STA409

Answer to Assignment 2

12112627
李乐平

1. Solution.

Barnard's exact test is a statistical method used to analyze 2 × 2 contingency tables, examining the association between two categorical variables. It's known to be more powerful than Fisher's exact test for such tables.

X	Column 1	Column 2	Count
Row 1	x_{11}	x_{12}	
Row 2	x_{21}	x_{22}	
Count	c_1	c_2	t

Say there is a 2 × 2 contingency table $\mathbf{X_0}$ like above to be tested. Define p_1, p_2 the theoretical binomial probabilities for x_{11} and x_{12}. Here are the steps for performing Barnard's exact test:

1. Define the null and alternative hypotheses:

Null hypothesis H_0 typically assumes no association between variables: $p_1 = p_2$.

Alternative hypotheses H_1 could be one of the follows:

- $H_1 : p_1 < p_2$
- $H_1 : p_1 > p_2$
- $H_1 : p_1 \neq p_2$

2. Compute the Wald statistic T(X):

Estimate the probabilities p_1 and p_2 from the observed sample.

Calculate the Wald statistic using pooled or unpooled variance:

- Pooled variance:

$$T(X) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{c_1} + \frac{1}{c_2})}}, \text{where } \hat{p} = \frac{c_1\hat{p}_1 + c_2\hat{p}_2}{c_1 + c_2}$$

- Unpooled variance:

$$T(X) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{c_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{c_2}}}$$

where $\hat{p}_1 = \frac{x_{11}}{c_1}$ and $\hat{p}_2 = \frac{x_{22}}{c_2}$ are the estimated probabilities, and c_1 and c_2 are the column sums in the contingency table.

3. Compute the p-value:

The p-value is computed as

$$\text{p - value} = \max_{0 \le \pi \le 1} \sum_X C_{c_1}^{x_{11}} C_{c_2}^{x_{22}} \pi^{x_{11}+x_{12}} (1-\pi)^{t-x_{11}-x_{12}} ,$$

where

- for $p_1 < p_2$ alternative: sum probabilities of all table Xs with $T(X) < T(X_0)$;
- for $p_1 > p_2$ alternative: sum probabilities of all table Xs with $T(X) > T(X_0)$;

- for $p_1 \neq p_2$ alternative: sum probabilities of all table Xs with $|T(X)| > |T(X_0)|$,

and $0 \leqslant \pi \leqslant 1$ is the nuisance parameter that maximizes the p-value.

4. Complexity:

The complexity of this function is $O(nc_1c_2)$, where n is the number of sample points used to optimize the parameter π.

2. Solution.

(1). We want to test H_0: $\sigma_1=\sigma_2$ against H_1: $\sigma_1 \neq \sigma_2$.

Under H_0, we have S_2/S_1 as test statistic:

$$\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} = \frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1) = F(16,18),$$

Because $S_1^2/S_2^2=1.4938>1$, the critical region is given by

$$\overline{[F(0.975,16,18), F(0.025,16,18)]} = (-\infty, 0.368) \cup (2.640, +\infty)$$

where $F(\alpha, n_1, n_2)$ is the upper α-th quantile of $F(n_1, n_2)$.

p-value=$2\Pr(F(16, 18)>1.4938|H_0)=0.41>0.05$, hence we cannot reject H_0 under the significance level of 95%.

(2). With $\sigma_1^2=\sigma_2^2$, denoted as σ^2, we have

$$\begin{aligned} &\bar{X} \sim N(\mu_1, \frac{\sigma^2}{n_1}), \bar{Y} \sim N(\mu_2, \frac{\sigma^2}{n_2}), \\ &\bar{X}-\bar{Y} \sim N(\mu_1-\mu_2, (\frac{1}{n_1}+\frac{1}{n_2})\sigma^2), \\ &Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}} \sim N(0,1) \end{aligned},$$

and

$$\begin{aligned} &S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}, \\ &W = \frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2) \end{aligned}$$

therefore,

$$\begin{aligned} &T = \frac{Z}{\sqrt{W/(n_1+n_2-2)}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2) \\ &\because S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} = 4.9956, \\ &\therefore T = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}} = \frac{3-(\mu_1-\mu_2)}{1.6678} \end{aligned}$$

Hence a 95% CI of μ_1-μ_2 is given by

$$[\bar{X}_1-\bar{X}_2-t(\alpha/2,34)S_p\sqrt{1/n_1+1/n_2}, \bar{X}-\bar{Y}+t(\alpha/2,34)S_p\sqrt{1/n_1+1/n_2}] = [-0.3894, 6.3894].$$

(3). Now we want to compute the Type II error rate of the test H_0: μ_1-μ_2=0 against H_1: μ_1-μ_2>0.

$$\because Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}} \sim N(0,1)$$

the critical region is given by

$$[\Phi(1-\alpha)\sqrt{1/n_1+1/n_2}\sigma,+\infty)=[2.744,+\infty)$$

Therefore, by using the fact of

$$Z=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\sim N(0,1)$$
,

we have

$$\beta=\Pr(\overline{X}-\overline{Y}<\Phi(1-\alpha)\sqrt{1/n_1+1/n_2}\sigma\mid H_1)$$
$$=\Pr(N(0,1)<z_\alpha-\frac{\mu_1-\mu_2}{\sqrt{1/n_1+1/n_2}\sigma})$$
$$=\Phi(z_\alpha-\frac{\mu_1-\mu_2}{\sqrt{1/n_1+1/n_2}\sigma})$$
$$=0.5584$$

where z_α is the upper α-th quantile of standard normal distribution, and $\Phi(\cdot)$ is the c.d.f. of standard normal distribution.

(4). Now we want $p = 1-\beta \geqslant 0.8$, that is to say,

$$\Phi(z_\alpha-\frac{\mu_1-\mu_2}{\sqrt{1/n_1+1/n_2}\sigma})\le 0.2\cdot$$

$$\therefore z_\alpha-\frac{\mu_1-\mu_2}{\sqrt{1/n_1+1/n_2}\sigma}\le z_{0.8}$$
$$\therefore 1/n_1+1/n_2=\frac{n_1+n_2}{n_1n_2}\le(\frac{\mu_1-\mu_2}{(z_\alpha-z_{0.8})\sigma})^2=0.040436$$
$$\because n_1+n_2\le 0.040436n_1n_2\le 0.040436\frac{(n_1+n_2)^2}{4}$$
$$\therefore n_1+n_2\ge\frac{4}{0.040436}>98$$

When $n_1 = 49$, $n_2 = 50$, $(n_1 + n_2) / (n_1n_2) = 0.040408 < 0.040436$, hence the minimum of $n_1 + n_2$ is 99.

3. Solution.

Now we want to test whether 2 beverages differ significantly with the hypothesis test

H_0: $m_D=0$ against H_1: $m_D\neq 0$,

where m_D is the population median of the difference between the scores of 2 beverages.

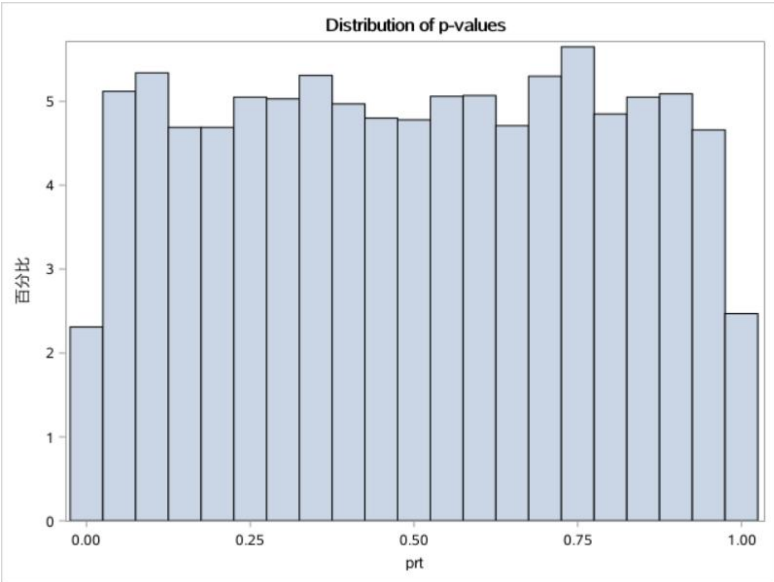
	1	2	3	4	5	6	7	8	9	10	11	12
A	10	8	6	8	7	5	1	3	9	7	6	9
B	6	5	2	2	4	6	4	5	9	8	5	7
Diff	4	3	4	6	3	-1	-3	-2	0	-1	1	2
AbsDiff	4	3	4	6	3	1	3	2	0	1	1	2
Rank	9.5	7	9.5	11	7	2	7	4.5	-	2	2	4.5
PosRank	9.5	7	9.5	11	7						2	4.5
NegRank						2	7	4.5		2		

$S^+=50.5$, $S^-=15.5$. We take $w_{obs}=S^+$ as test statistic. At the significance level of 99%, our criterion is to reject H_0 if $w_{obs}\leqslant 5$ or $w_{obs}\geqslant 61$.

Since $5<w_{obs}=50.5<61$, we cannot reject H_0 at the significance level of 99%.

4. Solution.

By plotting the distribution of p-values, we can observe that the p-value is somehow uniformly distributed.



5. Solution.

We tested two types of standard error estimations with the following results.

$$\hat{SE}_1=\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},\ \hat{p}=\frac{\sum_{i=1}^{n_1}x_{1i}+\sum_{i=1}^{n_2}x_{2i}}{n_1+n_2}$$

$$\hat{SE}_2=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}},\ p_1=\frac{\sum_{i=1}^{n_1}x_{1i}}{n_1},\ p_2=\frac{\sum_{i=1}^{n_2}x_{2i}}{n_2}$$

As we can see, the Type I error rate of 2 SEs satisfies

$$\alpha_1=0.0437<0.05<\alpha_2=0.0821$$

under the criterion of p-value < 0.05. That implies SE_1 is a better estimate than SE_2 because it is more consistent with the criterion.

6. Solution.

(1). The data table is as shown below.

表: WORK.A2Q6_HOMICIDE 视图: 列名 过滤器: (无)

列

- 全选
- convict_race
- victim_race
- death_penalty
- cnt

总行数: 8 总列数: 4 行 1-8

	convict_race	victim_race	death_penalty	cnt
1	white	white	1	18
2	white	white	0	133
3	white	black	1	1
4	white	black	0	10
5	black	white	1	11
6	black	white	0	51
7	black	black	1	6
8	black	black	0	96

(2). As the figure shown below, we can know that the proportion of the homicide convicts receive death penalty is about 0.1104, irrespective to the victims and convicts' race, and its 95% confidence intervals are

$$Wald:[0.0764,0.1445]$$
$$Exact:[0.0786,0.1496]$$

二项式比例的置信限		
比例 = 0.1104		
类型	95% 置信限	
Clopper-Pearson（精确）	0.0786	0.1496
Wald	0.0764	0.1445

(3). As the result shows, under the hypothesis test of H_0: $p \leqslant 0.1$ against H_1: $p>0.1$, where p is the proportion of homicide convicts receive death penalty, the test statistic Z = 0.6277.

And that leads to

$$\text{p}-\text{value}(Z)=0.2651>0.05$$
$$\text{p}-\text{value}(Exact)=0.2902>0.05$$

hence we cannot reject H_0 under the significance level of 95%.

H0: 比例 = 0.1 的检验	
H0 下的 ASE	0.0166
Z	0.6277
单侧 Pr > Z	0.2651
双侧 Pr > |Z|	0.5302
精确检验	
单侧 Pr >= P	0.2902
双侧 = 2 * 单侧	0.5804

(4). Under the hypothesis test of H_0: p_1-p_2=0 against H_1: p_1-$p_2 \neq 0$, where p_1 and p_2 are respectively the proportion of white and black homicide convicts receive death penalty, the test statistic Z=0.3925, leading to a p-value=0.6947>0.1, therefore, we cannot reject H_0 at the significance level of 0.1.

风险差值检验	
H0: P1 - P2 = 0	Wald 方法
风险差值	0.0136
ASE (H0)	0.0347
Z	0.3925
单侧 Pr > Z	0.3474
双侧 Pr > |Z|	0.6947
列 1 (death_penalty = 0)	

(5). Under the hypothesis test of H_0: $p_1 \leqslant p_2$ against H_1: $p_1>p_2$, where p_1 and p_2 are respectively the proportion of black and white homicide victims whose convicts receive death penalty, the test statistic Z=2.0343, leading to a p-value=0.0210<0.05, therefore, we have to reject H_0 at the significance level of 0.05.

风险差值检验	
H0: P1 - P2 = 0	Wald 方法
风险差值	0.0742
ASE (H0)	0.0365
Z	2.0343
单侧 Pr > Z	0.0210
双侧 Pr > |Z|	0.0419
列 1 (death_penalty = 0)	

(6). From the results above, we may infer that the convict is more likely receives a death penalty in the case of the victim is white in a homicide sentence, yet the race of the convict himself/herself is rather irrelevant.