Simple Statistical Inference II

This chapter introduces some further elementary statistical techniques and their application via S-Plus. Most of the material includes chi–square goodness of fit tests, tests for proportions and some manipulations with cross–classified data.

Goodness of Fit Tests

Goodness of fit (GOF) tests can be thought an another, more formal, technique to assess the distributional assumptions underlying the data generating mechanism. S-PLUS evaluates the two best known GOF tests:

- Chi-square (`chisq.gof`).
- Kolmogorov-Smirnov (`ks.gof` function).

The chi-square test applies only in the one-sample case; Kolmogorov- Smirnov can be used in both the one-sample and the two-sample cases. In addition S-Plus provides the function `cdf.compare` for visual inspection of the hypothesized distribution. The main argument is `distribution` which can assume all the well known distributions.

```
> z1 <- rnorm(50)
> par(mfrow=c(2,1))
> cdf.compare(z1, distribution="normal")
> z2 <- rnorm(50)
> cdf.compare(z1, z2)
> z3 <- runif(50)
> cdf.compare(z1, z3)
> cdf.compare(z2, z3)
```

Figures 1 and 2 display the results of the above calculations.

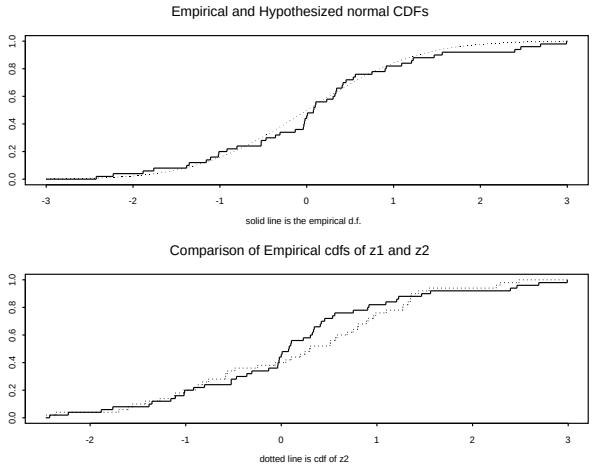

Figure 1: Output of cdf.compare function when assumptions hold.

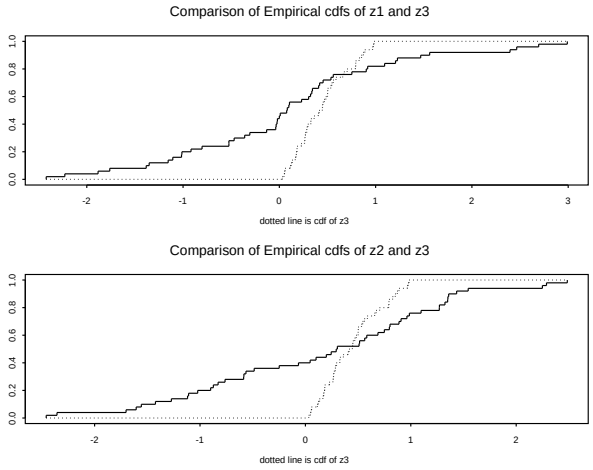

Figure 2: Output of cdf.compare function when assumptions do not hold.

Here is how we can use the function `chisq.gof` for performing the well known Pearsont's goodness of fit test. In the first example the true distribution is Gaussian while in the second is not.

```
> chisq1.out <- chisq.gof(z1, distribution="normal")
> chisq1.out

    Chi-square Goodness of Fit Test

data:  z1
Chi-square = 4.4, df = 9, p-value = 0.8832
alternative hypothesis: True cdf does not equal
the normal Distn. for at least one sample point.

> chisq1.out <- chisq.gof(z3, distribution="normal")
> chisq1.out <- chisq.gof(z1, distribution="normal")
> chisq1.out

    Chi-square Goodness of Fit Test

data:  z1
Chi-square = 4.4, df = 9, p-value = 0.8832
alternative hypothesis: True cdf does not equal the normal
Distn. for at least one sample point.
> chisq2.out <- chisq.gof(z3, distribution="normal")
> chisq2.out

    Chi-square Goodness of Fit Test

data:  z3
Chi-square = 95.2, df = 9, p-value = 0
alternative hypothesis: True cdf does not equal
the normal Distn. for at least one sample point.
```

The function `ks.gof` can be used for the one and two sample Kolmogorov–Smirnov test.

```
> ks.gof(z1, distribution="normal")    #tes if z1 is normal

    One sample Kolmogorov-Smirnov Test of Composite Normality

data:  z1
ks = 0.1057, p-value = 0.5
alternative hypothesis: True cdf is not the normal distn.
with estimated parameters sample estimates:
  mean of x standard deviation of x
 0.08944332                1.183127
```

```
Warning messages:
  The Dallal-Wilkinson approximation, used to calculate
     the p-value in testing composite normality,
     is most accurate for p-values <= 0.10 .
The calculated p-value is 0.168 and so is set to 0.5 . in: dall.wilk(test, nx)
> ks.gof(z1, distribution="t", df=3)  #test if z1 is t with 3 degrees of freedom

    One-sample Kolmogorov-Smirnov Test
    Hypothesized distribution = t

data:  z1
ks = 0.1259, p-value = 0.3745
alternative hypothesis: True cdf is not the t distn.
with the specified parameters

> ks.gof(z1,z2)  #test is z1 and z2 are identically distributed

    Two-Sample Kolmogorov-Smirnov Test

data:  z1 and z2
ks = 0.2, p-value = 0.2719
alternative hypothesis:
  cdf of z1 does not equal the
              cdf of z2 for at least one sample point.

> ks.gof(z1,z3)  #test is z1 and z3 are identically distributed

    Two-Sample Kolmogorov-Smirnov Test

data:  z1 and z3
ks = 0.48, p-value = 0
alternative hypothesis:
  cdf of z1 does not equal the
              cdf of z3 for at least one sample point.
```

Testing Hypotheses for Proporions

S-Plus provides the function `binom.test` for testing hypotheses about proportions. For instance consider tossing a coin 500 times resulting to a number of heads equal to 226 and suppose that we are interested on hypothesis $p = 0.5$.

```
> binom.test(226,500, p=0.5) #test p=0.5 against a two-sided alternative

    Exact binomial test
```

```
data:  226 out of 500
number of successes = 226, n = 500, p-value = 0.0355
alternative hypothesis: true p is not equal to 0.5

> binom.test(226,500, p=0.4) #test p=0.4 against a two-sided alternative

    Exact binomial test

data:  226 out of 500
number of successes = 226, n = 500, p-value = 0.0198
alternative hypothesis: true p is not equal to 0.4

> prop.test(266,500)   #large sample test and confidence interval for p=0.5

    1-sample proportions test with continuity correction

data:  266 out of 500, null probability 0.5
X-square = 1.922, df = 1, p-value = 0.1656
alternative hypothesis: true P(success) in Group 1 is not equal to 0.5
95 percent confidence interval:
 0.4871883 0.5763127
sample estimates:
 prop'n in Group 1
             0.532
```

Try the following for testing whether the median assumes a specific value which is equivalent to the sign test.

```
x <- rnorm(100)
y <- sum(x>0)
binom.test(y, 100) # median == 0 ?
y <- rnorm(100)
d <- x - y
binom.test(sum(d>0),length(d))  # sign test
```

Suppose now that there is interest on comparing two coins and assume that the first coin was tossed 200 resulted to 80 heads while the second coin was tosses 150 and resulted to 100 heads. A large sample test is given in the following:

```
> x <- c(80,100)
> n <- c(200,150)
> prop.test(x,n)

    2-sample test for equality of proportions with continuity correction

data:  x out of n
```

```
X-square = 23.345, df = 1, p-value = 0
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.3739929 -0.1593405
sample estimates:
 prop'n in Group 1 prop'n in Group 2
               0.4         0.6666667
```

Cross Classified Data

Consider the data.frame `solder` which is provided in `S-Plus`. The variables of interest first are Solder (factor with 5 levels) and Mask (with 2 levels). To make a 2x5 contigency table use the function `table` while the `chisq.test` yields the Pearson chi square test. In addition cross calssification of the data can be accomplished with the function `crosstabs`. The argument on the left is the variable that needs to be classified while the arguments on the right show the corresponding categories.

```
> attach(solder)
> names(solder)
[1] "Opening" "Solder"  "Mask"    "PadType" "Panel"   "skips"
> X <- table(Solder, Mask)
> X
      A1.5  A3 A6 B3 B6
 Thin   90 120 60 90 90
Thick   90 150 30 90 90
> chisq.test(X)

    Pearson's chi-square test without Yates' continuity correction

data:  X
X-square = 13.3333, df = 4, p-value = 0.0098

> crosstabs(skips~Solder+Mask)
Call:
crosstabs(skips ~ Solder + Mask)
4977 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
Solder |Mask
       |A1.5   |A3     |A6     |B3     |B6     |RowTotl|
-------+-------+-------+-------+-------+-------+-------+
```

```
Thin   | 248   | 438   | 928   | 703   |1353   |3670   |
       |0.0676 |0.1193 |0.2529 |0.1916 |0.3687 |0.74   |
       |0.8552 |0.6854 |0.7682 |0.7285 |0.7216 |       |
       |0.0498 |0.0880 |0.1865 |0.1412 |0.2719 |       |
-------+-------+-------+-------+-------+-------+-------+
Thick  |  42   | 201   | 280   | 262   | 522   |1307   |
       |0.0321 |0.1538 |0.2142 |0.2005 |0.3994 |0.26   |
       |0.1448 |0.3146 |0.2318 |0.2715 |0.2784 |       |
       |0.0084 |0.0404 |0.0563 |0.0526 |0.1049 |       |
-------+-------+-------+-------+-------+-------+-------+
ColTotl|290    |639    |1208   |965    |1875   |4977   |
       |0.058  |0.128  |0.243  |0.194  |0.377  |       |
-------+-------+-------+-------+-------+-------+-------+
Test for independence of all factors
    Chi^2 = 38.41358 d.f.= 4 (p=9.206478e-008)
    Yates' correction not used
```