The University of Sydney
STAT2012 Statistical Tests

Semester 2	Computer Practice Week 12	2015

Useful R commands

- Chi-square goodness-of-fit test on class frequencies:

  If y is a vector of k observed class frequencies $y_1, ..., y_k$, the test statistic

  $$\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{k} \frac{y_i^2}{np_{i0}} - n \sim \chi_{k-1}^2$$

  and its p-value for testing the hypothesis on class probabilities:

  $$H_0 : \ p_1 = p_{10}, \ \cdots, \ p_k = p_{k0}$$

  can be calculated using the R code

  ```
  chisq.test(y,p=p)
  ```

  where p is a vector of k expected probabilities $p_{10}, \ldots, p_{k0}$ under H_0.

  Note that p can be calculated under H_0 of certain hypothesized distribution for the data x and y is a vector of frequencies for each value or ranges of values of x. The degree of freedom for the test is $k - 1 - h$ where h is the number of distribution parameters to be estimated from the data. Note that the p-value in running chisq.test(y,p=p) does not account for the lose of degree of freedom and hence is *incorrect*.

Important points

- You will perform Chi-square goodness-of-fit test on a set of hypothesized proportions.
- You will perform Chi-square goodness-of-fit test on whether a simulated data set follows a binomial distribution.
- Lastly, you will perform Chi-square goodness-of-fit test on whether the residuals from a two-way ANOVA model follow a normal distribution.

Practice Problems

1. The data set `survey` contains measurements of the following variables from 95 students:

sex	1=male; 2=female
age	Year
height:	Inches
credit:	Number of credit cards in possession
pulse:	Number of heartbeats in one minute
pulse.ex:	Number of heartbeats in one minute after regular exercise over a period
exercise:	Number of hours during last week
smoke:	1=yes; 2=no
hand:	1=left-handed; 2=right-handed; 3=ambidextrous

Open the data set `survey`.

```
survey=read.csv("http://www.maths.usyd.edu.au/u/UG/IM/STAT2012/r/survey.csv")
attach(survey)
```

Test if the proportions of students in the four groups: FS (female smoker), FN (female non-smoker), MS (male smoker) and MN (male non-smoker) are 0.2, 0.3, 0.2, 0.3 respectively using the *Chi-square goodness-of-fit test*.

(a) State the null and alternative hypotheses.

(b) Count the number of students in the four groups: female students who smoke, female students who do not smoke, male students who smoke and male students who do not smoke. For example:

```
n1=length(pulse[smoke==1 & sex==2])
```

State if the rule of five is satisfied.

(c) Perform the test using `chisq.test(y,p=p)` and report the test statistic and p-value. Draw your conclusion about H_0 based on the p-value.

2. Conduct the following experiment on the *Chi-square goodness-of-fit test* to test if a data follows a binomial distribution.

(a) State the null and alternative hypotheses.

(b) Generate a data set of 200 observations which follows a binomial distribution with $n = 4$ and $p = 0.4$. Obtain a vector `y` of frequencies y_i for the simulated values 0,1,2,3,4.

```
set.seed(12345)   #set random no. generator, diff. seed gives diff. b
b=rbinom(200,4,0.4)
b
x=c(0:4)
y=c()
y[1]=length(b[b==0])
```

```
y[2]=length(b[b==1])
y[3]=length(b[b==2])
y[4]=length(b[b==3])
y[5]=length(b[b==4])
y
```

(c) Plot the data y using a bar chart.

(d) Estimate the probability prob of the binomial distribution and hence calculate the vectors of expected probabilities p and expected frequencies ey. State the probability estimate and if the rule of five is satisfied.

(e) Calculate the test statistic and p-value for the test. Hint: subtract the number of parameter estimates from the degrees of freedom. Draw your conclusion of the test based on the p-value.

3. The table below gives the prices for four levels of discount and four promotion programs. The price is repeatedly measured 10 times for each level of discount and promotion program resulting in $n = 160$ observations.

```
discount promotion pri
1 10% 1 4.10
2 20% 1 3.57
...
160 40% 7 4.70
```

Open the data set price.

```
price=read.csv("http://www.maths.usyd.edu.au/u/UG/IM/STAT2012/r/price.csv")
attach(price)
```

Test if the residuals of the *two-way ANOVA* model follow a normal distribution using the *Chi-square goodness-of-fit* test with $k = 5$ classes.

(a) State the null and alternative hypotheses.

(b) Set x to be the residuals of the two-way ANOVA model with interaction.

```
promotion=as.factor(promotion)
x=aov(pri˜promotion*discount)$res
```

Then plot a histogram for x using hist(x,col="lightgray") to view the distribution.

(c) Divide the range of x into 5 intervals and store the cut-off points in int. State the estimates of the unknown μ and σ^2 which are given by the sample mean and sample standard deviaton respectively.

(d) Count the observed frequencies y for the intervals. Hint: take some suitable R commands from the lecture notes to calculate y.

(e) Calculate the expected probabilities p for the intervals. Hint: take some suitable R commands from the lecture notes to calculate p.

(f) Calculate the test statistic and the p-value. Draw your conclusion about H_0 based on the p-value. Hint: the `chisq.test` command will not give you the right degrees of freedom. Take some suitable R commands from the lecture notes to calculate the expected frequencies `ey` and then the test statistic and p-value.

(g) Plot two bar charts for `y` and `ey` respectively side by side to view their differences. The codes for bar charts are given.