pstat120cfinal

Preeti Kulkarni

9/8/2022

1. A study to determine the effectiveness of a drug, or serum, for the treatment of arthritis resulted in the comparison of two groups, each consisting of 400 arthritic patients. One group was inoculated with the serum, whereas the other received a placebo (an inoculation that appears to contain serum but actually is not active). After a period of time, each person in the study was asked whether their arthritic condition had improved, and the observed results are presented in the accompanying table. The question of interest is: Do these data present evidence to indicate that the proportion of arthritic individuals who improved differs depending on whether or not they received the drug?

a. Conduct a hypothesis test using the X2 test statistic, with α = .05.

Report (i) the null and alternative hypotheses;

H_0 : The proportion of arthritic individuals who improved is the same regardless of taking drugs H_a : The proportion of arthritic individuals who improved is not the same regardless of taking drugs

ii. the expected cell counts;

$expected\ counts = \frac{Row\ total * Column\ total}{Grand\ Total}$ Row total = 234+148=382 Row total = 166+252= 418 Column total = 234+166=400 Grand total = 234+148+166+252=800

$E[n_{11}] = E[n_{12}]$= (382)(400)/800= 191 $E[n_{21}] = E[n_{22}]$= (418)(400)/800=209

iii. the test statistic;

$\chi^2 = \sum \frac{(O-E)^2}{E}$ where is the frequencies observed and E is the frequencies expected. We find this to be 36.198.

```
treated<-c(234,166)
not<-c(148,252)
data1<-data.frame(treated, not)
chisq.test(data1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data1
## X-squared = 36.198, df = 1, p-value = 1.782e-09
```

iv. the critical value;

```
qchisq(0.95,1)
```

```
## [1] 3.841459
```

v. the p-value; and

We can see our p value is 1.782e-09 from the table below.

vi. the conclusion.

We can see here our critical value is smaller than 36.198 and our p value is <0.05. Thus, we can reject the null hypothesis and say individuals who improved differs depending on whether they receive the drugs.

b. Using the Z-statistic, test the hypothesis that the proportion of treated persons who improved is equal to the proportion of untreated persons who improved, with α = .05. Hint:Express each proportion as a mean. See Section 10.3 of the textbook for a refresher.

Report (i) the null and alternative hypotheses; H_0 : The proportion of arthritic individuals who improved is the same regardless of taking drugs H_a : The proportion of arthritic individuals who improved is not the same regardless of taking drugs

ii. the test statistic;

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{(}\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}$$

$$p_1 = 234/400 = 0.585$$
$$p_2 = 148/400 = 0.37$$
$$p = (234+148)/900 = 0.4775$$
$$Z = \frac{0.585 - 0.37}{\sqrt{(}0.4775(1-0.4775)(1/200))}$$
$$Z = 6.087284$$

iii. the critical value;

Our critical value is seen here at 1.96, which is found through the table.

(iv)the p-value; Using our z table and finding out critical value, we get a p value of around 0. We can also find this value using the pnorm function.

```
(1-pnorm(6.087284))*2
```

```
## [1] 1.148422e-09
```

$p(Z \geq 6.087284) = 1.148422e-09$

and (v) the conclusion.

We can see we have an extremely small p value, <.05, and the critical value is smaller than our z score, so we can reject the null hypothesis that there is no difference in means.

c. Prove that (assuming α is the same for both tests) the χ2 statistic X2 is equivalent to the square of the test statistic Z (Z2). In other words, prove that the χ2 test used in part (a) is equivalent to the two-tailed Z-test used in part (b).

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{(}\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}$$

$$Z^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}$$

$$Z^2 = \frac{n_1 n_2(\hat{p}_1 - \hat{p}_2)^2}{(n_1 + n_2)\hat{p}\hat{q}}$$

$$\text{Now we can notice } \hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Now, by looking at the expected counts formula and contingency table rules

$$\hat{E}(n_{13}) = \frac{(n_{13} + n_{14})(n_{13} + n_{31})}{n_3 + n_4} =$$

$$\frac{(y_3 + y_4)(n_{13} + n_{31})}{n_3 + n_4} =$$

$$n_3\hat{p}$$

Following this rule, we can say expected values of others will be

$$\hat{E}(n_{12}) = n_2\hat{p},\ \hat{E}(n_{31}) = n_1\hat{q}$$

Now we can rewrite χ^2 statistic:

$$\chi^2 = \frac{n_1^2(\hat{p}_1 - \hat{p})^2}{n_1\hat{p}} + \frac{n_1^2(\hat{q}_1 - \hat{q})^2}{n_1\hat{q}} + \frac{n_2^2(\hat{p}_2 - \hat{p})^2}{n_2\hat{p}} + \frac{n_2^2(\hat{q}_2 - \hat{q})^2}{n_2\hat{q}}$$

$$\chi^2 = \frac{n_1(\hat{p}_1 - \hat{p})^2}{\hat{p}} + \frac{n_1[(1 - \hat{p}_1) - (1 - \hat{p})]^2}{\hat{q}} + \frac{n_2(\hat{p}_2 - \hat{p})^2}{\hat{p}} + \frac{n_2[(1 - \hat{p}_2) - (1 - \hat{p})]^2}{\hat{q}}$$

$$\chi^2 = \frac{n_1(\hat{p}_1 - \hat{p})}{\hat{p}\hat{q}} + \frac{n_2(\hat{p}_2 - \hat{p})}{\hat{p}\hat{q}}$$

Now we can plug in our expression for $\hat{p}$

$$\chi^2 = \frac{n_1}{\hat{p}\hat{q}}(\frac{n_1\hat{p}_1 + n_2\hat{p}_1 - n_1\hat{p}_1 - n_2\hat{p}_2}{n_1 + n_2})^2 + \frac{n_2}{\hat{p}\hat{q}}(\frac{n_1\hat{p}_2 + n_2\hat{p}_2 - n_1\hat{p}_1 - n_2\hat{p}_2}{n_1 + n_2})^2$$

$$\chi^2 = \frac{n_1 n_2(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}(n_1 + n_2)} = Z^2$$

We can see here our chi sqaured value is 36.198 and wse know our z value is 6.087284, and (6.087284)^2= 37.05503, which is extremely close to what we got.

2. Consider the following model for the responses measured in a randomized block design containing b blocks and k treatments: $Y_{ij} = \mu + \tau i + \beta_j + \epsilon_{ij}$,

a. Assuming the model just described is accurate, show that observations taken from different blocks are independent of one another. That is, show that Yij and Yij’ are independent if j = j’, as are Yij and Yi’j’ if i = i’ and j = j’.

$$\text{Given that } y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$
$$\text{We can say } E_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_C^2), B_j \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$$
$$\text{Where } E_{ij} \text{ and } B_j \text{ are independent}$$
$$E[E_{ij}] = \mu + \tau_i \text{ Where } \tau_i \text{ is a constant, thus}$$
$$E[\tau_i] = 0 \text{ and } var(Y_{ij}) = \sigma_B^2 + \sigma_C^2$$
$$\text{Assuming that } i \neq i'$$
$$Cov(y_{ij}, y_{ij'}) = Cov(\mu + \tau_i + B_j + E_{ij}, \mu + \tau_i + B_{j'} + E_{ij'})$$
$$= cov(B_j + E_{ij}, B_{j'} + E_{ij'})$$
$$= cov(B_j, B_{j'}) + cov(B_j, E_{i'j'}) + cov(E_{ij}, B_{j'} + cov(E_{ij}, E_{i'j'})$$
$$= 0 + 0 + 0 + 0 = 0$$
$$\text{Assuming j} \neq j'$$
$$Cov(y_{ij}, y_{ij'}) = Cov(\mu + \tau_i + B_j + E_{ij}, \mu + \tau_i + B_{j'} + E_{ij'})$$
$$= cov(B_j + E_{ij}, B_{j'} + E_{ij'})$$
$$= cov(B_j B_{j'}) + cov(B_j E_{ij'}) + cov(B_{j'}, E_{ij}) + cov(E_{ij}, E_{ij'})$$
$$= 0 + 0 + 0 + 0 = 0$$
$$\text{Thus, } Y_{ij} \text{ and } Y_{ij'} \text{ are indepdent if } i \neq i' and j \neq j'$$

b. Derive the covariance of two observations from the same block. That is, find Cov (Yij , Yi'j) if i = i'.

$$cov(y_{ij}, y_{i'j} = cov(\mu + \tau_i + \beta_j + E_{ij}, \mu + \tau_{i'} + \beta_j + E_{i'j})$$
$$\text{Since we know } \tau_i \text{is fixed, we can rewrite}$$
$$= cov(B_j + E_{ij}, B_j + E_{i'j})$$
$$= cov(B_j, B_j) + cov(B_j, E_{i'j}) + cov(E_{ij}, B_j) + cov(E_{ij}, E_{i'j})$$
$$= var(B_j) + 0 + 0 + 0$$
$$= \sigma_B^2$$

(c) Two random variables that have a joint normal distribution are independent if and only if their covariance is 0. Use the result from part (b) to determine the conditions under which two observations from the same block are independent of one another.

$$\text{For two observtions from the same block to be indepedent, } cov(Y_{ij}, Y_{i'j}) = 0$$
$$\text{Thus } \sigma_B^2 = 0$$
$$\text{Thus, } cov(Y_{ij}, Y_{i'j}) = 0$$

(d) Find the expected value and variance of Yij .

$$E[Y_{ij}] = \mu + \tau_i \text{ since } \mu, \tau_i \text{ are constant}$$
$$var(Y_{ij}) = v(B_j) + V(E_{ij}) + 2cov(B_j, E_{ij})$$
$$= \sigma_B^2 + \sigma_C^2 + 0$$
$$= \sigma_B^2 + \sigma_C^2$$
$$\text{This implies } Y_{ij} \sim N(\mu + \tau_i, \sigma_B^2 + \sigma_c^2)$$

(e) Let $\bar{Y}_{i*}$ denote the average of all responses to treatment i. Use the model to derive E[$\bar{Y}_{i*}$] and V[$\bar{Y}_{i*}$].

$$\text{We found } Y_{ij} \sim N(\mu + \tau_i, \sigma_B^2 + \sigma_c^2)$$
$$\text{Thus } \bar{Y}_{i*} = \frac{1}{n} \sum Y_{ij} \text{ from j=1 to n, to find the average}$$
$$= \frac{1}{n}(Y_{i1}, \ldots\ldots Y_{in})$$
$$\text{Now we can plug this into our } Y_{ij}$$
$$n\bar{Y}_{i*} \sim N(n(\mu + \tau_i), n^2(\sigma_B^2 + \sigma_C^2))$$
$$\text{We can see } E[n\bar{Y}_{i*}] = n(\mu + \tau_i), var(n\bar{Y}_{i*}) = n^2(\sigma_B^2 + \sigma_C^2))$$
$$\text{By isolating the } \bar{Y}_{i*}$$
$$E[\bar{Y}_{i*}] = \mu + \tau_i, \;\; var(\bar{Y}_{i*}) = \frac{(\sigma_B^2 + \sigma_C^2))}{n}$$

f. Calculate the bias of $\bar{Y}_{i*}$. Is it an unbiased estimator of the mean response to treatment i?

$$\text{We know } E[B_j] \text{ and } E[E_{ij}] = 0$$

$$\bar{Y}_{i*} = \frac{1}{n}\sum Y_{ij}$$

$$\bar{Y}_{i*} = \frac{1}{n}\sum(\mu + \tau_i + B_j + E_{ij})$$

$$= \frac{1}{n}\sum \mu + \frac{1}{n}\sum \tau_i + \frac{1}{n}\sum B_j + \frac{1}{n}\sum E_{ij} \text{ from j=1 to n}$$

$$= \frac{n\mu}{n} + \frac{n\tau_i}{n} + 0 + 0$$

$$= \mu + \tau_i$$

$$\text{The bias of } \bar{Y}_{i*} \text{ is 0 because } E[\bar{Y}_{i*}] = \bar{Y}_{i*}$$

$$\text{Since the bias is 0, we have an unbiased estimator}$$

3. For a comparison of the academic effectiveness of two junior high schools A and B, an experiment was designed using ten sets of identical twins, where each twin had just completed the sixth grade. In each case, the twins in the same set had obtained their previous schooling in the same classrooms at each grade level. One child was selected at random from each set and assigned to school A. The other was sent to school B. Near the end of the ninth grade, an achievement test was given to each child in the experiment. The results are shown in the accompanying table.

a. Using the sign test, test the hypothesis that the two schools are the same in academic effectiveness, as measured by scores on the achievement test, against the alternative that the schools are not equally effective. What would you conclude with α = .05?

```
groupA<-c(67,80,65,70,86,50,63,81,86,60)
groupB<-c(39,75,69,55,74,52,56,72,89,47)
d_i<-c(28,5,-4,15,12,-2,7,9,-3,13)
sign<-c('+','+','-','+', '+', '-', '+','+', '-', '+')
data2<-data.frame(groupA,groupB,d_i,sign)
data2
```

```
##    groupA groupB d_i sign
## 1      67     39  28    +
## 2      80     75   5    +
## 3      65     69  -4    -
## 4      70     55  15    +
## 5      86     74  12    +
## 6      50     52  -2    -
## 7      63     56   7    +
## 8      81     72   9    +
## 9      86     89  -3    -
## 10     60     47  13    +
```

Here, we have 3 negtive signs, and 7 positive signs. H_0 : the median difference is 0 H_a : the median difference is not 0

```
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
binom.test(x=7, n=10, p=0.5, alternative="two.sided")
```

```
##
##  Exact binomial test
##
## data:  7 and 10
## number of successes = 7, number of trials = 10, p-value = 0.3438
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3475471 0.9332605
## sample estimates:
## probability of success
##                    0.7
```

```
SIGN.test(d_i, md=0, alternative= 'two.sided', conf.level=0.95)
```

```
##
##  One-sample Sign-Test
##
## data:  d_i
## s = 7, p-value = 0.3437
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##  -2.675556 14.351111
## sample estimates:
## median of x
##           8
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level  L.E.pt  U.E.pt
## Lower Achieved CI     0.8906 -2.0000 13.0000
## Interpolated CI       0.9500 -2.6756 14.3511
## Upper Achieved CI     0.9785 -3.0000 15.0000
```

Here, we can see our p value is 0.3438 which is greater than 0.05 at an alpha level of 0.05. Thus, we cannot reject the null hypothesis that the schools are the same in academic effectiveness.

b. Suppose it is suspected that junior high school A has a superior faculty and better learning facilities. Test the hypothesis of equal academic effectiveness against the alternative that school A is superior. What is the p-value associated with this test?

We can use the Kruskal-Wallis test here to compare if the means are equal or if one is greater. Since our p value for both groups is >0.05, both groups of data are normally distributed.

```
shapiro.test(groupA)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  groupA
## W = 0.93686, p-value = 0.5187
```

```
shapiro.test(groupB)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  groupB
## W = 0.96297, p-value = 0.8191
```

```
kwTest<- kruskal.test(groupA,groupB)
kwTest
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  groupA and groupB
## Kruskal-Wallis chi-squared = 9, df = 9, p-value = 0.4373
```

We have a chi squared value of 9, and a p value of 0.4373.

$H = (12/(N(N+1)) * (\sum T2/n) - 3(N+1)\ H = (20(20+1)) * (\sum T2/20) - 3(20+1)\ H = 0.029 * 2244.2 - 63 = 1.12$

and at alpha=0.05, our critical value is 16.9190. Our p value is greater than .05, and our test value of 1.12 is less than 16.9, thus we fail to reject the null hypothesis that there is equal effectiveness between schools.

Another way to do this is by using the n choose k formula.

At M=7, we have $p(M \geq 7) = [(\frac{10}{7}) + (\frac{10}{8}) + (\frac{10}{8}) + (\frac{10}{10})]\ 0.5^{10} = 0.1718$

Again, our p value, 0.1718 > 0.05, so we fail to reject the null.

c. Repeat the test in (a), using the Wilcoxon signed-rank test. Compare your answers.

```
wilcox.test(groupA,groupB, paired=TRUE,conf.int=T, conf.level=0.95)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  groupA and groupB
## V = 49, p-value = 0.02734
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##   1.5 14.0
## sample estimates:
## (pseudo)median
##              7
```

```
wilcox.test(groupB,groupA, paired=TRUE,conf.int=T, conf.level=0.95)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  groupB and groupA
## V = 6, p-value = 0.02734
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -14.0  -1.5
## sample estimates:
## (pseudo)median
##             -7
```

We can see here we have a rank value of 49 max and 6 min and a p value of 0.02734. By looking at the table, we can find a critical value for n=10 and alpha=0.05 at 8. Since our min test statistic is lower that 8, at 6, and our p value is less than 0.05, we can reject the null hypothesis that the two schools are the same in academic effectiveness.

This answer is different from part a and b because we reject our null hypothesis this time.

4. Let Y1, Y2, …, Yn denote a random sample from an exponentially distributed population with density f (y|θ) = θe−θy, 0 < y. (Note that the mean of this population is μ =1θ.) Use the conjugate gamma (α, β) prior for θ to find the following:

a. The joint density, or f (y1, y2, …, yn, θ);

where Π is i=1 to n.

Using the definition 4.9 in the textbook, we are given the gamma distribution with parameters beta and alpha.

$$f(y_1, \ldots\ldots y_n, \theta) = (\ \Pi\theta e^{-\theta y_i})\frac{1}{\Gamma\alpha\beta^\alpha}\theta^{(\alpha-1)}e^{(-\theta/\beta)}$$

$$= ([\theta e^{-\theta y_1}]\frac{1}{\Gamma(\alpha)\beta^\alpha}\theta^{\alpha-1}e^{\frac{-\theta}{\beta}})([\theta e^{-\theta y_2}]\frac{1}{\Gamma(\alpha)}\beta^\alpha\theta^{\alpha-1}e^{\frac{-\theta}{\beta}})\ *\ \ldots.$$

$$= \frac{\theta^{n+\alpha-1}}{\Gamma\alpha\beta^\alpha}e^{-\theta\sum y_i-\frac{\theta}{\beta}}$$

$$= \frac{\theta^{n+\alpha-1}}{\Gamma\alpha\beta^\alpha}e^{-\theta(\sum(y_i)-1/\beta)}$$

$$= \frac{\theta^{n+\alpha-1}}{\Gamma\alpha\beta^\alpha}e^{-\theta/\frac{1}{\sum y_i}-\theta/\beta}$$

$$= \frac{\theta^{n+\alpha-1}}{\Gamma\alpha\beta^\alpha}e^{\frac{-\theta}{1/\sum y_i+\beta/1}}$$

$$= \frac{\theta^{n+\alpha-1}}{\Gamma\alpha\beta^\alpha}e^{\frac{\frac{-\theta}{\beta}}{\beta\sum y_i+1}}$$

b. The marginal density, or m (y1, y2, …, yn); where the integration is from 0 to infinity.

$$m(y_1, \ldots . y_n) = \int f(y_1, \ldots . y_n, \theta) d\theta$$

$$= \int \frac{\theta^{n+\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{\frac{\frac{-\theta}{\beta}}{\beta \sum y_i + 1}}$$

$$= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int \theta^{n+\alpha-1} e^{\frac{\frac{-\theta}{\beta}}{\beta \sum y_i + 1}} d\theta$$

We can notice now that this is similar to the gamma density therefore

$$m(y_1, \ldots . y_n) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \Gamma(n+\alpha)(\frac{\beta}{\beta \sum y_i + 1})$$

Now, by using the density of gamma distribution $n + \alpha$ and $\frac{\beta}{\beta \sum_{i=1}^{n} y_i + 1}$

$$\int_0^{\inf} f(x) dx = \int \frac{1}{\Gamma(n+\alpha)(\frac{\beta}{\beta \sum y_i + 1})^{n+\alpha}} \chi^{n+\alpha} e^{\frac{\frac{-\theta}{\beta}}{\beta \sum y_i + 1}} dx = 1$$

Thus, by using this and part A, we see

$$m(y_1, \ldots . y_n) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \Gamma(n+\alpha)(\frac{\beta}{\beta \sum y_i + 1})$$

c. The posterior density for θ|(y1, y2, …, yn).

The posterior density is found by dividing the marginal by the joint density.

$$g(\theta | y_1, \ldots . y_n) = \frac{f(y_1, \ldots . y_n, \theta)}{m(y_1, \ldots . y_n)}$$

$$= \frac{\frac{\theta^{n+\alpha-1}}{\Gamma\alpha\beta^{\alpha}} e^{\frac{\frac{-\theta}{\beta}}{\beta \sum y_i + 1}}}{\frac{1}{\Gamma(\alpha)\beta^{\alpha}} \Gamma(n+\alpha)(\frac{\beta}{\beta \sum y_i + 1})}$$

$$= \frac{1}{\Gamma(n+\alpha)(\frac{\beta}{\beta \sum y_i + 1})^{(n+\alpha)}} \theta^{n+\alpha} e^{\frac{\frac{-\theta}{\beta}}{\beta \sum y_i + 1}}$$