Applied Statistics – Homework 7

Sai Nikhil
001564864

T Sai Nikhil

1 14.8.9

(a)

```
> x <- 4
> n <- 45
> z <- 1.645
> p <- x / n
> CI_90 <- p + c(-1, 1) * z * sqrt(p * (1 - p) / n)
> CI_90
[1] 0.01910277 0.15867501
```

90% Confidence interval

(b)

$H_0 : p = 0.22$

$H_a : p \neq 0.22$

population proportion of children with special needs whose mothers had more than 12 years of schooling

(c)

```
> x <- 4
> n <- 45
> z <- 1.645
> p <- x / n
> CI_90 <- p + c(-1, 1) * z * sqrt(p * (1 - p) / n)
> binom.test(x, n, 0.22, conf.level = 0.95, alternative = "two.sided")

	Exact binomial test

data:  x and n
number of successes = 4, number of trials = 45, p-value = 0.03077
alternative hypothesis: true probability of success is not equal to 0.22
95 percent confidence interval:
 0.02475296 0.21221174
sample estimates:
probability of success
            0.08888889

>
```

(d) p-value $= 0.03077 < \alpha = 0.05$

$\Rightarrow$ reject H_0

Therefore, proportion of children with special needs whose mothers had more than 12 years of schooling $\neq 0.22$

(e)

$P_0 = 0.22 \quad \alpha = 0.05 \quad Z_\alpha = 1.96$

$P_1 = 0.10 \quad \beta = 0.10 \quad Z_\beta = 1.65$

$$n = \left[\frac{Z_\alpha \sqrt{P_0(1-P_0)} + Z_\beta \sqrt{P_1(1-P_1)}}{P_0 - P_1}\right]^2$$

$$= \left[\frac{1.96\sqrt{0.22(1-0.22)} + 1.65\sqrt{0.10(1-0.10)}}{0.10-0.22}\right]^2$$

≈ 119

(2)

14.8.12

(a)

Tuberculin skin test result

		Positive	Negative	Total
Sharing Needles	Admit	$97 \times 24.7\%$ ≈ 24	$97 \times 75.3\%$ $= 73.041 \approx 73$	97
	Deny	$161 \times 17.4\%$ ≈ 28	161×82.6 ≈ 133	161
	Total	52	206	258

$$\text{Common } p\text{-value} = \frac{52}{258} \approx 0.20$$

(b)

$H_0 : P_1 = P_2$

$H_A : P_1 \neq P_2$

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} = \frac{(0.247 - 0.174) - 0}{\sqrt{0.2 \times 0.8\left(\frac{1}{97} + \frac{1}{161}\right)}}$$

$$\approx 1.4202 < 1.96$$

p-value $\approx 0.078 \Rightarrow$ we cannot reject H_0

(c)

$P_1 = P_2$

$\Rightarrow$ Samples collected do not provide evidence that the proportions of people having positive tuberculin skin test result differ between those who had admitted to share needles and those who had denied

(d)

$$95\%\ CI = (\hat{P}_1 - \hat{P}_2) \pm 1.96 \sqrt{\frac{\hat{P}_1 \hat{q}_1}{n_1} + \frac{\hat{P}_2 \hat{q}_2}{n_2}}$$

$$= (0.247 - 0.174) \pm 1.96 \sqrt{\frac{(0.247)(0.753)}{97} + \frac{(0.174)(0.826)}{161}}$$

$$= (-0.0309,\ 0.1769)$$

(3) 15.6.8

Drove while drinking	1983	1987	Total
Yes	1250	991	2241
No	1387	1666	3053
Total	2637	2657	5294

(a) H_0 : proportion of students who drove while drinking are same in 1983 & 1987

H_1 : proportion of students who drove while drinking are different in 1983 & 1987

$\alpha = 0.05$

Expected	1983	1987	Total
Yes	$\frac{2241 \times 2637}{5294} \approx 1116.27$	$\frac{2241 \times 2657}{5294} \simeq 1124.73$	2241
No	$\frac{3053 \times 2637}{5294} \simeq 1520.73$	$\frac{3053 \times 2657}{5294} \simeq 1532.27$	3053
Total	2637	2657	5294

Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
1250	1116.27	16.02
991	1124.73	15.90
1387	1520.73	11.76
1666	1532.27	11.97
Total		55.35

$\chi^2 \approx 55.36$

$df = (r-1)(c-1) = (2-1)(2-1) = 1$

$\Rightarrow$ χ^2-critical value $\simeq 3.841$

(b) Test Statistic > χ^2- critical value

$\Rightarrow$ reject H_0

$\therefore$ proportion of students who drove while drinking are the same in two calender years.

(c) $H_0 : P_1 = P_2$

$H_A = P_1 \neq P_2$

$\alpha = 0.05$

$$\hat{P}_1 = \frac{1250}{2637} \approx 0.474$$

$$\hat{P}_2 = \frac{991}{2657} \approx 0.373$$

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

$$= \frac{1250 + 991}{2637 + 2657} \approx 0.423$$

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} = \frac{0.474 - 0.373}{\sqrt{\frac{0.423 \times 0.577}{2637} + \frac{0.423 \times 0.577}{2657}}}$$

$$\approx 7.44$$

$$P\text{-Value} = 2P(Z > Z_0)$$

$$= 2\left(1 - P(Z < 7.44)\right)$$

$$\simeq 0.000 < \alpha$$

$\Rightarrow$ reject H_0

$\therefore$ The conclusion is same $(P_1 \neq P_2)$

(d) $$95\%\ CI = \left(\hat{P}_1 - \hat{P}_2\right) \pm Z_{critical} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

$$= (0.075, 0.128)$$

(e) It doesn't contain "0".

It is obvious because we reject H_0 and $P_1 \neq P_2$.

(12)

H_0 : Time of Screening & Diagnosis are independent

H_1 : Time of Screening & Diagnosis are associated

(a)

First Screening		Second Screening: Present	Second Screening: Absent	Total
Present	observed (O)	1763	489	2252
	Expected (E)	1467.02	784.98	2252.00
	$(O-E)^2/E$	59.72	11.60	171.32
Absent	observed (O)	403	670	1073
	Expected (E)	698.98	374.02	1073.00
	$(O-E)^2/E$	125.33	234.23	359.56
Total	observed (O)	2166	1159	3325
	Expected (E)	2166.00	1159.00	3325.00
	$(O-E)^2/E$	185	345.83	530.88

$\therefore \chi^2_{calc} \simeq 530.88$ (p-Value $\simeq$ 0.0000)

$\alpha = 0.05$, $df = (2-1)\times(2-1) = 1$

$\chi^2_{(0.05,1)} = 3.841 < \chi^2_{calc}$

$\Rightarrow$ reject H_0 $\Rightarrow$ there is association between time of Screening & diagnosis

(b) The deta comes from paired observation. we need to do pair analysis.

H_0 : Time of Screening & Diagnosis are independent

H_1 : Time of Screeing & Diagnosis are associated

$$\chi^2 = \frac{(|r-s| - 1)^2}{r+s} = \frac{[(2166 - 1073) - 1]^2}{(2166 + 1073)}$$

$$\approx 368.158$$

$df = 1$

We get that $p < 0.01$. So, "p" is less than 0.05, thus we reject null hypothesis at 0.05 level of significance. We again conclude that time of screening & Diagnosis are associated.

5

```
> library("MKinfer")
> s1 <- rbinom(100000, 50, 0.16)
> x1 <- round(mean(s1), 0)
> CI_wald <- binomCI(x1, 50, 0.16, conf.level = 0.90, method = "wald")
> CI_wald

   wald confidence interval

90 percent confidence interval:
            5 %      95 %
prob 0.07472104 0.245279

sample estimate:
prob
0.16

additional information:
standard error of prob
            0.05184593

> CI_wilson <- binomCI(x1, 50, 0.16, conf.level = 0.90, method = "wilson")
> CI_wilson

   wilson confidence interval

90 percent confidence interval:
           5 %      95 %
prob 0.0925781 0.2623285

sample estimate:
prob
0.16

additional information:
standard error of prob
            0.05160045

>
```

← wald interval

← wilson interval

The standard error of probability for Wald and Wilson is 0.0518 & 0.0516 (rounded to 4 decimal places).

This implies that of the two, wilson interval is more robust for $p=0.16$