The Stata Journal (2003)
3, Number 2, pp. 109–132

Multiple-test procedures and smile plots

Roger Newson
King's College London, UK
roger.newson@kcl.ac.uk

The ALSPAC Study Team
University of Bristol, UK
http://www.alspac.bris.ac.uk

Abstract.

`multproc` carries out multiple-test procedures, taking as input a list of p-values and an uncorrected critical p-value, and calculating a corrected overall critical p-value for rejection of null hypotheses. These procedures define a confidence region for a set-valued parameter, namely the set of null hypotheses that are true. They aim to control either the family-wise error rate (FWER) or the false discovery rate (FDR) at a level no greater than the uncorrected critical p-value. `smileplot` calls `multproc` and then creates a smile plot, with data points corresponding to estimated parameters, the p-values (on a reverse log scale) on the y-axis, and the parameter estimates (or another variable) on the x-axis. There are y-axis reference lines at the uncorrected and corrected overall critical p-values. The reference line for the corrected overall critical p-value, known as the parapet line, is an informal "upper confidence limit" for the set of null hypotheses that are true and defines a boundary between data mining and data dredging. A smile plot summarizes a set of multiple analyses just as a Cochrane forest plot summarizes a meta-analysis.

Keywords: st0035, smile plot, multiple-test procedure, closed testing procedure, data mining, family-wise error rate, false discovery rate, Bonferroni, Šidák, Holm, Holland, Copenhaver, Hochberg, Rom, Simes, Benjamini, Yekutieli, Krieger, Liu

1 Introduction

A p-value is defined (informally) as the probability of observing a sample difference at least as large as the one in our sample, assuming that the *population* difference is zero. However, if we take a large number of samples or calculate a large number of confidence intervals for different parameters using the same sample, the probability of *not* observing at least one "significant" difference tends to fall, even if all null hypotheses are true and all population differences are zero. A skeptical public will inevitably ask whether a reported difference is "significant" when considered as one out of a large number of parameters estimated.

Common responses to this problem use the Bonferroni or Šidák inequalities. If $P_1, \ldots, P_m$ are observed p-values, and α is a critical p-value, the Bonferroni inequality states that

$$\Pr\{\min(P_j : 1 \leq j \leq m) \leq \alpha/m\} \leq \alpha$$

The Šidák inequality (Šidák 1967) is less conservative and applies whenever the events of Type I error for different tests are mutually nonnegatively correlated, which is the case if the P_j are derived from two-tailed tests based on normally distributed test statistics. The Šidák inequality states that

$$\Pr\left\{\min\left(P_j : 1 \leq j \leq m\right) \leq 1-(1-\alpha)^{1/m}\right\} \leq \alpha$$

Most statistically minded scientists view p-values as a means to the end of defining confidence intervals or other confidence regions. If there are m parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, we may derive $100(1-\alpha/m)\%$ confidence intervals (thanks to the Bonferroni inequality) or $100(1-\alpha)^{1/m}\%$ confidence intervals (thanks to the Šidák inequality) for each of the θ_i, and the Cartesian product of these confidence intervals is a conservative rectangular confidence region for $\boldsymbol{\theta}$. In other words, we are $100(1-\alpha)\%$ confident that all the θ_i are inside their respective confidence limits. The method of rectangular confidence regions is summarized in Miller (1966) and Šidák (1967). It is not easy to calculate $100(1-\alpha/m)\%$ or $100(1-\alpha)^{1/m}\%$ confidence intervals in official Stata, which requires the `level` option of an estimation command to be an integer. A possible solution to this problem is to use the `parmest` package (Newson 1999), which is downloadable from SSC, and which now allows the calculation of multiple pairs of confidence limits with possibly noninteger confidence levels.

Most scientists, most of the time, do not use corrected confidence intervals of this kind. It is more common to use multiple-test procedures, which reject a subset of the null hypotheses and enable us to be $100(1-\alpha)\%$ confident that all, or some, of the rejected null hypotheses are false. This is often more concise, and less conservative, than giving a full list of corrected confidence limits. Also, confidence interval formulas may be less reliable at confidence levels in excess of 99.5% than at confidence levels of 95%.

Multiple-test procedures, on their own, have the disadvantage that they give information only about the statistical significance of results, as measured by the p-values, and say nothing about their practical significance in affecting practical decisions, as measured by the parameter estimates. Also, the results are not often expressed graphically. It would be useful to have a plot that summarized a set of multiple analyses just as a Cochrane forest plot summarizes a meta-analysis, giving quantitative information, at a glance, about the statistical and practical significance of the estimated parameters. To create such a plot, we developed the `smileplot` package, which carries out multiple-test procedures and, optionally, plots the p-values on a reverse log scale against the corresponding parameter estimates, with a reference line (the parapet line) separating the rejected p-values from the acceptable p-values. The parapet line is so named because, informally, null hypotheses that raise their heads above it are shot down.

2 The smileplot package

2.1 Syntax

`multproc` [`if` *exp*] [`in` *range*] [`, puncor(`*# | scalarname | varname*`)`
`pcor(`*# | scalarname | varname*`) method(`*method_name*`) pvalue(`*varname*`)`
`rank(`*newvarname*`) gpuncor(`*newvarname*`) critical(`*newvarname*`)`
`gpcor(`*newvarname*`) nhcred(`*newvarname*`) reject(`*newvarname*`) float fast`]

`smileplot` [`if` *exp*] [`in` *range*] [`, estimate(`*varname*`) logbase(`*#*`)`
`maxylabs(`*#*`) xlog nline(`*#*`) ptsymbol(`*symbolstyle*`) ptlabel(`*varname*`)`
`scatteropts(`*scatter_options*`) refopts(`*added_line_options_1*`)`
`nrefopts(`*added_line_options_2*`) urefopts(`*added_line_options_3*`)`
`crefopts(`*added_line_options_4*`) plot(`*plot*`) by(`*varlist*[`,`*suboptions*]`)`
multproc_options twoway_options]

method_name may be one of a selection of method names (see below).

`by` ...: may be used with `multproc` and `smileplot`; see [R] **by**.

2.2 Description

`multproc` takes, as input, a dataset with one observation for each of a set of multiple statistical tests, including a variable containing p-values for these tests and an uncorrected overall critical p-value specified by the user, and carries out a multiple-test procedure. This procedure calculates a corrected overall critical p-value, which has the feature that an individual null hypothesis is considered to be acceptable only if its corresponding p-value is greater than the corrected overall critical p-value. `smileplot` takes, as input, a dataset with one observation for each of a set of estimated parameters and data on their estimates and p-values. `smileplot` calls `multproc` to carry out a multiple-test procedure and then creates a smile plot, with data points corresponding to estimated parameters, the p-values (on a reverse log scale) on the y-axis, and the parameter estimates (or another variable) on the x-axis. There are y-axis reference lines at the uncorrected and corrected overall critical p-values. The y-axis reference line at the corrected overall critical p-value is known as the parapet line, and data points on or above it correspond to rejected null hypotheses. There may be an x-axis reference line at the value of a parameter under a null hypothesis (defaulting to one if the x-axis is logged, or to zero otherwise). The user can therefore see, at a glance, both the statistical significance and the practical significance of each parameter estimate. Datasets suitable for input to `multproc` and `smileplot` may be created (directly or indirectly) by `statsby` or `postfile` (in official Stata) or by the `parmest` package previously mentioned (downloadable from SSC).

`smileplot` changed markedly in style in the transition from Stata 7 to Stata 8. Users who prefer to use the Stata 7 version can still do so by using the program `smileplot7`, which is distributed as part of the `smileplot` package.

2.3 Options for multproc and smileplot

`puncor(`*# | scalarname | varname*`)` specifies the uncorrected overall critical p-value for statistical significance. This option may be specified as a number, as a scalar, or as a variable (in which case the variable is expected to contain only one nonmissing value in the sample, or one nonmissing value in each by-group if `by` *varlist*`:` is used). If absent, this option is set to `1-$S_level/100`, where `$S_level` is the value of the currently set default confidence level (see [R] **level**).

`pcor(`*# | scalarname | varname*`)` specifies the corrected overall critical p-value for statistical significance. This option may be specified either as a number, or as a scalar, or as a variable (in which case the variable is expected to contain only one nonmissing value in the sample, or one nonmissing value in each by-group if `by` *varlist*`:` is used). If absent, this option is set by the method specified in the `method()` option (see below).

`method(`*method_name*`)` specifies the multiple-test procedure method to be used for deriving the corrected p-value threshold from the uncorrected p-value threshold. This option is ignored and set to `userspecified` if the `pcor()` option is specified and is in the range $0 \leq$ `pcor()` ≤ 1. Otherwise, if `method()` is absent, it is set to `bonferroni`.

`pvalue(`*varname*`)` is the name of the variable containing the p-values. If this option is absent, `multproc` looks for a variable named `p` (as created by `parmby` or `parmest`). `multproc` carries out a multiple-test procedure on all observations selected by the `if` or `in` qualifiers, which also have nonmissing values for the variable containing the p-values.

`rank(`*newvarname*`)` is the name of a new variable to be generated that contains, in each observation, the rank of the corresponding p-value, from the lowest to the highest. Tied p-values are ranked according to their position in the input dataset. If `by` *varlist*`:` is specified, then the ranks are defined within the by-group.

`gpuncor(`*newvarname*`)` is the name of a new variable to be generated, containing, in each observation, the uncorrected overall critical p-value, as specified by the `puncor()` option or by the standard default if the `puncor()` option is not specified. This new variable will have the same value for all observations in the sample of observations used by `multproc` or `smileplot`. If `by` *varlist*`:` is specified, the value of this new variable will be the same in all observations within each by-group but may be different for observations in different by-groups if the `puncor()` option is specified as a variable with different values in different by-groups.

`critical(`*newvarname*`)` is the name of a new variable to be generated that contains, in each observation, an individual critical p-value corresponding to the original p-value in the variable specified by `pvalue()`. The values of the individual critical p-values

are defined by a nondecreasing function (specified by the `method()` option) of the ranks of the corresponding original p-values (generated by the `rank()` option). The corrected overall critical p-value is selected from the individual critical p-values in a way specified by the `method()` option, depending on whether the method specified is a one-step method, a step-down method, or a step-up method.

`gpcor(`*newvarname*`)` is the name of a new variable to be generated that contains, in each observation, the corrected overall critical p-value, as specified by the `pcor()` option or by the `method()` option if the `pcor()` option is not specified. If `by` *varlist*`:` is specified, the value of this new variable will be the same in all observations within each by-group but may be different for observations in different by-groups.

`nhcred(`*newvarname*`)` is the name of a new variable to be generated that contains, in each observation, an indicator of the credibility of the corresponding null hypothesis under the method specified by the `method()` option. This indicator is 1 if the null hypothesis is acceptable and 0 otherwise. A null hypothesis is said to be acceptable if its p-value is greater than the corrected overall p-value threshold. The set of observations with a value of 1 corresponds to a set of acceptable null hypotheses. The exact interpretation of the set of acceptable null hypotheses depends on whether the method specified controls the family-wise error rate (FWER) or the false discovery rate (FDR).

`reject(`*newvarname*`)` is the name of a new variable to be generated that contains, for each observation, an indicator of the rejection of the corresponding null hypothesis under the method specified by the `method()` option. This indicator is 1 if the null hypothesis is rejected, and 0 otherwise. The new variable generated by the `reject()` option is therefore the negation of the new variable generated by the `nhcred()` option.

`float` specifies that the generated p-value variables specified by `gpcor()`, `critical()`, and `gpuncor()` (if requested) will be created as `float` variables. If `float` is absent, these generated variables are created as `double` variables. Whether or not `float` is specified, all generated variables are stored to the lowest precision possible without loss of information.

`fast` is an option for programmers. It specifies that `multproc` and `smileplot` will not take any action to restore the original data if the user presses *Break*.

2.4 Options for smileplot only

`estimate(`*varname*`)` specifies the name of the variable to be plotted on the x-axis, which usually contains the parameter estimates. If this option is absent, `smileplot` looks for a variable named `estimate` (as created by `parmby` or `parmest`). `smileplot` carries out a multiple-test procedure by calling `multproc` for observations with non-missing values for the variables specified by the `estimate()` and `pvalue()` options, using the `if` or `in` qualifiers if these are supplied by the user. Note that the variable specified by `estimate()` may contain values that are not parameter estimates. For

instance, the observations may correspond to genes in a genome scan, the p-values may be derived from tests for associations of those genes with a disease, and the x-axis variable specified by `estimate()` may contain the positions of those genes on a chromosome map.

`logbase(#)` specifies a log base used to define the y-axis labels. This log base is a factor by which each y-axis label is divided to arrive at the next y-axis label, where the y-axis labels are ordered from the highest p-value to the lowest p-value. If absent, this option is set to 10, so the y-axis labels are set to nonpositive powers of 10. If this rule defines too many y-axis labels, the y-axis labels are set to be every kth member of the logarithmic series, where k is the minimum positive integer such that the number of y-axis labels defined in this way is not too large.

`maxylabs(#)` specifies the maximum number of y-axis labels allowed. If this option is not specified, it is set to 25, so as to be similar to the Stata 7 version of `smileplot`, which can be used with `smileplot7`. `maxylabs()` is used with `logbase()` to decide the default sequence of labels on the left y-axis. These are chosen to be spaced exponentially, separated by a factor equal to the smallest possible power of `logbase()` such that the number of labels is no more than `maxylabs()`. This is usually a sensible default, but it can be overridden by the *twoway_options*.

`xlog` specifies that the x-axis must have a log scale. It is typically used if the parameters estimated are odds ratios or geometric mean ratios. It affects the default value of the `nline()` option (see below). It may be overridden by specifications in an `xscale()` option in the *twoway_options*.

`nline(#)` specifies the position, on the x-axis, of the reference line indicating the value of the estimated parameters under the null hypothesis. If unspecified, this option is set to 1 if `xlog` is specified and to 0 otherwise. This option allows the user to plot odds ratios and geometric mean ratios on a linear scale instead of the more usual log scale. If `nline()` is set to a missing value by specifying `nline(.)`, the null reference line is suppressed. This is useful for creating "smile plots" in which the x-axis variable specified by the `estimate()` option contains values other than parameter estimates, such as positions of genes on a chromosome map.

`ptsymbol(`*symbolstyle*`)` specifies a graph symbol for the data points of the smile plot (see [G] ***symbolstyle***). If absent, it is set to `Th` (hollow triangles).

`ptlabel(`*varname*`)` specifies a variable to be used to label the data points. If this option is absent, there are no data point labels, only unlabeled data points.

`scatteropts(`*scatter_options*`)` specifies a sequence of options for the `twoway scatter` plot type. These options may include `msymbol()` and `mlabel()` options, which override the `ptsymbol()` and `ptlabel()` options, respectively, and other options specifying nondefault attributes for the symbols or labels, such as size and color. (See [G] **graph twoway scatter**.) The user can specify any of these options except for `xaxis()` or `yaxis()` because `smileplot` automatically sets the first x-axis to be the x-axis of the smile plot (specified by the `estimate()` option) and the first and second y-axes to be the left and right y-axes used by the smile plot (corresponding

to the `pvalue()` option). The second y-axis is used to display the values of the uncorrected and corrected overall critical p-values.

`refopts(`*added_line_options_1*`)` specifies a list of added line suboptions, as allowed for the `xline()` or `yline()` options (see [G] ***added_line_options***). These suboptions control the style of the x-axis and y-axis reference lines of the smile plot, corresponding to the null hypothesis, the uncorrected overall critical p-value, and the corrected overall critical p-value, respectively. The suboptions apply to all 3 of these reference lines, except if overridden by the `nrefopts()`, `urefopts()` or `crefopts()` options (see below). If `refopts()` is absent, the lines styles depend on the scheme.

`nrefopts(`*added_line_options_2*`)` specifies a list of added line suboptions, which control the style of the x-axis reference line of the smile plot, corresponding to the null hypothesis.

`urefopts(`*added_line_options_3*`)` specifies a list of added line suboptions, which control the style of the y-axis reference line of the smile plot indicating the uncorrected overall critical p-value.

`crefopts(`*added_line_options_4*`)` specifies a list of added line suboptions, which control the style of the y-axis reference line of the smile plot indicating the corrected overall critical p-value.

`plot(`*plot*`)` provides a way to add other plots to the generated graph. See [G] ***plot_option***.

`by(`*varlist*`[,`*suboptions*`])` is a `graph twoway` option and works as in [G] ***by_option***, creating one subplot for each by-group, arranged in an array as specified by the user. The corrected overall critical p-value, indicated by a line at the same level on all the subplots, is calculated from all the p-values from all the by-groups pooled together, not for the subset of p-values in each by-group individually. (This is in contrast to the use of `by` *varlist*`:`, which causes corrected individual and overall critical p-values to be calculated only from the subset of p-values in each by-group.)

multproc_options is a set of options recognized by the `multproc` command.

twoway_options is a set of options recognized by the `graph twoway` command; see [G] ***twoway_options***.

2.5 Saved Results

`multproc` and `smileplot` save the following results in `r()`:

Scalars

`r(puncor)`	Uncorrected critical p-value	`r(pcor)`	Corrected critical p-value
`r(npvalues)`	Number of p-values	`r(nreject)`	Number of p-values rejected

Macros

`r(method)`	The `method()` option

3 Methods and Formulas

We assume that there is a sequence of m distinct parameters, $\theta_1, \ldots, \theta_m$, estimated using estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$ and having the values $\theta_1^{(0)}, \ldots, \theta_m^{(0)}$ under their respective null hypotheses. Typically, $\theta_i^{(0)}$ is 0 for difference parameters such as linear regression coefficients, or 1 for ratio parameters such as relative risks. $P_1, \ldots, P_m$ denote the observed p-values for testing the m null hypotheses. Each P_i has the property that, if $0 \leq \alpha \leq 1$,

$$\Pr\left(P_i \leq \alpha \mid \theta_i = \theta_i^{(0)} \right) \leq \alpha$$

$R_1, \ldots, R_m$ denote the ranks (in ascending order) of $P_1, \ldots, P_m$, and $Q_1, \ldots, Q_m$ denote the p-values in ascending order, so that, for each i, $Q_{R_i} = P_i$. `multproc` aims to define a "credible (or acceptable) subset" of indices $C \subseteq \{1 \ldots m\}$, such that the null hypotheses $\{\theta_i = \theta_i^{(0)} : i \in C\}$ are acceptable, and the complementary set of null hypotheses $\{\theta_i = \theta_i^{(0)} : i \notin C\}$ are rejected. This is done by defining an uncorrected p-value threshold p_{unc}, calculating a corrected p-value threshold, p_{cor}, from p_{unc} and $Q_1, \ldots, Q_m$ and defining the acceptable subset, C, to be the subset of indices i such that $P_i > p_{\text{cor}}$. The output variable generated by the `nhcred()` option has values of 1 for indices in C and 0 for indices not in C. Conversely, the output variable generated by the `reject()` option has values of 0 for indices in C and 1 for indices not in C.

`smileplot` calls `multproc` and then plots the $\widehat{\theta}_i$ (or another variable) on the x-axis against the corresponding P_i on the y-axis on a reverse log scale, so that the higher a data point is, the more statistically significant it is. The y-axis reference lines correspond to p_{unc} and p_{cor}, and the x-axis reference line corresponds to an assumed common value of $\theta_i^{(0)}$ for all i, set by the `nline()` option. The smile plot is so named because, if the standard errors of the various $\widehat{\theta}_i$ are similar, the data points lie around a smile-shaped line. The higher the corners of the smile, the more reason the investigators have to be happy.

The method for calculating the corrected p-value threshold, p_{cor}, is specified by the `method()` option. The methods available are listed in Table 1, and may be classified in three ways:

- By the form of the algorithm used to calculate the corrected p-value, p_{cor}. The three forms (or step types) are one-step, step-down, and step-up.
- By the interpretation of the uncorrected overall critical p-value, p_{unc}. This may be an upper bound for the family-wise error rate (FWER) or for the false discovery rate (FDR).
- By the correlation assumed between the P_i. A method may assume independence, nonnegative correlation, or arbitrary correlation.

The remaining subsections of this section explain the three modes of classification and present the formulas.

Table 1: Multiple-test procedures specified by the `method()` option of `multproc`.

`method()`	*Step type*	*FWER/FDR*	*Correlation assumed*
`userspecified`	One-step	User-specified	User-specified
`bonferroni`	One-step	FWER	Arbitrary
`sidak`	One-step	FWER	Non-negative
`holm`	Step-down	FWER	Arbitrary
`holland`	Step-down	FWER	Non-negative
`liu1`	Step-down	FDR	Non-negative
`liu2`	Step-down	FDR	Arbitrary
`hochberg`	Step-up	FWER	Independence
`rom`	Step-up	FWER	Independence
`simes`	Step-up	FDR	Non-negative
`yekutieli`	Step-up	FDR	Arbitrary
`krieger`	Step-up	FDR	Independence

3.1 Formulas for one-step, step-down, and step-up methods

Each method works by specifying a nondecreasing sequence of individual critical p-values $c_1, \ldots, c_m$, corresponding to the ordered p-values $Q_1, \ldots, Q_m$. These c_i can be output by the `critical()` option. Once these c_i have been specified, a method selects an overall corrected critical p-value, p_{cor}, from the c_i in one of three ways:

- One-step: The c_i are all equal to a common value, p_{cor}, defined by a rule not dependent on i.
- Step-down: p_{cor} is set to the minimum c_i, such that $Q_i > c_i$, if such a c_i exists, and to the maximum critical p-value c_m otherwise.
- Step-up: p_{cor} is set to the maximum c_i, such that $Q_i \leq c_i$, if such a c_i exists, and to the minimum critical p-value c_1 otherwise.

Therefore, a one-step procedure subjects all the Q_i to the same "significance hurdle"; a step-down procedure subjects the Q_i in ascending order to increasingly easy "hurdles" until the first one fails; and a step-up procedure subjects the Q_i in descending order to increasingly difficult "hurdles" until the first one succeeds. Different methods of each of the three step types differ by the methods for specifying the c_i. The rules, together with references justifying them, are as follows:

One-step methods

1. `userspecified`.

$$c_i = p_{\text{cor}}$$

   where p_{cor} is specified by the user as the `pcor()` option.

2. `bonferroni`.

$$c_i = p_{\text{cor}} = p_{\text{unc}}/m$$

3. `sidak` (Šidák 1967).

$$c_i = p_{\text{cor}} = 1 - (1 - p_{\text{unc}})^{1/m}$$

Step-down methods

1. `holm` (Holm 1979).

$$c_i = p_{\text{unc}}/(m - i + 1)$$

2. `holland` (Holland and Copenhaver 1987).

$$c_i = 1 - (1 - p_{\text{unc}})^{1/(m-i+1)}$$

   Note that the Holland–Copenhaver procedure used by `multproc` is the simplified version of the procedure in the original reference, which also specifies a more complicated version of the procedure, using logical dependencies between the null hypotheses.

3. `liu1` (Benjamini and Liu 1999a; Sarkar 2002).

$$c_i = 1 - \left\{ 1 - \min\left(1\,,\, \frac{m}{m-i+1} p_{\text{unc}}\right) \right\}^{1/(m-i+1)}$$

4. `liu2` (Benjamini and Liu 1999b).

$$c_i = \min\left\{ 1\,,\, \frac{m}{(m-i+1)^2} p_{\text{unc}} \right\}$$

   Note that the two Benjamini–Liu methods can, in principle, yield corrected p-values up to and including 1, and therefore p-values greater than the uncorrected p-value.

Step-up methods

1. `hochberg` (Hochberg 1988).

$$c_i = p_{\text{unc}}/(m - i + 1)$$

   Note that the c_i are the same as those for the step-down Holm procedure.

2. `rom` (Rom 1990).

   The c_i are defined by "backwards recursion", starting with c_m and defining the other c_i in terms of the c_k for $k > i$:

$$c_i = \begin{cases} p_{\mathrm{unc}}, & \text{if } i = m \\ (m-i+1)^{-1}\left\{\sum_{j=1}^{m-i} c_m^j - \sum_{j=2}^{m-i} \binom{m-i+1}{j} c_{i+j-1}^j\right\} & \text{if } i < m \end{cases}$$

3. `simes` (Simes 1986; Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001, first method).

$$c_i = \frac{i}{m} p_{\mathrm{unc}}$$

4. `yekutieli` (Benjamini and Yekutieli 2001, second method).

$$c_i = \frac{i}{m \sum_{j=1}^{m} j^{-1}} p_{\mathrm{unc}}$$

5. `krieger` (Benjamini, Krieger, and Yekutieli 2001).

$$c_i = \frac{i}{\widehat{m}_0} \frac{p_{\mathrm{unc}}}{(p_{\mathrm{unc}} + 1)}$$

   where $\widehat{m}_0$ is the number of acceptable null hypotheses calculated by substituting $p_{\mathrm{unc}}/(p_{\mathrm{unc}} + 1)$ for p_{unc} in the `simes` method. The `krieger` method is therefore a two-stage method, where the first stage involves using a modified `simes` method to calculate $\widehat{m}_0$ as an estimate of the number of true null hypotheses, and the second stage involves using a further-modified `simes` method to calculate the c_i.

3.2 FWER-controlling and FDR-controlling procedures

Traditionally, when scientists carry out multiple tests and wish to define an "upper bound" to the set of null hypotheses that are true, they control the family-wise error rate (FWER), defined as the probability that at least one true null hypothesis is rejected. If the uncorrected p-value threshold is p_{unc}, the corrected p-value threshold p_{cor} is chosen so that, if a subset of null hypotheses $\theta_i = \theta_i^{(0)}$ is true, the probability of at least one of the corresponding observed P_i being equal to or less than p_{cor} is no greater than p_{unc}. In practice, procedures controlling the FWER usually err on the side of conservatism, so that the true FWER is less than p_{unc}. In general, a FWER-controlling procedure defines a conservative $100(1 - p_{\mathrm{unc}})\%$ confidence region for a set-valued parameter, namely the set of null hypotheses that are true. This confidence region is a set of subsets of null hypotheses. Usually (but not always), this confidence region is the power set of a set of credible or acceptable null hypotheses. In this case, we are $100(1-p_{\mathrm{unc}})\%$ confident that the set of true null hypotheses is some subset (possibly empty) of the acceptable set. The FWER-controlling procedures implemented in `multproc` all generate confidence regions that are power sets of an acceptable set, which can be specified by the new variable

generated by the `nhcred()` option. Whether or not the confidence region is a power set, it usually has the hereditary property, which is to say that, if a subset of null hypotheses is in the confidence region, any subset of that subset is also in the confidence region. Multiple-test procedures generating confidence regions with the hereditary property are closed testing procedures, as discussed in Marcus, Peritz, and Gabriel (1976) and Wright (1992). A more recent textbook on traditional 20th century approaches to multiple comparisons is Hsu (1996).

Multiple-test procedures controlling the FWER have the disadvantage that they are often very conservative, leading to low power to detect real differences. Worse still, the power is lost progressively and tends to 0 for detection of true population differences of any given size, as the number of estimated parameters increases. If we use a FWER-controlling procedure on two disjoint sets of measured parameters and then use the same FWER-controlling procedure on the union of the two sets, the critical corrected p-value for the union will nearly always be lower than the critical p-values for either of the two component sets. This is because the corrected critical p-value is approximately inversely proportional to the number of measured parameters, or exactly inversely proportional in the case of the Bonferroni procedure. It follows that, with FWER-controlling procedures, it is possible to combine several apparently productive data mining expeditions to form a single apparently unproductive data mining expedition.

Benjamini and Hochberg (1995) proposed to remedy this difficulty by using less conservative multiple-test procedures, which control the false discovery rate (FDR) instead of the FWER. FDR-controlling procedures have the advantage of detecting more differences as "significant", at the price of being $100(1-p_{\mathrm{unc}})\%$ confident that *some* of these differences are real, instead of being $100(1-p_{\mathrm{unc}})\%$ confident that *all* of these differences are real. If we denote by R the number of null hypotheses rejected by a multiple-test procedure, denote by V the number of these rejected null hypotheses which are in fact true, and define

$$Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0, \end{cases} \tag{1}$$

the FDR is defined as the expectation $\mathrm{E}(Q)$. If all the null hypotheses are true, the FDR is the FWER. At the other extreme, if the number of measured parameters θ_i is large and a large proportion of them are appreciably different from the null-hypothesis values $\theta_i^{(0)}$, the probability that $R = 0$ will be very small, and $1 - \mathrm{FDR}$ will approximate to the expectation of the positive predictive power. (The positive predictive power is here defined as the proportion of rejected null hypotheses that are in fact false, equal to $1-Q$ if $R \neq 0$.) If we use `multproc` with a FDR-controlling procedure such as the `simes` method, the FDR will be no more than the value p_{unc} given by the `puncor()` option. FDR-controlling procedures are (rightly or wrongly) a fashionable area of statistics at present, and new methods are being developed all the time. A good place to follow recent developments is Yoav Benjamini's web site at *http://www.math.tau.ac.il/~ybenja/*.

The quantity Q has the attractive property that, if we combine two disjoint sets of measured parameters into one combined set, the value of Q for the combined set will be a mean of the values of Q for the component sets, weighted by the values of R for the component sets, provided that the component R-values are both positive. Therefore, if we use $1-Q$ as a measure of productivity of a data mining expedition (in terms of true discoveries per false discovery), a union of multiple productive data mining expeditions will be a single large productive data mining expedition.

The interpretation of the FDR is still controversial. However, one possible interpretation, in terms of confidence levels, is as follows. If $0 < \gamma \leq 1$, we have the inequality

$$\Pr(Q \geq \gamma) \leq \mathrm{E}(Q)/\gamma = \mathrm{FDR}/\gamma$$

and therefore

$$\Pr(Q < \gamma) = 1 - \Pr(Q \geq \gamma) \geq 1 - \mathrm{FDR}/\gamma$$

Therefore, if FDR$\leq p_{\mathrm{unc}}$, we can be $100(1 - p_{\mathrm{unc}}/\gamma)\%$ confident that Q will be strictly less than γ. In other words, if we control the FDR at $p_{\mathrm{unc}} = \beta\gamma$, we can be $100(1 - \beta)\%$ confident that over $100(1 - \gamma)\%$ of any rejected null hypotheses will be false. In particular, if we choose $\gamma = 1$, we can be $100(1 - p_{\mathrm{unc}})\%$ confident that $Q < 1$, or in other words, that, *if* any null hypotheses are rejected, *then* at least some of these rejected null hypotheses will be false. For instance, if we set $p_{\mathrm{unc}} = 0.05$, we can choose $\gamma = 1$ and $\beta = 0.05$ and be 95% confident that at least some of any detected differences will be real, or choose $\gamma = 0.5$ and $\beta = 0.1$ and be 90% confident that over half of any detected differences will be real. Alternatively, if we set $\beta = 0.05$, $\gamma = 0.05$ and $p_{\mathrm{unc}} = 0.05 \times 0.05 = 0.0025$, we can be 95% confident that over 95% of any detected differences will be real.

The FDR, like the FWER, can be used for generating confidence regions for a set-valued parameter, namely the set of null hypotheses that are true. Given a set of rejected null hypotheses, a confidence region for the set of true null hypotheses can be defined as follows. If we choose $\gamma = 1$, the confidence region is the set of all subsets of null hypotheses that do not contain the rejected set as a nonempty subset. If we choose any other γ, the confidence region is the set of all subsets of null hypotheses that do not contain at least 100γ percent of the rejected set as a non-empty subset. These confidence regions are not power sets but have the hereditary property, so FDR-controlling procedures are closed testing procedures. In particular, the empty set is always in the confidence region because it is impossible to prove a null hypothesis.

FDR-controlling procedures typically have more power to detect real differences than FWER-controlling procedures with the same value of p_{unc}, especially if the number, m, of measured parameters is large. The price of this increased power is that FDR-controlling procedures typically have a larger proportion of false discoveries than FWER-controlling procedures with the same value of p_{unc}. This is because, instead of aiming for the perfectionist goal of no false discoveries, a FDR-controlling procedure aims to control the number of false discoveries to an acceptable proportion of the number of true discoveries. It is not usually clear which of the discoveries are false, or even how many of the discoveries are false, because the FDR is the expectation of Q, not a deterministic value

of Q. If the number of measured parameters is extremely large, if a fairly large number of null hypotheses is expected to be rejected, and if the joint sampling distribution of the p-values is not highly correlated, some kind of consistency law might apply to R and Q, making Q very close to the FDR. In this case, we could therefore be 100% confident that at least $100(1-p_{\text{unc}})\%$ of detected differences are real, rather than being $100(1-p_{\text{unc}})\%$ confident that at least some of the differences are real. This ideal situation might be expected to hold if the set of multiple parameters measured is the total statistical output of a productive scientist over the scientist's career, or even the total statistical output of a community of scientists over a year. However, a typical scientific report might feature only a few tens or hundreds of measured parameters, of which the number of detected differences might be in single figures, or low double figures. In this case, the proportion of these detected differences that are false will not necessarily be bounded above by the FDR, even if we use an FDR-controlling procedure.

A further caution about the interpretation of the FDR arises from the fact that R may be 0. Zaykin, Young, and Westfall (2000) raised this caution in response to Weller et al. (1998), who advocated the use of FDR-controlling procedures in genomic analyses. By (1), the value of Q is 0 by definition if no null hypotheses are rejected. The FDR can therefore be expressed as

$$\text{FDR} = \Pr(R=0)\text{E}(Q|R=0) + \Pr(R>0)\text{E}(Q|R>0) = \Pr(R>0)\text{E}(Q|R>0),$$

where $1-\text{E}(Q|R>0)$ is the conditional mean of the positive predictive power, given that some null hypotheses are rejected. It follows that the conditional mean proportion of false discoveries, given that there are any discoveries at all, is greater than the advertised FDR by a factor of $1/\Pr(R>0)$. This is not much of a problem if this factor is very close to 1, as it will be if the number of measured parameters is large and a sizable proportion of the differences measured have a high probability of being detected. However, if the number of differences detected by a FDR-controlling procedure is only 1 or 2, the conditional mean proportion of those that are spurious might be appreciably more than the advertised FDR.

For all the above reasons, the interpretation of results from FDR-controlling procedures remains controversial. However, new FDR-controlling procedures continue to be developed and tested, with the result that this area of statistics is in a state of flux. Fortunately, `multproc` incorporates a wide choice of procedures, and new ones can easily be added as required. We might expect FDR-controlling procedures to be most useful in large-scale data mining expeditions where the prior probability that $R=0$ is very low. It is worth mentioning that Sterne, Davey Smith, and Cox (2001) recently calculated that the positive predictive power of published discoveries in the field of epidemiology might plausibly be as low as 0.53, corresponding to an FDR as high as 0.47. (This positive predictive power was derived using Bayes' theorem, assuming a probability of 90% that a tested null hypothesis is true, a typical study power of 50%, and a confidence level of 95%.) This suggests that, in epidemiological papers with large tables of results, the rate of "false alarms" might possibly be lowered if it became customary to use multiple-test procedures controlling the FDR at a lower level than 0.47.

3.3 Correlation between multiple p-values

The choice of a multiple-test procedure is also affected by assumptions about the joint sampling distribution of the individual p-values P_i, at least for those P_i that test true null hypotheses. Typically, negatively correlated p-values require a more conservative procedure than independent p-values, which in turn require a more conservative procedure than positively correlated p-values. This is because the likelihood of at least one of several p-values falling below a critical level is greater if they tend to fall below a critical level in different samples than if they fall below a critical level independently, and greater if they fall below a critical level independently than if they tend to fall below a critical level in the same samples. Therefore, other things being equal, procedures allowing arbitrary correlation are more conservative than procedures assuming independence or nonnegative correlation. These points are discussed rigorously by Šidák (1967) and Benjamini and Yekutieli (2001).

Methods assuming independence are appropriate if the p-values are calculated from independent sets of data. Methods assuming nonnegative correlation are appropriate if the p-values are from two-tailed tests using test statistics with a joint multivariate normal distribution or a joint multivariate t-distribution. Therefore, if it is appropriate to calculate confidence intervals and p-values using Stata estimation commands (which use standard errors calculated from an estimated dispersion matrix), then it is appropriate to use methods which assume nonnegative correlation. Methods allowing arbitrary correlation are appropriate if it is possible for different p-values to be negatively correlated when the null hypotheses are true, so that different tests tend to produce spuriously significant results in different samples. This might happen if the data points are patients with or without a disease, the sample size is small, and the multiple p-values are from multiple Fisher's exact tests for association between the disease and membership of multiple mutually exclusive categories (such as genotypes). It might also happen if the p-values are from one-tailed tests.

4 Examples

4.1 Oily fish consumption and fatty acids in red blood cells

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a multi-purpose birth cohort study based at Bristol University, England, involving over 14,000 pregnancies in the Avon area of England in the early 1990s, the children from which have been followed through childhood. For further information, refer to the study website at *http://www.alspac.bris.ac.uk*. At 32 weeks gestation, mothers were asked to complete a food frequency questionnaire (FFQ), asking about current consumption levels of a wide range of foods. Blood samples were taken from the mothers one or more times during pregnancy, and from the umbilical cord at birth, and the fatty acid composition of the cell membranes of the red blood cells (RBCs) was analyzed by measuring amounts of 40 fatty acids as a percentage of total cell membrane fatty acid.

One FFQ question asked about current consumption of oily fish (such as, pilchards, sardines, mackerel, tuna, herring, trout, or salmon) on an ordinal categorical scale (never/rarely, once every two weeks, 1–3 times per week or over 3 times per week). Correlations between oily fish consumption and RBC fatty acid percentages were assessed using Somers' D (clustered by pregnancy), which is discussed in detail in Newson (2002) and calculated using the `somersd` package (downloadable from SSC). Somers' D is here the difference between the probability that a randomly chosen blood sample from a higher-consuming group has a higher level of the specified fatty acid than a randomly chosen blood sample from the lower-consuming group and the probability that a randomly chosen blood sample from the lower-consuming group has a higher level of the specified fatty acid than a randomly chosen blood sample from the higher-consuming group. Therefore, Somers' D measures ordinal correlation on a scale from -1 to 1. 4,720 mothers had data on both fish consumption and maternal blood fatty acids, contributing a total of 7,159 maternal blood samples, whereas 1,733 mothers had data on both fish consumption and cord blood fatty acids, contributing a total of 1,753 cord blood samples. The Somers' D estimates and their p-values were output, using the `parmest` package (mentioned earlier and downloadable from SSC), to a Stata dataset with one observation per measured Somers' D parameter. This dataset contained a variable `somd`, containing Somers' D estimates; a variable `p`, containing the corresponding p-values; and a variable `fa`, containing an identifying label for the particular fatty acid involved. These variables were then plotted using `smileplot`. For maternal blood, the Stata output is below, and the results are shown as Figure 1.

```
. * Smile plot *
. smileplot,pvalue(p) estimate(somd) ptlabel(fa) method(holland) /*
>   */ scatteropts(mlabpos(12) mlabsize(small)) refopts(lpattern(shortdash)) /*
>   */ xtitle("Somers' D for trend with oily fish group") /*
>   */ ytitle("Uncorrected P-value") ylabel(,nogrid) /*
>   */ xsize(4) ysize(2.392) saving(ofishgp_m_1.gph,replace)
Method: holland
Uncorrected overall critical P-value: .05
Number of P-values: 40
Corrected overall critical P-value: .00183023
Number of rejected P-values: 12
(file ofishgp_m_1.gph saved)
```

The options `pvalue()`, `estimate()`, and `ptlabel()` are set to the variables `p`, `somd`, and `fa`, respectively. The `method()` option is set to `holland`, and `puncor()` defaults to 0.05. The other options set are `graph` options. Using the Holland–Copenhaver-corrected critical p-value of 0.00183023, we reject 12 of the 40 p-values. In Figure 1, we see the p-values plotted against the corresponding Somers' D estimates and labeled with a fatty acid label. The letter "w" in the fatty acid label represents a Greek omega (ω), so that, for instance, "205w3" and "226w3" represent the fatty acids 20:5 (ω-3) (or eicosapentaenoic acid) and 22:6 (ω-3) (or decosahexaenoic acid), commonly derived from fish oils, whereas "182w6" represents 18:2 (ω-6) (or α-linoleic acid), commonly derived from vegetable oils. Typical Somers' D values range from -0.1 to 0.1, so there is a lot of overlap between the distributions of RBC membrane composition in frequent fish eaters and in infrequent fish eaters. The x-axis reference line represents the value of 0 expected

for Somers' D under the null hypothesis of no association between fish consumption and RBC fatty acid level. The lower and upper y-axis reference lines represent the uncorrected and corrected critical p-values, respectively. The upper y-axis reference line (or parapet line) represents an upper bound for the set of null hypotheses that are true. We are (conservatively) 95% confident that the set of fatty acids unassociated with oily fish consumption is some subset, possibly empty, of the set of fatty acids below the parapet line. Therefore, it seems that, for whatever reason, a pregnant woman's fish consumption level is associated with the fatty acid composition of her own RBC membranes, especially with their content of fish-derived fatty acids. The importance of this association is discussed in Williams et al. (2001).

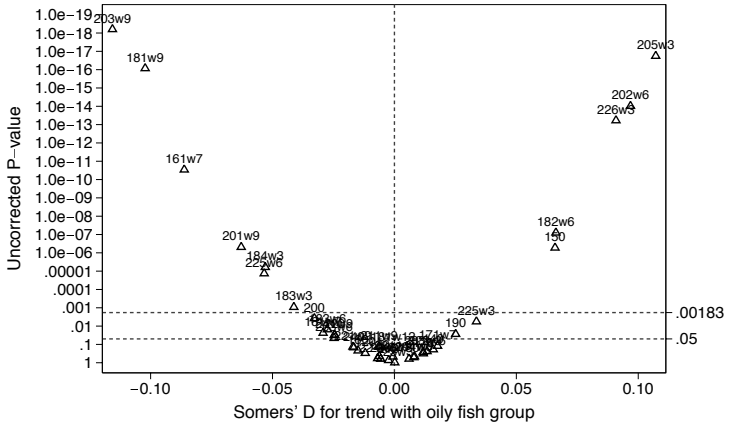

Figure 1: Smile plot of Somers' D for oily fish group and maternal RBC fatty acids.

Figure 2 shows the smile plot for associations between maternal fish consumption and cord blood fatty acids. This time, typical Somers' D values range from -0.05 to 0.05, and 2 out of the 40 fatty acids show a "nominally significant" negative association with oily fish consumption ($P \leq 0.05$). However, both are below the parapet line. This suggests that these "significant" associations could easily be the 5 percent that we would expect to be significant at the 5 percent level by chance, assuming all null hypotheses to be true. There is therefore little evidence that a pregnant woman's fish consumption predicts the fatty acid composition of her baby's RBC membranes.

Plotting the p-values on a reverse log scale implies that the higher a data point is, the more significant it is, and draws attention to points above the parapet line by giving them more than their share of space. It also gives the skeptical reader an idea of what might or might not have been achieved by publication bias or by the notorious practice of "salami science", whereby scientists distribute a given quantity of results over as many papers as possible. For instance, cynical readers might suspect that scientists would submit the results for cord blood and maternal blood fatty acids as two separate papers, rather than as one large paper, or even publish only the "significant" maternal

blood results and discard the "nonsignificant" cord blood results. These practices, if followed, will make the corrected critical p-value less conservative. However, splitting a group of results into two equal subgroups will double the Bonferroni-corrected critical p-value, whereas pooling two equal subgroups will halve the Bonferroni-corrected critical p-value. Other multiple-test procedures that control family-wise error rate (FWER), including the Holland–Copenhaver procedure used here, are slightly less conservative than the Bonferroni correction, but similar scaling laws seem to apply to the individual critical p-values c_i as we halve or double the number of measured parameters. On a reverse log scale, this implies that splitting (or pooling) a group of results will lower (or raise) the parapet line by approximately 0.3 $\log_{10}$ units, where a $\log_{10}$ unit is the space between two of the y-axis tick marks on Figures 1 and 2. Therefore, pooling the two smile plots is only likely to lose one "significant" association (involving acid 18:3 (ω-3), or α-linolenic acid).

On the other hand, the reverse log scale is not the only possible scale for plotting p-values, and other scales may be better in some ways and worse in others. Possible alternatives include power transformations and the lods transformation traditionally used by geneticists (see Sham 1998). Readers interested in investigating alternative transformations may find that a useful tool is Patrick Royston's `tgraph`, which plots data using specified monotonic transformations. See the original paper by Royston (1996) for discussion, but use the later version of the code, which is downloadable from SSC.

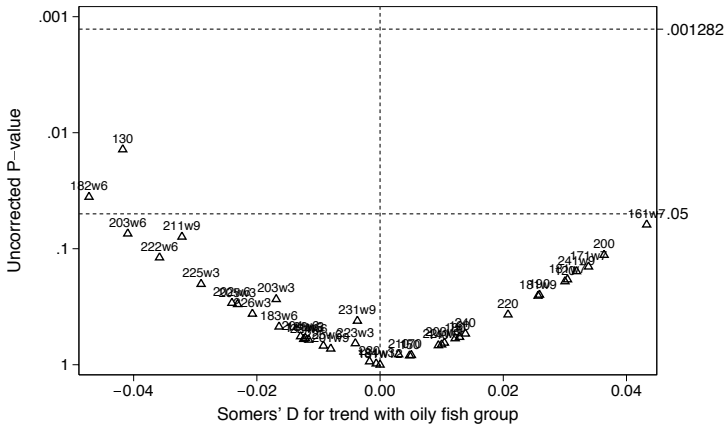

Figure 2: Smile plot of Somers' D for oily fish group and cord RBC fatty acids.

(*Continued on next page*)

4.2 Data mining using the by option

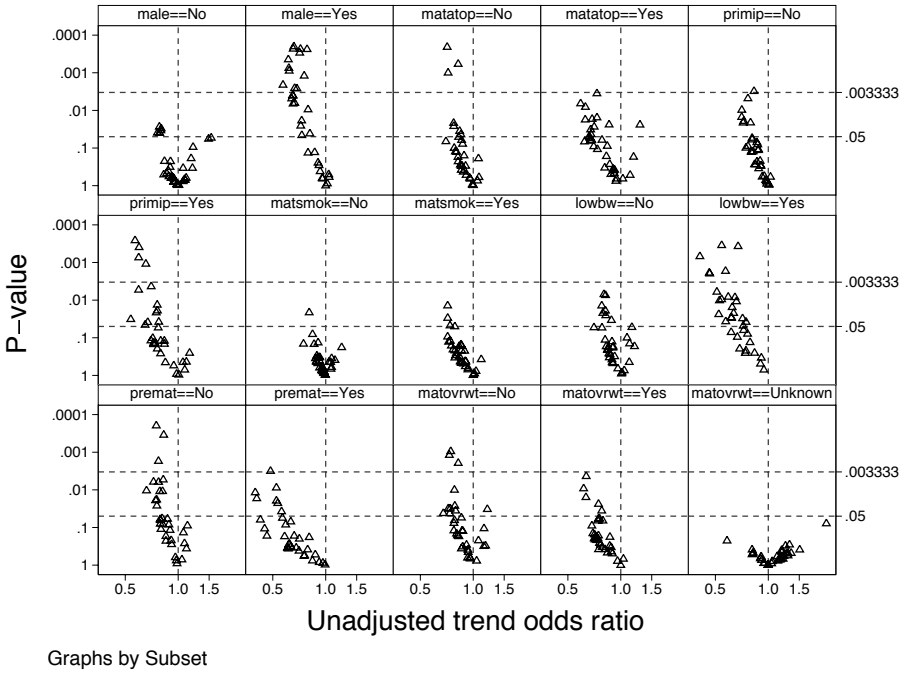

Figure 3: 495 subset-specific unadjusted ORs for persistent wheezing.

In practice, scientists are often called upon to measure more than 40 associations at a time. An example from the ALSPAC cohort involved assessing 33 FFQ-based candidate risk factors for binomial and multinomial disease outcomes, using logistic regression. The risk factors were either continuous factors, such as dietary selenium, in which case the trend was measured by a per-doubling odds ratio, or ordinal categorical factors, such as oily fish consumption, in which case the trend was measured by a per-category odds ratio. The main analyses are to be published elsewhere. However, a subsidiary analysis is presented here. The disease outcome was persistent wheezing, defined as wheezing reported at ages 0–6 months and again at ages 30–42 months. Trends for the 33 risk factors were measured in each of 15 subsets of children, defined by sex (male or female), maternal atopic disease history, primiparity, maternal smoking, low birthweight and prematurity (no or yes), and maternal overweight status (no, yes, or unknown). This implies $33 \times 15 = 495$ subset-specific odds ratios. These odds ratios were calculated, stored in a dataset with their confidence limits and p-values using the `parmest` package, and then entered into `smileplot`, using the `method(simes)` and `by(subset)` options and the default uncorrected p-value of 0.05. The resulting array of smile plots (for unadjusted odds ratios) is presented as Figure 3 and is more informative if it is enlarged and the data points are labeled by exposure.

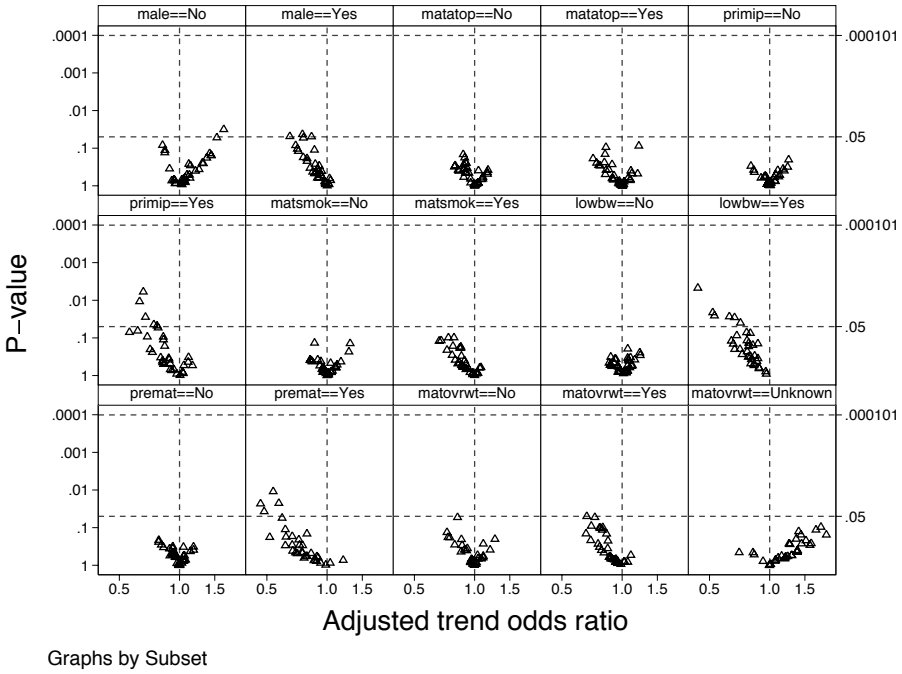

Figure 4: 495 subset-specific adjusted ORs for persistent wheezing.

The Simes procedure is an FDR-controlling procedure and rejected 33 null hypotheses of the 495. We can therefore be 95% confident that some of these 33 odds ratios are not due to chance, or 90% confident that most of them are not due to chance. None of the null hypotheses were rejected by the Bonferroni or Holland–Copenhaver procedures. The second smile plot from the left in the top row contains subset-specific odds ratios less than 1 for boys, and a few of these odds ratios are above the parapet line of $P = 0.003333$. We, therefore, have reason to believe that some foods and nutrients consumed during pregnancy by mothers are negatively associated with persistent wheezing in their sons, even though the odds ratios are part of an enormous data mining expedition. (Note that we are *not* claiming evidence of an "interaction", however defined, and we are *definitely not* claiming that the associations are restricted to subsets. We are claiming only that some associations are present, at least in subsets.)

Unfortunately, as can be seen from the smile plots, the size of the odds ratios is typically between 0.5 and 1.5. Such modest associations might not all be due to chance, but they might be due to confounding. We recalculated the 495 odds ratios, adjusting for a list of confounders, including maternal housing tenure and maternal education as proxies for "socioeconomic status". These adjusted odds ratios are smile-plotted in Figure 4. The Simes parapet line is now higher at 0.000101 because the Simes

procedure is a step-up procedure. The lowest p-values are typically higher, mostly because adjusting for confounders has widened the confidence intervals. A minority of p-values are between the uncorrected and corrected critical p-values, but drawing attention to these might possibly be viewed as "data dredging", rather than data mining. The advantage of smile plots is that these points can be seen at a glance.

4.3 Psoriasis genetics

Genetics is a field in which multiple-test procedures are increasingly required because of the unprecedented availability of so many genetic markers (Weller et al. 1998). Mallon et al. (1998) carried out a small, unmatched case-control study, with 14 cases and 147 controls, to measure the association, in HIV-positive patients, between psoriasis and the Cw-0602 gene. Using polymerase chain reaction (PCR), it was possible to distinguish 22 alleles (versions) of the Cw locus (gene). The authors estimated odds ratios between each allele and psoriasis, using Fisher's exact tests and the corresponding so-called "exact" confidence intervals (see [ST] **epitab** or Mehta, Patel, and Gray 1985). The authors predicted, *a priori*, that the Cw-0602 allele would be associated with psoriasis, whereas the other 21 alleles would not. However, it might be unreasonable to expect a skeptical public to believe this, so a Bonferroni correction was used.

We have re-analyzed the data using `smileplot` and `multproc`. The data were reformatted into a dataset with one observation for each of the 22 alleles, and with variables `label`, `or`, and `p_exact`, containing, respectively, the allele name, the odds ratio with psoriasis, and the Fisher exact p-value. The small number of cases implied that the log-odds ratios would be far from Normally distributed, and, indeed, some odds ratios were zero. Although the alleles are not mutually exclusive (as each patient has one from each parent), we would expect that, if all null hypotheses are true, the events of Type I error for different alleles might be negatively associated. Therefore, it makes sense to use the Holm procedure (to control the FWER) or, possibly, the Yekutieli procedure (to control the FDR). The program output, in part, was as follows:

```
. smileplot,pv(p_exact) esti(or) ptl(tlabel) me(holm) nline(1) /*
>  */ refopts(lpattern(shortdash)) scatteropts(mlabsize(medium) mlabpos(12)) /*
>  */ xlab(0(1)12) ytitle("Fisher's exact P-value") ylab(,nogrid) /*
>  */ xsize(4) ysize(2.392) saving(smplot1.gph,replace)
Method: holm
Uncorrected overall critical P-value: .05
Number of P-values: 22
Corrected overall critical P-value: .00238095
Number of rejected P-values: 1
(file smplot1.gph saved)
. more
. multproc,pv(p_exact) me(bonferroni)
Method: bonferroni
Uncorrected overall critical P-value: .05
Number of P-values: 22
Corrected overall critical P-value: .00227273
Number of rejected P-values: 1
```

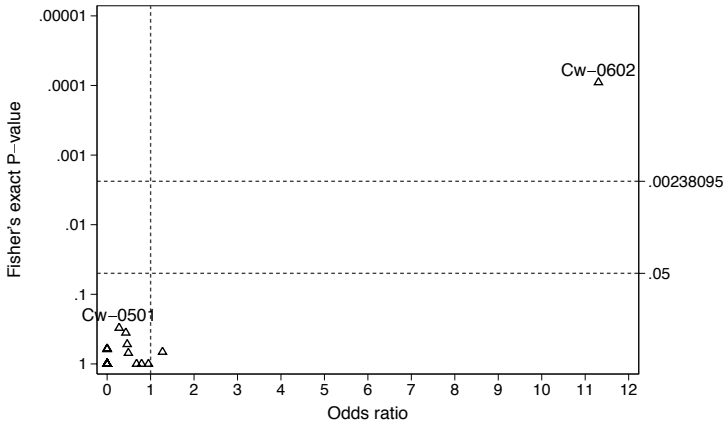

Figure 5: Odds ratios for association of 22 Cw alleles with psoriasis.

The smile plot is given as Figure 5 and was made using the Holm procedure. Note that the `ptlabel()` option has been set to a new variable, `tlabel`, so that only the data points with the two lowest p-values are labeled. The `nline()` option ensures that the null-hypothesis line is at 1 rather than 0 (the default if `xlog` is not specified). After creating the smile plot, the program called `multproc`, which produces similar output to `smileplot` without the plot, to do a *post hoc* analysis using the `bonferroni` method. Note that the `holm` parapet line is lower than the `bonferroni` parapet line would be because the Holm procedure is a step-down procedure, and the parapet line is therefore the lowest critical p-value c_i such that $Q_i > c_i$, in this case c_2. The data point with the smallest p-value (allele Cw-0602) is clearly above the parapet. The data point with the second-smallest p-value (allele Cw-0501) is clearly below the parapet.

5 Acknowledgments

We would like to thank our collaborators in the ALSPAC Study Team (Institute of Child Health, University of Bristol, Bristol, UK) for allowing the use of their data in this paper. The whole ALSPAC Study Team comprises interviewers, computer technicians, laboratory technicians, clerical workers, research scientists, volunteers, and managers who continue to make the study possible. The ALSPAC study could not have been undertaken without the cooperation and support of the mothers and midwives who took part, or the financial support of the Medical Research Council, the Department of Health, the Department of the Environment, the Wellcome Trust, and other funders. The ALSPAC study is part of the WHO-initiated European Longitudinal Study of Pregnancy and Childhood (ELSPAC). Roger Newson's work at King's College London is financed by the UK Department of Health. Particular thanks are due to the late David Horrobin, of Laxdale Ltd., Stirling, Scotland, UK, who had the idea on which the analyses presented in Subsection 4.1 were based; Scotia Pharmaceuticals, Canada, for measuring the blood-cell fatty acid levels used in these analyses; our colleague Seif Shaheen, of King's College London, UK, who had the idea on which the analyses presented in Subsection 4.2 were based; and Nicholas J. Cox, of Durham University, UK, for a lot of very helpful and thought-provoking advice.

6 References

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57: 289–300.

Benjamini, Y., A. Krieger, and D. Yekutieli. 2001. Two staged linear step-up FDR controlling procedure. Downloadable from Yoav Benjamini's web site at *http://www.math.tau.ac.il/~ybenja/*.

Benjamini, Y. and W. Liu. 1999a. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 82: 163–170. Also downloadable from Yoav Benjamini's web site at *http://www.math.tau.ac.il/~ybenja/*.

—. 1999b. A distribution-free multiple-test procedure that controls the false discovery rate. Tech. rep., Department of Statistics and OR, Tel Aviv University, RP-SOR-99-3. Also downloadable from Yoav Benjamini's web site at *http://www.math.tau.ac.il/~ybenja/*.

Benjamini, Y. and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29: 1165–1188. Also downloadable from Yoav Benjamini's web site at *http://www.math.tau.ac.il/~ybenja/*.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800–802.

Holland, B. S. and M. D. Copenhaver. 1987. An improved sequentially rejective Bonferroni test procedure. *Biometrics* 43: 417–423.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6: 65–70.

Hsu, J. C. 1996. *Multiple Comparisons: Theory and Methods*. London: Chapman & Hall/CRC.

Mallon, E., D. Young, M. Bunce, F. Gotch, P. J. Easterbrook, R. Newson, and C. B. Bunker. 1998. HLA-Cw*0602 and HIV associated psoriasis. *British Journal of Dermatology* 139: 527–533.

Marcus, R., E. Peritz, and K. R. Gabriel. 1976. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63: 655–660.

Mehta, C. R., N. R. Patel, and R. Gray. 1985. Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables. *Journal of the American Statistical Association* 80: 969–973.

Miller, R. G. 1966. *Simultaneous Statistical Inference*. New York: McGraw–Hill.

Newson, R. 1999. dm65: A program for saving a model fit as a dataset. *Stata Technical Bulletin* 49: 2–6. In *Stata Technical Bulletin Reprints*, vol. 9, 19–23. College Station, TX: Stata Press.

—. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal* 2(1): 45–64.

Rom, D. M. 1990. A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* 77: 663–665.

Royston, P. 1996. gr21: Flexible axis scaling. *Stata Technical Bulletin* 34: 9–10. In *Stata Technical Bulletin Reprints*, vol. 6, 34–36. College Station, TX: Stata Press.

Sarkar, S. K. 2002. Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics* 30: 239–257.

Sham, P. 1998. *Statistics in Human Genetics*. London: Arnold.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 62: 626–633.

Simes, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73: 751–754.

Sterne, J. A. C., G. Davey Smith, and D. R. Cox. 2001. Sifting the evidence—what's wrong with significance tests? *British Medical Journal* 322: 226–231.

Weller, J. I., J. Z. Song, D. W. Heyen, H. A. Lewin, and M. Ron. 1998. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* 150: 1699–1706.

Williams, C., E. E. Birch, P. M. Emmett, K. Northstone, and the Avon Longitudinal Study of Pregnancy and Childhood (ALSPAC) Study Team. 2001. Stereoacuity at age 3.5 y in children born full-term is associated with prenatal and postnatal dietary factors: a report from a population-based cohort study. *American Journal of Clinical Nutrition* 73: 316–322.

Wright, S. P. 1992. Adjusted p-values for simultaneous inference. *Biometrics* 48: 1005–1013.

Zaykin, D. V., S. S. Young, and P. H. Westfall. 2000. Using the false discovery rate approach in the genetic dissection of complex traits: A response to Weller et al. *Genetics* 154: 1917–1918.

About the Author

Roger Newson is a Lecturer in Medical Statistics at King's College, London, UK, working principally on research projects in asthma epidemiology. He wrote the `smileplot` package.