

Relation to the two proportions Z-test

Recall the two sample problem for proportions: suppose we wish to test equality of two proportions p_1 and p_2 , based on two independent samples from $\text{Ber}(p_1)$ and $\text{Ber}(p_2)$, with respective sample size n_1 and n_2 . Let the samples be X_1, \dots, X_{n_1} where $X_i \sim \text{Ber}(p_1)$ and Y_1, \dots, Y_{n_2} where $Y_i \sim \text{Ber}(p_2)$. We consider hypotheses

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2.$$

The test statistic was

$$\hat{Z}_{n,\text{pool}} = \frac{\hat{p}_{(1)} - \hat{p}_{(2)}}{\text{SE}(\hat{p}_{(1)} - \hat{p}_{(2)})} = \frac{\hat{p}_{(1)} - \hat{p}_{(2)}}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (106)$$

where the two sample proportions are $\bar{X}_{n_1} =: \hat{p}_{(1)}$ and $\bar{Y}_{n_2} =: \hat{p}_{(2)}$; and with $n = n_1 + n_2$ the pooled sample proportion is

$$\hat{p} = \frac{1}{n} \left(\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i \right) = \frac{n_1}{n} \hat{p}_{(1)} + \frac{n_2}{n} \hat{p}_{(2)}.$$

It was argued that $\hat{Z}_{n,\text{pool}} \rightsquigarrow Z$ under the null hypothesis provided $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$ in such a way that $n_1/n_2 \rightarrow c$ where $c > 0$.

The problem looks very similar to the independence problem for a 2×2 contingency table. Indeed we could write the data in a table

	Popul. 1	Popul. 2	Total
success	$n_1 \hat{p}_{(1)}$	$n_2 \hat{p}_{(2)}$	$n \hat{p}$
failure	$n_1 (1 - \hat{p}_{(1)})$	$n_2 (1 - \hat{p}_{(2)})$	$n (1 - \hat{p})$
Total	n_1	n_2	n

Thus the matrix of counts M is

$$M = \begin{pmatrix} n_1 \hat{p}_{(1)} & n_2 \hat{p}_{(2)} \\ n_1 (1 - \hat{p}_{(1)}) & n_2 (1 - \hat{p}_{(2)}) \end{pmatrix}.$$

Let us write down the chi square statistic for this table, without worrying at first about the slightly different nature of the data (in the two sample problem n_1 and n_2 are nonrandom, in the independence problem they are random). The expected counts are

$$\begin{aligned} \widehat{EM}_{11} &= \frac{\text{row total} \cdot \text{column total}}{\text{table total}} = \frac{n \hat{p} \cdot n_1}{n} = \hat{p} \cdot n_1, & \widehat{EM}_{12} &= \frac{n \hat{p} \cdot n_2}{n} = \hat{p} \cdot n_2 \\ \widehat{EM}_{21} &= \frac{n (1 - \hat{p}) \cdot n_1}{n} = (1 - \hat{p}) \cdot n_1, & \widehat{EM}_{22} &= \frac{n (1 - \hat{p}) \cdot n_2}{n} = (1 - \hat{p}) \cdot n_2. \end{aligned}$$

Thus the chi square statistic is

$$\begin{aligned}
\chi^2(M) &= \sum_{\text{all cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \\
&= \frac{(n_1 \hat{p}_{(1)} - n_1 \hat{p})^2}{n_1 \hat{p}} + \frac{(n_2 \hat{p}_{(2)} - n_2 \hat{p})^2}{n_2 \hat{p}} + \frac{(n_1 (1 - \hat{p}_{(1)}) - n_1 (1 - \hat{p}))^2}{n_1 (1 - \hat{p})} + \frac{(n_2 (1 - \hat{p}_{(2)}) - n_2 (1 - \hat{p}))^2}{n_2 (1 - \hat{p})} \\
&= \frac{n_1 (\hat{p}_{(1)} - \hat{p})^2}{\hat{p}} + \frac{n_2 (\hat{p}_{(2)} - \hat{p})^2}{\hat{p}} + \frac{n_1 (\hat{p}_{(1)} - \hat{p})^2}{(1 - \hat{p})} + \frac{n_2 (\hat{p}_{(2)} - \hat{p})^2}{(1 - \hat{p})} \\
&= \left(n_1 (\hat{p}_{(1)} - \hat{p})^2 + n_2 (\hat{p}_{(2)} - \hat{p})^2 \right) \left(\frac{1}{\hat{p}} + \frac{1}{(1 - \hat{p})} \right).
\end{aligned}$$

Now

$$\begin{aligned}
\hat{p}_{(1)} - \hat{p} &= \hat{p}_{(1)} - \frac{n_1}{n} \hat{p}_{(1)} - \frac{n_2}{n} \hat{p}_{(2)} = \frac{n_2}{n} (\hat{p}_{(1)} - \hat{p}_{(2)}), \\
\hat{p}_{(2)} - \hat{p} &= \frac{n_1}{n} (\hat{p}_{(1)} - \hat{p}_{(2)})
\end{aligned}$$

hence

$$\left(n_1 (\hat{p}_{(1)} - \hat{p})^2 + n_2 (\hat{p}_{(2)} - \hat{p})^2 \right) = \frac{n_1 n_2 (n_1 + n_2)}{n^2} (\hat{p}_{(1)} - \hat{p}_{(2)})^2 = \frac{n_1 n_2}{n} (\hat{p}_{(1)} - \hat{p}_{(2)})^2$$

and we obtain

$$\chi^2(M) = \frac{n_1 n_2 (\hat{p}_{(1)} - \hat{p}_{(2)})^2}{n \hat{p} (1 - \hat{p})} = \left(\hat{Z}_{n, \text{pool}} \right)^2.$$

This can serve as a proof that $\chi^2(M) \rightsquigarrow \chi_1^2$ for the 2×2 table M : it is a consequence of $\hat{Z}_{n, \text{pool}} \rightsquigarrow Z$ and the fact that $Z^2 \sim \chi_1^2$. Recall that here M was slightly different from the matrix of counts appearing in Theorem 4.7: now the columns of M come from two different multinomial populations $\mathfrak{M}_r(n_1, \mathbf{q}_1)$ and $\mathfrak{M}_r(n_2, \mathbf{q}_2)$ (where $\mathbf{q}_1 = (p_1, 1 - p_1)^\top$ and $\mathbf{q}_2 = (p_2, 1 - p_2)^\top$) and the null hypothesis is $p_1 = p_2$ which is equivalent to $\mathbf{q}_1 = \mathbf{q}_2$. Since this can also be expressed as

$$\text{rank} \begin{pmatrix} p_1 & p_2 \\ 1 - p_1 & 1 - p_2 \end{pmatrix} = 1$$

we see that the null hypothesis about EM is the same as in the independence problem. It turns out that the only difference is nonrandomness/randomness of the column totals. We have shown (in conjunction with Theorem 4.7) that this difference can be neglected in the asymptotic distribution of $\chi^2(M)$, for $r = s = 2$.

It can be shown that in general $\chi^2(\mathbf{M}) \rightsquigarrow \chi_{(r-1)(s-1)}^2$ for the problem of observing s multinomial distributions $\mathfrak{M}_r(n_l, \mathbf{q}_l)$, $l = 1, \dots, s$ and testing the null hypothesis $H_0 : \mathbf{q}_1 = \dots = \mathbf{q}_s$. The test for this is known as the **chi square test for homogeneity**. It works exactly the same as the test for independence, once the matrix \mathbf{M} of counts is given. *Above we have seen that the chi square test for homogeneity in a 2×2 table is equivalent to the two sided Z-test for equality of proportions.*

Also the independence test in a 2×2 table is equivalent to a Z-test based on $\left| \hat{Z}_{n, \text{pool}} \right|$.

Programming simulations

The following represents a simple program for simulating the Z-test in Matlab. To write a program, type "edit". A new window appears. Type the program below line for line (we write line numbers, but they should not be typed):

```
1: k=0
2: rejec=0
3: while k<1000
4:     x=normrnd(0,1,1,100)
5:     rejec=rejec+ztest(x,0,1)
6:     k=k+1
7: end
```

Note that when you type "while" then the following text is automatically indented until you type "end". The code in between represents a loop. Line 1 sets the counter variable k=1; inside the loop it is increased by 1 (line 6) until the loop ends. Line 3 ensures that the loop ends at k=1000. Inside the loop, a random sample from $N(0, 1)$ of size 100 is generated (as a 1×100 vector). Line 5 is they key line: the function "zest" carries out a two sided Z-test for null hypothesis $H_0 : \mu = 0$, the output is 0 or 1-the decision. In line 5 these decisions are added; the previous value of "rejec" is increased by 1 if the test rejects, and not increased otherwise. Thus, at the end, "rejec" gives the number of rejections in 1000 simulations. Since the standard significance level is 5% and the data are generated under the null hypothesis, we expect 50 rejections in k=1000 runs. The program thus serves to illustrate the significance level of the test. It also allows to check the power, when the data are generated under some alternative (e.g. add a shift, e.g. $\mu = 0.5$ to x):

```
4: x=normrnd(0,1,1,100)+0.5
```

When you typed the program, save it under some name e.g. "ztest1.m". Don't use "ztest" since that name is reserved in Matlab for the test itself as used in the program. Then you can simply call it at the command line in Matlab, typing

```
>>ztest1
```

and the program will run, giving output k=1000 and rejec=50 (possibly) at the end.

A file path problem: At first Matlab may not find the program you just typed and saved. In this case, first find the path under which your program was saved:

```
>>which ztest1
```

to which you may get the output

```
>>C:\Documents and Settings\Yourname\My Documents\MATLAB\ztest1.m
```

Then type

```
>>addpath('C:\Documents and Settings\Yourname\My Documents\MATLAB')
```

and then Matlab knows where your file "ztest1.m" is. After that you can run "ztest1" from the command line.

The signed rank test again. The program below takes up the signed rank test as discussed before, for the "half Cauchy" data. Before you can use it, you have to input the sample size n and the shift m at the command line:

```
>>n=100
```

```
>>m=0
```

The shift can then be modified to positive values. Recall that the half Cauchy density

$$f(x) = \begin{cases} \frac{1}{\sigma} \varphi\left(\frac{x}{\sigma}\right), & x \leq 0 \\ \frac{1}{\pi(1+x^2)}, & x \geq 0 \end{cases} \quad \text{where } \sigma = \sqrt{\pi/2}$$

has median 0 but is heavily skewed to the right, so the signed rank test should detect it as an alternative (i.e. as nonsymmetric) even for $m = 0$. When $m > 0$ then detection should be much easier, since in that case the median of our data is $m > 0$. Here is the program "signrank1"; lines 5 and 6 generate a sample "d" of size n from the shifted half Cauchy density $f(\cdot - m)$:

```
1: k=0
2: rejec=0
3: sigma=sqrt(pi/2)
4: while k<1000
5:     b=binornd(1,0.5,1,n)
6:     d=m+abs(trnd(1,1,n)).*b+abs(normrnd(0,sigma,1,n)).*(b-1)
7:     [p,h]=signrank(d)
8:     rejec=rejec+h
9:     k=k+1
10: end
```

Again "rejec" at the end gives the number of 5% (two sided) rejections. We can thus explore the power of the signed rank test for various alternatives (m) and various sample sizes (n).