More on Tests of 2 $\times$ 2 Tables

Statistics 149

Spring 2006

Randomization Test in 2 $\times$ 2 Tables

If you believe that normality assumptions underlying the z and Chi-Square tests are invalid, one approach is to do a randomization/permutation test.

So lets assume that the null hypothesis is true, that is, the explanatory variable is not associated with the response variable. For example, aspirin has no effect on stroke.

If this is the case, then the deviation seen from the null hypothesis in the data is due to the random mechanism for assigning subjects to the explanatory variable.

So the idea is to consider what would happen under the different allocations, assuming the the response variable didn't change.

Lets consider the example to right. In this case there are

$$\binom{6}{3} = 20$$

different ways to allocate the 6 subjects to the 2 groups. So we need to examine how many successes and failures fall into these 20 different possible assignments, assuming that changing the group assignment doesn't affect the response variable.

Subject Code	Actual Group	Response
A	1	1
B	1	1
C	1	1
D	2	0
E	2	1
F	2	0

Subject Code	Actual Group	Response
A	1	1
B	1	1
C	1	1
D	2	0
E	2	1
F	2	0

	Success	Failure
Group 1	3	0
Group 2	1	2
Total	4	2

Hypothetical Grouping		Number of 1's		
Group 1	Group 2	in 1	in 2	$\hat{\pi}_1 - \hat{\pi}_2$
ADF	BCE	1	3	-0.5
BDF	ACE	1	3	-0.5
CDF	ABE	1	3	-0.5
DEF	ABC	1	3	-0.5
ABD	CEF	2	2	0.0
ABF	CDE	2	2	0.0
ACD	BEF	2	2	0.0
ACF	BDE	2	2	0.0
BCF	ADE	2	2	0.0
ADE	BCF	2	2	0.0
AEF	BCD	2	2	0.0
BCD	AEF	2	2	0.0
BDE	ACF	2	2	0.0
BEF	ACD	2	2	0.0
CDE	ABF	2	2	0.0
CEF	ABD	2	2	0.0
ABC	**DEF**	**3**	**1**	**0.5**
ABE	CDF	3	1	0.5
ACE	BDF	3	1	0.5
BCE	ADF	3	1	0.5

In this example 8 of the 20 possibilities are as or more extreme as seen in observed table, giving a 2-sided p-value of 0.4.

In theory this could be done for any 2×2 table. Lets consider the aspirin example

Treatment	No Stroke	Stroke	Total
Aspirin	63	15	78
Placebo	43	34	77
Total	106	49	155

The number of different possible allocations of subjects to treatment is $\binom{155}{78} = 2.91 \times 10^{45}$. So enumerating all the possible allocations will take a while :).

However instead of doing an exact calculation, we can approximate it by simulation. So instead of generating all $\binom{T}{R_1}$ possible allocations, we can simulate B of these uniformly from the set of all possible ones.

For each simulated configuration, calculate the test statistic of interest (say X^2 or $\hat{\pi}_1 - \hat{\pi}_2$) giving statistics $z_1, \ldots, z_B$. Also calculate the statistic for the observed data (call it z_c). Then the p-value can be approximated by

$$\hat{p}\text{-value} = \frac{\#z_i \text{ more extreme than } z_c}{B}$$

The **R** function `chisq.test` has this ability built in with the option `simulate.p.value=T`.

For the aspirin study the results look like

```
> chisq.test(aspirin, correct=F)

        Pearson's Chi-squared test

data:  aspirin X-squared = 11.1349, df = 1, p-value = 0.0008472

> chisq.test(aspirin, simulate.p.value = T)  # default B = 2000)

        Pearson's Chi-squared test with simulated p-value
        (based on 2000 replicates)

data:  aspirin X-squared = 11.1349, df = NA, p-value = 0.001499

> chisq.test(aspirin, simulate.p.value = T, B = 10^6)

        Pearson's Chi-squared test with simulated p-value
        (based on 1e+06 replicates)

data:  aspirin X-squared = 11.1349, df = NA, p-value = 0.001016
```

The true randomization p-value is 0.001010

The accuracy of the simulation approximation depends on the choice of B. The standard error of $\hat{p}$-value is

$$SE(\hat{p}\text{-value}) = \sqrt{\frac{p(1-p)}{B}}$$

so choices of B can be based on bounding this for reasonable guesses for p-value p.

Fisher's Exact Test

There is a different approach to calculating the randomization p-value that requires much less computation.

In the toy example, while there were 20 different possible configurations based on the explanatory variable, there were only 3 different tables generated

	Success	Failure	Total
Group 1	3	0	3
Group 2	1	2	3
Total	4	2	6

	Success	Failure	Total
Group 1	2	1	3
Group 2	2	1	3
Total	4	2	6

	Success	Failure	Total
Group 1	1	2	3
Group 2	3	0	3
Total	4	2	6

In the aspirin example, while there are $O(10^{45})$ different treatment allocations, there are only 50 tables that need to be considered.

In addition to only being very few tables that need to be considered, they all have a special property, they all have the same marginal totals.

So calculating a p-value involves figuring out what is the probability of seeing different possible tables under the null hypothesis.

Actually we don't need to do it for all tables.

We only need to calculate the probabilities of tables as or more extreme than the one observed.

Consider the possible table

	Variable 2		
Variable 1	Success	Failure	Total
Success	n_{11}	n_{12}	R_1
Failure	n_{21}	n_{22}	R_2
Total	C_1	C_2	T

Under the null hypothesis of

$$H_0 : \omega_1 = \omega_2$$

$n_{11} \sim Hyper(R_1, C_1, T)$.

So we need to be able to calculate hypergeometric probabilities for some tables.

$$P[n_{11} = k] = \frac{\binom{C_1}{k}\binom{C_2}{R_1-k}}{\binom{T}{R_1}}$$

$$= \frac{R_1!R_2!C_1!C_2!}{T!k!(R_1-k)!(C_1-k)!(R_2-C_1+k)!}$$

for $k = 0, \ldots, \min(R_1, C_1)$.

assuming that $R_1 \leq R_2$ and $C_1 \leq C_2$ (otherwise you need to fiddle the bounds on k).

Actually any cell in the table has the same effective hypergeometric distribution so it doesn't matter what cell you pick.

What is considered as or more extreme?

For one sided tests its easy. Suppose the alternative you are interested in is

$$H_A : \pi_1 > \pi_2$$

In this case, any table with $n_1 1 > n_{11}(obs)$ is more consistent with the alternative, so the p-valueis

$$p\text{-value} = \sum_{k=n_{11}(obs)}^{k_{max}} P[n_{11} = k]$$

where k_{max} is the largest possible value of n_{11} for the given margins.

Similarly for

$$H_A : \pi_1 < \pi_2$$

$$p\text{-value} = \sum_{k=k_{min}}^{n_{11}(obs)} P[n_{11} = k]$$

where k_{min} is the smallest possible value of n_{11} for the given margins.

For two sided alternatives it a bit more complicated. There are a couple of possibilities

- Include tables where $|\log \phi| \geq |\log \hat{\phi}|$. This should be equivalent to finding tables where $X^2 \geq X^2(obs)$.
- Include tables where $P[n_{11} = k] \leq P[n_{11} = n_{11}(obs)]$.

  This is what **R** does in its function `fisher.test`

Most of the time these should give the same answer. If not they should be similar.

This is what is known as Fisher's Exact test.

Lets see how it works for a couple of datasets.

For the small dataset, with the randomization p-value of 0.4

```
> fisher.test(permute)

        Fisher's Exact Test for Count Data

data:  permute
p-value = 0.4
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2031288       Inf
sample estimates:
odds ratio
       Inf
```

For the aspirin study

```
> chisq.test(aspirin, correct=F)

        Pearson's Chi-squared test

data:  aspirin
X-squared = 11.1349, df = 1, p-value = 0.0008472

> chisq.test(aspirin)

        Pearson's Chi-squared test with Yates'
        continuity correction

data:  aspirin
X-squared = 10.0119, df = 1, p-value = 0.001555
```

```
> chisq.test(aspirin, simulate.p.value = T, B = 10^6)

        Pearson's Chi-squared test with simulated p-value
        (based on 1e+06 replicates)

data:  aspirin
X-squared = 11.1349, df = NA, p-value = 0.000978

> fisher.test(aspirin)

        Fisher's Exact Test for Count Data

data:  aspirin
p-value = 0.001010
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval: 1.531635 7.356786
sample estimates:
odds ratio
  3.294432
```

In this case, the simulated p-value is close to the Fisher Exact Test p-value (as it should be). In addition, the Chi-Square test p-value is also similar.

This example also illustrates the problem with the Yates' correction. It tends to be too conservative, giving larger p-values than necessary. For smaller sample sizes, it tends to match well with Fisher's Exact test. However for larger samples, the correction isn't needed.

One additional comment about `fisher.test`. As you might have noticed that it gives a confidence interval for ϕ, the odds ratio.

This is based on the fact that the distribution of n_{11} can be determined under different alternatives described by the odds ratio. It has what is known as non-central hypergeometric distribution with parameters R_1, C_1, T, and ϕ ($Hyper\text{-}nc(R_1, C_1, T, \phi)$). The idea behind this interval is to determine for which values of ϕ_0, the hypothesis test for

$$H_0 : \frac{\omega_2}{\omega_1} = \phi_0 \qquad \text{vs} \qquad H_0 : \frac{\omega_2}{\omega_1} \neq \phi_0$$

is not rejected by Fisher's Exact test.

For the aspirin example

```
> fisher.test(aspirin)

        Fisher's Exact Test for Count Data

data:  aspirin
p-value = 0.001010
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.531635 7.356786
sample estimates:
odds ratio
  3.294432

> or.ci(63,78,43,77)
      prop1     prop2      OR    lower    upper conf.level
1 0.8076923 0.5584416 3.32093 1.615356 6.827334       0.95
```

Note that the estimate of ϕ is a bit different $\hat{\phi}$ as it is based on the MLE for the non-central hypergeometric model.

Don't worry about the difference as in most cases it will be small.

Paired Binary Data

In the cases considered so far, it is assumed that each binary/multinomial trial is independent. However there are situations where this isn't true.

In some cases you will have paired data such as

- Observations on twins
- Patients are their own control (right eye vs left eye)
- Two raters examining the same objects

In these situations, the previous analyzes will not work as they do not account for the correlation in the data.

Example: Presidential Approval

Approval of the President's performance in office in two surveys, one month apart, for a random sample of 1600 voting-age Americans.

	2nd Survey		
1st Survey	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Is there any evidence that the number of American supporting the president is changing.

	Variable 2		
Variable 1	Success	Failure	Total
Success	n_{11}	n_{12}	R_1
Failure	n_{21}	n_{22}	R_2
Total	C_1	C_2	T

The underlying probability model can be describe by the following table. Each observation must fall into one of the four cells.

	Variable 2		
Variable 1	Success	Failure	Total
Success	p_{11}	p_{12}	p_{1+}
Failure	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

- p_{1+} = proportion success in 1st variable

$$\hat{p}_{1+} = \frac{R_1}{T} = \frac{944}{1600} = 0.59$$

- p_{+1} = proportion supporting in 1st variable

$$\hat{p}_{+1} = \frac{C_1}{T} = \frac{880}{1600} = 0.55$$

So the difference in approval is given by

$$p_{1+} - p_{+1}$$

In the example, the estimate is

$$\hat{p}_{1+} - \hat{p}_{+1} = 0.59 - 0.55 = 0.04$$

So the presidential approval estimated to have gone down by 4%.

However most people in this survey really don't tell much about how preferences are changing as they don't change there mind ($\frac{1364}{1600} = 0.85$).

The information about what is happening is coming from the people in the off diagonals ($\frac{236}{1600} = 0.15$).

We can see that this must be the case as

$$\begin{aligned} p_{1+} - p_{+1} &= (p_{11} + p_{12}) - (p_{11} + P_{21}) \\ &= p_{12} - p_{21} \end{aligned}$$

The difference in probabilities only depends on the probabilities in the off diagonal cells.

So this can be estimated by

$$\begin{aligned} \hat{p}_{1+} - \hat{p}_{+1} &= \hat{p}_{12} - \hat{p}_{21} \\ &= \frac{n_{12} - n_{21}}{T} \end{aligned}$$

McNemar's Test

To examine the hypotheses

$$H_0 : p_{12} - p_{21} = 0 \qquad H_0 : p_{12} - p_{21} \neq 0$$

or equivalently

$$H_0 : p_{1+} - p_{+1} = 0 \qquad H_0 : p_{1+} - p_{+1} \neq 0$$

the following test statistic can be used

$$X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

which should be compared to a χ^2_1 distribution.

This is based on the result, that under the null hypothesis

$$\mathrm{Var}(\hat{p}_{12} - \hat{p}_{21}) = \frac{p_{12} + p_{21}}{T}$$

Thus the standard error is estimated by

$$SE(\hat{p}_{12} - \hat{p}_{21}) = \frac{\sqrt{n_{12} + n_{21}}}{T}$$

For the example

$$X^2 = \frac{(150 - 86)^2}{150 + 86} = 17.36$$

which has a corresponding p-value of 0.00003.

So there is strong evidence that the President's support is going down as $\hat{p}_{1+} > \hat{p}_{+1}$

This can be done in **R** as follows

```
> Performance <- matrix(c(794, 86, 150, 570), nr = 2,
     dimnames = list("1st Survey" = c("Approve", "Disapprove"),
                     "2nd Survey" = c("Approve", "Disapprove")))
> Performance
             2nd Survey
1st Survey    Approve Disapprove
  Approve         794         150
  Disapprove       86         570
> mcnemar.test(Performance, correct=F)

        McNemar's Chi-squared test

data:  Performance
McNemar's chi-squared = 17.3559, df = 1, p-value = 3.099e-05
```

Note that p-value of this test is based on an asymptotic approximation. One rule of thumb I have seen is that the χ^2_1 approximation should be ok if $n_{12} + n_{21} \geq 10$

There is a continuity corrected version of McNemar's test if you are worried about this. In the case the test statistic is modified to

$$X^2_c = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}$$

This version is the default in **R**, which is gotten by `correct=T`.

If you wish to do a 1-sided test, you can compute the test statistic

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

(which is the square root of X^2) and compare to a $N(0, 1)$ distribution.