

## Math 350 - Homework 11 - Solutions

1. According to the Mendelian theory of genetics, a certain garden pea plant should produce white, pink, or red flowers, with respective probabilities  $\frac{1}{4}$ ,  $\frac{1}{2}$ ,  $\frac{1}{4}$ . To test this theory a sample of 564 peas was studied with the result that 141 produced white, 291 produced pink, and 132 produced red flowers. Approximate the  $p$ -value of this data set

(a) by using the chi-square approximation, and

(b) by using a simulation.

Let the probabilities corresponding to the null-hypothesis be  $p_w = 1/4, p_p = 1/2, p_r = 1/4$ . Let  $n = 564$  and represent the data by the numbers:  $n_w = 141, n_p = 291, n_r = 132$ . Then the test statistic  $T$  has the value

$$\begin{aligned} t &= \frac{(n_w - np_w)^2}{np_w} + \frac{(n_p - np_p)^2}{np_p} + \frac{(n_r - np_r)^2}{np_r} \\ &= \frac{(141 - \frac{564}{4})^2}{\frac{564}{4}} + \frac{(291 - \frac{564}{2})^2}{\frac{564}{2}} + \frac{(132 - \frac{564}{4})^2}{\frac{564}{4}} \\ &= 0.8617. \end{aligned}$$

The  $p$ -value we wish to find is the quantity

$$p\text{-value} = P_{H_0}(T \geq 0.8617).$$

(a) The quantity  $P_{H_0}(T \leq 0.8617)$  is approximately the value at 0.8617 of the cumulative distribution function of a  $\chi^2$ -random variable with 2 degrees of freedom. In Matlab, this can be obtained using the command

$$\text{chi2cdf}(0.8617, 2)$$

which gave the value 0.35. Thus

$$p\text{-value} = 1 - P_{H_0}(T \leq 0.8617) = 1 - 0.35 = 0.65.$$

This value is large enough that, in the present situation, we may regard it as being consistent with the null-hypothesis.

(b) The  $p$ -value can also be obtained by simulating the following experiment: Let  $X_1, \dots, X_n$ , for  $n = 564$ , be i.i.d. random variables taking on the possible values  $\{1, 2, 3\}$  (representing, respectively,  $w, p$ , and  $r$ ), with probabilities  $1/4, 1/2, 1/4$ . Let  $N_1, N_2, N_3$  be, respectively, the number of occurrences of 1, 2, 3 among the  $X_i$ . On each run of the experiment, we calculate the test statistic

$$T = \frac{(N_1 - \frac{564}{4})^2}{\frac{564}{4}} + \frac{(N_2 - \frac{564}{2})^2}{\frac{564}{2}} + \frac{(N_3 - \frac{564}{4})^2}{\frac{564}{4}},$$

then over many runs we calculate the relative number of times that  $T \geq 0.8617$ . By the law of large numbers, this should approximate the  $p$ -value we want.

The following Matlab script does this. I have chosen  $r = 1000$  for the number of times the experiment is repeated. One run of the below script gave the approximate  $p$ -value: 0.6420.

```
r=1000; %number of runs of the experiment
n=564;
t=0.8617;
k=0; %k counts the number of times T is greater than t.
for i=1:r
    X = floor(4*rand(1,n));
    %By dividing the interval [0,1] into 4 equal length intervals,
    %and then associating the first to 1 (w), the second and third
    %to 2 (p), and the fourth to 3 (r), then 1, 2, 3 will
    %have the required probabilities 1/4, 1/2, 1/4.
    N1 = sum(X==0);
    N2 = sum(X==1 | X==2);
    N3 = sum(X==3);
    T = (N1-n/4)^2/(n/4)+(N2-n/2)^2/(n/2)+(N3-n/4)^2/(n/4);
    k = k + (T>=t);
end
p_value=k/r
```

2. Approximate the  $p$ -value of the hypothesis that the following 10 values are random numbers:

0.12, 0.18, 0.06, 0.33, 0.72, 0.83, 0.36, 0.27, 0.77, 0.74.

Let us first put the given data values in increasing order. (In Matlab, we can order the entries of a vector  $v = [y_1, \dots, y_m]$  by using the command `sort(v)`.)

0.06, 0.12, 0.18, 0.27, 0.33, 0.36, 0.72, 0.74, 0.77, 0.83.

We refer to these ordered values by  $y_1, y_2, \dots, y_{10}$ .

The first step is to obtain the value of the Kolmogorov-Smirnov test statistic,  $D$ , for the data. Since the hypothetical distribution is uniform over  $(0, 1)$ , the cumulative distribution function is  $F(x) = x$ . By formula (9.4) on page 224,  $D$  is

$$\max \left\{ \frac{j}{10} - y_j, y_j - \frac{j-1}{10} \text{ for } j = 1, \dots, 10 \right\}.$$

A possible way to compute this  $D$  in Matlab is as follows:

```
y=sort([.12 .18 .06 .33 .72 .83 .36 .27 .77 .74]);
D=max([(1:10)/10 - y, y - (0:9)/10])
D=0.2400
```

We can now proceed to obtain the  $p$ -value  $= P_F\{D \geq 0.24\}$  for this data by simulation. Due to the proposition on page 225, this can be done as follows: We repeat many times the above script that calculates the value of  $D$  but now using simulated data consisting of 10 random numbers uniformly distributed over  $(0, 1)$ . Let  $d_1, \dots, d_r$  be the values of  $D$  for  $r$  runs. Then by the law of large numbers, the fraction

$$\frac{\#\{i : d_i \geq 0.24\}}{r}$$

approximates  $P_F\{D \geq 0.24\}$ .

```
r=1000; %number of runs of experiment of drawing n random numbers
n=10;
d=0.24; %value of D test statistic from given data
k=0;
for i=1:r
    y=sort(rand(1,n));
    d_sim=max([(1:n)/n-y,y-(0:n-1)/n]);
    k=k+(d_sim>=d);
end
p_value=k/r
```

One run of this program produced the  $p$ -value 0.5270.

3. Approximate the  $p$ -value of the test that the following data set comes from an exponentially distributed population: 122, 133, 106, 128, 135, 126.

In this case, the parameter of the exponential distribution is not specified. We can estimate it from the data by taking the arithmetic mean:  $\hat{\theta} = 125$ . This corresponds to the exponential parameter  $\lambda = 1/125$ .

We first calculate the value of the Kolmogorov-Smirnov statistic:

```
%Obtaining the value of the Kolmogorov-Smirnov statistic:
y = sort([122 133 106 128 135 126]);
theta = sum(y)/6;
Fy = 1-exp(-y/theta);
d = max([(1:6)/6-Fy,Fy-(0:5)/6])
d =
    0.5717
```

We now use simulation to approximate the  $p$ -value  $P_F(D > d)$ , where  $F = F_{\hat{\theta}}$ .

```
r=10000; %number of runs of the simulated experiment
k=0;
for j=1:r
    ysim = -theta*log(rand(1,6));
    ysim = sort(ysim);
    msim = sum(ysim)/6;
    Fsim = 1-exp(-ysim/msim);
    dsim = max([(1:6)/6-Fsim, Fsim-(0:5)/6]);
```

```

    k    = k + (dsim>=d);
end
p_value=k/r
p_value =
    3.0000e-04

```

This is a very small value, so the null-hypothesis should be rejected.

By looking at a list of exponentially distributed random numbers with the same mean 125 we can see that the given list is very atypical (the variance is too small):

```

    31.5196  228.1835  376.1226  32.2104  774.0066  336.6557
   196.1580  435.7191  144.6801  125.1169  302.0655  40.5057
   165.5592  25.3201  53.5120  317.0821  213.3519  306.9325
    70.6846  210.0443  150.9375  48.6124  71.9132  139.5258
    30.6059  18.8831  137.5893  63.0777  35.0333  13.4612

```

4. *Generate the values of 10 independent exponential random variables each having mean 1. Then, based on the Kolmogorov-Smirnov test quantity, approximate the p-value of the test that the data do indeed come from an exponential distribution with mean 1.*

This is done by the following Matlab script.

```

n=10;
%Generate the values of n independent exponential
%random variables each having mean 1:
Y=-log(rand(1,n));
%Value of the Kolmogorov-Smirnov test statistic:
y=sort(Y);
w=1-exp(-y);
d=max([(1:n)/n-w,w-(0:n-1)/n]);
%Now we approximate the p-value by simulation:
r=1000;
k=0;
for i=1:r
    u=sort(rand(1,n));
    d_s=max([(1:n)/n-u,u-(0:n-1)/n]);
    k=k+(d_s>=d);
end
p_value=k/r
p_value =
    0.3740

```