

## Homework Set 9 , given Fri 4/25, due Friday 5/2

Problem 9.1 below requires a simulation using the MATLAB program (for Problem 9.2 the program is also recommended). On Campus, Matlab is available for students in the Academic Computing Center, Engineering Library (ACCEL) in Carpenter Hall ([www.accel.cornell.edu](http://www.accel.cornell.edu)). The Matlab program runs on many computers there; it contains the statistics package and is installed under Windows (not Unix).

The present homework will appear additionally as a text file on the website so that data and code can be copied and pasted.

**Problem 9.1.** *Power of the signed rank test.* The Matlab simulation program below is for exploring the power of the signed rank test for the "half Cauchy" data treated before (cf. p. xvii Appendix, notes).

Before you can use it, you have to input the sample size  $n$  and the shift  $m$  at the command line:

```
>>n=100
>>m=0
```

The shift can then be modified to positive values. Recall that the half Cauchy density has median 0 but is heavily skewed to the right, so the signed rank test should detect it as an alternative (i.e. as nonsymmetric) even for  $m = 0$ . Here is the program "signrank1":

```
1: k=0
2: rejec=0
3: sigma=sqrt(pi/2)
4: while k<1000
5:     b=binornd(1,0.5,1,n)
6:     d=m+abs(trnd(1,1,n)).*b+abs(normrnd(0,sigma,1,n)).*(b-1)
7:     [p,h]=signrank(d)
8:     rejec=rejec+h
9:     k=k+1
10: end
```

The output "rejec" at the end gives the number of 5% (two sided) rejections in 1000 runs. We can thus explore the power of the signed rank test for various alternatives ( $m$ ) and various sample sizes ( $n$ ).

a) Estimate the power (probability of rejection) of the signed rank test using the above simulation (one program run suffices), for sample size  $n = 100$  and values  $m = 0$ ,  $m = 0.1$  and  $m = 0.2$ .

b) Repeat a) for sample size  $n = 1000$ .

c) Write a sentence interpreting the behaviour of the power under varying shift  $m$  and sample size  $n$ .

**Problem 9.2.** *Are birthdays of people uniformly distributed over the year ?* This problem is treated in the article: G. Berresford, "The uniformity assumption in the birthday problem", Math. Mag. 53 1980, no. 5, 286-288. The data used are on the website

[http://www.dartmouth.edu/~Echance/teaching\\_aids/data/birthday.txt](http://www.dartmouth.edu/~Echance/teaching_aids/data/birthday.txt)

they give the the number of births in the U.S. for each day of the year 1978. The following summary for each month is based on these data. The numbers are in thousand, rounded to one decimal.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
270.7	249.9	276.6	254.6	270.8	270.8	294.7	302.8	293.9	289.0	274.7	284.9

a) Carry out a chi square test of goodness of fit for uniform distribution of birthdays. Report the value of the chi square statistic, p-value and decision.

b) The births may be uniformly distributed over days of the year, but not over months due to the different lengths of the months. Recall that the lengths of the months in days are; 31 28 31 30 31 30 31 31 30 31 30 31. Carry out a test for uniformity over days, using the above data for months and adjusting the expected counts for the different lengths. Report the value of the chi square statistic, p-value and decision.

**Hint:** in Matlab a chi square godness of fit for uniformity, with these data, is performed by the series of commands

```
>> bins=1:12;
>>obsCounts=[270.7 249.9 276.6 254.6 270.8 270.8 294.7 302.8 293.9 289.0 274.7 284.9]*1000
>>n = sum(obsCounts);
>>expCounts=ones(1,12)*n/12;
>>[h,p,st] = chi2gof(bins,'frequency',obsCounts,'expected',expCounts)
```

Here "bins" defines indices of bins, 1, . . . ,12. Next, "obsCounts" defines our vector of observed counts; notify that the values have to be multiplied by 1000 since we have to use the actual counts (not in units of 1000). Then n gives the total sum of births, and "expCounts" defines the vector of expected counts for uniformity, using the command "ones(1,12)" to obtain the vector [1 . . . 1 ] and multiply it with n/12. Finally, "[h,p,st] = " carries out the test, where output "h" is the decision at 5% significance, output "p" is the P-value and output "st" gives as first item chi2stat=, the value of the chi square statistic. Ignore the remaining output.

For b) modify "expCounts" appropriately (see "obsCounts" how to input vectors directly).