addition, or even the set $\mathcal{C}_0(\mathcal{R})$ of continuous functions on the real line. But, in general, parsimony favours smaller groups.

For an example of a slightly unusual group action, let $\mathcal{G} = \mathcal{R}^+$ with multiplication, acting on distributions on the real line as follows:

$$(g^\dagger F)(x) = \frac{g\, F(x)}{1 - F(x) + g\, F(x)}.$$

Here $F(x) = F((-\infty, x])$ is the cumulative distribution function, $g\, F(x)$ is scalar multiplication, and $(g^\dagger F)(x)$ is the transformed cumulative distribution function.

Exercise 26.1: Let $\mathcal{U}$ be the set of natural numbers, and let $x(u) = u \pmod 2$ be the parity, zero for even numbers and one for odd numbers. Let Y be a real-valued process with independent components such that, for $\theta = (\mu_0, \mu_1, \sigma)$, the 1-D marginal distribution F_θ is either $C(\mu_0, \sigma)$ if $x_u = 0$ or $N(\mu_1, \sigma^2)$ if $x_u = 1$. The parameter space

is $\Theta = \mathcal{R}^2 \times \mathcal{R}^+$, and $C(\mu, \sigma)$ is the Cauchy distribution with median μ and probable error σ, or inter-quartile range 2σ.

Which of the following are group actions $\mathcal{F} \to \mathcal{F}$ for $\mathcal{G} = \mathcal{R}$?

$$
\begin{aligned}
&\text{(i)} && g^\dagger F_\theta = (1-x)C(\mu_0 + g, \sigma) + xN(\mu_1 + g, \sigma^2);\\
&\text{(ii)} && g^\dagger F_\theta = (1-x)C(\mu_0 - g, \sigma) + xN(\mu_1 + 2g, \sigma^2);\\
&\text{(iii)} && g^\dagger F_\theta = (1-x)C(\mu_0, \sigma e^g) + xN(\mu_1, \sigma^2 e^g);\\
&\text{(iv)} && g^\dagger F_\theta = (1-x)C(\mu_0, \sigma e^g) + xN(\mu_1, \sigma^2 e^{2g});\\
&\text{(v)} && g^\dagger F_\theta = (1-x)C(\mu_0 + g, \sigma e^g) + xN(\mu_1, \sigma^2 e^{2g});\\
&\text{(vi)} && g^\dagger F_\theta = (1-x)C(\mu_1, \sigma) + xN(\mu_0, \sigma^2);\\
&\text{(vii)} && g^\dagger F_\theta = (1-x)C(\mu_0, \sigma) + xN(\mu_1, \sigma^2).
\end{aligned}
$$

Identify the group actions that are induced by an action on the state space either covariate independent or covariate dependent? Which group actions are faithful?

26.5. Interference. Interference, or no interference, is a property of the family of conditional distributions of the response $Y \equiv Y[S]$ given the treatment assignment $T[S] = \mathbf{t}$. Intuitively, we should expect that the conditional distribution of Y_u should depend only on the treatment t_u applied to unit u, in which case there is *no interference.* Otherwise, if the conditional distribution of Y_u depends also on the treatment applied to other units $u' \neq u$, we say that there is interference.

Lack of interference is a mathematical property of the probability model. But certain specific forms of interference—commonly pairwise interference by contamination or by competition from neighbouring plots or closely related pairs of units—are known to occur, and the phenomenon is detectable in data.

Interference is not related to correlation or lack of independence of responses on different observational units or experimental units. In the simplest probability models, the responses for distinct units may be conditionally independent given $\mathbf{t}$. But typical models used for the analysis of block designs and split-plot designs allow for correlated responses for distinct experimental units. Indeed, a great part of the theory of block designs is geared towards the exploitation of response correlation to improve the efficiency of treatment estimates. Similar remarks hold for standard statistical models used in the analysis of agricultural field trials, which are geared to accommodate isotropic spatial correlation.

The mathematical definition of *no interference* is a little more complicated than the informal description given above. Let S be the sample, and let $S' \subset S$ be a sub-sample. If the response distribution on S' given the treatment assignment $T[S] = \mathbf{t}$ for all units depends only on the treatment assignment on S', we have conditional independence: $Y[S'] \perp\!\!\!\perp T[S] \,|\, T[S']$. *Lack of interference* means that the conditional independence condition $Y[S'] \perp\!\!\!\perp T[S] \,|\, T[S']$ holds for every subset $S' \subset S$. *No singleton interference* is a weaker version of the same condition, holding for each singleton subset $S' = \{u\}$ consisting of one unit.

No singleton interference implies that the conditional distribution of Y_i given $\mathbf{t}$ depends only on t_i, not on the assignments t_j for experimental units $j \neq i$. The description given in the first paragraph above refers to singleton interference.

This definition of no interference is based on Cox (1958, section 2.4) who stated the condition in the form *the observation on one unit should be unaffected by the particular assignment of treatment to other units.* In the Neyman-Rubin formulation of

causal inference, (Rubin, 2005), this phrase is interpreted as an arithmetic identity concerning counterfactual outcomes as real numbers, called the stable unit-treatment value assumption, or SUTVA. The present interpretation in terms of probability distributions and conditional independence makes no appeal to counterfactuals, and does not require the response to be real-valued. Nonetheless, if the counterfactual scaffolding could be overlooked, no singleton interference and SUTVA are similar in spirit if not in specifics.

Example 26.1 A/B treatment for scour. Newborn lambs are prone to scour, which is not immediately life-threatening, but substantially retards growth. Two treatments are available for testing, one a drench given orally, the other an antibiotic delivered by injection, both given in the 24-hour period following birth. The units for experimentation purposes are all lambs born to a particular flock of 50 blackface ewes in the spring of 2015. Immediately after birth, each lamb was assigned independently with equal probability to one or other treatment level. Each lamb was weighed at birth and subsequently at six weeks.

The observational units are lambs, baseline is set at birth, birth weight is a baseline variable which may be used as a covariate, and the response is either 6-week weight or 6-week weight gain. Sex and sibling count are also baseline variables. Each pair of lambs, twins included, has probability 0.5 of receiving different treatment levels, so each lamb is one experimental unit according to the definition. Finally, ewe or mother is an equivalence relation on observational units, each block being one family.

Two alternative randomization schemes are worth mentioning. For the efficient administration scheme in which twins are automatically given the same treatment level, each ewe or family group is one experimental unit. For a randomized blocks design in which twins are automatically given different treatment levels, each ewe is a block and each lamb is an experimental unit.

Male lambs tend to be slightly heavier than females, both at birth and afterwards, but this source of variation is ignored in the discussion that follows.

Consider first the twin lambs alone. Each twin i has a sib $j = s(i)$ such that $s(j) = s^2(i) = i$. There are four treatment classes $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$, and the ordered pair (i,j) belongs to treatment class (t_i, t_j). Suppose that the response distribution for each focal lamb i in treatment class (r,s) is Gaussian:

$$(t_i, t_{s(i)}) = (r,s) \Longrightarrow Y_i \sim N(\mu_{rs}, \sigma^2).$$

No interference implies that the response distribution for experimental unit i does not depend on the treatment applied to any other unit such as $s(i)$. For this version of the Gaussian model, no interference means $\mu_{00} = \mu_{01}$ and $\mu_{10} = \mu_{11}$, or $\mu_{rs} = \alpha_r$ for some $\alpha = (\alpha_0, \alpha_1) \in \mathcal{R}^2$.

In practice, one can examine the sample average weights of focal lambs in each of the four treatment classes to test whether or not the no-interference condition is violated. Comparison of sample averages is reasonably straightforward even though sib values Y_i and $Y_{s(i)}$ must not be assumed independent. Nevertheless, once interference is identified, a natural biological explanation can be found in shared infection or cross-contamination from udder-sharing.

The mathematical remedy is also straightforward. Given the possibility of interference, it is best to regard each family as one observational unit, which is also an experimental unit. As they are applied to experimental units, the six treatment levels are 0, and 1 for singletons, and $(0,0), (0,1), (1,0), (1,1)$ for twins. Sibling count is a covariate, which governs not only the randomization distribution, but also the state space—$Y_u \in \mathcal{R}$ for singletons, $Y_u \in \mathcal{R}^2$ for twins. For each singleton lamb i, the response

distribution is either $F_0 = N(\mu_0, \sigma^2)$ *if* $t_i = 0$ *or* $F_1 = N(\mu_1, \sigma^2)$ *if* $t_i = 1$, *and the treatment effect is the scalar mean difference*

$$\Delta^{(1)} = \mu_1 - \mu_0.$$

The response for twins $u = (i, j)$ *(or* $u' = (j, i)$*) is a two-component Gaussian vector* $Y(u)$ *whose conditional mean* μ_t *depends on the treatment assignment* $t = (r, s)$. *If* $\mu \mapsto \mu'$ *is a transposition of components,* $Y'(u) = Y(u')$ *is also Gaussian with conditional mean* $\mu'_t = \mu_{t'}$, *implying* $\mu'_{rs} = \mu_{sr}$, *or equivalently* $\mu_{rs} = (\alpha_{rs}, \alpha_{sr})$ *for some* 2×2 *array* α. *The space of the mean-value vectors* μ_{rs} *has dimension four, so there are three linearly independent treatment effects,*

$$\Delta_0^{(2)} = \mu_{10} - \mu_{00}, \quad \Delta_1^{(2)} = \mu_{11} - \mu_{01}, \quad \textit{and} \quad \Delta^{(2)} = \mu_{11} - \mu_{00}.$$

Each of these combinations is vector-valued, but the third is a scalar, symmetric with respect to transposition, and the linear combination $\mu_{10} - \mu_{01}$ *is a skew-symmetric scalar. In the absence of sibling interference, the mean value vectors satisfy* $\mu_{rs} = (\alpha_r, \alpha_s)$, *for some two-component vector* α, *and each version of the treatment effect reduces to the scalar* $\Delta^{(2)} = \alpha_1 - \alpha_0$.

Lack of sibling interference can be tested only if every treatment combination occurs at least once in the design. For the efficient administration scheme only $\Delta^{(2)}$ *is identifiable. For the randomized blocks design in matched pairs, only* $\mu_{10} - \mu_{01}$ *is identifiable. These treatment effects are equal under lack of sibling interference, but not, in general, otherwise.*

Lack of sibling interference does not imply that the treatment effect for twins is the same as the treatment effect for single lambs. In general, the treatment effect may vary from one unit to another depending on the covariate value, which in this example includes sex and mother's age. Sib count is also a covariate. If there is no sib count by treatment interaction, then $\Delta^{(1)} = \Delta^{(2)}$*; otherwise the values are different.*

The preceding discussion of interference presumes for simplicity that treatment has no effect on variances or covariances. It is certainly possible that the response variance may depend on treatment, and, in the case of twin pairs, the covariances also.

26.6. Commutativity of treatment action. The effect of treatment is a group action on processes, sending F to $g^\dagger F$. To see what this means for finite-dimensional distributions, let S, S' be two samples, and let $F_S, F_{S'}$ be the finite-dimensional null distributions associated with F. If $S \subset S'$, the Kolmogorov condition implies that F_S is the marginal distribution of $F_{S'}$ under the natural projection $\mathcal{S}^{S'} \to \mathcal{S}^S$. The group action must be such that $g^\dagger F_S$ is also the marginal distribution of $g^\dagger F_{S'}$ under the same projection. This is a commutativity condition: marginal projection of distributions commutes with group action.

26.7. Covariate versus treatment. The treatment level is assigned to each unit; the covariate value is a property of the unit. In computational work, it is all too easy to overlook this logical distinction, but $\mathbf{x}$ and $\mathbf{t}$ do not occur on an equal footing in the probability model. The sample S determines the covariate configuration $\mathbf{x} \equiv x[S]$ by subset restriction, and the same applies to relationships, so there is only one covariate configuration or block structure for a given sample. However, there are multiple possibilities for treatment assignment, and each probability model specifies a probability distribution $F(\cdot \,|\, \mathbf{t}; \theta)$ for every possible $\mathbf{t}$, not just the one that happened to occur by chance in a particular experiment. The difference

$$F_{\mathbf{x}}(A \,|\, \mathbf{t}; \theta) - F_{\mathbf{x}}(A \,|\, \mathbf{t}'; \theta)$$

is the effect of treatment assignment $\mathbf{t}$ versus $\mathbf{t}'$ on the probability of the event $Y[S] \in A$ *for this particular sample* S. Any similar comparison of probabilities for $\mathbf{x}$ versus $\mathbf{x}'$ necessarily involves different samples $S' \neq S$.

26.8. Linguistic effects. It is good to be reminded of the logical difference between covariate and treatment by careful use of language. A statement such as *The effect of treatment is to increase survival times by 30%* carries with it an implication of intervention—that the treatment level for patient i may be set at either $t_i = 0$ or $t_i = 1$, not only for patients who participated in the study but also for subsequent eligible individuals.

On the most literal level, the statement appears to be an assertion about survival *times*, which could be interpreted as a statement about random variables or points in the state space. However, if it is made on the strength of a fitted probability model, the statement can be interpreted only as an assertion about probabilities or probability distributions. In that sense, it is best regarded as a figure of speech—more metaphor than mathematics. One interpretation is as follows. For every patient, the probability of surviving to time s with $T = 0$ is the same as the the probability of surviving to time $1.3s$ with $T = 1$.

A probability model for one patient in a randomized experiment, is a joint distribution for (T, Y) on the product space *treatment levels* $\times$ *state space*. That implies two conditional survival distributions $F_0(\cdot; x_i)$ and $F_1(\cdot; x_i)$, one for the control level $t_i = 0$ and one for the active level $t_i = 1$. In that context, the statement means that the probability assigned by F_0 to the event or time interval $(0, s)$ is the same as that assigned by F_1 to the longer interval $(0, 1.3s)$, i.e., $F_1((0, 1.3s); x) = F_0((0, s); x)$ for every $s > 0$. The probability model does not imply two (counterfactual) survival times $Y_0(i)$ and $Y_1(i)$ such that $Y_1(i) = 1.3Y_0(i)$, as an extreme literal interpretation of the statement might suggest.

A similar statement such as *The effect of treatment is to decrease the hazard rate by 30%* also carries an implication of intervention, and has a similar statistical interpretation. The model specifies two distributions with survivor functions $F_0((s, \infty))$ and $F_1((s, \infty))$, and the proportional-hazards treatment effect is a real number g, which acts on survival distributions $g\colon F_0 \mapsto F_1$ by

$$\log F_1((S, \infty)) = e^g \log F_0((s, \infty)) = 0.7 \log F_0((s, \infty))$$

for every $s > 0$. The first version of the treatment effect implies a multiplicative action $y \mapsto 1.3y$ on the state space $\mathcal{R}^+$, which has an induced action on probability distributions. The second version implies hazard multiplication, which is a group action on survival distributions. However, hazard multiplication is not induced from an action on the state space, so it is not possible to specify the treatment effect through a pair of counterfactual times Y_0, Y_1 related to one another by hazard reduction. The issue here is related to SUTVA (the stable unit-treatment value assumption), which holds for lifetime multiplicaton but not for hazard multiplication.

If there are inhomogeneities of a similar magnitude associated with a covariate such as sex, a parallel statement *the effect of sex (F versus M) is to increase the survival time by 30%* is semantically inappropriate. Instead, one might report that the survival distributions have the same shape for both sexes, but they differ by a multiplicative factor of 1.3, favouring longer times for women. Alternatively, one could make a more guarded or restrictive assertion about mean or median survival times.

In summary, the treatment effect is a comparison of two probabilities for the same patient; the covariate effect is a comparison of probabilities for different patients.

26.9. Effects of treatment. Suppose that treatment has two levels and that there is no interference among units. Fix one experimental unit $i = 0$, say, and $x = x_0$. For this unit, the response distribution is either F_0 if $T = 0$ or F_1 if $T = 1$. In principle, any additive skew-symmetric functional $\Delta_{10} = U(F_1) - U(F_0)$ of the two conditional distributions is *an effect of treatment.* For example, if $U(F)$ is the mean of the distribution, Δ_{10} is the difference between the means of the two distributions, which is the *effect of treatment on the mean response*; if $U(F)$ is the log variance, Δ_{10} is the effect on the response variance of switching from level 0 to level 1. Likewise, if the response is a survival time, $U(F)$ could be the log mean, the five-year survival odds, or the log hazard function at two years post-baseline. In the latter case, $U(F; s) = \log(F'(s)/(1 - F(s)))$ is the hazard function at time $s > 0$, and Δ is the log hazard ratio after two years.

If treatment has k levels with conditional response distribution F_r for treatment level r, the functional $\Delta_{rs} = U(F_r) - U(F_s)$ is the effect on the response distribution as measured by U of assigning treatment level r versus s to this particular unit. The effect matrix Δ is additive and skew-symmetric.

Under the standard exchangeability assumption as used in regression models, equality of covariates $x_i = x_j$ for units i, j implies equality of distributions $(T_i, Y_i) \sim (T_j, Y_j)$, so that

$$F_r^{(i)}(A) = \mathrm{pr}(Y_i \in A \,|\, T_i = r) = \mathrm{pr}(Y_j \in A \,|\, T_j = r) = F_r^{(j)}(A).$$

The treatment effects matrix Δ_{rs} may vary from one unit to another, but exchangeability implies that it is constant over units having the same covariate value.

26.10. Parameterization of treatment effects. A typical statistical model for one experimental unit consists of a *family of distributions* $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$ such that each of the conditional distributions F_r belongs to $\mathcal{F}$. Switching from treatment level s to level r is a group action $g_{rs} : \mathcal{F} \to \mathcal{F}$ on probability distributions; if F_s is the response distribution for treatment level s, then $F_r = g_{rs}F_s$ is the response distribution for level r. The group action is subject to two consistency conditions (i) $g_{rs} = g_{sr}^{-1}$ (the inverse condition), and (ii) $g_{rs}g_{st} = g_{rt}$ (the transitivity condition). Transitivity and skew-symmetry imply that $g_{rs} = g_r g_s^{-1}$; in other words, there exist group elements $\{g_r\}$, one for each treatment level, such that $g_{rs} = g_r g_s^{-1}$. Standard practice is to choose a reference level, or control level, and set $g_0 = I$. Whatever the reference distribution $F_0 \in \mathcal{F}$ may be, the treatment distribution for level r is $F_r = g_r F_0$, and $g_r \in \mathcal{G}$ is the treatment effect relative to the control.

In this general setting, the treatment levels may be qualitative, such as distinct varieties in a cereal variety trial, or they may be quantitative, such as dose level in a pharmaceutical trial. In the simplest settings where the group of treatment effects is the real numbers, the difference between F_r and F_s is encapsulated in a single real number g_{rs}, whose interpretation is most commonly a difference of means or a difference of transformed means such as a log odds ratio. Each treatment orbit is a one-dimensional curve in $\mathcal{F}$. But if the response distributions also have unequal variances, it may be necessary to set $\mathcal{G} = \mathcal{R}^2$, so that the first component of g is the mean difference, and the second component is the log variance ratio. Likewise, for a bivariate or multivariate response.

The group action is a homomorphism associating with each $g \in \mathcal{G}$ an invertible transformation hg on probability distributions. Prior to this point, we have not distinguished the group element $g \in \mathcal{G}$ from the transformation $hg : \mathcal{F} \to \mathcal{F}$, implying that the homomorphism $g \mapsto hg$ is 1–1, or faithful. In essentially all applications, $\mathcal{G}$ is a topological group such as $\mathcal{R}$ or $\mathcal{R}^d$ with addition, and the homomorphism is invariably continuous and differentiable, so each orbit is a manifold in $\mathcal{F}$ of the same dimension as the group.

In a more complicated situation it may be necessary to choose a much larger space to accommodate the range of treatment effects that is anticipated. For a longitudinal design or a study involving growth curves, each distribution $F \in \mathcal{F}$ is a real-valued temporal stochastic process, so F_r and F_s have mean trajectories $\mu_r(t), \mu_s(t)$. If the distributions F_r and F_s are related by state-space translation, it suffices to set $\mathcal{G} = \mathcal{C}_0(\mathcal{R}^+)$ (continuous functions $\mathcal{R}^+ \to \mathcal{R}$), implying $\mu_r(t) - \mu_s(t) = g_r(t) - g_s(t)$. Finite-dimensional sub-groups include the constant functions, polynomials, and cubic splines having specified knots. If F_r and F_s are not related by state-space translation, an alternative group action must be found. Some simple examples are as follows.

Example 26.2 Bernoulli model If the response distribution is Bernoulli, the state space for one unit is $\{0,1\}$, *and* $\mathcal{F} = (0,1)$ *is the unit interval. The control probability* $\mathrm{pr}(Y_i = 1 \,|\, T_i = 0) = \pi_0(x_i)$ *and the treatment probability* $\mathrm{pr}(Y_i = 1 \,|\, T_i = 1) = \pi_1(x_i)$ *are two points in* $(0,1)$, *both functions of the covariate vector. According to the standard recipe, the group is the real line* $\mathcal{G} = \mathcal{R}$ *acting on the space of Bernoulli distributions, which means that each* $g \in \mathcal{G}$ *is associated with a transformation* $g\colon (0,1) \to (0,1)$ *that is a group action on the unit interval. For technical reasons, the end points are excluded. If the treatment effect relative to the reference level is the real number* $g \in \mathcal{G}$, *the pair* $(F_0, F_1) \equiv (\pi_0, \pi_1)$ *is related by group action:* $\pi_1 = g\pi_0$. *Although the group element is a real number,* $\pi \mapsto g\pi$ *is not addition or scalar multiplication, which could yield a negative number. In a generalized linear model,* $\eta\colon (0,1) \to \mathcal{R}$ *is a given monotone invertible transformation, and the group action is real addition on the transformed scale, so that*

$$\eta(g\pi) = \eta(\pi) + g, \quad \text{or} \quad \eta(\pi_r) = \eta(\pi_0) + g_r,$$

with $g_0 = 0$ *as the treatment value for the reference level.*

In order to define a group action on $(0,1)$, *the transformation* $\eta\colon (0,1) \to \mathcal{R}$ *is necessarily 1–1 and invertible. All instances arising in applications are also continuous, the logit, probit, log-log and complementary log-log functions being most natural.*

Multinomial response models have the same structure. For the cumulative logistic and related models, $\mathcal{G} = \mathcal{R}$; *the action on the probability simplex is similar to the Bernoulli case, but it is not transitive (there are infinitely many orbits). For the standard* k*-class multinomial logistic model, the group is* $\mathcal{G} = \mathcal{R}^k$ *or* $\mathcal{R}^k/\mathbf{1}$, *and the action on the simplex is transitive.*

Example 26.3 PH model In the proportional-hazards model for survival times, $\mathcal{F}$ *is the set of Borel distributions on the positive real line, and* $\mathcal{G} = \mathcal{R}^+$ *is the positive real line with multiplication. Every distribution* $F \in \mathcal{F}$ *has a hazard function or hazard measure*

$$H(0,t] = -\log F\big((t,\infty)\big),$$

and g *acts on* $\mathcal{F}$ *by hazard multiplication. Certain subsets of* $\mathcal{F}$ *such as the exponential and Weibull distributions, which are closed under this action, may be used as sub-models. The accelerated-lifetimes model uses the same family and the same group: only the action* $gF(A) = F(g^{-1}A)$ *induced by lifetime multiplication is different.*

26.11. Illustration of covariate and treatment effects. Let $x_i \in \{0,1\}$ be a binary covariate, such as sex, which is registered at baseline. In the absence of treatment intervention, the parameter space is $\Theta = \mathcal{R}^2 \times \mathcal{R}^+$ with points $\theta = (\mu_0, \mu_1, \sigma)$. In both models that follow, the components of Y are independent. In model I, the distribution of Y_i is $N(\mu_{x_i}, \sigma^2(1 + x_i))$, i.e., $N(\mu_0, \sigma^2)$ for all i with $x_i = 0$ and $N(\mu_1, 2\sigma^2)$ otherwise. In model II, the distribution is $N(\mu_0, \sigma^2)$ if $x_i = 0$ or Cauchy $C(\mu_1, \sigma)$ with probable error σ if $x_i = 1$. Both specifications satisfy the Kolmogorov consistency condition, so they are both acceptable as real-valued stochastic processes.

For model II, the family of distributions $N(\star,\star)$ for $x_i = 0$, is not the same as the family $C(\star,\star)$ of distributions for $x_i = 1$, so the 'covariate effect' cannot be expressed as a group action on probability distributions. For model I, the two families are equal: $\mathcal{F} = N(\star,\star)$. However, it is not possible to specify the association $N(\mu_0,\sigma^2) \mapsto N(\mu_1,2\sigma^2)$ by a group action $g\colon \mathcal{F} \to \mathcal{F}$ because $\sigma^2 \mapsto 2\sigma^2$ for every g implies that there can be no identity element. While both models are entirely satisfactory for covariate effects, neither formulation satisfies the conditions required for a treatment effect.

For each of the preceding models, we consider three ways in which the distribution might be modulated by a two-level treatment factor. In each case, the space of treatment effects is a group, either $\mathcal{G} = \mathcal{R}$ or $\mathcal{G} = \mathcal{R}^2$, acting on distributions:

$$\begin{aligned}
&(a) \quad \mathcal{G} = \mathcal{R} \ : && g\theta = (\mu_0 + g, \mu_1 + g, \sigma)\\
&(b) \quad \mathcal{G} = \mathcal{R}^2 : && g\theta = (\mu_0 + g_0, \mu_1 + g_1, \sigma)\\
&(c) \quad \mathcal{G} = \mathcal{R}^2 : && g\theta = (\mu_0 + g_0, \mu_1 + g_0, \sigma e^{g_1})
\end{aligned}$$

For model IIb, $\mathcal{F}$ is three-dimensional, $\mathcal{G}$ is two-dimensional, and the response distribution for each component is as follows

	$t=0$	$t=1$
$x=0$	$N(\mu_0,\sigma^2)$	$N(\mu_0+g_0,\sigma^2)$
$x=1$	$C(\mu_1,\sigma)$	$C(\mu_1+g_1,\sigma)$

For model Ic, $\mathcal{F}$ is three-dimensional, $\mathcal{G}$ is two-dimensional, and the response distribution for each component is as follows

	$t=0$	$t=1$
$x=0$	$N(\mu_0,\sigma^2)$	$N(\mu_0+g_0,\, \sigma^2 e^{2g_1})$
$x=1$	$N(\mu_1,2\sigma^2)$	$N(\mu_1+g_0,\, 2\sigma^2 e^{2g_1})$

26.12. Treatment effect for a vital process. A stochastic process in continuous time is called *vital* if the state space contains an absorbing state $\flat$ such that $Y(0) \neq \flat$, and the survival time $T = \inf\{t : Y(t) = \flat\}$ is finite with probability one. For $s > 0$, let $Z_s = Y_{T-s}$ be the reverse-time process, so that the transformation $Y \mapsto (T, Z)$ is one-to-one. For ease of exposition, Z is assumed to be a real-valued process, so the state space for Y is the disjoint union $\mathcal{R} \cup \{\flat\}$.

Let $\mathcal{F}_{(0,\infty)}(\mathcal{R})$ be the set of real-valued stochastic processes with index set $t > 0$. The statement that $F \in \mathcal{F}$ is a stochastic process means that, for each finite list of time points $\tau = (\tau_1, \ldots, \tau_k)$, F_τ is the probability distribution on $\mathcal{R}^k$ of the values at the given time points. Moreover, these distributions are mutually consistent in the sense that if $\tau' = (\tau_1, \ldots, \tau_k, \tau_{k+1})$ is an extended list, $F_{\tau'}(A \times \mathcal{R}) = F_\tau(A)$. The reverse-time transformation $Y \mapsto (T, Z)$ means that every vital process has a factorization in which the first factor is the survival distribution on $(0,\infty)$, and the second factor is a function $t \mapsto F^t$ where $F^t \in \mathcal{F}_{(0,\infty)}(\mathcal{R})$ is a real-valued stochastic process. To say the same thing in another way, the second factor associates with each $t > 0$ the conditional distribution of Z given $T = t$.

Assume that the function $t \mapsto F^t$ is continuous in $t > 0$, and that each one-dimensional distribution F^t_τ has finite mean. Formally, the set of vital processes with state space $\mathcal{R} \cup \{\flat\}$ is equal to the product set

$$\mathcal{V}(\mathcal{R} \cup \{\flat\}) = \mathcal{F}(\mathcal{R}^+) \times \mathcal{C}(\mathcal{R}^+)$$

where $\mathcal{C}(\mathcal{R}^+)$ is the set of continuous functions $t \mapsto F^t$. To each $t > 0$ there corresponds a one-dimensional distribution F_s^t and a mean function

$$E\big(Z(s) \,|\, T = t\big) = \int_{-\infty}^{\infty} z \, dF_s^t(z) = \mu(s,t)$$

which is a real number defined for each $s, t > 0$.

We are now in a position to define several versions of "the effect of treatment on a vital process by group action" as follows. Each $g \in \mathcal{G}$ has two components (g^0, g^1). The first component is a positive number, which acts on survival distributions $\mathcal{F}(\mathcal{R}^+)$ either by hazard multiplication or by lifetime multiplication. The remaining component is a continuous function $\mathcal{R}^+ \to \mathcal{R}$, possibly constant, which acts on the state space by translation $Z(s) \mapsto Z(s) + g^1(s)$, implying that

$$E\big(Z(s) \,|\, T = t, \mathrm{Trt} = r\big) = \mu(s,t) + g_r^1(s).$$

In particular, if μ is additive and each group element $g_r^1(s) = \gamma(r)$ is constant, we have

$$E\big(Z(s) \,|\, T = t, \mathrm{Trt} = r\big) = \alpha(s) + \beta(t) + \gamma(r)$$

for some temporal functions α, β and treatment effect γ. For the vital process Y in the forward direction, the conditional mean

$$E\big(Y(s') \,|\, T = t, \mathrm{Trt} = r\big) = \alpha(t - s') + \beta(t) + \gamma(r)$$

for $s' < t$ is not additive in the temporal variables s', t.

27. Role of counterfactuals

27.1. Background. The concept of counterfactual or potential outcomes, and the associated philosophical notion of causality and causal effects, constitute a sort of Rorschach test for philosophers and statisticians. Counterfactuals are an unavoidable part of everyday speech and speculation: *if only Hillary Clinton had campaigned more forcefully in Pennsylvania and Michigan, the U.S. electorate would not be facing the embarrassment that is in store for the next few years.* When we talk in these nby A.P. Dawid.

The attitude adopted in these notes is studiously mathematical and probabilistic. In the end, counterfactuals do not arise, so there is nothing to say—either about counterfactuals or about causality. Yet it is necessary at times to compare and contrast approaches, or at least to compare modes of expression. Some authors *define* the treatment effect as the difference between unobservable, counterfactual or potential outcomes for the same unit; others are content to work with the estimator, which is necessarily an average difference between observed responses for different units, possibly matched in pairs. The view taken in these notes, that statistical thinking begins with a probability model, means that each unit may be assigned one or other treatment level by randomization, and the conditional distribution necessarily associates with each unit one probability distribution $F_0, F_1, \ldots$ for each of the treatment levels. The probabilistic definition that emerges for a treatment effect is not a difference between counterfactual values or outcomes for the same unit, but a difference between response probabilities for the same unit. A covariate effect is a difference between response distributions for different units.

To a great extent, the differences alluded to in the preceding paragraph are more differences in mode of expression than they are differences of substance. In the most common situation where the state space is the real line and the treatment effect is additive, the differences are not great. In other situations where the response space is binary, or the treatment effect is not expressible as an action on the state space, direct comparison is considerably more difficult.

27.2. Counterfactual variables. There is a long history in the statistical literature of defining treatment effects in terms of counterfactual random variables, or potential outcomes $Y(i,t)$, one counterfactual value for each unit i and treatment level $t \in \{1,\ldots,k\}$. Notionally, $Y(i,t)$ is the response that would be observed if treatment t were assigned to unit i, so the counterfactual array is a matrix of order $n \times k$. The n-component vector $Y[\mathbf{t}]$ with components $Y(i,t_i)$ for $i \in S$ is the joint response that would be observed, or is actually observed, for treatment assignment vector $\mathbf{t}$. The counterfactual array is constructed in such a way that $Y[\mathbf{t}]$ is distributed according to the model $F_{\mathbf{x}}(\cdot \mid \mathbf{t}; \theta)$, so the introduction of counterfactuals has no implications for probability distributions, conditional distributions given $\mathbf{t}$ or likelihood functions.

It appears that counterfactuals were first introduced by Neyman for didactic or rhetorical purposes, in discussions related to the role of randomization. For that purpose, it it would be pointless to object to a concept having no apparent statistical or probabilistic implications. There is, however, a linguistic implication. If the response is real-valued, the difference $Y(i,t) - Y(i,t')$, even though it is an admitted figment of counterfactual imagination, may sometimes be described as the *actual effect caused by assigning treatment level t rather than t' on this particular unit.* The casual introduction of the word *cause* in the phrase *effect caused by* is regrettable. Ultimately, it is no more than a rhetorical flourish based on a counterfactual figment, but *causal* has a strong emotional appeal, which appears to transcend reason, particularly in medical applications. If a counterfactual difference is accepted literally as a legitimate target of estimation, it changes the focus of the investigation in ways that are beyond the scope of probability models.

No concept of counterfactual exists in the probability model described here. Instead, for unit i the probability model implies a marginal response distribution $\text{pr}(Y_i \in A \,|\, T = \mathbf{t}) = F(A; t_i)$ on the real line or other observation space. There is one probability distribution for each treatment level, and the difference $F(A; t) - F(A; t')$ is defined to be the effect of treatment assignment t versus t' on the probability of the event $Y_i \in A$. At this point, the only difference between the probabilistic and counterfactual approaches lies in the usage or avoidance of the word causal. Covariate dependence has been suppressed in the notation, so the effect of treatment on the probabilities, on the expected values and other distributional summaries, may depend also on registered baseline variables.

The more common summary functional for treatment effect is the difference of expected values, but other functionals such as the log odds ratio or the log hazard ratio are better suited for binary models and survival models respectively.

In the standard statistical model, the treatment effect t versus t' is defined for each unit by the difference between two probability distributions, one distribution corresponding to the assignment $i \mapsto t$ and one corresponding to $i \mapsto t'$. In the counterfactual world, the treatment effect is defined for each unit as a difference between notional or potential or counterfactual outcomes, $Y(i, t) - Y(i, t')$, not as a difference between two probability distributions.

Appeals to causality and counterfactuals are avoided in these notes.

27.3. SUTVA.