[HTML]DB5800

Fatemeh Nadi
810101285
fatemehnadi@ut.ac.ir

Homework 5
Statistical Inference , Fall 1401

December 28, 2022

1

- a
  FALSE: We are 100% confident that 82% Americans in this sample think it's the government's responsibility to promote equality between men and women.
  Confidence intervals are calculated based on population proportions, not samples.

- b
  TRUE.

- c

  TRUE

- d
  TRUE

$$ME = \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} \to \frac{ME}{2} = \frac{1}{2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}$$
$$= \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{4n}}$$

- e
  TRUE: . Because the entire interval is above 0.50, we do have evidence, at the 95% confidence level, that more than half of Americans(at the time of this poll) think it's the government's responsibility to promote equality between men and women.

2

$\hat{p} = 0.53, \qquad n = 450, \quad \alpha = 0.05$

1. independence: $450 < 10\%$ of voters in the district.
2. sample size / skew: $4500.5 = 225 > 10$ S-F condition met $\rightarrow$ nearly normal sampling distribution.

With these conditions verified, the normal model may be applied to $\hat{p}$.

Based on the probability distribution, establish the null hypothesis and alternative hypothesis.

$$H_0: \ p = 0.5$$
$$H_a: \ p > 0.5$$

$$\hat{p} \sim N\left(mean = 0.5, SE = \sqrt{\frac{0.5 \times 0.5}{450}} \simeq 0.023\right)$$

$Z = \frac{\hat{p}-p}{SE} = \frac{0.53-0.5}{0.023} = 1.27$

$p - value = pnorm(1.27, lower.tail = FALSE) = 0.102 \nleq 0.05$

It is impossible to reject the null hypothesis, and any convincing evidence does not support the campaign manager's claim, because the p-value is larger than 0.05, we cannot reject it.

3

- a

  Two-way table

Observed	Nevirapine	Lopinavir	Total
Fail	26	10	36
Treatment	94	110	204
Total	120	120	240

- b

$$H_0: \; There\ is\ no\ difference\ in\ the\ proportion\ of\ virologic\ failure\ between\ treatment\ groups.$$
$$H_a: \; There\ is\ difference\ in\ the\ proportion\ of\ virologic\ failure\ between\ treatment\ groups.$$

- c

  First, the conditions must be verified.

  1. Independence:
  1.1. Within groups: random sample $\rightarrow$ The women are randomly sampled and assigned to different treatment groups and $n <$ 10% population $\rightarrow$ 240 $<$ 10% all women with HIV after childbirth.
  1.2. Between groups: treatment of a woman with Nevirapine or Lopinavir. so each case only happens in one group.

  2. Sample size / skew: samples should meet the success-failure condition (at least 10 successes and 10 failures), according to table in part 'a' each groups has at least 10 expected cases.

  All is met, so we can assume that the sampling distribution of the difference between two proportions is nearly normal.

  Based on the probability distribution, establish the null hypothesis and alternative hypothesis.

$$H_0: \; \hat{p}_{Nevirapine} - \hat{p}_{Lopinavir} = 0$$
$$H_a: \; \hat{p}_{Nevirapine} - \hat{p}_{Lopinavir} \neq 0$$

Observed	Nevirapine	Lopinavir
Total	120	120
$\hat{p}$	$\frac{26}{120} = 0.216$	$\frac{10}{120} = 0.08$

$$
\begin{cases}
\hat{p}_{pool} = \frac{\#\ success\ n_{Nevirapine} + \#\ success\ n_{Lopinavir}}{n_{Nevirapine} + n_{Lopinavir}} = \frac{36}{120} = 0.15 \\
\\
SE_{pool} = \sqrt{\frac{\hat{p}_{Nevirapine}(1-\hat{p}_{Nevirapine})}{n_{Nevirapine}} + \frac{\hat{p}_{Lopinavir}(1-\hat{p}_{Lopinavir})}{n_{Lopinavir}}} = 0.0452
\end{cases}
$$

$\hat{p}_{Nevirapine} - \hat{p}_{Lopinavir} \sim N\left(mean = 0, SE = 0.0452\right)$

$point\ estimate\ = \hat{p}_{Nevirapine} - \hat{p}_{Lopinavir} = 0.216 - 0.08 = 0.133$

$$Z = \frac{0.133 - 0}{0.0452} = 2.944$$

$$
\begin{aligned}
p - value &= p\left(|Z| > 2.944\right) \\
&= 2 * p\left(Z > 2.944\right) \\
&= 2 * pnorm(2.9441, lower.tail = FALSE) \\
&= 0.0032 < 0.05
\end{aligned}
$$

Reject H_0, $\rightarrow$ There is a difference in the proportion of virologic failures between people who take Nevaripine and Lopinavir.

4

First, the conditions must be verified.

1. Independence:
1.1. Within groups: random sample $\rightarrow$ Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare was randomized, and $n < 10\%$ population $\rightarrow 1503 < 10\%$ all the population.
1.2. Between groups: Each respondent answers only one type of questions.

2. Sample size / skew: samples should meet the success-failure condition (at least 10 successes and 10 failures).

$$\begin{cases} 771 \times 0.47 = 362, \quad 771 \times 0.49 = 377, \quad 771 \times 0.04 = 30 \\ 732 \times 0.34 = 248, \quad 732 \times 0.63 = 461, \quad 732 \times 0.03 = 21 \end{cases}$$

Each groups has at least 10 expected cases.
All is met, the normal model can be used for the point estimate of the difference in support, where p_1 corresponds to the original ordering and p_2 to the reversed ordering:

$$Point\ Estimate =$$

p$_1$-p_2 = 0.47 - 0.34 = 0.13

$$SE = \sqrt{\frac{p_1\,(1-p_1)}{n_1} + \frac{p_2\,(1-p_2)}{n_2}} = 0.025$$

$CI = 90\% \rightarrow > qnorm((1-.9)/2, lower.tail = FALSE) = 1.644$

Calculate 90% confidence interval:

$$Point\ Estimate \pm z^{*}SE \rightarrow 0.13 \pm 1.644 * 0.025$$

Approximate 90% CI for difference proportion: (0.0889, 0.1711)

8.89% to 17.11%, depending on the ordering of the two statements in the survey question, could change the approval rating for the 2010 healthcare law.

5

Observed Values:

Observed	Rock	Scissors	Paper	Total
n	43	21	35	99
$\hat{p}$	0.434	0.212	0.353	

Expected Values:

Expected	Rock	Scissors	Paper	Total
n	33	33	33	99
$\hat{p}$	0.33	0.33	0.33	

Based on the probability distribution, establish the null hypothesis and alternative hypothesis.

$$\begin{aligned} H_0 &: The\ distribution\ of\ players'\ preferences\ between\ these\ three\ options\ is\ random. \\ &or\ if\ specific\ options\ are\ favored\ above\ others \\ H_a &: The\ distribution\ of\ players'\ preferences\ between\ these\ three\ options\ is\ not\ random. \\ &and\ specific\ opportunities\ are\ favored\ above\ others. \end{aligned}$$

Checking conditions:

1. Independence: Sampled observations must be independent.
1.1. random sample/assignment.$\rightarrow$ We assume that each person makes ones choices randomly.
1.2. if sampling without replacement, $n < 10\%$ of population
1.3. each case only contributes to one cell in the table.$\rightarrow$ Each time we choose one type of them.

2. Sample size / skew: Each particular scenario must have at least 5 expected cases, according to the last table we can see this condition is met.

All is met.
Chi-square test statistic formula:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

χ^2 Test Statistic

Types	Rock	Scissors	Paper	Total
Observed - Expected	10	-12	12	10
$(Observed - Expected)^2$	100	144	144	388
$\frac{(Observed - Expected)^2}{Expected}$	3.03	4.36	4.36	11.7575

$\chi^2 = 11.7575$
$df = k - 1 = 3 - 1 = 2$

$> \ pchisq(11.7575, 2, lower.tail = FALSE) = 0.002798281 < 0.05$

Reject H_0 $\rightarrow$ The distribution of players' preferences between these three options is not random and specific opportunities are favored above others. Chances of this Rock Paper Scissors uneven distribution are less than 0.2 percent.

6

Establish the null hypothesis and alternative hypothesis.

$$H_0: \ Gender\ (male\ or\ female)\ and\ the\ preferred\ condiment\ on\ a\ sandwich\ are\ independent.$$
$$H_a: \ Gender\ (male\ or\ female)\ and\ the\ preferred\ condiment\ on\ a\ sandwich\ are\ dependent$$

First, create a two-way table:

	Ketchup	Mustard	Relish	Total
Male-obsereved	15	23	10	48
Male-expected	$\left(\frac{48\times40}{100}\right) =$ 19.2	20.16	8.64	48
Female-obsereved	25	19	8	52
Female-expected	20.8	21.84	9.36	52
Total	40	42	18	100

Checking conditions:

1. Independence: Sampled observations must be independent.
1.1. random sample/assignment.→ We assume that this selections are random.
1.2. if sampling without replacement, $n < 10\%$ of population who order
1.3. each case only contributes to one cell in the table.

2. Sample size / skew: Each particular scenario must have at least 5 expected cases, according to the last table we can see this condition is met.

All is met.

Chi-square test statistic formula:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

χ^2 Test Statistic

Types	Ketchup	Mustard	Relish	Total
Male: Observed - Expected	-4.2	2.84	1.36	
Female: Observed - Expected	4.2	-2.84	-1.36	
Male: $(Observed - Expected)^2$	17.64	8.0656	1.8496	
Feale: $(Observed - Expected)^2$	17.64	8.0656	1.8496	
Male: $\frac{(Observed-Expected)^2}{Expected}$	0.91875	0.4	0.214	1.5328
Female: $\frac{(Observed-Expected)^2}{Expected}$	0.848	0.3693	0.1974	1.4147
				2.9475

$\chi^2 = 1.4147$
$df = (R-1)(C-1) = 2$

$> \ pchisq(1.4147, 2, lower.tail = FALSE) = 0.4929 > 0.05$

Can't reject H_0 $\rightarrow$ There is not a relationship between gender (male or female) and the preferred condiment on a sandwich.

7 R

Please see this file: "RCodes.R"

Establish the null hypothesis and alternative hypothesis.

$$H_0: \textit{ Employees are absent equally throughout the week.}$$
$$H_a: \textit{ Employees are not absent equally throughout the week.}$$

First, create a two-way table:

	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Number of Absences - observed	18	16	10	10	16	70
Number of Absences - expected	14	14	14	14	14	70
Proportion of Absences - expected	0.2	0.2	0.2	0.2	0.2	1

Checking conditions:

1. Independence: Sampled observations must be independent.
1.1. random sample/assignment.$\rightarrow$ it's a random sample of 70 managers
1.2. if sampling without replacement, $n < 10\%$ of population (all managers)
1.3. each case only contributes to one cell in the table.

2. Sample size / skew: Each particular scenario must have at least 5 expected cases, according to the last table we can see this condition is met.

All is met, so, we can run this test:

```
> null.probs = c(1/5,1/5,1/5,1/5,1/5)
> freqs = c(18,16,10,10,16)
> chisq.test(freqs, p=null.probs)

        Chi-squared test for given probabilities

data:  freqs
X-squared = 4, df = 4, p-value = 0.406
```

Figure 1: Q7 - code and result of Chi squared test

$p_{value} \nless 0.05 \rightarrow$ can't reject H_0, which means employees are absent equally throughout the week.

8 R

Please see this file: "RCodes.R"

Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.
First, create a two-way table:

```
> # read data and create two-way table
> t1 = table(survey$Smoke, survey$Exer)
>
> # view table
> t1

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
```

Figure 2: Q8 -Two-way table of survay dataset

Checking conditions:

1. Independence: Sampled observations must be independent.
1.1. random sample/assignment.→ Assume that this dataset was collected in a random way.
1.2. if sampling without replacement, $n < 10\%$ of population (all smoker student)
1.3. each case only contributes to one cell in the table.

2. Sample size / skew: Each particular scenario must have at least 5 expected cases, according to the last table this condition is not met.
We can merge the None and the Some columns for this problem and then run the test.

```
> # change table
> t2 = cbind(t1[,"Freq"], t1[,"None"] + t1[,"Some"])
>
> # Apply colnames
> colnames(t2) <- c("Freq", "None & Some")
> # viewtable
> t2
      Freq None & Some
Heavy    7           4
Never   87         102
Occas   12           7
Regul    9           8
```

Figure 3: Q8 -Merging two columns

A column's value is less than 5, but we ignore this and assume that all conditions are met.

So, we can run Chi square test

```
> # run chi test
> chisq.test(t2)

	Pearson's Chi-squared test

data:  t2
X-squared = 3.2328, df = 3, p-value = 0.3571
```

Figure 4: Q8 -Chi square test

We cannot reject the null hypothesis that students' smoking habits are independent of their exercise levels due to the p-value of 0.4828.