A Cheat Sheet for Statistical Consultants

Chong Gu
Purdue University

In statistical consulting, we help the clients to formulate and solve research problems involving data. The crucial part is proper problem formulation, for which we need to ask the right questions to query experimental settings, data collection/preprocessing details, and the research questions the clients wish to address; there is no cheat sheet on this part.

Once the problems are properly formulated, actual data analysis could be routine albeit laborious. This cheat sheet enlists some commonly used techniques and notes on some useful insights and niche settings. The key is to *apply the right tools in the right settings.*

Different research questions may require different data analytical approaches using different parts of the data. If you can address the clients' needs using simple tools, don't attempt sophisticated procedures.

1 Basic Techniques

The bread-and-butter tools compiled here should cover most of the everyday data analysis needs. These tools are conceptually elementary, but some are technically and/or computationally sophisticated, such as the rank-based tests, the polychoric correlation, and the proportional odds models. With the readily available software facilities, one may simply focus on the conceptual settings and the proper interpretations of the analysis results.

1.1 Graphical Displays

A picture is worth a thousand words. Before formal analysis, it is often a good idea to plot the data to gain some insights and get a feel about what you are getting into.

For categorical data, standard plots include **bar charts** and **pie charts**.

For continuous data, **histograms** and **scatter plots** are commonly used. Colors and/or plotting symbols could be used to superimpose scatter plots of multiple groups of data. Side-by-side **boxplots** are effective in contrasting multiple 1-D distributions.

Scatter plot matrix should be looked at before multiple regression is attempted.

Transformations of continuous data might be needed to spread out the scatter more evenly.

1.2 Simple Tests

Two-sample *t*-tests and **one-way ANOVA** are commonly used for the analysis of treatment effects (of one factor) using *independent samples.* A common population variance is usually assumed, to be estimated by the pooled sample variances. The default two-sample t-test implemented in the R function `t.test()` however does *not* assume a common population variance.

Replacing the original data by their ranks in the pooled data, one has the **Wilcoxon rank-sum test** (two-sample) and the **Kruskal-Wallace test** (general one-way ANOVA); the Wilcoxon rank-sum test is equivalent to the **Mann-Whitney test**.

To test against a known population mean or median, one has the **one-sample *t*-test** and the **Wilcoxon signed-rank test**.

With *paired samples*, take intra-pair differences and perform the one-sample t-test using the differences; this is the **paired *t*-test**. Intra-pair ranking leads to the **sign test**, treating the signs of the differences as Bernoulli observations.

The t-tests do assume normality, but they are robust; theoretical properties still hold "to an extent" when normality is violated. Non-normality alone should *not* be the basis for abandoning t-tests, especially with small to moderate sample sizes. For highly skewed data, proper transformations typically help.

With the rank-based tests, one trusts the ordering of the values more than the values themselves, and one loses a lot of power relative to the t-tests. The rank-based tests are practically useful only when the sample size is large.

The rank-sum test and signed-rank test are implemented in the R function `wilcox.test()` using the exact (discrete) null distributions. The Kruskal-Wallace test is implemented in the R function `kruskal.test()`, which reduces to the rank-sum test in the two-sample setting but uses χ^2 approximation.

The rank-based tests actually assume continuous random variables with zero probability for ties. In practice, ties do occur, and tie handling is usually intuitive but *ad hoc.*

1.3 Correlations

To assess the association between a pair of random variables (X, Y), one has a few versions of correlations. Write $\sigma_x^2 = \mathrm{Var}(X)$, $\sigma_y^2 = \mathrm{Var}(Y)$, and $\sigma_{xy} = \mathrm{Cov}(X, Y)$. **Pearson correlation** $\rho = \sigma_{xy}/(\sigma_x \sigma_y)$, or simply the correlation, is defined for all pairs with finite variances; it makes more sense for continuous pairs, and is most interpretable for bivariate normal. Observing (x_i, y_i), the sample version is $r = \sum_i (x_i - \bar{x})(y_i - \bar{y})/\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}$.

For (X, Y) both continuous, with cdf's $F_X(x)$ and $F_Y(y)$, respectively, consider $U_X = F_X^{-1}(X)$ and $U_Y = F_Y^{-1}(Y)$; $U_X, U_Y \sim U(0, 1)$. **Spearman's $\boldsymbol{\rho}$** is the Pearson correlation between U_X and U_Y. Observing (x_i, y_i), one may order x_i's and y_i's separately and replace the original observations by their respective ranks, then calculate the sample correlation of the two sets of ranks.

Consider (X_1, Y_1), (X_2, Y_2) *i.i.d.* from a bivariate distribution, **Kendall's $\boldsymbol{\tau}$** is defined as

$$P\big((X_1 - X_2)(Y_1 - Y_2) > 0\big) - P\big((X_1 - X_2)(Y_1 - Y_2) < 0\big).$$

Observing (x_i, y_i), $i = 1, \ldots, n$, there are $N = \binom{n}{2}$ (i, j) pairs, $N_c = \#\{(x_i - x_j)(y_i - y_j) > 0\}$ *concordant* and $N_d = \#\{(x_i - x_j)(y_i - y_j) < 0\}$ *discordant*, and the sample τ is $(N_c - N_d)/N$.

Spearman's ρ and Kendall's τ both assume continuous random variables with zero probability for ties. Practical tie handling is intuitive but *ad hoc.*

One may calculate all three versions of correlations using the R function `cor()`, with argument `method` set to `"pearson"` (the default), `"spearman"`, or `"kendall"`.

For ordinal variables such as Likert scales, one may perceive them as interval-censored from latent continuous variables, with the intervals defined through unknown cut-points. Assuming bivariate normal for the latent variables, one may maximize the likelihood of the data with respect to parameters in the cut-points and the correlation coefficient, and the resulting MLE of correlation coefficient yields the **polychoric correlation**; without loss of generality, the variances of latent variables could be set to 1. Similar idea can be used to calculate the sample correlation between a continuous variable and an ordinal variable. Check out R package `polycor`.

1.4 Response Modeling

For *continuous responses*, possibly after proper transformations, one may use **standard linear models** with normal errors, $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, as implemented in R function `lm()`.

For *binary responses*, one may use **logistic regression**, as implemented in R function `glm()`. For *ordinal responses*, one may use **proportional odds models**, where the ordinal response is taken as interval-censored from a latent continuous variable $\tilde{Y}$ following a logistic distribution, with cdf $e^z/(1+e^z)$ for $z = (\tilde{y} - \mathbf{x}^T \boldsymbol{\beta})/\sigma$; without loss of generality, one may set $\sigma = 1$ and estimate the cut-points (multiple "intercepts") along with $\boldsymbol{\beta}$ (common "slopes"). Proportional odds models are implemented in R package `MASS` (function `polr()`) and package `ordinal` (function `clm()`). Technically, logistic regression is a special case of proportional odds models, with one cut-point.

For *event count responses*, the default choice is **Poisson regression**, as implemented in `glm()`, and if the purpose is to model the *event rate*, an *offset* term is often needed to adjust for the amount of exposure. As an alternative, negative binomial regression is sometimes used to accommodate over-dispersion, but the model is not as interpretable.

For *nominal categorical responses* with three or more categories, one may use **multinomial regression**; an implementation is in R function `multinom()` in package `nnet`. The response here is technically *multivariate*, whereas the models listed earlier all deal with univariate responses.

2 Sample Size Planning and Power Calculation

To ensure certain "minimal performance level" of inferential procedures, sample size planning is often desired at the design stage *before* data are being collected. It is analytically tractable only in a few simple settings.

Confidence intervals (CIs) and hypothesis testing are primary inferential tools. For CIs, one would like them to be no wider than some prespecified width *given coverage*. For tests, one would like the power (rejection probability) be above some prespecified level *for specific alternatives*.

The key word here is *planning*, so power calculation is meaningless when the data have already been collected, or after the study design has been set.

2.1 One/Two-Sample Settings

Consider $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, for σ^2 *known*. Inferences concerning μ are based on $\bar{Y} \sim N(\mu, \sigma^2/n)$. The width of CI for μ is governed by $\sigma/\sqrt{n}$, and the powers of tests for $H_0 : \mu = \mu_0$ are known monotone functions of $\sqrt{n}|\mu - \mu_0|/\sigma$.

For two samples $Y_{ij} \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$, $j = 1, \ldots, n_i$, for σ_i^2's *known*, inferences concerning $\mu_1 - \mu_2$ are based on $\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. The width of CI for $\mu_1 - \mu_2$ is governed by $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$, and the powers of tests for $H_0 : \mu_1 = \mu_2$ are known monotone functions of $|\mu_1 - \mu_2|/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

For σ^2's unknown, one needs prior knowledge or pilot studies to obtain their ballpark values. The actual inferences would be using t-statistics instead of z-statistics, and the performance guarantee based on the "z-calculations" only holds approximately for n large.

For $Y \sim \text{Bin}(n,p)$, where $Y = \sum_i Y_i$ with $Y_i \sim \text{Bin}(1,p)$, $\sigma^2 = \text{Var}(Y_i) = p(1-p) \leq (0.5)^2$. Using the normal approximation of binomial, one may plug in $\sigma = 0.5$ for conservative sample size planning for inferences concerning p.

2.2 General Case

For anything beyond a two-sample setting, there could be many tests/CIs of interest, of which each could have their own required sample sizes, so a generic notion of sample size planning or power calculation may not be clearly defined.

For an example, consider balanced one-way ANOVA with $Y_{ij} \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, I$, $j = 1, \ldots, n$, for σ^2 *known.* Potentially of interest are the test for $H_0 : \mu_1 = \cdots = \mu_I$, and inferences concerning *contrasts* $\theta = \sum_i c_i\mu_i$ for given c_i's satisfying $\sum_i c_i = 0$.

For a single contrast $\theta = \sum_i c_i\mu_i$, $\hat{\theta} = \sum_i c_i\bar{Y}_{i\cdot} \sim N\big(\theta, (\sigma^2/n)\sum_i c_i^2\big)$, and sample size planning is straightforward. One may even set common performance requirement for "standardized" c_i's, say $\sum_i c_i^2 = 1$, to target a common sample size for all such contrasts.

The test for $H_0 : \mu_1 = \ldots = \mu_I$ is based on $X^2 = n\sum_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2/\sigma^2$, which follows a noncentral χ^2 distribution with $(I-1)$ df and noncentrality parameter $\lambda = n\sum_i(\mu_i - \bar{\mu})^2/\sigma^2$; $X^2/(I-1) \sim F_{(I-1),\infty}$ for $\lambda = 0$. The power of the test might be monotone in λ, but who has the mental capacity to unpack $\sum_i(\mu_i - \bar{\mu})^2$ into individual μ_i's for intuitive perception, or do different μ_i configurations sharing the same $\sum_i(\mu_i - \bar{\mu})^2$ value represent equivalent alternatives in practice?

For σ^2 unknown, t-tests and F-test will be used. One needs a ballpark value of σ^2 for sample size planning, and the performance guarantee only holds approximately.

More complicated cases are less tractable, but given *any* specific setting and a specific test, one can always generate data from specific alternatives of interest, to simulate the power *one sample size, one alternative at a time.*

2.3 Further Notes

Sample size planning is realistic for single-purpose studies involving quantitative measures.

Sample size planning is generally *unrealistic* for survey based studies, the main reason being that such studies typically have multiple purposes. Also, while one may treat sums/averages of Likert scales as semi-continuous, their inter-subject variability is *not* your typical physical measurement error, unlikely to remain constant for different cohorts of subjects; this would invalidate possible σ^2 values otherwise obtainable from previous/pilot studies.

Of course, sample size planning is always doable for binary polling using simple random sampling, as noted at the end of §2.1, but the n typically runs into four figures or more.

3 Mixed-Effect Models

Consider $Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\mathbf{b} + \epsilon_i$, or in matrix terms, $\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{\epsilon}$, where $\mathbf{b} \sim N(\mathbf{0}, B)$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$, independent. One has $E[\mathbf{Y}] = X\boldsymbol{\beta}$ and $\text{Var}(\mathbf{Y}) = \sigma^2 I + ZBZ^T$. This is a mixed-effect model with the *fixed effects* in $\mathbf{x}^T\boldsymbol{\beta}$ and the *random effects* in $\mathbf{z}^T\mathbf{b}$. The matrix B is structured involving a few parameters, and terms in $\sigma^2 I + ZBZ^T$ are known as *variance components.*

Fixed effects should be reproducible "in the future," such as lab settings and physical measurements, whereas random effects are "sampled from the general population" in which the analysis results are to be applied, say the machine operators or batches of supplies.

3.1 Random Intercepts

Random effects provide a convenient device for modeling correlations among observations. Two common examples are repeated measures in longitudinal studies and observations taken from clustered subjects.

In longitudinal studies, a subject is followed over time and multiple observations are taken from the same subject along the way. Rewrite the model as $Y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b} + \epsilon_{ij}$ for the jth observation from subject i, one may set $\mathbf{z}_{ij}^T\mathbf{b} = b_i$ for $b_i \sim N(0,\tau^2)$, independent; b_i's are the *subject effect* and $B = \tau^2 I$, and the intra-subject correlation is seen to be $\tau^2/(\tau^2+\sigma^2)$.

With clustered subjects, say in medical studies involving multiple clinics, one may assume inter-cluster independence but often has to entertain intra-cluster correlations. Write $Y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i + \epsilon_{ij}$ for the observation taken from the jth subject in cluster i, for $b_i \sim N(0,\tau^2)$; $\tau^2/(\tau^2+\sigma^2)$ is then the intra-cluster correlation.

As "mean components," b_i's modify the intercept to be subject/cluster-specific.

3.2 Parameter Estimation

It is straightforward to write down the joint likelihood of $(\boldsymbol{\beta}, \sigma^2, B)$, but the MLE of the variance components are typically biased. The preferred method is to estimate $\boldsymbol{\beta}$ and (σ^2, B) separately.

Write $I - X(X^TX)^{-1}X^T = UU^T$, where U is of full column rank, $U^TU = I$. One may use $U^T\mathbf{Y} \sim N(\mathbf{0}, \sigma^2 I + U^T ZBZ^T U)$ to estimate (σ^2, B) via MLE; the method is known as restricted ML (REML), and typically delivers unbiased estimates.

Given (σ^2, B), one may minimize $(\mathbf{Y} - X\boldsymbol{\beta})^T(\sigma^2 I + ZBZ^T)^{-1}(\mathbf{Y} - X\boldsymbol{\beta})$ to estimate $\boldsymbol{\beta}$, or minimize $(\mathbf{Y} - X\boldsymbol{\beta} - Z\mathbf{b})^T(\mathbf{Y} - X\boldsymbol{\beta} - Z\mathbf{b}) + \sigma^2\mathbf{b}^TB^{-1}\mathbf{b}$ to estimate $(\boldsymbol{\beta}, \mathbf{b})$ jointly. The $\hat{\boldsymbol{\beta}}$ from the two approaches are identical.

R package `lme4` implements normal-error mixed-effect models in function `lmer()` with syntax mimicking that of `lm()`. Random intercepts can be entered as `fit=lmer(y~...+(1|id),...)`, where `id` is a factor of subject/cluster IDs; `summary(fit)` will report the REML estimates of (τ^2, σ^2) along with $\hat{\boldsymbol{\beta}}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}})$, but `fitted(fit)` returns $X\hat{\boldsymbol{\beta}} + Z\hat{\mathbf{b}}$.

3.3 Non-Gaussian Regression

Inserting $\mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\mathbf{b}$ in the place of $\mathbf{x}^T\boldsymbol{\beta}$, one may incorporate random-effects in non-Gaussian regression. REML is no longer possible. The joint likelihood of $(\boldsymbol{\beta}, \mathbf{b}, B)$ should be easy to write down, but the marginal likelihood of $(\boldsymbol{\beta}, B)$ is generally intractable.

Function `glmer()` in package `lme4` implements non-Gaussian mixed-effect models for the families implemented in `glm()`; $(\boldsymbol{\beta}, \mathbf{b})$ appear to be estimated jointly given B, and B is estimated via some numerical approximation of the marginal likelihood of $(\boldsymbol{\beta}, B)$.

Package `ordinal` implements mixed-effect proportional odds models in function `clmm()`.

4 Blocking

Blocks are some physical entities on which experiments are conducted, and an additive block effect is typically a nuisance. Blocks crossed with treatment levels often help to enhance statistical power, whereas blocks nested under treatment levels are typically the experimental units.

4.1 Paired t-Test and Crossed Blocks

Consider $Y_{ij} = \mu_i + b_j + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma^2)$, $b_j \sim N(0, \tau^2)$, with treatment levels $i = 1, 2$ and subjects $j = 1, \ldots, n$; b_j's are the subject effect as in §3.1 and ϵ_{ij}'s are measurement errors. One is interested in inferences concerning $\delta = \mu_1 - \mu_2$. In practice, the inter-subject variability is often much larger than the measurement error, so if feasible, paired data allowing the cancellation of b_j's are usually preferred over independent data.

For paired t-test, one works with $d_j = Y_{1j} - Y_{2j} = \delta + (\epsilon_{1j} - \epsilon_{2j}) = \delta + e_j$. The test for $H_0 : \delta = 0$ is based on $t = \sqrt{n}\bar{d}/s_d$ with $(n-1)$ df, where $s_d^2 = \frac{1}{n-1}\sum_j (d_j - \bar{d})^2 = \frac{1}{n-1}\sum_j (e_j - \bar{e})^2 = s_e^2$.

Working with the standard additive ANOVA $Y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij}$ with both α_i and b_j as fixed effects, the test for $H_0 : \alpha_1 = \alpha_2 = 0$ is based on $F = \text{MSA}/\text{MSE}$ with $(1, n-1)$ df's, where $\text{MSA} = \sum_{i,j}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2/(2-1) = n\bar{d}^2/2$ and $\text{MSE} = \frac{1}{(2-1)(n-1)}\sum_{i,j}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2 = \frac{1}{n-1}\sum_{i,j}(\epsilon_{ij} - \bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot j} + \bar{\epsilon}_{\cdot\cdot})^2 = s_e^2/2$. The F-test here duplicates the paired t-test.

Keep the same setting as above but allow more than 2 treatment levels, say I. For any contrast of μ_i's, $\theta = \sum_i c_i\mu_i$ with $\sum_i c_i = 0$, $\hat{\theta} = \sum_i c_i\bar{Y}_{i\cdot} = \theta + \sum_i c_i(\bar{b} + \bar{\epsilon}_{i\cdot}) = \theta + \sum_i c_i\bar{\epsilon}_{i\cdot}$, so inferences concerning the contrasts of μ_i's should remain the same regardless whether b_j's are taken as fixed or random. This remains true when one has more than one observation per cell, as long as the design remains balanced. The standard error for a general linear combination of μ_i's, say the intercept, would however differ for b_j's treated as fixed or random, as $\bar{b}$ will not cancel out.

In the balanced case, the REML estimate of σ^2 in the mixed-effect model should be identical to the MSE in the standard additive ANOVA *unless* $\hat{\tau}^2 = 0$, so the standard errors for the contrasts of μ_i's should be the same coming from `lm(y~trt+blk)` or `lmer(y~trt+(1|blk))`; possible numerical differences should be due to differences in $\hat{\sigma}^2$.

By the way, a nonparametric test based on intra-block ranking is **Friedman's test**, which reduces to the sign test for pairs; it is implemented in R function `friedman.test()`.

4.2 More Paired Tests

Inserting the structure $\mu_i + b_j$ into non-Gaussian regression frameworks, one may induce paired tests for binary, ordinal, or event count data. For inferences concerning the contrasts of μ_i's, fixed and random b_j's may no longer be technically equivalent, but treating b_j's as fixed effect does not feel too much off with "balanced designs."

What amounts to a "balanced design" should be straightforward for Bernoulli or ordinal responses, but binomial and Poisson responses can be perceived as sums of smaller parts; $Y_1 + Y_2 \sim \text{Bin}(m_1 + m_2, p)$ for $Y_i \sim \text{Bin}(m_i, p)$, $Y_1 + Y_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$ for $Y_i \sim \text{Poisson}(\lambda_i)$. For $Y_{ij} \sim \text{Bin}(m_{ij}, p_{ij})$, where $\log p_{ij}/(1 + p_{ij}) = \mu_i + b_j$, a balanced design should be $m_{ij} = m$. For $Y_{ij} \sim \text{Poisson}(\delta_{ij}e^{\mu_i + b_j})$ with *known* exposure δ_{ij}, a balanced design should be $\delta_{ij} = \delta$; an offset $\log(\delta_{ij})$ can balance the estimation but not the design.

For a concrete case, consider Bernoulli pairs $Y_{ij} \sim \text{Bin}(1, p_{ij})$, where $\log p_{ij}/(1 + p_{ij}) = \alpha_i + b_j$ for $\alpha_1 = -\alpha_2 = \alpha$; b_j's are parameters of fixed effect, free of constraint. The log likelihood of a pair (Y_{1j}, Y_{2j}) is seen to be

$$l(\alpha, b) = Y_1(\alpha + b) - \log(1 + e^{\alpha + b}) + Y_2(-\alpha + b) - \log(1 + e^{-\alpha + b}),$$

where the subscript j is omitted in the notation. For the $(1, 1)$ and $(0, 0)$ pairs, $l(\alpha, b)$ is maximized at $b = \pm\infty$, contributing no information on α. For the $(0, 1)$ and $(1, 0)$ pairs, $l(\alpha, b)$ is maximized at $b = 0$. The profile log likelihood of α is thus

$$(n_{1,0}-n_{0,1})\alpha-(n_{1,0}+n_{0,1})\log\{(1+e^{\alpha})(1+e^{-\alpha})\}=2\{n_{1,0}\alpha-(n_{1,0}+n_{0,1})\log(1+e^{\alpha})\},$$

simply the log likelihood of $n_{1,0}\sim\text{Bin}(n_{1,0}+n_{0,1},\frac{e^{\alpha}}{1+e^{\alpha}})$, where $n_{0,1}$, $n_{1,0}$ are the respective number of $(0,1)$, $(1,0)$ pairs. The induced paired test for $H_0:\alpha=0$ thus reduces to a sign test, or the McNemar test for a 2×2 square table (see §5.2).

4.3 Nested Random Blocks

For random blocks nested under treatment levels, one may write $Y_{ijk}=\mu_i+b_{j(i)}+\epsilon_{ijk}$, where $\epsilon_{ijk}\sim N(0,\sigma^2)$, $b_{j(i)}\sim N(0,\tau^2)$. We assume the balanced case, with $i=1,\ldots,I$, $j=1,\ldots,J$, $k=1,\ldots,K$, which allows clean formulas.

Inferences concerning μ_i are based on $\bar{Y}_{i\cdot\cdot}\sim N\big(\mu_i,\frac{1}{J}(\tau^2+\sigma^2/K)\big)$, and $\sum_{i,j}(\bar{Y}_{ij\cdot}-\bar{Y}_{i\cdot\cdot})^2=\sum_{i,j}(b_{j(i)}-\bar{b}_{\cdot(i)}+\bar{\epsilon}_{ij\cdot}-\bar{\epsilon}_{i\cdot\cdot})^2$ has expectation $I(J-1)(\tau^2+\sigma^2/K)$, so one may simply work with the block means $\bar{Y}_{ij\cdot}=\mu_i+b_{j(i)}+\bar{\epsilon}_{ij\cdot}=\mu_i+e_{ij}$; remember that the blocks here are the experimental units. This works as long as the block size K is fixed, regardless whether $j(i)$'s are balanced.

One may also use $\mu_i+b_{j(i)}$ in non-Gaussian regression, where one could entertain one observation per block. In fact, even in the Gaussian case, $b_{j(i)}$ and ϵ_{ij} could be identifiable if $\epsilon_{ij}\sim N(0,w_{ij}\sigma^2)$ for some known w_{ij}'s unequal; this would be the case for $\bar{Y}_{ij\cdot}$ if the block sizes K_{ij} vary.

4.4 Split-Plots

Now suppose the blocks are nested under level i of one factor but crossed with level k of another factor. One may write $Y_{ijk}=\mu+\alpha_i+b_{j(i)}+\gamma_k+\epsilon_{ijk}$, where $\sum_i\alpha_i=0=\sum_k\gamma_k$; the effects of the two factors are *additive* here.

With $\hat{\mu}+\hat{\alpha}_i=\bar{Y}_{i\cdot\cdot}=\mu+\alpha_i+\bar{b}_{\cdot(i)}+\bar{\epsilon}_{i\cdot\cdot}$, inferences concerning α_i's involve $\bar{b}_{\cdot(i)}+\bar{\epsilon}_{i\cdot\cdot}$, so an estimate of $(\tau^2+\sigma^2/K)$ will go to the denominator in a test. The experimental units for α_i are the blocks indexed by $j(i)$.

With $\hat{\mu}+\hat{\gamma}_k=\bar{Y}_{\cdot\cdot k}=\mu+\gamma_k+\bar{b}_{\cdot(\cdot)}+\bar{\epsilon}_{\cdot\cdot k}$, contrasts of $\hat{\gamma}_k$'s involve only $\bar{\epsilon}_{\cdot\cdot k}$ as $\bar{b}_{\cdot(\cdot)}$ will cancel out, so only an estimate of σ^2 is needed. The experimental units for γ_k are the individuals indexed by ijk.

4.5 Miscellaneous

With the repeated measures $Y_{ij}=\mathbf{x}_{ij}^T\boldsymbol{\beta}+b_i+\epsilon_{ij}$ as in §3.1, $\mathbf{x}_{ij}$'s for the same subject i may only differ in a factor covariate. Assuming *additivity* for the effect of that factor and those of the rest covariates, one may write $\mathbf{x}_{ij}^T\boldsymbol{\beta}=\tilde{\mathbf{x}}_i^T\tilde{\boldsymbol{\beta}}+\mu_k$, where k denotes the factor level. To assess the treatment effect in the μ_k's, one may merge $\tilde{b}_i=\tilde{\mathbf{x}}_i^T\tilde{\boldsymbol{\beta}}+b_i$ as a nuisance, yielding $Y_{ij}=\mu_k+\tilde{b}_i+\epsilon_{ij}$; the term $\tilde{\mathbf{x}}^T\tilde{\boldsymbol{\beta}}$ here can be replaced by arbitrary function of the common covariates $\tilde{\mathbf{x}}$.

In unbalanced cases with $Y_{ij}=\mu_i+b_j+\epsilon_{ij}$ as in §4.1, one should treat the random b_i's as they are using a mixed-effect model. An example of such is a *mixture of paired and independent observations*, where paired data were expected but some of the pairs had one leg missing; if all pairs are crippled, b_j and ϵ_{ij} are not identifiable, and one goes back to the two-sample t-test with independent data.

Things get complicated when the blocks are neither nested nor crossed, say if the interaction $(\alpha\gamma)_{ik}$ is to be entertained in the setting of §4.4, and/or with whatever imbalance in the design. It should be a safe exercise to enter the variables as they are in the `lmer()` or `glmer()` fits using the original data.

5 Contingency Tables

Consider categorical variables $X_1, \ldots, X_m$ with $K_1, \ldots, K_m$ categories, respectively. An observation $(x_1, \ldots, x_m)$ tallies one count to a cell in a $K_1 \times \cdots \times K_m$ array. Adding over all observations, one has an m-way contingency table of dimension $K_1 \times \cdots \times K_m$.

The joint distribution of $(X_1, \ldots, X_m)$ is multinomial with $K_1 \times \cdots \times K_m$ cells, and of interest are possible structures among the cell probabilities, such as the (conditional) independence relations among the margins.

For an example, consider a 2×2 table with joint distribution $\left(\begin{smallmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{smallmatrix}\right)$ and observed counts $\left(\begin{smallmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{smallmatrix}\right)$; $(n_{00}, n_{01}, n_{10}, n_{11}) \sim \text{Multinomial}(n_{\cdot\cdot}; p_{00}, p_{01}, p_{10}, p_{11})$, where $n_{\cdot\cdot} = n_{00} + n_{01} + n_{10} + n_{11}$. The two margins are independent when $p_{ij} = p_{i\cdot} p_{\cdot j}$, where $p_{i\cdot} = p_{i0} + p_{i1}$ and $p_{\cdot j} = p_{0j} + p_{1j}$ are marginal probabilities.

5.1 Surrogate Poisson Regression

For $Y_i \sim \text{Poisson}(\lambda_i)$, $i = 1, \ldots, k$, independent, it is known that

$$(Y_1, \ldots, Y_k) | \textstyle\sum_i Y_i \sim \text{Multinomial}(\sum_i Y_i; p_1, \ldots, p_k),$$

where $p_i = \lambda_i / \sum_j \lambda_j$. Drawing on this, one may treat the cell counts in a contingency table as independent Poisson responses and fit surrogate Poisson regression with cell characteristics as covariates; the fitted cell counts always add up to the observed total, and yield the estimated multinomial probabilities with simple scaling. *Structures among the λ_i's are identical to structures among the p_i's*, but the λ_i's are free of any "unity constraint" in a form like $\sum_i p_i = 1$.

For an example, consider a 3-way table with cell counts $Y_{ijk} \sim \text{Poisson}(\lambda_{ijk})$. One may write

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk},$$

where the intercept μ *must* be included to ensure the equivalence of surrogate Poisson regression with the intended multinomial fitting. If all terms are included, one has the saturated fit $\hat{\lambda}_{ijk} = Y_{ijk}$. Keeping only the main effects, the three margins are independent of each other. Dropping terms $(\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$, one has the conditional independence of j and k given i. Conditional independence structures can be represented by undirected graphs, and this line of models are in the toolbox for graphical modeling.

For a 2×2 table, the association between the two margins is often measured by the *log odds ratio* $\eta = \log \frac{p_{00} p_{11}}{p_{01} p_{10}}$, with $\eta = 0$ at independence. Now suppose

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$$

for j, k binary. Normalizing the 2×2 "slices" of the 3-way array λ_{ijk} with fixed i, one has the conditional distributions of (j, k) given i, with a constant log odds ratio $\eta = (\beta\gamma)_{00} + (\beta\gamma)_{11} - (\beta\gamma)_{01} - (\beta\gamma)_{10}$ not varying with i.

Remember that this is *not* response modeling, but a device to explore structures among multinomial probabilities. One may explore any structure of interest for the application at hand, not just conditional independence.

Now consider a one-way table with $Y_i \sim \text{Poisson}(\lambda_i)$, $i = 0, \ldots, k$, and one is to fit a binomial model $p_i = \binom{k}{i} p^i (1-p)^{k-i}$. This can be achieved via

$$\log \lambda_i = \mu + i \log \tfrac{p}{1-p} + \log \tbinom{k}{i},$$

where $k\log(1-p)$ is absorbed into μ; the i here is numeric, its coefficient is the logit of the binomial p, and there is an offset term in $\log\binom{k}{i}$.

The classical goodness-of-fit statistic $\chi^2=\sum_i \frac{(O_i-\hat{E}_i)^2}{\hat{E}_i}$, for i a generic subscript adding over all cells, can be obtained in the sum of squared Pearson residuals; its numerical value should be close to the residual deviance of the fit.

5.2 Square Tables

Square tables result from paired observations on the same categorical variable, such as the residence of people at two time points, product ratings by two raters, etc.; not all $k \times k$ tables are square tables despite the appearance.

Square tables are generally diagonal heavy, so independence is out of question. Of potential interest is the level of diagonal dominance. Perceiving two layers of table entries, one with independent margins and the other diagonal exclusive, one may consider

$$\log\lambda_{ij}=\mu+\alpha_i+\beta_j+\gamma_1 I_{[i=j=1]}+\cdots+\gamma_k I_{[i=j=k]},$$

where $\alpha_i+\beta_j$ are for the independence layer and the $\gamma_l I_{[i=j=l]}$ terms yield saturated fits on the diagonal; this is the *mover-stayer model* of James Lindsey. The mass of the independence layer is given by $\sum_{i,j} e^{\mu+\alpha_i+\beta_j}$, and a simple calculation yields the *percent diagonal-exclusive* as an intuitive measure for diagonal dominance.

If a pair of off-diagonal cells (i,j), (j,i) demonstrate serious lack of fit to the mover-stayer model, then maybe the two categories are easily confused, and a possible merger might be considered.

A structure of common interest for a square table is its *symmetry*, $H_0: p_{ij}=p_{ji}, \forall i,j$, for which one has the **McNemar test**; check out R function `mcnemar.test()`.

5.3 Non-Integer Entries

In some applications, an observation may not be fully committing to a single category on a variable X, in which case non-integer entries could be formed in a contingency table.

For an example, consider performance ratings on a 3-level Likert scale received by a person on a team from his teammates; the team size varies so some normalization is necessary. On a 4-person team, the ratings received could be $(0,2,1)$, and on a 7-person team, the ratings received might be $(3,1,2)$. "Taking the average" is an intuitive choice, but for categorical variables, an average is not a number but a *composition* $\mathbf{p}$ (see §7), $(0,2,1)/3$ and $(3,1,2)/6$ for the hypothetical ratings given above; the usual fully committing scenario is a special case, with $\mathbf{p}$ a unit vector.

In general, an observation on a categorical variable X_i can be taken as a composition $\mathbf{p}_i$, and an observation of $(X_1,\ldots,X_m)$ is an array formed by the outer product of the $\mathbf{p}_i$'s, with total mass one. Adding over all observations, one gets a contingency table with possibly non-integer entries, which could be analyzed the same way as the usual tables with integer entries.

6 Dimension Reduction

Multiple continuous variables are generally correlated, and the variability is often concentrated in some lower dimensional spaces. Dimension reduction techniques help one to zero in onto those dimensions of most interest. The methods compiled here assume normality to various extent, are sensitive to transformations, and some are even sensitive to linear scaling.

6.1 Principal Component Analysis

Consider $\mathbf{X} = (X_1, \ldots, X_p)^T$ with $\mathrm{Var}(\mathbf{X}) = \Sigma$. Let $\Sigma = U\Lambda U^T$ be the eigenvalue decomposition of Σ, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ with λ_j's in decreasing order. $U^T\mathbf{X} = \tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_p)^T$ are the *principal components* (PCs), uncorrelated, with $\mathrm{Var}(\tilde{X}_j) = \lambda_j$. The "total" variability of $\mathbf{X}$ is quantified by $\mathrm{trace}(\Sigma) = \sum_j \lambda_j$ in the setting, invariant under orthogonal transformations of $\mathbf{X}$, and one would focus on a few leading PCs that capture a major portion of the "total" variability. Clearly, the results are sensitive to individual scalings of the X_j's, and linear combinations of the X_j's should make practical sense.

In practice, one uses the sample version of Σ in the analysis, and for that to be reliable, the sample size n should be much larger than p. Check out R functions `prcomp()` and `princomp()`.

Using the leading PCs as covariates in regression analysis instead of the original X_j's, one may eliminate multicollinearity, reduce estimation variance, but could suffer on model interpretability.

6.2 Factor Analysis

Consider $\mathbf{Y} = \boldsymbol{\mu} + L\mathbf{F} + \boldsymbol{\epsilon}$, where L is $p \times m$, $\mathbf{F} \sim N(\mathbf{0}, I)$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Psi)$ for $\Psi = \mathrm{diag}(\psi_1 \ldots, \psi_p)$; $\mathrm{Var}(\mathbf{Y}) = \Sigma = LL^T + \Psi$. $F_1, \ldots, F_m$ are the *common factors*, $\epsilon_1, \ldots, \epsilon_p$ are the *errors* or *specific factors*, and L is the *loading matrix* with l_{ij} the loading of Y_i on F_j. Note that F_j's are latent and $L\mathbf{F} = (LP)(P^T\mathbf{F}) = \tilde{L}\tilde{\mathbf{F}}$ for any $P_{m\times m}$ orthogonal, so the common factors are well defined only up to an orthogonal transformation, or *rotation*.

Technically, $\Sigma = LL^T + \Psi$ may not have a solution except for $m = p$, but minor discrepancies are part of practical estimations using empirical data. The number of common factors, m, is often obtained from a principal component analysis of $\mathbf{Y}$; likelihood ratio tests also help. Check out R function `factanal()` and R package `psych`.

Factor analysis is widely used in social sciences to uncover hidden patterns among large number of survey questions, and the analysis results may be used to group questions into subscales.

Unfortunately, it has been a common practice to apply factor analysis directly on Likert scale questions, effectively treating ordinal variables as multivariate normal. A remedy to this is to make use of the *polychoric correlation* (see §1.3), attempting a decomposition of $\Sigma = LL^T + \Psi$ for Σ a correlation matrix obtained from R package `polycor`.

6.3 Canonical Correlations

Consider $\mathbf{X}$ with $\mathrm{Var}(\mathbf{X}) = \Sigma_{xx}$, $\mathbf{Y}$ with $\mathrm{Var}(\mathbf{Y}) = \Sigma_{yy}$, and $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \Sigma_{xy}$. One seeks $\mathbf{a}$, $\mathbf{b}$ to maximize $\mathrm{Cor}(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y})$, where $\mathbf{a}^T\mathbf{a} = \mathbf{b}^T\mathbf{b} = 1$ for definiteness; the solution yields the first pair of *canonical variables* $(\tilde{X}_1, \tilde{Y}_1) = (\mathbf{a}_1^T\mathbf{X}, \mathbf{b}_1^T\mathbf{Y})$ with *canonical correlation* $\rho_1 = \mathrm{Cor}(\tilde{X}_1, \tilde{Y}_1)$. Repeating the process but keeping $\tilde{X}_j$ uncorrelated with $\tilde{X}_1, \ldots, \tilde{X}_{j-1}$ and $\tilde{Y}_j$ uncorrelated with $\tilde{Y}_1, \ldots, \tilde{Y}_{j-1}$, one gets the second pair, the third pair, etc.

Technically, canonical correlation analysis is readily available from the singular value decomposition $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2} = U\Lambda V^T$, with $\mathbf{a}_j$'s in the columns of $\Sigma_{xx}^{-1/2}U$, $\mathbf{b}_j$'s in the columns of $\Sigma_{yy}^{-1/2}V$, and ρ_j's in the diagonals of Λ. Check out R function `cancor()`.

6.4 Partial Least Squares Regression

Consider $Y_i = \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, or $\mathbf{Y} = \alpha + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. With $X_{n\times p}$ less than full column rank, which is the case when $p > n$ or with severe collinearity, $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$ no longer works.

In the spirit of forward regression, one may select the "best" linear predictor, the next "best", etc., one term at a time, *not in the columns of X but in some linear combinations thereof.*

First centralize $\mathbf{Y}$ and the columns of X to obtain $\mathbf{Y}^{(0)} = P_{\mathbf{1}}^{\perp}\mathbf{Y}$ and $X^{(0)} = P_{\mathbf{1}}^{\perp}X$, where $P_{\mathbf{1}}^{\perp} = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$. For a linear combination of the original predictors, $z = \sum_{j=1}^{p} c_j x_j = \mathbf{x}^T\mathbf{c}$ with samples in $X\mathbf{c}$, its sample covariance with the response Y is $\mathbf{Y}^{(0)T}X^{(0)}\mathbf{c}/(n-1)$. By Cauchy-Schwarz, $(\mathbf{Y}^{(0)T}X^{(0)}\mathbf{c})^2 \leq (\mathbf{Y}^{(0)T}X^{(0)}X^{(0)T}\mathbf{Y}^{(0)})(\mathbf{c}^T\mathbf{c})$, with the maximum attained at $\mathbf{c} \propto X^{(0)T}\mathbf{Y}^{(0)}$, so the "best" linear predictor of $\mathbf{Y}$ is in $\mathbf{z}_1 \propto X^{(0)}X^{(0)T}\mathbf{Y}^{(0)}$. Working with $\mathbf{Y}^{(1)} = P_{\mathbf{z}_1}^{\perp}\mathbf{Y}^{(0)}$ and $X^{(1)} = P_{\mathbf{z}_1}^{\perp}X^{(0)}$, for $P_{\mathbf{z}_1}^{\perp} = I - \mathbf{z}_1\mathbf{z}_1^T/(\mathbf{z}_1^T\mathbf{z}_1)$, one obtains the next "best" linear predictor in $\mathbf{z}_2 \propto X^{(1)}X^{(1)T}\mathbf{Y}^{(1)}$, and so on so forth, up to say $\mathbf{z}_l$. One then fit $\mathbf{Y} = \alpha + (\mathbf{z}_1, \ldots, \mathbf{z}_l)\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$.

Clearly, $\mathbf{1}^T\mathbf{z}_j = \mathbf{z}_j^T\mathbf{z}_k = 0, \forall j \neq k$, so $\hat{\alpha} = \bar{Y}$, $\hat{\tilde{\beta}}_j = (\mathbf{z}_j^T\mathbf{Y})/(\mathbf{z}_j^T\mathbf{z}_j)$. $P_{\mathbf{z}}^{\perp}$'s are idempotent so $X^{(j)T}\mathbf{Y}^{(j)} = X^{(j)T}\mathbf{Y}$, thus $\mathbf{Y}$ needs no updating. The process can be stopped adaptively as with standard forward regression, or up to a prespecified l, but it ends automatically once $\mathbf{z}_j^T\mathbf{Y} = 0$.

Actually, the extra sum of squares (SS) gained by adding $X\mathbf{c}$ is $(\mathbf{Y}^TX\mathbf{c})^2/\mathbf{c}^T(X^TX)\mathbf{c}$, not $(\mathbf{Y}^TX\mathbf{c})^2/\mathbf{c}^T\mathbf{c}$, but the $X\mathbf{c}$ maximizing the extra SS is the projection of $\mathbf{Y}$ in the column space of X, solving the least squares problem directly.

While similar to the PCs of §6.1 that decompose the variability of X variables, the $\mathbf{z}_j$'s here decompose the linear association of X variables with Y.

R implementations of partial least squares regression can be found in many packages, such as `caret` and `pls`.

6.5 Discriminant Analysis

Bayes Rule under Multivariate Normal

Consider $\mathbf{X}_i \sim N(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, 2$. Observing a future $\mathbf{X}$, one is to identify which population it might have come from. The Bayes rule compares the log likelihood ratio

$$-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)+K = -\tfrac{1}{2}\mathbf{x}^T(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{x}+(\boldsymbol{\mu}_1^T\Sigma_1^{-1}-\boldsymbol{\mu}_2^T\Sigma_2^{-1})\mathbf{x}+K$$

against a threshold that is determined by the prior of i and the misclassification costs; the *discriminant function* $-\frac{1}{2}\mathbf{x}^T(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1^T\Sigma_1^{-1} - \boldsymbol{\mu}_2^T\Sigma_2^{-1})\mathbf{x}$ is *quadratic* in $\mathbf{x}$.

If $\Sigma_1 = \Sigma_2$, the discriminant function reduces to a *linear* function in $\mathbf{x}$, $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}\mathbf{x}$.

With training sample sizes much larger than the dimension of $\mathbf{X}$, one simply substitutes the sample versions of $\boldsymbol{\mu}_i$ and Σ_i.

Fisher's Linear Discriminants

Consider a one-way ANOVA structure with multivariate responses, $\mathbf{X}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij}$, $i = 1, \ldots, g$, $j = 1, \ldots, n_i$, $\boldsymbol{\epsilon}_{ij} \sim N(\mathbf{0}, \Sigma)$. One seeks linear combinations of $\mathbf{X}$ that best separate the groups.

Write $W = \sum_{i,j}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T$, $B = \sum_{i,j}(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})^T$, where $\bar{\mathbf{X}}_i = \sum_j \mathbf{X}_{ij}/n_i$, $\bar{\mathbf{X}} = \sum_{i,j}\mathbf{X}_{ij}/\sum_i n_i$. One looks to maximize $F_{\mathbf{c}} = (\mathbf{c}^TB\mathbf{c})/(\mathbf{c}^TW\mathbf{c})$ with respect to $\mathbf{c}$; $F_{\mathbf{c}}$ is proportional to the overall F-test in a standard one-way ANOVA with observations $\tilde{X}_{ij} = \mathbf{c}^T\mathbf{X}_{ij}$.

Given the eigenvalue decomposition $W^{-1/2}BW^{-1/2} = U\Lambda U^T$ with eigen vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{g-1}$ associated with the non-zero eigenvalues $\lambda_1 > \cdots > \lambda_{g-1}$, $F_{\mathbf{c}}$ is maximized at $\mathbf{c}_1 \propto W^{-1/2}\mathbf{u}_1$ with value λ_1, and $\mathbf{c}_1^T\mathbf{x}$ is the *first discriminant.* Imposing constraints $\mathbf{c}^TW\mathbf{c}_i = 0, \forall i < k$, $F_{\mathbf{c}}$ is maximized at $\mathbf{c}_k \propto W^{-1/2}\mathbf{u}_k$ with value λ_k, yielding the *k-th discriminant* $\mathbf{c}_k^T\mathbf{x}$. For W singular, one may substitute the Moore-Penrose inverse W^+ in the place of W^{-1}.

For $g = 2$, $B \propto (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T$, $\mathbf{u}_1 \propto W^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, so Fisher's discriminant is given by $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T W^{-1}\mathbf{x}$, the same as the sample version of $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1}\mathbf{x}$ since $W \propto \hat{\Sigma}$.

R implementations of linear and quadratic discriminant analysis can be found in package `MASS`, in the `lda()` and `qda()` functions.

7 Compositions

A composition refers to the proportions of a set of parts that make up a whole. For example, local residential trash are collected as recyclables and non-recyclables, and per household observations could come in as $(73.5\%, 26.5\%)$, $(37.4\%, 62.6\%)$, etc.

Compositions are closely related to multinomial observations, say $\mathbf{y} \sim \text{Multinomial}(m, \mathbf{p})$, where $\mathbf{y} = (y_1, \ldots, y_k)^T$, $\sum_j y_j = m$, $\mathbf{p} = (p_1, \ldots, p_k)^T$, $\sum_j p_j = 1$. One may record a multinomial observation as $(m, \mathbf{y}/m)$, where the composition $\mathbf{y}/m$ carries just part of the information.

7.1 Distances Between Compositions

A composition can be perceived as a discrete probability distribution, and a discrepancy measure between compositions $\mathbf{p}$ and $\mathbf{q}$ could be the Kullback-Leibler $\text{KL}(\mathbf{p}, \mathbf{q}) = \sum_j p_j \log(p_j/q_j)$ or its symmetrized version $\text{SKL}(\mathbf{p}, \mathbf{q}) = \text{KL}(\mathbf{p}, \mathbf{q}) + \text{KL}(\mathbf{q}, \mathbf{p}) = \sum_j (p_j - q_j)(\log p_j - \log q_j)$. To assess discrepancies from a base line, say $\mathbf{p}_0$, $\text{KL}(\mathbf{p}_0, \mathbf{p})$ might be appropriate. In general, $\text{SKL}(\mathbf{p}, \mathbf{q})$ would be more adequate. $\text{SKL}(\mathbf{p}, \mathbf{q})$ is *not* a normed distance like the Euclidean distance.

To visualize the relative proximity of a set of compositions, one may compute the pair-wise SKL's to form a distance matrix, then feed it into *multi-dimensional scaling* (MDS). Many R packages have MDS implementations, including `cmdscale()` in package `stats`, and `isoMDS()` and `sammon()` in package `MASS`.

7.2 Regression with Compositions as Responses

In applications involving compositions, one may wish to regress observed compositions on covariates, or simply to use a "sample mean" to estimate the "population mean." Unfortunately, *compositions alone contain insufficient information to support the task.*

For logistic regression with binomial response $y \sim \text{Bin}(m, p)$, one may enter the response in two categories as $(y, m - y)$, or as y/m coupled with m as weight; y/m is the same as $(y/m, 1 - y/m)$ for two categories, a composition. With y/m alone and m missing, logistic regression can not be done properly.

Let $\mathbf{p}$ be the "sample" composition and $\boldsymbol{\pi}$ be the "population" composition, it is intuitive to argue that $E[\mathbf{p}] = \boldsymbol{\pi}$. The "weight" m however controls the "variance" of $\mathbf{p}$, without which one can not properly quantify the "lack-of-fit" of the fitted $\hat{\boldsymbol{\pi}}$.

A "sample" composition $\mathbf{p}$ should be accompanied by a "weight" m to quantify how reliable it is to portray the "population" composition $\boldsymbol{\pi}$. With $(m, \mathbf{p})$, one may simply extend logistic/multinomial regression to non-integer $(m, m\mathbf{p})$.

For $\mathbf{p}_i$ from a homogeneous population without covariate, the proper sample mean can only be defined with the "weights" m_i supplied, $\bar{\mathbf{p}} = (\sum_i m_i \mathbf{p}_i)/(\sum_i m_i)$. The m_i's in the $\bar{\mathbf{p}}$ definition could be relative, but one does need absolute m_i's to quantify the likely deviation of $\bar{\mathbf{p}}$ from $\boldsymbol{\pi}$.