STAT20 Homework #11

Seth Metcalf

Table of Contents

Introduction

This is Homework #10, which contains questions from Chapters 28 and 29.

Due 3 May 2021.

Chapter 28

1. Ch 28 A5:

Suppose that a die is rolled 600 times with the following results.

Make a chi-test of the null hypothesis that the die is fair.

null: The die is fair alt: The die is not fair

```
obs=c(90, 110, 100, 80, 120, 100)
exp=c(100, 100, 100, 100, 100, 100)

1 - pchisq(sum((obs-exp)^2/exp),df=5)

## [1] 0.07523525
```

We fail to reject the null. This is completely due to chance and it is most likely that the die is fair.

2. Ch 28 C3:

The table below shows the distribution of marital status by sex for persons age 25- 34 in Wyoming. Question: Are the distributions really different for men and women? You may assume the data are from a simple random sample of 299 persons, of whom 143 were men and 156 were women.

Make a chi-test of the null hypothesis that marital status and gender are independent.

```
obs=c(31.5, 60.1, 8.4)
exp=c(143*c(315/1000, 601/1000, 84/1000), 156*c(192/1000, 673/1000,
135/1000))

chisq=sum((obs-exp)^2/exp)
1-pchisq(chisq,df=2)

## [1] 2.263701e-09
```

According to this it seems to not be independent at all.

If they are not independent, who are the women marrying?

They would be marrying other women.

3. Ch 28 Rev 2:

As part of a study on the selection of grand juries in Alameda county, the educational level of grand jurors was compared with the county distribution. Could a simple random sample of 62 people from the county show a distribution of educational level so different from the county-wide one?

Do the appropriate hypothesis test.

```
obs=c(28.4, 48.5, 11.9, 11.2)
exp=c(1*c(84/378,294/378), 10*c(84/378,294/378), 16*c(84/378,294/378),
35*c(84/378,294/378))

1 - pchisq(sum((obs-exp)^2/exp),df=3)

## [1] 0
```

4. Ch 28 Rev 9:

Each respondent in the Current Population Survey of March 2005 can be classified by age and marital status. The table below shows results (counts) for the women who were age 20-29 in Montana. Read questions i) and ii), then answer a) and b).

i) Women of different ages seem to have different distributions of marital status. Or is this just chance variation?

ii) If the difference is real, what accounts for it?

a) Can you answer these questions with the information given? If so, answer them. If not, why not?

No because simple random sample was not stated.

b) Can you answer these questions if the data in the table resulted from a simple random sample of women age 20–29 in Montana? If so, answer them. If not, why not?

Yes you can.They seem to have different distriubutions of marital status, not chance variation. It accounts for it by being in different stages in life.

Chapter 29

5. Ch 29 B1:

One hundred investigators each set out to test a different null hypothesis. Unknown to them, all the null hypotheses happen to be true. Investigator #1 gets a p-value of 58%, plotted in the graph below as the point (1, 58). Investigator #2 gets a p-value of 42%, plotted as (2, 42). And so forth. The 5%-line is shown. (The y axis should really be labeled p-value.)

a) How many investigators should get a statistically significant result?

6

b) How many do?

8 investigators should get statistically significant results

c) How many should get a result which is highly significant?

12

6. Ch 29 B8:

An investigator has independent samples from box A and from box B. Her null hypothesis says that the two boxes have the same average. She looks at the difference

average of sample from A - average of sample from B

The two-sample z-test gives z1.79.

a) bigger than the average of box B?

b) smaller than the average of box B?

c) different from the average of box B?

Is the difference statistically significant if the alternative hypothesis says that the average of box A is...

7. Ch 29 D1:

One term, there were 600 students who took the final in Statistics 2 at the University of California, Berkeley. The average score was 65, and the SD was 20 points. At the beginning of the next academic year, the 25 teaching assistants assigned to the course took exactly the same test. The TAs averaged 72, and their SD was 20 points too. Did the TAs do significantly better than the students?

Tf appropriate, make a two-sample z-test. If this isn't appropriate, explain why not.

No, it is only appropriate if the samples are simple random samples. In this scenario they are not.

8. Ch 12 A3:

In employment discrimination cases, some courts have held that there is proof of discrimination when the percentage of blacks among a firm's employees is lower than the percentage of blacks in the surrounding geographical region, provided the difference is "statistically significant" by the z-test. Suppose that in one city, 10% of the people are black. Suppose too that every firm in the city hires employees by a process which, as far as race is concerned, is equivalent to simple random sampling. Would any of these firms ever be found guilty of discrimination by the z-test?

Explain briefly.

9. Ch 29 Rev 9bc (relabeled a and b):

In 1970, 36% of first-year college students thought that "being very well off financially is very important or essential." By 2000, the percentage had increased to 74%. These percentages are based on nationwide multistage cluster samples.

a) Does it make sense to ask if the difference is statistically significant? Can you answer on the basis of the information given?

No, the samples are not simple random samples and therefore asking for a test of significance does not make sense in this scenario.

b) Repeat a), assuming the percentages are based on independent simple random samples of 1,000 first-year college students drawn each year.

Given the fact that:

```
pCol=(360 + 740)/(1000 + 1000)
estsdCol=(1-0)*sqrt(pCol*(1-pCol))

# 1970
col1970=360/1000
```

```r
sesum11970=estsdCol*sqrt(1000)
se11970=sesum11970/1000*100*sqrt((2000-1000)/1999) # correction factor

# 2000
col2000=740/1000
sesum22000=estsdCol*sqrt(1000)
se22000=sesum22000/1000*100*sqrt((2000-1000)/1999) # correction factor

sediffCol=sqrt(se11970^2+se22000^2)
zCol=(col1970*100-col2000*100-0)/sediffCol
pnorm(zCol)
```

```
## [1] 3.885695e-129
```

Based off of this, the difference is statistically significant.