Taming Inferential Statistics

Let's perform some inferential statistics on data contained in a .csv file. The first row of the .csv file contains variable names for each column; the remaining rows contain the data which may be numeric or text. The first 13 rows of the SampleData.csv are shown below:

STUDENT	SEX	STATE	EYES	SHOESIZE	HEIGHT	SIBLINGS	PARTY
1	Female	Connecticut	Blue	10	67	1	Independent
2	Male	New Jersey	Blue	10.5	74	2	Democrat
3	Male	New Jersey	Green	11.5	72	0	Libertarian
4	Male	New Jersey	Blue	12	72	2	Democrat
5	Male	New York	Blue	10.5	69	4	Republican
6	Male	New Jersey	Green	13	74	1	Libertarian
7	Female	New Jersey	Brown	7	62	2	Democrat
8	Male	Pennsylvania	Brown	11	70	1	Democrat
9	Male	Pennsylvania	Blue	10.5	70	1	Democrat
10	Male		Brown	11.5	73	4	Republican
11	Female	New York	Brown	8	65	4	Independent
12	Female	New Jersey	Brown	8.5	67	3	Republican
13	Female	Pennsylvania	Green	8	69	2	Independent

We will first import the data into TamingStatistics :

```
D←import '[Pathname]\SampleData.csv'
```

If you're using MiServer you can import this data or any other data from menu.

(to start go to apl.jerrymbrennan.com and choose Practice Using Live Apl)

The result is a namespace containing variables corresponding to the columns of the spreadsheet.

```
D
#.Util.[Namespace]
```

To reference a variable in the database we prefix it with the namespace. Thus to obtain statistics about a particular field, we can enter for example:

```
      mean D.HEIGHT
69.611
     mode D.STATE
New Jersey
```

Confidence Intervals

The operator `confidenceInterval` generates a confidence interval for the sample mean, proportion, variance or standard deviation. It may be abbreviated to `confInt.`

```
      To obtain a 95% confidence interval for the mean height:

      mean confidenceInterval D.HEIGHT
68.187 71.036
```

To obtain a 90% confidence interval for the proportion of students from New York one must provide the confidence interval as the left argument:

```
    .90 proportion confidenceInterval D.STATE in 'New York'
0.062221 0.30815
```

To obtain a 95% confidence interval for the standard deviation of student heights:

```
      0.95 sdev confInt D.HEIGHT
2.8359 4.9351
```

Hypothesis Testing

Hypothesis testing is a little more complex. Suppose we wish to test whether the average height of students is 68 inches. We could create a "hypothesis" object (namespace) and obtain the p-value:

```
     H←D.HEIGHT mean hypothesis = 68
     H.pValue
0.028163
```

Or we could obtain a full report using a significance level of .01:

```
     .01 report H
```

H_0: μ=68	H_1: $\mu\neq$68
Test Statistic: t=2.3247	P-Value: p=0.028163
Critical Value: t(α/2;df=26)=2.7787	Significance Level: α=0.01

Conclusion: Fail to reject H_0

Other individual items can be extracted from the namespace H:

```
   H.TestStatistic
2.3247
   H.DegreesOfFreedom
26
```

We can also perform hypothesis testing on proportions. Note that if the left argument for report is omitted the significance level is 0.05:

```
report (D.STATE in 'New York') proportion hypothesis < .2
```

H_0: $p \geq 0.2$	H_1: $p < 0.2$
Test Statistic: $Z = 0.19245$	P-Value: $p = 0.42369$
Critical Value: $Z(\alpha) = 1.6449$	Significance Level: $\alpha = 0.05$

Conclusion: Fail to reject H_0

A hypothesis test on the variance (or standard deviation) can also be performed using the hypothesis operator:

```
report D.HEIGHT variance hypothesis ≤ 5
```

H_0: $\sigma^2 \leq 5$	H_1: $\sigma^2 > 5$
Test Statistic: $\chi^2 = 67.433$	P-Value: $p = 0.000015567$
Critical Value: $\chi^2(\alpha; df=26) = 38.885$	Significance Level: $\alpha = 0.05$

Conclusion: Reject H_0

Two-Sample Tests

To perform a two-sample test, we simply substitute a vector containing the sample data for the second population for the hypothesized value:

```
MALE_HEIGHT←D.(SEX eq'Male')/D.HEIGHT
FEMALE_HEIGHT←D.(SEX eq 'Female')/D.HEIGHT
report  MALE_HEIGHT mean hypothesis > FEMALE_HEIGHT
```

H_0: $\mu_1 \leq \mu_2$	H_1: $\mu_1 > \mu_2$
Test Statistic: $t = 4.4514$	P-Value: $p = 0.00032644$
Critical Value: $t(\alpha; df=13) = 1.7709$	Significance Level: $\alpha = 0.05$

Conclusion: Reject H_0

GoodnessOfFit Tests

The `goodnessOfFit` operator determines if data represent a sample from a particular distribution. To see if heights of students are normally distributed, use the following example:

```
   report normal goodnessOfFit D.HEIGHT
  FROM      TO  OBS      EXP           ERR       CHISQ
62.409  64.209    4   1.8038   2.1962E0     2.674
64.209  66.01     1   2.4799  -1.4799E0     0.88314
66.01   67.811    2   4.0468  -2.0468E0     1.0353
67.811  69.611    5   5.1695  -1.6949E-1    0.0055572
69.611  71.412    6   5.1695   8.3052E-1    0.13343
71.412  73.212    5   4.0468   9.5318E-1    0.22451
73.212  75.013    3   2.4799   5.2010E-1    0.10908
75.013  76.813    1   1.8038  -8.0379E-1    0.35818
         Total   27  27        4.4409E-16   5.4231
```

H_0: Normal	H_1: not Normal
Test Statistic: $\chi^2=5.4231$	P-Value: p=0.60847
Critical Value: $\chi^2(\alpha;df=7\ )=14.067$	Significance Level: $\alpha=0.05$

Conclusion: Fail to reject H_0

In the discrete case, lets see if states are equally represented in the student body:

```
   report uniform goodnessOfFit D.STATE
VALUE          OBS   EXP   ERR      CHISQ
Connecticut      2   4.5  -2.5   1.3889
New Jersey      11   4.5   6.5   9.3889
New York         5   4.5   0.5   0.055556
Pennsylvania     7   4.5   2.5   1.3889
Θ                1   4.5  -3.5   2.7222
Maryland         1   4.5  -3.5   2.7222
Total           27  27     0    17.667
```

H_0: Uniform	H_1: not Uniform
Test Statistic: $\chi^2=17.667$	P-Value: p=0.0033945
Critical Value: $\chi^2(\alpha;df=5\ )=11.070$	Significance Level: $\alpha=0.05$

```
Conclusion: Reject H₀
```

We can perform a multinomial goodness-of-fit test as well:

```
      XX←'Democrat,Republican,Independent,Libertarian,Other'
      PP←0.4,0.3,0.1,0.1,0.1
      H←XX PP multinomial goodnessOfFit D.PARTY
      report H
VALUE         OBS   EXP          ERR      CHISQ
Democrat        8  10.8  ¯2.8000E0      0.72593
Republican      7   8.1  ¯1.1000E0      0.14938
Independent     8   2.7   5.3000E0     10.404
Libertarian     2   2.7  ¯7.0000E¯1     0.18148
Other           2   2.7  ¯7.0000E¯1     0.18148
Total          27  27    ¯8.8818E¯16   11.642
```

H_0: Multinomial	H_1: not Multinomial
Test Statistic: $\chi^2 = 11.642$	P-Value: $p = 0.020222$
Critical Value: $\chi^2(\alpha; df = 4) = 9.488$	Significance Level: $\alpha = 0.05$

```
Conclusion: Reject H₀
```

TESTS OF INDEPENDENCE

We perform a test of independence to determine if there is a relationship between political party and the sex of an individual:

```
     report  D.SEX  independent D.PARTY
CATEGORY1  CATEGORY2    OBS       EXP          ERR       CHISQ
Female     Democrat       2   2.3704   ¯3.7037E¯1   0.05787
Female     Independent    4   2.3704    1.6296E0    1.1204
Female     Libertarian    0   0.59259  ¯5.9259E¯1   0.59259
Female     Other          0   0.59259  ¯5.9259E¯1   0.59259
Female     Republican     2   2.0741   ¯7.4074E¯2   0.0026455
Male       Democrat       6   5.6296    3.7037E¯1   0.024366
Male       Independent    4   5.6296   ¯1.6296E0    0.47173
Male       Libertarian    2   1.4074    5.9259E¯1   0.24951
Male       Other          2   1.4074    5.9259E¯1   0.24951
Male       Republican     5   4.9259    7.4074E¯2   0.0011139
           Total         27  27         4.4409E¯16  3.3623
```

H_0: Independent	H_1: not Independent
Test Statistic: χ^2=3.3623	P-Value: p=0.49912
Critical Value: $\chi^2(\alpha;df=4\ \ )=9.488$	Significance Level: $\alpha=0.05$

Conclusion: Fail to reject H_0

REGRESSION

Suppose we are investigating a murder. Near the body is a bloody footprint size 10-1/2. How tall is the suspect? Let's obtain a model for estimating height from shoe size.

```
      RR←regress D.HEIGHT D.SHOESIZE
      report RR
  Source            SS  DF         MS        F          P
  Regression   254.63    1   254.63    77.131 4.1145E¯9
  Error         82.533  25     3.3013   0      0.0000E0
  Total        337.17   26     0        0      0.0000E0

         Coeff         SE          T          P
  B0     50.03       2.26   22.17137    0.00000
  B1      1.90       0.22    8.78240    0.00000
```

From this we conclude that the model is `Y=50.03 + 1.90 X.`
Applying the model to the shoesize `10.5:`

```
    RR.b
50.03 1.9
    10.5⊥⌽RR.b
70.034
```

Now we can estimate the prediction interval:

```
    R predictionInterval 10.5
66.2224 73.8465
⍝ Got VALUE ERROR for R and RR in all versions
```

One Way ANOVA

Suppose we want to see if political party affiliation has an impact on the number of children in a family:

```
      A← oneWayAnova D.SIBLINGS D.PARTY
      report A
GROUP         N  XBAR     SDEV        LCB     UCB
Independent   8   2     1.069     1.2836    2.7164
Democrat      8   1.5   0.53452   0.78363   2.2164
Libertarian   2   0.5   0.70711  ¯0.93273   1.9327
Republican    7   3     1         2.2342    3.7658
Other         2   1.5   2.1213    0.067266  2.9327
Source           SS  DF       MS       F         P
Treatments   13.963   4  3.4907    3.657 0.019795
Error        21      22  0.95455   0      0
Total        34.963  26  0         0      0
```