Spreadsheet Statistics & Regression

PURPOSE
The purpose of these examples is to help users of AS-EASY-AS familiarize themselves with some basic descriptive statistical concepts, and its regression analysis capabilities, while using built-in functions to calculate them.

It should be clearly noted that this example is NOT meant to provide the user with training in and understanding of statistics and statistical concepts. It's only meant to be an example of some of the statistical tools available in AS-EASY-AS.

Rainfall in Bright City

PART #1, STATISTICS EXAMPLE
Consider the following table of the monthly rainfall, in inches, in Bright City:

Month	J	F	M	A	M	J	J	A	S	O	N	D
Rainfall in inches	2	0	12	14	9	0	0	3	0	4	0	0

Taking a quick look at the data above, would you say the weather in Bright City is generally dry or wet? This is the type of answer that descriptive statistics can help you answer.

SPREADSHEET SETUP
Although much of how you set up a worksheet has to do with style, in many cases it's more efficient to setup the data in a columnar fashion. So, let's set up the spreadsheet with

a. The first column (A) listing the months.

b. The second (B) column listing the rainfall for each month, as shown in RS1.WKS, below.

Sheet: RS1.WKS

	A	B
1		
2	**Month**	**Rain**
3	Jan	2
4	Feb	0
5	Mar	12
6	Apr	14
7	May	9
8	Jun	0
9	Jul	0
10	Aug	3
11	Sep	0
12	Oct	4
13	Nov	0
14	Dec	0
15		

The first descriptive statistic measure we'll use is the MEDIAN. To find the median, we need to sort the rainfall from low to high. This can done using the Data Sort command, and visually determining the MEDIAN, i.e., the value with as many values higher than **it,** as lower, or better yet using the @MEDIAN function.

If we sort the data, as shown below, we can visually determine the median.

0 0 0 0 0 0 2 3 4 9 12 14

since there is an even number of measurements, the median is the average of two midpoint rainfalls,
i.e., median = (0+2)/2 = 1.0 inches.

RS2.WKS, below, shows the same result using the built-in @MEDIAN function of AS-EASY-AS. i.e., cell b16 contains the formula @MEDIAN(B3..B14).

Sheet: RS2.WKS

	A	B
1		
2	**Month**	**Rain**
3	Jan	2
4	Feb	0
5	Mar	12
6	Apr	14
7	May	9
8	Jun	0
9	Jul	0
10	Aug	3
11	Sep	0
12	Oct	4
13	Nov	0
14	Dec	0
15		
16	Median	1

Another important statistic, in the type of analysis we are looking at, is the MEAN. Analytically, the mean is the sum of all values of the observations multiplied by their frequency of occurrence. In a simplified form, however, and if no frequency tables are constructed, but each measurement is tabulated individually, as in this case, Mean = Average, which is the ***(Sum of All measurements)/(Number of measurements)***. Luckily, AS-EASY-AS has built-in formulas to handle Averages. As shown in RS3.WKS, below, the @AVG function is used to calculate the required statistic, i.e. cell B17 contains the formula @AVG(B3..B14).

Sheet: RS3.WKS

	A	B
1		
2	**Month**	**Rain**
3	Jan	2
4	Feb	0
5	Mar	12
6	Apr	14
7	May	9
8	Jun	0
9	Jul	0
10	Aug	3
11	Sep	0
12	Oct	4
13	Nov	0
14	Dec	0
15		
16	Median	1
17	Mean	3.66

As shown in RS4.WKS, basic principles may also be used to calculate the MEAN (or AVERAGE), as follows:

1. Type "Total" at cell A18.

2. In cell B18, use the function "SUM" to calculate the total rainfall of the whole year, i.e., type @SUM(B3..B14) and press ENTER.

3. In cell A19, type "Count" and press ENTER.

4. In cell B19, type @COUNT(B3..B14) and press ENTER (you could always type in 12, instead, since you already know the count for this very simple problem.)

5. Type "Average" at cell A20

6. In cell B20 type +B18/B19 and press ENTER.

The calculated result in cell B20 should be exactly the same as the contents of cell B17.

Sheet RS4.WKS

	A	B
1		
2	**Month**	**Rain**
3	Jan	2
4	Feb	0
5	Mar	12
6	Apr	14
7	May	9
8	Jun	0
9	Jul	0
10	Aug	3
11	Sep	0
12	Oct	4
13	Nov	0
14	Dec	0
15		
16	Median	1
17	Mean	3.66
18	Total	44
19	Count	12
20	Average	3.66

CONCLUSION

We started this example by asking the question if the weather in Bright City is generally Dry or Wet. The indicators we calculated so far, however, cannot answer the question, because "Generally Dry or Wet" is not a directly measurable quantity. The Median and Average values we calculated are meaningful when compared with similar values for other cities. For example, if we knew that the average rainfall in Dark City was about 7.5 inches, then, based on the statistic indicators we calculated here, we could say that Dark City is "Generally about Twice as Wet as Bright City".

PART #2, REGRESSION

To further see how the analytical capabilities of AS-EASY-AS can help in day-to-day analyses, let us consider that we have the annual average Summer temperature and average annual snowfall in Bright City for the last 10 years. And, let us assume for a moment that there is some correlation between these two averages, i.e., the Winter snowfall depends on the

average temperature for the months of June, July and August. We want to find a way to possibly predict the amount of snowfall in a given year by examining the average temperature. Please, remember that this is just an example demonstrating the analytical capabilities of the program. It is not to be taken as a valid prediction model of snowfall.

Our Data for the last 10 years is as follows (shown in worksheet form, as ST1.WKS):

	A	B	C
1	**Year**	**Summer**	**Snowfall**
2		**Temp(F)**	**(in)**
3	1	84.0	33.5
4	2	82.5	32.2
5	3	79.0	30.0
6	4	78.3	30.2
7	5	81.2	31.1
8	6	86.5	35.0
9	7	77.0	28.5
10	8	75.2	26.9
11	9	80.0	30.5
12	10	83.5	33.1
13			
14			
15			

We will use the AS-EASY-AS built-in regression analysis tools to examine the suspected relationship between the two measured quantities, and possibly come up with a model so that we can predict future snowfalls based on the Summer temperatures. Since there is only one independent variable (Summer temperatures), we'll seek a simple linear regression model.

Select Data Regress. Specify B3..B12 as the independent range, C3..C12 as the dependant range and D1 as the output, then click on OK. The worksheet shown below ST2.WKS, shown the parameters calculated by AS-EASY-AS. Let's examine them.

	A	B	C	D	E
1	Year	Summer	Snowfall	Intercept	-24.6888
2		Temp(F)	(in)	Slope	0.691139
3	1	84.0	33.5	R^2	0.982551
4	2	82.5	32.2	Sum X^2	65266.12
5	3	79.0	30.0	Sum Y^2	9725.06
6	4	78.3	30.2	Sum X*Y	25179.21
7	5	81.2	31.1	Count	10
8	6	86.5	35.0	SIGMAx	10.89360
9	7	77.0	28.5	SIGMAy	5.296000
10	8	75.2	26.9	RegErr	0.339869
11	9	80.0	30.5	SlopeErr	0.032563
12	10	83.5	33.1	Formula:	-24.6888+0.691139*@X
13					
14					

Note: Cell E12 will actually display -24.6888, but what is shown above are the actual contents of the cell.

The analysis of every single factor calculated by AS-EASY-AS is beyond the scope of this example. Detailed descriptions of these factors and terms can be found in any statistics textbook. However, what we are interested in is first, what is the prediction model calculated by AS-EASY-AS, and second, how good a predictor is it. The model is given by the formula in cell E12,

```
Y = -24.6888 + 0.691139*X

where Y = Snowfall in inches, and
      X = Average Summer temperature
```

In deciding how good the prediction model is, we take a look at the R^2 value (correlation coefficient). The closer this value is to 1.0, the better the calculated model can describe the average dependence of snowfall on average temperature. With a value of 0.982551, it looks like the model is a very good one.

In order to further validate the model, let's pick a couple of the years we analyzed, and see how well the model would have predicted the snowfall.

Year #1,
Summer Temperature (X) = 84
Snowfall (Y) = -24.6888 + 0.691139*X = -24.6888 + 0.691139*84 = 33.37
Which agrees fairly well with the measured value of 33.5 inches of snow in that year.

Year #6,
Summer Temperature (X) = 86.5
Snowfall (Y) = -24.6888 + 0.691139*X = -24.6888 + 0.691139*86.5 = 35.1
Which agrees fairly well with the measured value of 35.0 inches of snow in that year.

CONCLUSION

The prediction model calculated by the regression analysis module of AS-EASY-AS does a fairly good job of predicting snowfall based on average Summer temperatures, at least within the range of measured data.