Stat 21 Test 1 Solutions

Due: March 4, 2020 by noon EST

Gender discrimination

For problems 1-2[1], consider the following setting. A large, global supermarket chain is facing a class action lawsuit for gender discrimination in their job promotion practices. Each store location employs people at two levels: a general floor worker level and a supervisory level. The claim of the lawsuit is that whether or not an employee is promoted to a supervisory level depends on that employee's gender identity. To investigate this claim, a random sample of $100,006$ supermarket employees is obtained from all stores in the US.

Problem 1

One of the prosecutors suggests collecting the data on the sample of $100,006$ employees by categorizing whether or not the employees identified as women. Out of the $27,604$ supervisors surveyed, $12,005$ identified as women and out of the $72,402$ floor workers surveyed, only $31,133$ identified as women.

Determine whether or not there is evidence of discrimination. Use an $\alpha = 0.01$ level of significance. Show all your work and interpret your conclusion within the context of the problem and the appropriateness of the required assumptions.

Solution:

You can use either a difference in proportions test

```
success <- c(12005, 31133)  ##supervisors first, floor workers second
n <- c(27604, 72402)
prop.test(success, n)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  success out of n
## X-squared = 1.9353, df = 1, p-value = 0.1642
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.001996832  0.011794443
## sample estimates:
##    prop 1    prop 2
## 0.4349007 0.4300019
```

or a Chi-squared test for homogeneity:

```
gender.data1 <- matrix(c(12005, 27604-12005, 31133, 72402-31133), byrow=TRUE, ncol=2)
chisq.test(gender.data1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gender.data1
## X-squared = 1.9353, df = 1, p-value = 0.1642
```

[1] The data for these problems is hypothetical.

Either way, you'll should find a p-value of 0.1642 which is larger than most typical choices for α.

If you did a Chi-square test of independence instead, that is also ok and you will get the same p-value. *The difference is in whether or not it's reasonable to assume the two levels of employees are independent.*

If membership in one group excludes you from membership in another, then those groups are **not** independent.

Assumptions:

- Sample size - cell counts are larger than 5 or observed successes and failures are more than 10
- Representative sample - if the population is employees in the US then yes sample is representative, if population is all employees across the globe then no, not representative
- Independent sample - yes, sample was selected randomly

Hypotheses:

$$H_0 : p_{supervisor} - p_{floor} = 0 H_A : p_{supervisor} - p_{floor} \neq 0$$

or

H_0 : job level is independent of gender identity

or

H_0 : the distribution of genders represented in the supervisors matches the distribution of genders represented in the floor workers

Conclusion/interpretation:

Either "with a p-value this large" or "with a 95% CI" that contains zero, we fail to reject the null hypothesis. Technically, in looking for evidence of discrimination we care about discrimination in either direction. However, realistically, we expect discrimination to typically occur in the direction of promotions so if students did a one-sided test that is fine *provided they are testing that* $p_{supervisor} > p_{floor}$.

They don't have to show the null and alternative but they do have to clearly interpret the conclusion.

If students use any abbreviations, they should clearly define what they mean; e.g. $p_{supervisor}$ = probability that a randomly chosen employee is a supervisor.

Problem 2

Another prosecutor has recently taken a statistics class and has a different idea. They suggest that instead of categorizing gender according to a binary system, they take extra privacy precautions and invest some additional money into interviewing the random sample of 100, 006 employees to instead ask participants if their gender is best described as "cisgender male", "cisgender female", or "transgender or gender non-conforming (GNC)". (For clarification on the terms cisgender and transgender, you may refer to this description.)

Gender identity	Floor worker	Supervisor
Trans or GNC	1994	6
Cis woman	40354	8809
Cis man	43372	5471

The table above summarizes their findings based on this alternative classification of gender identities. Determine whether or not there is evidence of discrimination by testing the claim that the distribution of gender identities across each level of workers is the same. Use an $\alpha = 0.01$ level of significance. Show all your work and interpret your conclusion within the context of the problem and the appropriateness of the required assumptions.

```
gender.data2 <- matrix(c(1994, 6, 40354, 8809, 43372, 5471), byrow=TRUE, ncol=2)
chisq.test(gender.data2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gender.data2
## X-squared = 1228.7, df = 2, p-value < 2.2e-16
```

Same assumptions as above. Correct test and interpretation. Must use **chi-square test of homogeneity** or a **test of independence**.

Hypotheses:

H_0 : the distribution of genders represented in the supervisors matches the distribution of genders represented in the floor workers

or

H_0 : job level is independent of gender identity

School segregation

For problems 3-5[2], consider the following setting.

In 2000 a particular school district had a total of 7982 students enrolled in school. At this time, the percent of white students who were exposed to school poverty[3] was 30% whereas this was 34.5% for Black students. The number of white students in this district in 2000 was 5184 and the number of Black students was 1035. (Note these are the demographics for students who identify as one of these two races alone.)

In 2015, the number of students enrolled in school in this district increased to 11977. The percent of white students exposed to school poverty increased to 47.8% and the percent for Black students increased to 76.1%. The number of white students in this district in 2015 was 7306 and the number of Black students was 2037.

Assumptions:

- Sample size - observed successes and failures more than 10
- Representative sample - population could be all students past, present, and future in this district, or population could be all US/state schools etc. hard to know without clearly specifying the population.
- Independent sample - no strictly speaking

groups of white students in 2000 and 2015 must be independent

Problem 3

Suppose we want to determine if, relative to the increased size of the school district from 2000 to 2015, is the difference between the proportion of white students exposed to poverty likely due to something besides chance. Perform an appropriate statistical hypothesis test at an $\alpha = 0.05$ confidence level to answer this question. Show all your work and interpret your conclusion within the context of the problem and the appropriateness of the required assumptions.

```
n3 <- c(5184, 7306)
success3 <- c(.3*5184, .478*7306)
prop.test(success3, n3)
```

[2] The data for these problems is also hypothetical but are informed by the summary statistics the 2020 paper *Racial Segregation and School Poverty in the United States, 1999–2016* by E. Fahle, S. Reardon, D. Kalogrides, E. Weathers, and H. Jang.

[3] Exposure to school poverty was determined by whether or not a student attended a school where where more than 75% of students were eligible for the Federal free and reduced-price lunch program.

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  success3 out of n3
## X-squared = 398.24, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1951003 -0.1608997
## sample estimates:
## prop 1 prop 2
##  0.300  0.478
```

```
##
n3.2 <- c(7306,5184)
success3.2 <- c(.478*7306,.3*5184)
prop.test(success3.2, n3.2)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  success3.2 out of n3.2
## X-squared = 398.24, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1608997 0.1951003
## sample estimates:
## prop 1 prop 2
##  0.478  0.300
```

Hypotheses:

$$H_0 : p_{w,2000} - p_{w,2015} = 0 H_A : p_{w,2000} - p_{w,2015} \neq 0$$

groups of black students in 2000 and 2015 must be independent

two sided-alternative or **one-sided** because asked "relative to increase in population

Problem 4

Perform the same analysis as you did in Problem 3 but this time to determine if there is statistical evidence of an actual difference between the proportion of Black students exposed to poverty between 2000 and 2015.

```
n4 <- c(1035, 2037)
success4 <- c(.345*1035, .761*2037)
prop.test(success4, n4)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  success4 out of n4
## X-squared = 502.78, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.4511046 -0.3808954
## sample estimates:
## prop 1 prop 2
##  0.345  0.761
```

Hypotheses:

$$H_0 : p_{b,2000} - p_{b,2015} = 0 H_A : p_{b,2000} - p_{b,2015} \neq 0$$

two sided-alternative or **one-sided** because asked "relative to increase in population

Problem 5

Now suppose we want to estimate the difference in the proportion of students who are exposed to school poverty along racial categories. Find a 95% confidence interval for the difference between the proportion of Black students exposed to school poverty and the proportion of white students exposed to school poverty in the year 2015. Make sure you specify the point estimate, the margin of error, and the estimate of the standard deviation of the difference in sample proportions. Then interpret your interval in the context of the problem.

```
n5 <- c(2037, 7306)
success5 <- c(.761*2037, .478*7306)
prop.test(success5, n5)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  success5 out of n5
## X-squared = 512.39, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2609102 0.3050898
## sample estimates:
## prop 1 prop 2
##  0.761  0.478
```

```
.761-.478  ##point estimate
```

```
## [1] 0.283
```

```
(ME <- (0.3050898 - 0.2609102)/2)  ## margin of error
```

```
## [1] 0.0220898
```

```
z_crit <- qnorm(0.025)
-1*ME/z_crit ## SE
```

```
## [1] 0.01127051
```

Hypotheses:

$$H_0 : p_{w,2015} - p_{b,2015} = 0 H_A : p_{w,2015} - p_{b,2015} \neq 0$$

Must at least mention assumptions, groups of black and white students must be independent

two sided-alternative only