MidTerm1 Statistics

Mohsen Nabian

7/1/2015

========== Question 1 ==========

Using R, write a program to calculate all the prime numbers less than 100. A prime number is a positive integer divisible (without remainder) only by 1 and itself. Create the program by testing each number from 1 to 100 against all integers less than it using %%. Your function should return a vector of all the primes $<$ 100. (15 pt)

```
prime<-2;
n<-100
for (i in 1:n)
{
    for (j in 2:(i-1))
    {
        if (i%%j==0)
        {
          break;   #That means it is devidable.
        }
        if (j==(i-1))

        {
          prime<-c(prime,i)

        }

    }
}
print(prime)
```

```
##  [1]  2  3  5  7 11 13 17 19 23 29 31 37 41 43 47 53 59 61 67 71 73 79 83
## [24] 89 97
```

========== Question 2 ==========

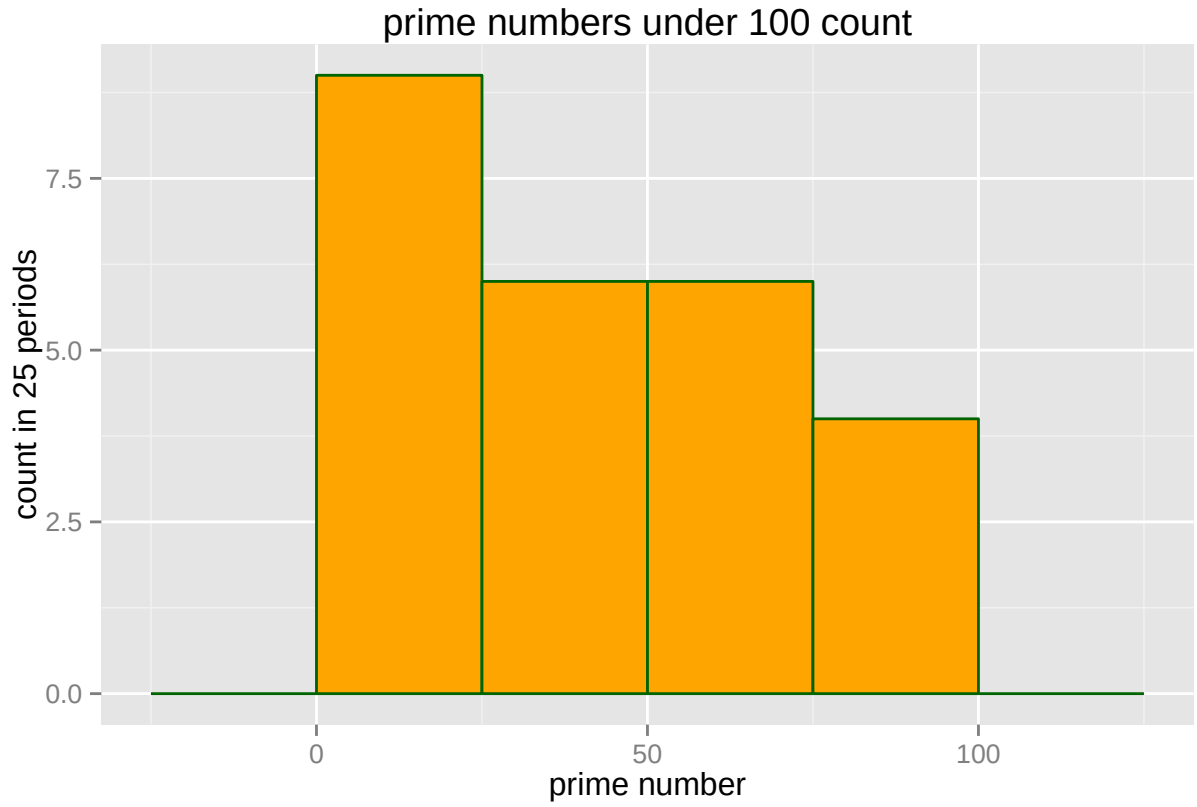
Using R, create a histogram of the result from 1 using ggplot. Be sure to nicely label your axes and title the graph. (5pt)

answer:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```
df<-data.frame(pm=prime)
plotting<-ggplot(df,aes(x=pm))
plotting+geom_histogram(binwidth=25,fill="orange", colour="darkgreen")+xlab("prime number")+ylab("count
```



========== Question 3 ==========

You flip a coin five times. a. What's the chance of getting three or more heads in a row? (5 pt)

chance of getting three or more heads in a row: Based on the queston the following occurances is our desiered:

$$[HHHTT], [THHHT], [TTHHH], [HHHHT], [THHHH], [HHHHH]$$

Total number of all occurances:

$$N = 2^5 = 32$$

So the probability of getting three or more heads in a row is:

$$\frac{6}{32} = 0.19$$

b. What's the chance of getting three or more heads in a row conditional on knowing the first flip was a heads? (5 pt)

Based on the queston the following occurances is our desiered: knowing the first is Head:

a)Desiered with 3 heads in a raw :

$$[HHHTT], [HHHTH], [HTHHH]$$

b)Desiered with 4 heads in a raw:

$$[HHHHT]$$

c)Desiered with 5 heads in a raw:

$$[HHHHH]$$

Total number of all occurances:

$$N = 2^5 = 32$$

So the probability of getting three or more heads in a row knowing the first flip is Head is:

$$\frac{5}{32} = 0.16$$

========== Question 4 ==========

NASA has declared that the Earth is likely to be hit by an asteroid this year based on an astronomical observation it has made. These things are hard to judge for certain, but it is known that the test NASA used is pretty good - it has an accuracy (sensitivity) of 99% and a false positive rate of only 1%. It is further known that the general probability of an asteroid hitting earth in any given year is 1 in 100,000. What is the probability we will actually be hit by an asteroid this year given NASA's test? (10 pt)

P(H):Probability of Earth being hit by an asteroid.

$$P(H) = \frac{1}{100000} = 0.00001$$

P(+):Probability of positive experiment result.

$$P(+|H) = 0.99$$

$$P(+|NH) = 0.01$$

$$P(+) = P(+|H)P(H) + P(+|NH)P(NH) = (0.99)(0.00001) + (0.01)(1 - 0.00001) = 0.010$$

$$P(H|+) = \frac{P(+|H)P(H)}{P(+)} = \frac{(0.99)(0.00001)}{0.010} = 0.00099$$

Which shows that the porobability is below 0.1 percent. That is why it is not gotten serious!

========== Question 5 ==========

The average number of snow days in Boston in a winter month is 1. Assuming these events follow a poisson distribution, calculate (using R) the probability of getting 5 or more snow days in a month. (5 pt)

```
snow_prob<-0;
for (i in 5:30)
{
  snow_prob<-snow_prob+dpois(x=i,lambda=1)
}
print(snow_prob)
```

```
## [1] 0.003659847
```

So the probability of having snow 5 or more days a month is 0.0037. Which is very low.

========== Question 6 ==========

You want to know how many hours of sleep the average college student gets. You start out with a preliminary survey of 10 people, and get the following data (in hours): 7,6,5,8,6,6,4,5,8,7. You hypothesize that despite what the doctors say, the average college student does not get 7 hours of sleep a night. What does your survey say? State your null hypothesis, research hypothesis (two tailed), and calculate your threshold value, test statistic, and p value. Do you reject the null or not? (10 pt)

H0: the average college student does get 7 hours of sleepa night. Ha: the average college student does NOT get 7 hours of sleepa night.

$$n = 10$$

$$\bar{x} = \frac{7+6+5+8+6+6+4+5+8+7}{10} = 6.2$$

$$smpleSD = \sqrt{((1/(n-1)) \times (((7-\bar{x})^2 + (6-\bar{x})^2 + (5-\bar{x})^2 + (8-\bar{x})^2 + (6-\bar{x})^2 + (6-\bar{x})^2 + (4-\bar{x})^2 + (5-\bar{x})^2 + (8-\bar{x}$$

$$\mu = 7$$

$$se = sd/\sqrt{n} = 1.317/3.16 = 0.417$$

$$df = n - 1 = 9$$

$$T_{statistics} = \frac{\bar{x} - \mu}{se}$$

$$T_{statistics} = \frac{6.2 - 7}{0.417} = -1.92$$

next step is to calculate thresholds for 95 percent:

```
T<-qt(0.975,9)
print(T)
```

```
## [1] 2.262157
```

$$Threshold = \pm 2.26$$

Since T-Statistics does not exceed the threshold, we can not reject the null hypothesis. In fact may be null hypotheis is true and the True Mean is 7.

========== Question 7 ==========

Despite the disappointing results in 6, you are confident in your hypothesis. Assuming your sample standard deviation and mean do not change and you want to survey as few people as possible, how many additional people would you have to survey to reject the null at the 0.05 level? (5 pt)

$$T_{statistics} = \frac{\sqrt{n} \times (\bar{x} - \mu)}{sd}$$

$$Thresholds(p_{value} = 0.95) = \pm T_{distribution}(0.975, n-1)$$

By increseing n we increase T_{statistics} to exceed the thresholds to reject the Null-Hypothesis. Knowing that increasing n would change thresholds very slightly, we may approximate n as follows and then check if our assumption works.In fact we need to do itteration over n: n_{old}=10

$$n = (\frac{T_{distribution}(0.975, n_{old} - 1) \times sd}{\bar{x} - \mu})^2$$

$$n = (\frac{2.26 \times 1.317}{6.2 - 7})^2 = 13.84$$

So n=14 would be a good choice. Now we substitute df=n-1=13 into the T_distribution.

$$n = (\frac{T_{distribution}(0.975, n-1) \times sd}{\bar{x} - \mu})^2$$

```r
T<-qt(0.975,13)
print(T)
```

```
## [1] 2.160369
```

$$n = (\frac{2.16 \times 1.317}{6.2 - 7})^2 = 12.64$$

If we do the same procedure with n=13 we will end up again to n=12.76, that means the itteration converged to n=13. As a result, n=13 could be a minimum sufficient amount of sampling if the mean and standard deviation stay the same as now. Test: if n=13:

$$T_{Statistics} = -2.19$$

$$Threshold = \pm 2.17$$

So n=13 is correct.

========== Question 8 ==========

You survey the same 10 people during finals period, and get the following hours: 5,4,5,7,5,4,5,4,6,5 . Do people get significantly less sleep during finals? (10 pt)

SO we have two dependent sample and the null hypothesis is

$$H_0 : \mu_1 = \mu_2$$

and

$$H_0 : \mu_1 \neq \mu_2$$

$$x_{difference} = x_1 - x_2$$

so the new sample is the following: x={2,2,0,1,1,2,-1,1,2,2}

$$n = 10$$

$$\bar{x} = 1.2$$

$$sd = 1.03$$

$$\mu = 0$$

$$se = \frac{sd}{\sqrt{n}} = \frac{1.03}{3.16} = 0.33$$

$$T_{statistics} = \frac{\bar{x} - \mu}{se}$$

$$T_{statistics} = \frac{1.2}{0.33} = 3.64$$

$$Thresholds(p_{value} = 0.95) = \pm T_{distribution}(0.975, 9) = 2.26$$

```r
qt(0.975,9)
```

```
## [1] 2.262157
```

Since T_statsitics 3.64 exceeds the threshold value 2.26 the null hypothesis is rejected with Type I error=0.05. So we can say with 95% certainity that students have different sleeping time near finals than regular time.

========== Question 9 ==========

You are a very bad gardener, and hypothesize that feeding houseplants vodka might help them relax and grow better. You perform an experiment to test your hypothesis, giving 15 houseplants water spiked with vodka, and 15 houseplants water alone. These are your results: condition live die treatment 4 11 control 8 7 This looks pretty bad for the treatment, but being as good at statistics as you are bad at gardening, you test it using the chi-square test. What are your results? (15 pt)

ANS_9: If we have two classifications each spans the whole population, in order to prove the independency of classifications, we need to have a table of populations like above as well as doing the Chi-Square analysis. Here are the hypothesis: H0: the age and political view are INDEPENDENT. H1: the age and political view are NOT totally independent.

Assuming null hypothesis, we can find expected value for each elemet of the sample table. Lets calculate the number of participants in our sample.

$$n = (4 + 11) + (8 + 7) = 30$$

Finding expected values:

$$fe = \frac{(rowtotal) \times (columntotal)}{overalltotal}$$

so here are the expected values table assuming independecy:

++condition+ live die
Treatment 6 9
Control 6 9

Calculating

$$X^2 = \sum \frac{(fo - fe)^2}{fe}$$

Thus

$$X^2 = \frac{(6-4)^2}{6} + \frac{(6-8)^2}{6} + \frac{(9-11)^2}{9} + \frac{(9-7)^2}{9}$$

$$x^2 = 0.67 + 0.67 + 0.44 + 0.44 = 2.22$$

Degree of freedom

$$df = (r-1)(c-1) = 1 \times 1 = 1$$

Now lets calcuate the thresholds for p-value=0.95

```
qchisq(0.95, 1)
```

```
## [1] 3.841459
```

Since xˆ2=2.22 value does NOT exceed the threshold=3.84, we are unable to reject the Null Hypothesis. As a result, our null hypothesis (Independency) could be true. In fact we can not say there is a relation between watering alchohol or pure water and death of the plants in this sample experiment.

========== Question 10 ==========

Perhaps you got things backwards, and plants need more stimulation to thrive. So you adjust your experiment into three treatment groups: water, vodka, and coffee. These are your results: condition mean days alive sd n water 50 10 20 vodka 45 7 10 coffee 55 4 10

The overall mean is 50 days (as we said, you're a bad gardener). Use an F test to determine if there is any significant difference among these three groups. (15 pt)

Here is the summary of the F-Test analysis

N total samples in G different groups of one same parameter x :

$$(\bar{x_1}, s_1, n_1), (\bar{x_1}, s_1, n_1), ..., (\bar{x_g}, s_g, n_g)$$

Null Hypothesis:

$$H_0 = \mu_1 = \mu_2 = ... = \mu_g = \bar{x}(average\, of\, all\, samples)$$

(That means that the paramter x in all those differnt group are the same, no dependency in x to any group.)

$$H_a = at\, least\, one\, group\, has\, different\, response\, to\, x.$$

(That means There could be some dependency between the parameter and anyone of the goups.)

$$F - statistic = \frac{average\, variance\, between\, groups}{average\, variance\, within\, groups}$$

N=total number of all data G=the number of groups

$$average\, variance\, between\, groups = \frac{n_1(\bar{x_1} - \bar{x})^2 + ... + n_G(\bar{x_G} - \bar{x})^2}{G - 1}$$

$$average\, variance\, within\, groups = \frac{(n_1 - 1) \times (s_1^2) + ... + (n_G - 1) \times (s_G^2)}{N - G}$$

degree of freedoms:

$$df1 = G - 1$$

$$df2 = N - G$$

water(mean=50,s=10,n=20);

Independent(mean=45,s=7,n=10);

Republicans(mean=55,s=4,n=10);

Overal Mean=(20*50+10*45+10*55)/40=(1000+450+550)/40=50;

N=40;

G=3;

$$average\, variance\, between\, groups = \frac{20 \times (50 - 50)^2 + 10 \times (45 - 50)^2 + 10 \times (55 - 50)^2}{3 - 1} = 250$$

$$average\, variance\, within\, groups = \frac{(20 - 1) \times (10^2) + (10 - 1) \times (7^2) + (10 - 1) \times (4^2)}{40 - 3} = 67.16$$

$$F - statistic = \frac{250}{67.16} = 3.72$$

$$df1 = 3 - 1 = 2$$

$$df2 = 40 - 3 = 37$$

```r
qf(0.95,2,37)
```

```
## [1] 3.251924
```

Since F-statistic=3.72 exceeds the threshold=3.25, we reject the Null Hypothesis with error 0.05. As a result,we accept the alternative hypothesis which is the fact that there is some relation between what we water the plants and plants living period.