UNIVERSITY OF TEXAS AT AUSTIN

Problem Set # 15

Goodness of fit.

Problem 15.1. Gregor Almond, the local horticulturalist, grows 400 progeny from a cross of peas. The cross is hypothesised to have a ratio of 1 green to 7 yellow seeds. Suppose that the cross actually produces 360 yellow and 40 green seeded plants.

(i) Calculate the observed value of the test statistic.
(ii) Using the χ^2-tables, what would your decision be at the significance level α.
(iii) Using **R** with the observed value of the test statistic, find the $p-$value.
(iv) Using the command

```
chisq.test()
```

perform the χ^2-test and provide the summary.
(v) In this case, you can test the same hypotheses using the $z-$test. Do this for practice!

Solution:

(i) The total sample size is $n = 400$, while the number of categories is $k = 2$. We are testing, in our usual notation,

$$H_0 : p_g = 1/8, p_y = 7/8 \quad vs. \quad H_a : \text{the color distribution is different from the null}$$

The observed counts are $O_y = 360$ and $O_g = 40$. The expected counts are $E_y = 350$ and $E_g = 50$. Note that both expected counts exceed 5, so we can use the χ^2-distribution with $k - 1 = 2 - 1 = 1$ degrees of freedom as the approximate distribution of our test statistic. The observed value of the test statistic is

$$q^2 = \frac{(360 - 350)^2}{350} + \frac{(40 - 50)^2}{50} = 2.285714.$$

(ii) At the 5%-significance level, the critical value of the χ^2-distribution with 1 degree of freedom is $\chi^2_{0.05}(df = 1) = 3.84$. Since this value is larger than the observed value of the test statistic, we **fail to reject**.

(iii) In **R**, the command line to use to get the upper-tail probability of a χ^2 random variable with one degree of freedom is

```
1-pchisq(2.285714, df=1)
```

We get 0.13057.

(iv) In **R**, we use the following commands:

```
> peas<-c(360,40)
> test<-chisq.test(peas,p=c(7/8,1/8))
> test
```

The output is:

```
        Chi-squared test for given probabilities

data:  peas
X-squared = 2.2857, df = 1, p-value =
0.1306
```

(v) Let p denote the population proportion of yellow-seeded peas. Then, we are testing

$$H_0 : p = 7/8 \quad vs. \quad H_a : p \neq 7/8.$$

The observed proportion of yellow-seeded peas is $\hat{p} = 0.9$. The $z-$score is

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.875 - 0.9}{\sqrt{\frac{0.875(0.125)}{400}}} = -1.511858.$$

With $Z \sim N(0, 1)$, we can express the $p-$value as

$$\mathbb{P}[Z < -1.511858] + \mathbb{P}[Z > 1.511858] = 2\mathbb{P}[Z < -1.511858].$$

In **R**, we can use the command

```
2*pnorm(-1.511858)
```

Of course, we get 0.13057, the same $p-$value as in the χ^2-test.

Problem 15.2. (8 points) You suspect that a die has been altered so that the outcomes of a roll (the numbers 1 through 6) are not equally likely. You roll the die 600 times and observe the following counts:

Outcome	1	2	3	4	5	6
Count	85	86	120	118	91	100

(i) Using the $\chi 2-$tables, at the significance level of 0.05 perform the goodness-of-fit test and report your conclusions.
(ii) Using **R**, perform the goodness-of-fit test.

Solution:

(i) The null hypothesis is that the die is fair, i.e., that the probability of every outcome is 1/6. Hence, under the null, the expected count for every possible outcome on the die out of 600 trials is 100. The observed value of our test statistics is

$$\begin{aligned} Q^2 &= \frac{(85-100)^2}{100} + \frac{(86-100)^2}{100} + \frac{(120-100)^2}{100} + \frac{(118-100)^2}{100} + \frac{(91-100)^2}{100} + \frac{(100-100)^2}{100} \\ &= \frac{1}{100}(15^2+14^2+20^2+18^2+9^2+0^2) = \frac{1226}{100} = 12.26 \end{aligned} \tag{15.1}$$

At the 0.05 significance level, the critical value of the χ^2-distribution with $6-1=5$ degrees of freedom is 11.07. Since the obserced value of the test statistic is greater than the critical value, we **reject** the null hypothesis at the significance level of 0.05.

(ii) In **R**, we just do

```
> die<-c(85,86, 120, 118, 91, 100)
> fair<-c(1/6,1/6,1/6,1/6,1/6,1/6)
> chisq.test(die,p=fair)
```

The output is

```
        Chi-squared test for given probabilities

data:  die
X-squared = 12.26, df = 5, p-value = 0.03139
```

Problem 15.3. The early education department of the local community college conducts a survey of 1000 randomly chosen children on their favorite among certain offered holidays. Here are the results of this survey:

Halloween	Thanksgiving	Arbor Day	Other
60%	10%	2%	28%

Then, the children were shown an inspirational movie on ecology and horticulture. After that, they were asked, again, to choose their favorite holiday among those offered. Here are the results of the renewed survey:

Halloween	Thanksgiving	Arbor Day	Other
58%	8%	8%	26%

Using the χ^2-goodness-of-fit test, say whether there is sufficient evidence that the chidrens' opinion was changed by the movie at the 0.01 significance level. Solve this problem both ways, i.e., using the χ^2-tables and using **R**.

Solution: We are testing

$$H_0 : p_1 = 0.60, p_2 = 0.10, p_3 = 0.02, p_4 = 0.28$$

$$vs.$$

$$H_a : \text{At least one of the probabilities is different from the null value}$$

Here, the observed counts are exactly the ones we obtain with the probabilities given in the second table:

$$O_1 = 1000(0.58) = 580,$$
$$O_2 = 1000(0.08) = 80,$$
$$O_3 = 1000(0.08) = 80,$$
$$O_4 = 1000(0.26) = 260.$$

The expected counts will be obtained using the proportions from the first tables. We get

$$E_1 = 1000(0.60) = 600,$$
$$E_2 = 1000(0.10) = 100,$$
$$E_3 = 1000(0.02) = 20,$$
$$E_4 = 1000(0.28) = 280.$$

The observed value of the test statistic is, then,

$$q^2 = \frac{(580-600)^2}{600} + \frac{(80-100)^2}{100} + \frac{(80-20)^2}{20} + \frac{(260-280)^2}{280} = 186.0952.$$

This observed value of the test statistic is obviously extraordinarily large so that there must be an effect of the inspirational film.

Using **R**, we would do the following:

```
> exp<-c(600,100,20,280)
> obs<-c(580,80,80,260)
> exp<-c(600,100,20,280)
> null.probs<-exp/1000
> chisq.test(obs,p=null.probs)
```

This is the output we get

```
        Chi-squared test for given probabilities

data:  obs
X-squared = 186.1, df = 3, p-value < 2.2e-16
```