Using the hypergeometric distribution to test for gene set enrichment in a set of "distinguished" genes

Alan E. Berger May 31, 2022

available at https://github.com/AlanBerger/Statistical_Inference_GitHub_files

An introduction to the hypergeometric distribution

Much of this material is taken from / modified from material I have posted in the Discussion Forums for the Statistical Inference Course in the Johns Hopkins Data Science Specialization on Coursera.

First, a conceptual example showing a common type of use of the hypergeometric distribution (from a tutorial I gave at NIH on gene set enrichment analysis): Suppose there were 10,000 candies in a container, 200 of which were Ghirardelli chocolates (Ghirardelli is one of a number of brands of fine chocolate), and being given a random batch of 50 candies taken from the bin. If you got 6 or more of the Ghirardelli chocolates, were you unusually lucky? (**Indeed yes** !) On average, one would only get **6 or more** of the Ghirardelli chocolates by chance 45 times out of 100,000 random draws of 50 candies (p = 0.00045, rounded to 5 digits to the right of the decimal point).

The hypergeometric distribution h(x; m, n , k) (using the notation in the functions R provides for dealing with the hypergeometric distribution) addresses the following situation. There is a population of N objects, m of which have some specified attribute, and so n = N-m of them do not have the specified attribute; and one takes a random sample of k of the N objects (without replacement, so the same object can be chosen at most once). Then h(x; m, n, k) is the probability of there being exactly x of the objects with the specified attribute in the random sample of k of the N objects. So in the example above, N is 10,000; m is 200; n is N-m; k is 50; and the probability p = 0.00045 of getting 6 or more of the chocolates (distinguished objects) in the sample is the sum of h(x; m, n = N-m, k) over x = 6, 7, . . . , 50 (since there were only 50 samples taken, h(x; m, n = N-m, k) here is 0 for x > 50).

Below is R code for doing this calculation. Note if you want to copy code lines for use in R, it is best to copy from the .Rmd file, since sometimes text in the pdf file can contain formatting characters that R does not accept. To get a copy of the Rmd file as a text file (so in a form that one can copy usable R code from); go to the https://github.com/AlanBerger/Statistical_Inference_GitHub_files web page, click on the desired Rmd file name listed there, then the Rmd file will display - click on the "raw" button - a text version will display, then (in Windows) right click and click on "Save as" to save a text version of the Rmd file. To download a pdf file from GitHub (note web links inside pdf files ***displayed in GitHub*** don't work), open the pdf file in GitHub, and then right click on "Download" and then click on "Save link as"

```r
N <- 10000
m <- 200
k <- 50
phyper(q = 5, m, n = N-m, k, lower.tail = FALSE)  # [1] 0.000453366
```

```
## [1] 0.000453366
```

```r
##### Note we set q = 5 NOT 6 since phyper with lower.tail = FALSE
##### BY ITS DEFINITION gives the probability of the right tail of the
##### hypergeometric distribution for values GREATER THAN q

##### so phyper(q = 6, m, n = N-m, k, lower.tail = FALSE)
##### would NOT include the probabilty
##### h(6; m, n = N-m, k) which via R is

dhyper(x = 6, m, n = N-m, k)  # [1] 0.000397832
```

```
## [1] 0.000397832
```

```r
# and a check that h(x; m,  n = N-m , k) is 0 for x > min(m, k)
dhyper(x = 51, m, n = N-m, k)
```

```
## [1] 0
```

The hypergeometric distribution is symmetric with respect to the number m of objects with the specified attribute and the sample size k, meaning

h(x; m, n = N-m, k) = h(x; k, n = N-k, m)

One can check this directly (with some algebra) from the algebraic formula for the hypergeometric distribution (as given in the R help for dhyper (or phyper, qhyper, rhyper), and in https://en.wikipedia.org/wiki/Hypergeometric_distribution (which also notes this symmetry) and in many statistics textbooks. This symmetry is natural, since the hypergeometric distribution also answers the question: if one has a population of N objects and one extracts (without replacement) a subset A of size m from the population; and also (after "putting back" the elements of A) draws (without replacement) a subset B of size k from the population, then what is the probability that there will be x elements in common between the two subsets A and B. The symmetry naturally follows since one can view either A or B as the set of "distinguished" objects (the objects having a specified attribute) and the other as the random sample.

This way (using the hypergeometric distribution) of getting the probability for having obtained q or more of the objects with the specified property in the random sample is also known as the ***Fisher exact test***.

Using the hypergeometric distribution to test for gene set enrichment in a set of "distinguished" genes: preliminary discussion and notation

An important application of statistical procedures is in analyzing data consisting of the expression levels of, in many cases, on the order of 20,000 human genes (a ***genome-wide*** data set) (or expression levels of the genes in another organism). Typically one is comparing levels of gene expression in samples from two groups, for example patients with some disease vs. normal controls; or samples from patients with different tumor subtypes; or from patients who did or did not respond to some treatment.

The multiple comparisons issue

When one does many statistical comparisons (here 1 for each gene examined), as noted in the Statistical Inference class and in many statistics articles and texts, one needs to deal with the fact that one will get a number of small p-values just by random chance. It is well known that one needs to address this ***multiple testing*** issue (also referred to as the ***multiple comparisons*** problem) when doing analysis of gene expression data.

Criteria for a gene to be considered "distinguished"

The initial analysis of a gene expression data set comparing expression levels between 2 groups will in general consist of obtaining, for each gene, a p-value from an appropriate statistical test, and, to address the multiple testing issue, for analysis of gene expression data one commonly calculates the **Benjamini-Hochberg False Discovery Rate** (**BH-FDR**) (details on the definition of the BH-FDR are provided in the Appendex at the end of this article).

It is also important to have a measure of the **effect size** for each gene, addressing the question of whether the difference in average expression levels between the 2 groups is biologically important. For gene expression data, the effect size is often taken to be the ratio **r** of the average expression levels of the gene in each group, with the larger average expression level placed in the numerator (**r** is called the **fold change**; one keeps track of which group had the higher average expression level for each gene).

The first step in analyzing a gene expression data set is usually to produce a list **L** of "***distinguished***" genes (using list in the customary sense, not as an R object type) that pass specified criteria on their p-value, BH-FDR, and fold change. Note to get a "reasonably small" BH-FDR, the p-value will have to be quite small, so the BH-FDR is the "binding" criterion on the level of statistical significance required. What is an appropriate condition on the BH-FDR depends on how the list of distinguidhed genes is going to be used (this is discussed further below). A common condition on the fold change is that it (as defined above) be at least 2. There is the added requirement for a gene to be considered distinguished that the larger average expression level is large enough to be biologically meaningful, so one

avoids calling genes distinguished whose average expression levels would have no biological effect).

Extracting biological information from gene expression data

An important way to extract biological meaning from a gene expression data set is to do a follow-on analysis to see if known biologically meaningful sets of genes (called ***gene sets***, such as the genes in a known biological pathway) are up (or down) regulated as an ensemble in one of the groups being compared. This has the advantage of in effect using a sort of average expression level over the genes in the gene set (an average in one form or another tends to give more statistical power), and also there are fewer gene sets than individual genes, so the multiple testing issue (which must still be accounted for) is less severe.

A straightforward way to do this is to determine a list L of distinguished genes, and see if that list is statistically significantly "enriched" in the genes that are part of each gene set being looked at. Often it is adventageous to use 2 lists - the distinguished genes that are upregulated (have higher average expression level) in each of the 2 groups being compared (note sometimes there may be few if any distinguished genes that are upregulated in one of the two groups).

For a given list L of distinguished genes, and for each gene set S being considered, one "asks" whether there are more genes in L that are also in the gene set S than one would expect by random chance, which is precisely what the hypergeometric distribution (Fisher exact test) addresses. There are a number of tools that carry out this type of analysis; for example the **DAVID** web site at NIH: https://david.ncifcrf.gov/ In this setting an appropriate choice for the required bound on the BH-FDR for a gene to be in the distinguished list would be 0.1 or even 0.2 since the "averaging" effect of looking at biologically meaningful sets of genes implies that having 10% or even 20% false discoveries in the list L still provides enough "true signal" to get meaningful results for this type of gene set analysis. Note if one were deciding whether to invest substantial resources doing research on one or a few particular genes that showed up as statistically significantly differentially expressed between the two groups (and with a sufficiently large effect size), one would want a quite small FDR, perhaps even below 0.01 (for a real situation one would also take into account any additional information known about the gene and the biology under study).

A "nice, manageable" list size for DAVID is roughly 100 to 400 genes but we have obtained useful results from DAVID with lists of size on the order of 50 to 2000 genes. Small lists have less statistical power while large lists might potentially attenuate the "signal" from a small group of biologically significant genes occurring toward the top of a ranked list of genes.

There are also a large number of gene set enrichment analysis algorithms that use the expression levels of **all** the genes (not just a selected list of distinguished genes which depends on the particular criteria used) to determine which sets of genes, as an ensemble, are up (or down) regulated in one of the two groups being compared (a primary example is the GSEA algorithm and software and the MSigDB library of gene set collections: https://www.gsea-msigdb.org/gsea/index.jsp).

Using the hypergeometric distribution to determine if a list L of distinguished genes is enriched in genes from a gene set S

The analysis parallels the example with the Ghirardelli chocolates at the beginning of this article: Suppose the list L has k genes in it, and that x genes in the list L are in the given gene set S (e.g., the genes in a given biological pathway) having m genes in it, and the total number of genes whose expression levels were measured and analyzed was N (and that included all of the m genes in S). If one took randomly chosen sets of k distinct genes from the N genes whose expression levels were measured, then if N is much larger than k and m, on average (only) a few genes (f) of the randomly chosen k genes would in the biological pathway S. One can use the hypergeometric distribution to decide of x is statistically significantly larger than f.

For example, suppose we had a whole genome study with human subjects, so the total number of genes in the data set was (on the order of) N = 20,000; the list L of distinguished genes that were upregulated in one of the two groups had k = 270 genes in it; the gene set S had m = 128 genes in it; and there were x = 8 genes in common (8 genes in L were also in the gene set S). What is the probability that there would be 8 or more genes in common between L and S (just by random chance)? The following R code answers that question (gives the p-value for L and S having 8 or more genes in common):

```
N <- 20000
k <- 270
m <- 128
x <- 8
# get the p-value for 8 or more genes in common
# Note as covered above, we need to set q equal x-1 when lower.tail = FALSE

phyper(q = x-1, m, n = N-m, k, lower.tail = FALSE)  # [1] 0.000352061
```

```
## [1] 0.000352061
```

```
# Let's see what the p-value would be if there were only 6 genes in common
x <- 6
phyper(q = x-1, m, n = N-m, k, lower.tail = FALSE)  # [1] 0.007893501
```

```
## [1] 0.007893501
```

```
# Let's see what the p-value would be if there were only 4 genes in common
phyper(q = 4-1, m, n = N-m, k, lower.tail = FALSE)  # [1] 0.09546785
```

```
## [1] 0.09546785
```

```r
# Note how the probabilities at each individual x rapidly decrease (using dhyper):
p.at.4.5.6.7.8.9.10 <- dhyper(x = 4:10, m, n = N-m, k)
pvalues <- data.frame(x = 4:10, p = p.at.4.5.6.7.8.9.10)
pvalues
```

```
##    x            p
## 1  4 6.552753e-02
## 2  5 2.204682e-02
## 3  6 6.108188e-03
## 4  7 1.433252e-03
## 5  8 2.907343e-04
## 6  9 5.178888e-05
## 7 10 8.201666e-06
```

Appendix The definition of the Benjamini-Hochberg False Discovery Rate (BH-FDR)

If we have done m hypothesis tests, the Benjamini-Hochberg (BH) procedure for estimating the false discovery rate (FDR) is the following:

1. Sort the p-values in ascending order. In a real application the p-values would likely be a column in a data frame also containing information on what was being examined in each of the m statistical tests, so one would use the **order** function on the column (vector) of p-values and use the result to sort the rows of the data frame.

2. For each row r, $r = 1, ..., m$, of the sorted data frame, let p_r denote the p-value for the statistical test in that row. We then, for each row r, calculate (estimate) the FDR that results, denote it by FDR(r), if we consider the first r statistical tests to be the ones that are statistically significant, i.e., we consider the statistical tests whose p-values are $\leq p_r$ to be the ones that are statistically significant. An aside: note statistical significance does not necessarily mean practical significance. One also should consider whether the difference that has been measured is of practical importance which is a judgement based on knowing about the science/subject area underlying what is being tested (often phrased in terms of whether an appropriately chosen measure of the "effect size" is large enough to "really matter".)

By definition, the exact false discovery rate (having declared the first r tests having p-value $\leq p_r$ to be the ones that are statistically significant) is the number F of these for which the null hypothesis is true (so they are false positives) divided by r (the number of tests we have taken to be the ones that have given a significant result).

We don't know what F is, but we can provide an estimate for it. Recall the definition of a p-value p, which is that the probability of getting a score from a statistical test that results

in a p-value $\leq p$ if the null hypothesis is true is precisely p (assuming the statistical test is appropriate for the situation being studied and that the conditions for validity of the test are satisfied). If the m statistical tests being done are independent, and the null hypothesis was true for all m tests, then we would expect on average to get $m * p_r$ p-values that are $\leq p_r$ just by random chance. This is the estimate for F. If the alternative hypothesis is true for some of the tests, then using m in the estimate $m * p_r$ for F is conservative, but this is often a good enough approximation.

3. The Benjamini-Hochberg estimate for FDR(r) is: FDR(r) = F/r which is estimated by $m * p_r/r$

Note as the row number r increases, this can go up or down, since p_r in the numerator has been sorted to be non-decreasing as r increases, while r in the denominator increases by 1 from row to row. So if in a stretch of rows, p_r doesn't change much, the FDR will decrease in that stretch of rows, but if in a stretch of rows p_r increases a fair amount, then FDR(r) will increase. In general the statistical tests won't all be independent, for example in gene expression data there will be correlations between "related" genes, but (in my view) the Benjamini-Hochberg procedure for estimating the FDR gives a good "reality check".

The value FDR(r) is what I call an "unadjusted" BH-FDR. The "full" estimate for the BH-FDR is actually

BH-FDR(r) = minimum for $s \geq r$ of FDR(s)

The idea is if FDR(s) is smaller than FDR(r) for some row $s > r$, then the r^{th} test belongs to a (larger) set of tests that have a smaller estimated FDR, so the r^{th} test fairly "inherits" that smaller FDR value. The common convention is to set any FDR(r) values that calculate to be larger than 1 to be 1, but I find it informative to keep the values as calculated. If there are a "considerable" number of FDR values distinctly > 1 that may indicate a "problem" with the statistical test and/or with the data (such as a confounding factor or a group that has distinct subgroups within it, which could cause problems with the statistical testing). If you are doing such a calculation for someone else, ask them what their preference is, since just "handing in" FDR values that are > 1 might not otherwise "go over well".

Note (with m the total number of statistical tests that were done) FDR(m) equals $m{*}p_m/m = p_m$ which will be at most 1, so BH-FDR(r) as defined above will always be at most 1. If one wants to select the tests that satisfy BH-FDR(r) $\leq \alpha$, one finds the largest index s for which FDR(s) is $\leq \alpha$, and then tests 1 through s are the ones for which the Benjamini-Hochberg false discovery rate is $\leq \alpha$.

Note a derivation of the **Bonferroni correction** for controlling the **family-wise error rate** (**FWER**) is given here: https://github.com/AlanBerger/Statistical_Inference_GitHub_files/blob/master/derivation-and-explanation-of-the-Bonferroni-Correction.pdf

= =

As noted above much of this material is copied or derived from posts of mine in the Discussion Forums for the Statistical Inference Course in the Johns Hopkins Data Science Specialization on Coursera. As such Coursera and Coursera authorized Partners retain additional rights to this material as described in their "Terms of Use" https://www.coursera.org/about/terms

Note the reader should not infer any endorsement or recommendation or approval for the material in this article from any of the sources or persons cited above or any other entities mentioned in this article.