genomicsclass / labs Public

Code Issues 16 Pull requests 1 Actions Projects Wiki Security

master

labs / inference / association_tests.Rmd

mikelove "finish inference" History

3 contributors

229 lines (186 sloc) | 9.11 KB

~~~
---
title: "Association tests"
output: html_document
layout: page
---

```{r options, echo=FALSE}
library(knitr)
opts_chunk$set(fig.path=paste0("figure/", sub("(.*).Rmd","\\1",basename(knitr:::knit_concord$get('
~~~

```{r,include=FALSE}
set.seed(1)
```

Association Tests

The statistical tests we have covered up to now leave out a
substantial portion of life science projects. Specifically, we are
referring to data that is binary, categorical and ordinal. To give a
very specific example, consider genetic data where you have two groups
of genotypes (AA/Aa or aa) for cases and controls for a given
disease. The statistical question is if genotype and disease are
associated. As in the examples we have been studying previously, we have two
populations (AA/Aa and aa) and then numeric data for each, where disease
status can be coded as 0 or 1. So why can't we
perform a t-test? Note that the data is either 0 (control) or 1
(cases). It is pretty clear that this data is not normally distributed
so the t-distribution approximation is certainly out of the
```
```

~~~
question. We could use CLT if the sample size is large enough;
otherwise, we can use *association tests*.

#### Lady Tasting Tea

One of the most famous examples of hypothesis testing was performed by
[R.A. Fisher](https://en.wikipedia.org/wiki/Ronald_Fisher).
An acquaintance of Fisher's claimed that she could tell if milk was added
before or after tea was poured. Fisher gave her four pairs of
cups of tea: one with milk poured first, the other after. The order
was randomized. Say she picked 3 out of 4 correctly, do we believe
she has a special ability? Hypothesis testing helps answer this
question by quantifying what happens by chance. This example is called
the "Lady Tasting Tea" experiment (and, as it turns out, Fisher's friend
was a scientist herself, [Muriel Bristol](https://en.wikipedia.org/wiki/Muriel_Bristol)).

The basic question we ask is: if the tester is actually guessing, what
are the chances that she gets 3 or more correct? Just as we have done
before, we can compute a probability under the null hypothesis that she
is guessing 4 of each. If we assume this null hypothesis, we can
think of this particular example as picking 4 balls out of an urn
with 4 green (correct answer) and 4 red (incorrect answer) balls.

Under the null hypothesis that she is simply guessing, each ball
has the same chance of being picked. We can then use combinatorics to
figure out each probability. The probability of picking 3 is
${4 \choose 3} {4 \choose 1} / {8 \choose 4} = 16/70$. The probability of
picking all 4 correct is
${4 \choose 4} {4 \choose 0}/{8 \choose 4}= 1/70$.
Thus, the chance of observing a 3 or something more extreme,
under the null hypothesis, is $\approx 0.24$. This is the p-value. The
procedure that produced this p-value is called _Fisher's exact test_ and
it uses the *hypergeometric distribution*.

#### Two By Two Tables

The data from the experiment above can be summarized by a two by two table:

```{r}
tab <- matrix(c(3,1,1,3),2,2)
rownames(tab)<-c("Poured Before","Poured After")
colnames(tab)<-c("Guessed before","Guessed after")
tab
```


The function `fisher.test` performs the calculations above and can be obtained like this:

```{r}
fisher.test(tab,alternative="greater")
~~~

~~~
```

#### Chi-square Test

Genome-wide association studies (GWAS) have become ubiquitous in
biology. One of the main statistical summaries used in these studies
are Manhattan plots. The y-axis of a Manhattan plot typically
represents the negative of log (base 10) of the p-values obtained for
association tests applied at millions of single nucleotide
polymorphisms (SNP). The x-axis is typically organized by chromosome
(chromosome 1 to 22, X, Y, etc.).
These p-values are obtained in a similar way to
the test performed on the tea taster. However, in that example the
number of green and red balls is experimentally fixed and the number
of answers given for each category is also fixed. Another way to say
this is that the sum of the rows and the sum of the columns are
fixed. This defines constraints on the possible ways we can fill the two
by two table and also permits us to use the hypergeometric
distribution. In general, this is not the case. Nonetheless, there is
another approach, the Chi-squared test, which is described below.

Imagine we have 250 individuals, where some of them have a given disease
and the rest do not. We observe that 20% of the individuals that are
homozygous for the minor allele (aa) have the disease compared to 10%
of the rest. Would we see this again if we picked another 250
individuals?

Let's create a dataset with these percentages:

```{r}
disease=factor(c(rep(0,180),rep(1,20),rep(0,40),rep(1,10)),
               labels=c("control","cases"))
genotype=factor(c(rep("AA/Aa",200),rep("aa",50)),
                levels=c("AA/Aa","aa"))
dat <- data.frame(disease, genotype)
dat <- dat[sample(nrow(dat)),] #shuffle them up
head(dat)
```

To create the appropriate two by two table, we will use the function
`table`. This function tabulates the frequency of each level in a
factor. For example:

```{r}
table(genotype)
table(disease)
```

If you provide the function with two factors, it will tabulate all possible pairs and thus create
~~~

~~~

```{r}
tab <- table(genotype,disease)
tab
```

Note that you can feed `table` n factors and it will tabulate all n-tables.

The typical statistics we use to summarize these results is the odds ratio (OR). We compute the od

```{r}
(tab[2,2]/tab[2,1]) / (tab[1,2]/tab[1,1])
```

To compute a p-value, we don't use the OR directly. We instead assume
that there is no association between genotype and disease, and then
compute what we expect to see in each *cell* of the table (note: this use of
the word "cell" refers to elements in a matrix or table and has
nothing to do with biological cells).
Under the null hypothesis,
the group with 200 individuals and the group with 50 individuals were
each randomly assigned the disease with the same probability. If this
is the case, then the probability of disease is:

```{r}
p=mean(disease=="cases")
p
```

The expected table is therefore:

```{r}
expected <- rbind(c(1-p,p)*sum(genotype=="AA/Aa"),
                  c(1-p,p)*sum(genotype=="aa"))
dimnames(expected)<-dimnames(tab)
expected
```

The Chi-square test uses an asymptotic result (similar to the CLT)
related to the sums of independent binary outcomes. Using this
approximation, we can compute the probability of seeing a deviation
from the expected table as big as the one we saw. The p-value for this
table is:

```{r}
chisq.test(tab)$p.value
```

Large Samples, Small p-values
~~~

As mentioned earlier, reporting only p-values is not an appropriate way to report the results of your experiment. Many genetic association studies seem to overemphasize p-values. They have large sample sizes and report impressively small p-values. Yet when one looks closely at the results, we realize odds ratios are quite modest: barely bigger than 1. In this case the difference of having genotype AA/Aa or aa might not change an individual's risk for a disease in an amount which is *practically significant*, in that one might not change one's behavior based on the small increase in risk.

There is not a one-to-one relationship between the odds ratio and the p-value. To demonstrate, we recalculate the p-value keeping all the proportions identical, but increasing the sample size by 10, which reduces the p-value substantially (as we saw with the t-test under the alternative hypothesis):

```{r}
tab<-tab*10
chisq.test(tab)$p.value
```

Confidence Intervals for the Odds Ratio

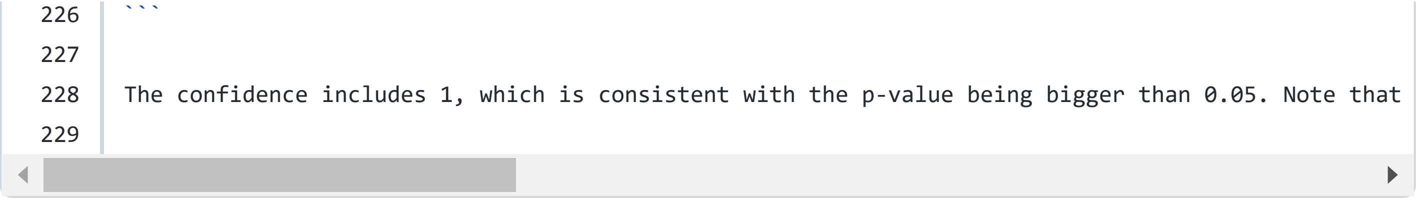
Computing confidence intervals for the OR is not mathematically straightforward. Unlike other statistics, for which we can derive useful approximations of their distributions, the OR is not only a ratio, but a ratio of ratios. Therefore, there is no simple way of using, for example, the CLT.

One approach is to use the theory of *generalized linear models* which provides estimates of the log odds ratio, rather than the OR itself, that can be shown to be asymptotically normal. Here we provide R code without presenting the theoretical details (for further details please see a reference on generalized linear models such as [Wikipedia](https://en.wikipedia.org/wiki/Generalized_linear_model) or [McCullagh and Nelder, 1989](https://books.google.com/books?hl=en&lr=&id=h9kFH2_FfBkC)):

```{r}
fit <- glm(disease~genotype,family="binomial",data=dat)
coeftab<- summary(fit)$coef
coeftab
```

The second row of the table shown above gives you the estimate and SE of the log odds ratio. Mathe

```{r}
ci <- coeftab[2,1] + c(-2,2)*coeftab[2,2]
exp(ci)
```



~~~
```

The confidence includes 1, which is consistent with the p-value being bigger than 0.05. Note that

~~~