Using the *cherry* R package

Jelle Goeman Aldo Solari

Package version 0.2-6
Date 2012-04-04

Contents

1 Citing *cherry*

If you use the *cherry* package, please cite the paper J. J. Goeman and A. Solari, Multiple testing for exploratory research, *Statistical Science*, 26 (4) 584–597.

2 Introduction

The *cherry* package is a package for multiple hypothesis testing. The approach used by *cherry* is specially designed for exploratory inference.

2.1 Exploratory inference

Suppose a researcher has performed an experiment, possibly a genomics experiment or some other experiment in which a large or small number of statistical hypotheses have been tested. From the results of the experiment the user wants to select a number of 'promising' results. Which results are considered promising may depend on any mixture of considerations, such as (unadjusted) significance, effect size, and domain knowledge. One question the researcher may ask is how many false positive findings are present in the selected list. This is the question the *cherry* package is designed to answer.

The suggested way of working with this package is as follows. Before the data are gathered, two choices have to be made. The first choice is what hypotheses are of potential interest. This is the working collection of hypotheses. The second choice is what statistical test is to be done for each hypothesis, and for each intersection (or combination) of hypotheses. Examples of such choices are given in the rest of this vignette. Nothing else has to be decided before data collection. After data collection, the researcher can study the data as much as he or she likes, before deciding on a collection of rejected hypotheses of interest. This choice, together with the working collection and the tests is fed into the *cherry* package, which will return a confidence statement on the maximum number of true null hypotheses, i.e. false rejections, among the selected set. On the basis of this assessment, the researcher may reevaluate and come up with a different selection of interesting hypotheses, for which *cherry* will again give a confidence statement. These confidence statements are not compromised by previous looks at the data, but remain valid however many times the researcher comes up with a new set.

In *cherry* the traditional roles of user and algorithm in multiple testing have been reversed. In classical multiple testing procedures the user's task is to set an error rate to be controlled, and the task of the multiple testing procedure is to decide what hypotheses to reject. In *cherry*, the user chooses what hypotheses to reject, and the multiple testing procedure calculates the error rate. This way of working allows the user much more freedom and control. Most importantly, the error rates that are calculated are not invalidated by multiple looks at the data, and the user is free to study the data in every possible way before finally deciding what rejections to make. This makes the method ideally suited for exploratory research.

Throughout the package and the vignette, the words "true hypothesis" and "false rejection" are used interchangeably, as rejection of a true hypothesis amounts to a false rejection. Similarly, "false hypothesis" and "correct rejection" often used as synonyms.

The theory behind the methods is explained in detail in the papers Goeman and Solari (2011a) and Goeman and Solari (2011b) and we refer readers to these papers for

details. In this document, we present some worked out examples to demonstrate how *cherry* can be used and how its results should be interpreted. We start in Section 3 with methods based on p-values that use either Fisher combinations or Simes inequality to combine p-values. These methods have the advantage that they are quick to use even if tens of thousands of hypotheses have been tested, but they do depend on assumptions and cannot always be used. Next, Section 4 presents the general method in its full flexibility. This general method is computationally quite intensive and should only be used if the total number of hypotheses in the multiple testing problem is not much greater than 20. More methods of different types will be added to the *cherry* package in the future.

For an explanation of the theory behind the methods, and for more explanation and examples, see the papers Goeman and Solari (2011a) and Goeman and Solari (2011b).

2.2 Intersection hypotheses and local tests

The *cherry* package assumes that before data collection the user is able to make a complete list of all hypotheses that are of potential interest. For each of these hypotheses, a statistical hypothesis test must be formulated. Moreover, statistical tests must also be formulated for all possible intersections of these chosen hypotheses. An intersection hypothesis of a collection of hypotheses is a hypothesis that is true if and only if every hypothesis in the collection of hypotheses is true. For example, if null hypothesis H_A asserts that the mean treatment effect of drug A is zero, and the hypothesis H_B asserts that the mean treatment effect of drug B is zero, then the intersection hypothesis $H_{AB} = H_A \cap H_B$ asserts that the treatment effects of both drugs are zero. A statistical test for an intersection hypothesis is known as a *global test* or a *local test*. Examples of frequently used global and local tests are F-tests in ANOVA models or regression models, or gene set tests in microarray data analysis.

The user of the *cherry* package is free to choose any local test that is valid for the data at hand. The methods in Section 4 allow the user to work with any self-defined local test. Unfortunately, these methods are computationally very expensive, and should only be used when the total number of hypotheses in the multiple testing problem is not much greater than 20. For specific choices of the local test, that use either Fisher combinations or Simes inequality, quicker algorithms are available. Special functions for those local tests are describe in Section 3

3 Methods based on p-values

In this section we present some simple methods that can be used when each hypothesis in the working collection has been tested and given a p-value, and when intersection hypotheses are tested with simple p-value-based global tests such as Simes inequality or Fisher combinations. We will illustrate these methods using a data set NAEP taken from Benjamini and Hochberg (2000)

```
> library(cherry)
> data(NAEP)
```

These are p-values for 34 null hypotheses for 34 American states. In each state, a test was performed for the hypothesis that there was no change in the average eighth-grade mathematics achievement scores between 1990 and 1992. We can assume the

data for different states, and therefore the p-values, to be independent. The p-values are sorted here, but that is not necessary for *cherry*.

The fact that we have p-values means that the choice of a statistical test for the individual hypotheses has already been made. The only remaining choice is the test for intersection hypotheses. Two options to test these intersection hypotheses are presented in this section.

3.1 Fisher combinations

Fisher combinations are based on the fact that if a p-value p_i is under the null hypothesis, it is uniformly distributed (or stochastically smaller than that). Consequently, $-\log(p_i)$ is exponentially distributed with parameter $\lambda = 1$, and $-2\log(p_i)$ is χ^2 distributed with 2 degrees of freedom. If a number of r p-values is independent, then $-2\sum_{i=1}^{r}\log(p_i)$ is χ^2 distributed with $2r$ degrees of freedom. This suggests the Fisher combination test, a test for intersection hypotheses based on negative sums of logarithms of p-values that uses critical values from a χ^2-distribution. This test is valid only when p-values for hypotheses that are under the null are independent. Fisher combination tests are powerful for detecting the presence of many small effects, but not so powerful for detecting few larger ones. Functions in *cherry* that use this test are the `pickFisher` and `curveFisher` functions.

Suppose the researcher has chosen Fisher combinations and, after looking at the NAEP data, picks the hypotheses HI, MN and IA, and wants to know how many correct rejections he or she would make when rejecting these null hypotheses. This can be found with

```
> pickFisher(NAEP, c("HI","MN","IA"))

Rejected 3 hypotheses at confidence level 0.95.
Correct rejections >= 2; False rejections <= 1.
```

The hypotheses are referred to by name in this function call, but they can be referred to by any other selection method method, such as a logical vector, index or negative index. We can conclude at the default 95% confidence that among the hypotheses HI, MN and IA there are at least 2 false hypotheses and at most one true one.

Leaving out the second argument, *select*, means rejecting all 34 hypotheses. This gives us an upper confidence bound to the number of true hypotheses in the complete working collection

```
> pickFisher(NAEP)

Rejected 34 hypotheses at confidence level 0.95.
Correct rejections >= 19; False rejections <= 15.
```

There are at least 19 correct rejections, i.e. false null hypotheses, and at most 15 false rejections, i.e. true null hypotheses among the 34 hypotheses. The 95% confidence set for the number of true null hypotheses among the 34 goes from 0 (lower bound) to 15 (upper bound), and consequently the same confidence set for the number of false hypothese from 19 to 34. For the selected set HI, MN and IA, the 95% confidence set for the number of true hypotheses goes from 0 to 1. It is important to know (and central to the method underlying the *cherry* package) that all these confidence intervals are simultaneous. There is no need to correct for multiple testing when selecting the most interesting one from all these confidence intervals.

Several options can be set in `pickFisher`. Setting the type I error rate *alpha* (default 0.05) changes the confidence level of the statements made. Setting *silent* to `TRUE` suppresses printing to the screen of the result. The `pickFisher` function's return value is the lower bound of the number of false null hypotheses, i.e. correct rejections.

```
> res <- pickFisher(NAEP, silent=TRUE)
> res

[1] 19
```

The `curveFisher` function can give some additional information over `pickFisher`. Called without further arguments, the function returns lower bounds for the number of false null hypotheses, like `pickFisher`, but simultaneously for selecting the hypotheses with the smallest 1,2,3,..., p-values. The results are displayed in a graph unless the *plot* argument is set to `FALSE`. From these results, we see with 95% confidence that 19 false null hypotheses are are present among all 34 hypotheses, as we saw before, but also that these 19 false null hypotheses must be among the 25 hypotheses with smallest p-values.

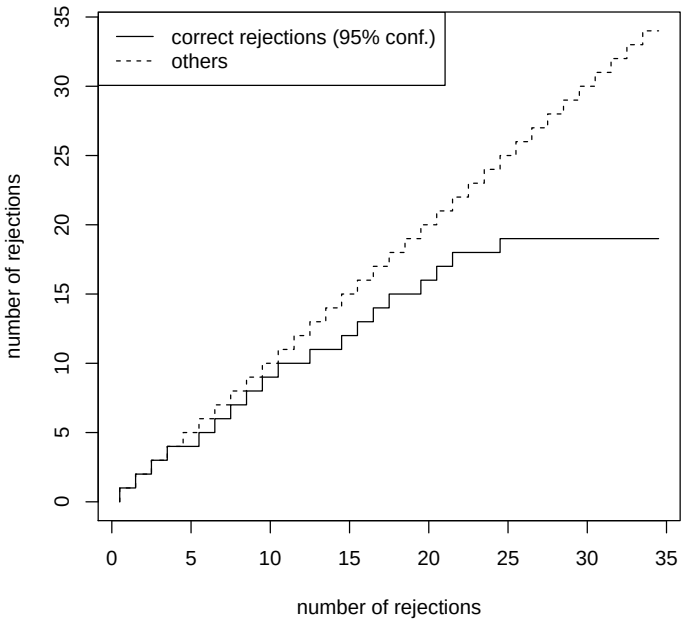
```
> res <- curveFisher(NAEP)
> res

RI NC HI MN NH IA CO TX ID AZ KY OK CT NM WY FL PA NY OH CA
 1  2  3  4  4  5  6  7  8  9 10 10 11 11 12 13 14 15 15 16
MD WV VA WI IN LA MI DE ND NE NJ AL AR GA
17 18 18 18 19 19 19 19 19 19 19 19 19 19
```

The curveFisher function may be further tailored. The *select* argument can be used to consider only a subset of the hypotheses, and the function will return the number of false null hypotheses among the 1,2,... smallest p-values in the selected set. Alternatively, the *order* argument can be used to set the order in which hypotheses should be rejected, rather than taking the order of increasing p-values. Compare

```
> curveFisher(NAEP, select=c(8,3,4,2), plot=FALSE)

NC HI MN TX
 1  2  3  3

> curveFisher(NAEP, order=c(8,3,4,2), plot=FALSE)

TX HI MN NC
 0  1  2  3
```

Here, the first call uses the rejection order NC, HI, MN, TX, determined by the p-values; the second uses the rejection order TX, HI, MN, NC, given by the input. We interpret the second result as follows. Choosing TX only we have a result of 0, which means that we do not have 95% confidence that TX corresponds to a false null hypothesis. The second result is 1, which means that choosing both TX and HI, we have detected at least one false null hypothesis with 95% confidence. The third result, 2, refers too rejection of TX, HI and MN, the fourth result to the collection of all four hypotheses.

```
> curveFisher(NAEP)
```



3.2 The Simes inequality

The Simes inequality says that for a sequence of ordered p-values $p_{(1)}, \ldots, p_{(r)}$, under the null hypothesis we have $p_{(i)} \geq i\alpha/r$ simultaneously for all i with probability at least $1 - \alpha$. This inequality holds if p-values are independent, as shown by Simes, but also under some forms of positive correlation, as shown by Sarkar and Chang (1997). Simes' inequality suggests Simes' test, a test that rejects an intersection hypothesis of r hypotheses if either the smallest p-value is smaller than α/r or the second smallest is smaller than $2\alpha/r$, or ..., or the largest p-value is smaller than $r\alpha/r = \alpha$. This test can be used as a local test in the *cherry* package. It is valid if p-values of true null hypotheses are always either independent or positively correlated. An alternative, more conservative test was formulated by Hommel; this test is valid whatever the distribution of the p-values.

The functions `pickSimes` and `curveSimes` work in exactly the same way as `pickFisher` and `curveFisher` above, and they use the same options and arguments.

```
> pickSimes(NAEP, c("HI","MN","IA"))

Rejected 3 hypotheses. At confidence level 0.95:
Correct rejections >= 2; False rejections <= 1.

> curveSimes(NAEP, plot=FALSE)

RI NC HI MN NH IA CO TX ID AZ KY OK CT NM WY FL PA NY OH CA
 1  2  3  4  4  5  6  6  6  6  6  6  6  6  6  6  6  6  6
```

```
MD WV VA WI IN LA MI DE ND NE NJ AL AR GA
 6  6  6  6  6  6  6  6  6  6  6  6  6  6
```

Comparing these results with the ones obtained for the Fisher combinations above, we see that the Simes test allows fewer rejections. This is not generally true, and Simes may be more powerful in other data sets. In general, Fisher combinations can be said to have more power if there are many small effect sizes, whereas Simes has more power in the presence of a few stronger effects.

The more conservative Hommel variant that makes no assumptions on the p-value distribution can be obtained by setting *hommel=TRUE*. This variant is more conservative than one based on the regular Simes test.

```
> pickSimes(NAEP, c("HI","MN","IA"), hommel=TRUE)

Rejected 3 hypotheses. At confidence level 0.95:
Correct rejections >= 2; False rejections <= 1.

> curveSimes(NAEP, plot=FALSE, hommel=TRUE)

RI NC HI MN NH IA CO TX ID AZ KY OK CT NM WY FL PA NY OH CA
 1  2  3  4  4  4  4  4  4  4  4  4  4  4  4  4  4  4  4  4
MD WV VA WI IN LA MI DE ND NE NJ AL AR GA
 4  4  4  4  4  4  4  4  4  4  4  4  4  4
```

4 The general method

There are many more possibilities than Fisher combinations or Simes inequality to make local tests. The `closed` function and its derivatives in *cherry* allow users to work with any type of local test. When working with such a user-defined local test, the output possibilities are greater than with the simple `pickFisher` and `curveFisher` functions above.

4.1 Defining a local test

We first illustrate the general method with the same local test as was used there: the Fisher combinations.

To illustrate these data we cannot take the NAEP data, as for 34 hypotheses the calculations will take too long. In fact, the code will not run for a collection of more than 31 hypotheses. We shall illustrate the general approach with a very small data set of 4 p-values, taken from Huang and Hsu (2007), but we take them out of the context they were presented in in that paper.

```
> ps <- c(A = 0.051, B = 0.064, C = 0.097, D = 0.108)
```

To define a local test, we must create an `R` function, such as the following

```
> mytest <- function(hypotheses) {
+   p.vals <- ps[hypotheses]
+   m <- length(p.vals)
+   statistic <- -2 * sum(log(p.vals))
+   p.out <- pchisq(statistic, df=2*m, lower.tail=FALSE)
+   return(p.out)
+ }
```

This function takes as input the names of the hypotheses that the intersection hypothesis is an intersection of, and returns a p-value, as calculated by the Fisher combinations test. Note that the function references the `ps` data we've just created. We can now call our test with

```
> mytest("A")

[1] 0.051

> mytest(c("B","C","D"))

[1] 0.02347135

> mytest(names(ps))

[1] 0.008391265
```

Note that calling `mytest` on a single hypothesis name just returns the p-value of that hypothesis.

A user can define their own favorite local test in exactly this way, by creating an R function that takes a vector of hypothesis names as input and gives a valid p-value for the corresponding intersection null hypothesis as output.

The `mytest` function used p-values as input data. This is not necessary or even typical. As a second example, we present a function for a local test based on the F-test in a multiple linear regression using the `LifeCycleSavings` example data. The null hypothesis of this test is the hypothesis that the covariates in its *hyps* argument all have regression coefficient zero in the multiple regression model fitted in `fullfit` below.

```
> hypotheses <- c("pop15", "pop75", "dpi", "ddpi")
> fullfit <- lm(sr~., data=LifeCycleSavings)
> myFtest <- function(hyps) {
+   others <- setdiff(hypotheses, hyps)
+   form <- formula(paste(c("sr~",  paste(c("1", others), collapse="+"))))
+   anov <- anova(lm(form, data=LifeCycleSavings), fullfit, test="F")
+   pvalue <- anov$"Pr("[2]        # NB replace Pr by P for for R < 2.14.0
+   return(pvalue)
+ }
> myFtest(c("pop15", "pop75"))

[1] 0.004834923
```

4.2 Performing closed testing

Next, we can perform the closed testing procedure using this definition of the local test. This is done using the function `closed`, as follows

```
> ct <- closed(mytest, names(ps))
```

Note that there is no need anymore to specify the data set that is used, because the reference to the data set is contained in the definition of `mytest`. The calculations can take a large amount of time, especially if the number of hypotheses is large.

By default, all tests are performed at level *alpha = 0.05* and only rejection or acceptance of hypotheses is stored, not adjusted p-values. It is possible to change the *alpha* argument to a different value, or to do the calculations using adjusted p-values so that the choice of the significance level may be postponed. This will be explained below in Section 4.4.

Using the `ct` object created by the call to `closed` we can now perform analysis by asking for the number of true and false hypotheses among certain sets of hypotheses of interest. Just displaying the object,

```
> ct

Closed testing result on 4 elementary hypotheses.
At confidence level 0.95: False hypotheses >= 2; True hypotheses <= 2.
```

gives a lower confidence bound on the number of false hypotheses among the collection of all four tested hypotheses. We conclude that there are at least two false null hypotheses among the four tested ones. If we are interested in a subset of hypotheses, we can use the `pick` function.

```
> pick(ct, c("A", "B"))

Rejected 2 hypotheses.
At confidence level 0.95: Correct rejections >= 1; False rejections <= 1.

> pick(ct, c("C", "D"))

Rejected 2 hypotheses.
At confidence level 0.95: Correct rejections >= 0; False rejections <= 2.
```

This gives us confidence limits for the number of false hypotheses in each chosen subset. The `pick` function returns the lower bound on the number of false hypotheses. It also displays the information on the screen, but this can be switched off if desired by setting the *silent* argument to `TRUE`. If the second argument to `pick` is left out, it is assumed that the set of all hypotheses is meant. To get back the names of the hypotheses, type

```
> hypotheses(ct)

[1] "A" "B" "C" "D"
```

From the results of `pick` above we see that there is evidence for one false hypothesis among A and B, but no such evidence among C and D. This seems to contradict the earlier statement that there was evidence for at least two false null hypotheses among the total set of A, B, C and D. However, this is only an apparent contradiction: the amount of evidence for a second false null hypothesis is not sufficient in the observed data of A and B alone, nor in the data of C and D alone, but the combined evidence of the data from all four hypotheses is sufficient.

Just like `pickSimes` and `pickFisher`, `pick` returns the lower confidence bound for the number of false null hypotheses as a number.

```
> res <- pick(ct, c("C", "D"), silent=TRUE)
> res

[1] 0
```

4.3 Defining rejections and the shortlist

The `pick` function allows users to check out any desired set of hypotheses, but gives no guidance as to what sets of hypotheses to probe. Two other functions can help to see structure in the results of the closed testing procedure.

The first of these is the `defining` function, which calculates the *defining rejections*. The defining rejections are a collection of sets of hypotheses with the property that for each set in the collection we can be confident that it contains at least one false null hypothesis. The collection of defining hypotheses is minimal in the sense that there are no smaller sets for which the same statement holds. In our `ct` object, the defining rejections are

```
> defining(ct)

[[1]]
[1] "A" "B"

[[2]]
[1] "A" "C"

[[3]]
[1] "B" "C"

[[4]]
[1] "A" "D"

[[5]]
[1] "B" "D"
```

For each of the listed defining sets, we can conclude that they contain at least one false hypothesis. For example, at least one of A and B must be false, but also at least one of A and C, and at least one of B and C. If any of the defining sets is a singleton, we can confidently conclude that the corresponding hypothesis is a false one.

The dual of the defining sets is the *shortlist*, introduced by Meinshausen (2011). Just like the defining sets, the shortlist is a collection of sets of hypotheses, but for the shortlist collection we can make the statement that at least one of the sets in the collection contains only false hypotheses. In the example,

```
> shortlist(ct)

[[1]]
[1] "A" "B"

[[2]]
[1] "B" "C" "D"

[[3]]
[1] "A" "C" "D"
```

the shortlist contains three sets. We conclude that either both hypotheses A and B are false, or all three hypotheses A and C and D or all three hypotheses B and C and D. Just as with the defining rejections, the shortlist only gives a minimum at the chosen

significance level. The possibility that all four hypotheses are false, for example, is equally compatible with the results of the closed testing procedure. However, the true set must contain at least one of the shortlist sets completely.

4.4 Adjusted p-values

Earlier calculations were all done at a significance level *alpha* of 0.05, and the results only used the rejection status of hypotheses. There is additional information is the p-values, however, and the `closed` function may also be used with adjusted p-values rather than a hard rejection yes/no. By definition, an adjusted p-value is the smallest alpha-level at which a certain hypothesis can be rejected in the multiple testing procedure. Using adjusted p-values is comparable to not setting the alpha level in advance, but simultaneously doing the same test procedure at all alpha levels. To use adjusted p-values, use

```
> cta <- closed(mytest, names(ps), alpha = NA)
```

Calculation times based on adjusted p-values can be substantially longer because more tests need to be calculated. If the user is not interested in adjusted p-values above a certain threshold, say 0.1, an alternative call is

```
> ctb <- closed(mytest, names(ps), alpha = 0.1, adjust = TRUE)
```

In this case, all adjusted p-values greater than the chosen threshold will be set to 1.

The `pick` function works slightly differently if adjusted p-values were used. It simultaneously presents results for all levels of alpha.

```
> pick(cta)
```

```
        alpha confidence true<= false>=
1 0.008391265  0.9916087      3       1
2 0.023471346  0.9765287      2       2
3 0.058232610  0.9417674      1       3
4 0.108000000  0.8920000      0       4
```

The results should be read as follows. Up to $\alpha = 0.0084$ we only have the trivial result that the number of true hypotheses is ≤ 4. From $\alpha = 0.0084$ to $\alpha = 0.023$ we get the result that at least one hypothesis is false; from $\alpha = 0.023$ to $\alpha = 0.058$, we get at least false two hypotheses, etcetera. In terms of adjusted p-values, the adjusted p-value for the statement that there are ate least two false null hypotheses among the four can be read off as 0.02347. The confidence column is simply $1 - \alpha$.

To extract adjusted p-values directly, there is the `adjusted` function. Calling `adjusted` on a set of hypotheses without extra arguments returns the adjusted p-value of the corresponding intersection hypothesis. An additional third argument *n* can be given to get the adjusted p-value for making the statement that at least n false null hypotheses occur in the chosen set, i.e. corresponding to the null hypothesis that at most $n - 1$ false null hypotheses are present.

```
> adjusted(cta, c("A", "B", "C"))
```

```
[1] 0.01314629
```

```
> adjusted(cta, c("A", "B", "C"), n=2)

[1] 0.03775654
```

We conclude that the compound hypothesis that no hypothesis among A, B and C is false, and the compound hypothesis that at most one among A, B, and C is false are both rejected at $\alpha = 0.05$, with adjusted p-values 0.013 and 0.038, respectively.

The `pick` function has one other additional feature in that the number in the table can be visualized in a plot. Setting *plot = TRUE* in `pick` results in a plot such as in Figure 1. In this plot, the adjusted p-values can be read off as tail probabilities in the plotted "distribution". These values are displayed at the top of the plot. The *plot* argument is ignored if adjusted p-values were not calculated.

```
> pick(cta, names(ps), plot=TRUE)
```

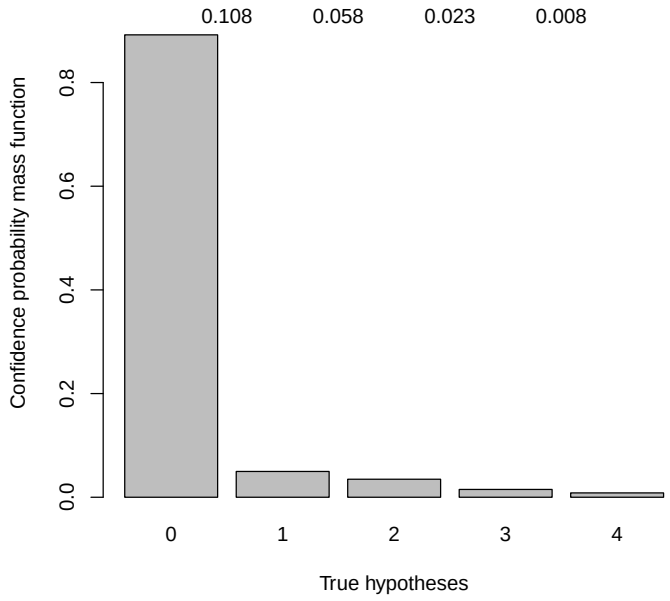


Figure 1: `pick(cta, names(ps), plot=TRUE)`

The `defining` and `shortlist` functions can be used for objects with adjusted p-values, but a specific value of alpha must be specified.

```
> shortlist(cta, alpha=0.05)

[[1]]
[1] "A" "B"

[[2]]
[1] "B" "C" "D"
```

```
[[3]]
[1] "A" "C" "D"
```

It is also possible to set the alpha level to be used in such cases in advance with the `alpha` function. Setting

```
> alpha(cta) <- 0.05
```

will cause functions such as `pick`, `defining`, `shortlist` to work as if alpha = 0.05 was set in advance. The object can be reset to adjusted p-values with

```
> alpha(cta) <- NA
```

References

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83.

Goeman, J. and Solari, A. (2011a). Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597.

Goeman, J. and Solari, A. (2011b). Rejoinder. *Statistical Science*, 26(4):608–612.

Huang, Y. and Hsu, J. (2007). Hochberg's step-up method: cutting corners off holm's step-down method. *Biometrika*, 94(4):965–975.

Meinshausen, N. (2011). Discussion of multiple testing for exploratory research by J. J. Goeman and A. Solari. *Statistical Science*, 26(4):601–603.

Sarkar, S. and Chang, C. (1997). The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, pages 1601–1608.