Dr Isabella Gollini

Probability and Statistics

Lab – Goodness of fit

- Section 1 contains a summary of the goodness of fit procedure.
- Section 2 [page 5] contains a guided example on then 2014 FIFA World Cup data.
- Sections 3 and 4 [pages 10 and 11] contain other two exercises. [Try to do them on your own, but do not hesitate to ask any questions!]

Remember to write you code into the `R` script, in order to be able to save it.

1 Goodness of Fit test procedure

1. **Specify the theoretical distribution**

   Specify a theoretical distribution that assign a fixed probability that a member of the population will fall into one of the categories [this is often given in the text of the exercise]. Distributions that we have seen during the course:

   - Uniform probability $p_r = \frac{1}{k}$ where k is the number of categories.
   - Fixed probabilities not from a uniform distribution [like the Mendel example of Section 12.3.]
   - Poisson distribution with μ estimated by the sample mean: $p_r = e^{-\bar{x}} \frac{\bar{x}^r}{r!}$ $(r = 0, 1, 2, \ldots)$

2. H_0 **and** H_1

   H_0 : The data comes from a population that fits the specified distribution

   H_1 : The data comes from a population that has a different distribution

3. **Test Statistic**

   Requirement: The sample size n should be large enough such that $E_r \geq 5$ in each category. We can merge different categories if that's meaningful. Otherwise we can not proceed with this test.

   To test the null hypothesis H_0 we use the test statistic:

   $$X^2 = \sum_{r=1}^{k} \frac{(O_r - E_r)^2}{E_r} \quad \sim \chi^2_{k-1-d}$$

   where O_r represents the observed frequency of the score r and E_r represents the expected frequency of the score r under the null hypothesis. Where k is the number of cells.
   d is the number of fitted parameters. For the three cases we have seen during the course:

   - Uniform probability $d = 0$
   - Fixed probabilities $d = 0$
   - Poisson distribution with μ estimated by the sample mean $d = 1$

4. **Rejection region Significance-level-α test**

   Reject H_0 if $X^2 \geq \chi^2_{k-1-d,\alpha}$. Do not reject H_0 otherwise.

5. p**-value**

   $$p-\text{value} = 1 - F(X^2)$$

   where $F(X^2)$ is the c.d.f. of a χ^2-distribution with $k - 1 - d$ degrees of freedom.

   Reject H_0 if $p-\text{value} \leq \alpha$. Do not reject H_0 otherwise.

`R` function

`Or` is a vector containing the observed frequencies, and `pr` is a vector of expected probabilities.

```
test <- chisq.test(Or, p = pr)
test
```

ATTENTION: The function `chisq.test` calculates the p-value assuming the test statistic follows a χ^2 distribution with $(k-1)$ degrees of freedom. If we want to test if the data follows a Poisson distribution, and the parameter of the Poisson distribution has been estimated by using the data the test statistic follows a χ^2 distribution with $(k-1-1)$ degrees of freedom. The correct p-value is obtained by using:

```
1 - pchisq(test$statistic, df = length(Or) - 2)
```

The dispersion test for the Poisson distribution

1. An alternative approach to testing the goodness of fit of a Poisson distribution to a sample of data is based upon the fact that, as shown in Section 5.5, for a Poisson distribution the variance is equal to the mean, i.e., $\sigma^2 = \mu$.

2. H_0 **and** H_1

   H_0: The data are a random sample from a Poisson distribution

   H_1: The data are a random sample from another distribution where $\sigma^2 > \mu$.

3. **Test Statistic**

   The *index of dispersion* is defined as,

   $$I = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}} \ ,$$

   which, under our Poisson hypothesis, has approximately the χ^2_{n-1} distribution.

4. **Rejection region Significance-level-α test**

   Reject H_0 if $I \geq \chi^2_{n-1,\alpha}$. Do not reject H_0 otherwise.

5. p**-value**

   $$p\text{−value} = 1 - F(I)$$

   where $F(I)$ is the c.d.f. of a χ^2-distribution with $n-1$ degrees of freedom.

   Reject H_0 if p−value $\leq \alpha$. Do not reject H_0 otherwise.

R code

```
n <- sum(Or)
xbar <- sum(r * Or) / n
s2 <- (sum(Or * r^2) - n * xbar^2) / (n - 1)
xbar
s2
```

```
ss <- s2 * (n - 1)
```

```
index <- ss / xbar
pvalue <- 1 - pchisq(index, df = n - 1)
pvalue
```

2 2014 FIFA World Cup

Is the number of goals scored by teams during the first stage matches of the 2014 FIFA World Cup a realization of a well known distribution?

The data are available from the FIFA official website `http://www.fifa.com/worldcup/archive/brazil2014/matches/`.

We are interested in analysing the frequency on the number of goals scored by teams during the first stage matches of the 2014 FIFA World Cup football championship that took place in Brazil in summer 2014:

Number of Goals	Frequency
0	26
1	31
2	21
3	11
4	5
5	2
≥ 6	0

Descriptive analysis of the data

Create a new vector `Goals` containing the number of goals, that in this case ranges from 0 to 5.

```
Goals <- 0:5
```

Create a new vector `Or` containing the observed frequencies of goals.

```
Or <- c(26, 31, 21, 11, 5, 2)
cbind(Goals, Or)
```

Check the sample size `n`.

```
n <- sum(Or)
n
```

You can notice that `n` is equal to twice the number of matches played in the first stage of the world cup, as each match is played by two teams.

Plot the data and decides to which distribution do you want to compare the data to.

```r
barplot(Or,
  names = Goals,
  xlab = "Goals",
  ylab = expression(O[r]),
  col = "red")
```

From the barplot, it seems reasonable to perform a test to investigate if the data comes from a population that follows a Poisson distribution. Another check that we may want to do perform before proceeding with the test is to check if the sample mean and sample variance are close to each other.

```r
xbar <- sum(Goals * Or) / n
xbar
s2 <- (sum(Goals^2 * Or)  - n * xbar^2) / (n - 1)
s2
```

As the sample mean and the sample variance are quite similar, this is another indication that proceed to the test for goodness of fit to a Poisson distribution seems reasonable.

Goodness of Fit for the Poisson Distribution

Let's calculate the expected probabilities under the null hypothesis that the data follows a Poisson distribution.

Because we want the cell probabilities to sum to 1 and the expected frequencies to sum to `n`, we modify the description of the final cell to be number of goals "≥ 5" and the corresponding probability to be $1 - F_5 = 1 - \sum_{r=0}^{4} p_r$.

```r
pr <- numeric(6)
pr[1:5] <- dpois(Goals[-6], xbar)
pr[6] <- 1 - sum(pr[1:5])
names(pr) <- c(0:4, ">= 5")
pr
```

The expected frequencies are given by:

```r
Er <- pr * n
Er
cbind(Or, Er)
```

We can do a barplot with the observed and the expected frequencies:

```
barplot(rbind(Or, Er),
  names = c(0:4, ">= 5"),
  xlab = "Goals",
  ylab = "Frequency",
  beside = TRUE,
  col = c("red", "turquoise"))
```

In order to perform the chi-square test we have to use the function `chisq.test` and specify the vector containing the observed frequencies, and `p`, a vector containing the probabilities under H_0.

```
test <- chisq.test(Or, p = pr)
test
```

ATTENTION! There is a warning message from R! There are some expected frequencies that are < 5, and the results are unreliable!

In order to get the right results we have to create a new class that contains the 4 and 5 goals. And we recalculate the observed frequencies, the expected probabilities under H_0, and the expected frequencies.

```
Or2 <- c(Or[1:4], sum(Or[5:6]))
pr2 <- c(pr[1:4], sum(pr[5:6]))
names(pr2) <- c(0:3, ">= 4")
Er2 <- pr2 * n
cbind(Or2, Er2)
```

We can now perform the chi-square test:

```
test <- chisq.test(Or2, p = pr2)
test
```

ATTENTION – Again! The p-value is wrong!

The function `chisq.test` calculates the p-value assuming the test statistic follows a χ^2 distribution with $(k-1)$ degrees of freedom, as it does not allow to include the information that the parameter of the Poisson distribution has been estimated by using the data, and so, the test statistic follows a χ^2 distribution with $(k-1-1)$ degrees of freedom. The correct p-value is obtained by using:

```
1 - pchisq(test$statistic, df = length(Or2) - 2)
```

where `test$statistic` is the value of the test statistic. Type `test$statistic` to check!

Draw conclusions:

The dispersion test for the Poisson distribution

A second approach to test the hypothesis that the data are a random sample from a Poisson distribution is the dispersion test.

We want to test if the sample variance is so much greater then the sample mean that it provides significant evidence to reject the hypothesis that the data are a random sample from a Poisson distribution?

Let's calculate the index of dispersion. The numerator is given by the corrected sum of squares of the data:

```
ss <- sum((Or * Goals^2)) - n * xbar^2
ss
```

So, the index of dispersion is:

```
index <- ss / xbar
index
```

The p-value is:

```
pvalue <- 1 - pchisq(index, df = n - 1)
pvalue
```

Draw conclusions:

Interpret your conclusion in the context of the application:

3 Birthdays – Adapted from 2013 Exam

The birthdays of a random sample of 200 students in a college were found to fall in the quarters of the year as follows:

Quarter	1	2	3	4
Frequency	62	48	44	46

(a) Plot the data.

(b) Write down the expected frequencies under the null hypothesis that the students are drawn from a population in which birthdays are uniformly distributed throughout the year.

(c) State the approximate distribution of the test statistic under the null hypothesis.

(d) Carry out the goodness of fit test and draw conclusions.

(e) Interpret your conclusion in the context of the application.

4 Homicides in London

Spiegelhalter and Barnett (2009)[1] in a paper in the statistics magazine *Significance* (`https://www.statslife.org.uk/significance`) examined data on the daily numbers of homicides in London for the 1095 days over the three year period from April 2004 to March 2007. The frequency distribution of the daily numbers of homicides is given in the table below.

Number of homicides	Frequency
0	713
1	299
2	66
3	16
4	1
≥ 5	0

Frequency distribution of daily numbers of homicides

(a) Calculate the sample mean and sample variance of the daily number of homicides.

(b) Carry out a chi-square goodness of fit test to test the hypothesis that the data may be regarded as a random sample from a Poisson distribution. Draw conclusions. [Pay attention to the number of degrees of freedom of the approximate distribution of the test statistic under the null hypothesis].

(c) As an alternative to the method of part (b), carry out a dispersion test to test the same hypothesis. Draw conclusions.

(d) Plot the observed and the expected frequencies.

(e) Interpret your conclusion in the context of the application.

[1] `http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2009.00334.x/abstract`