Data analysis and distribution fitting

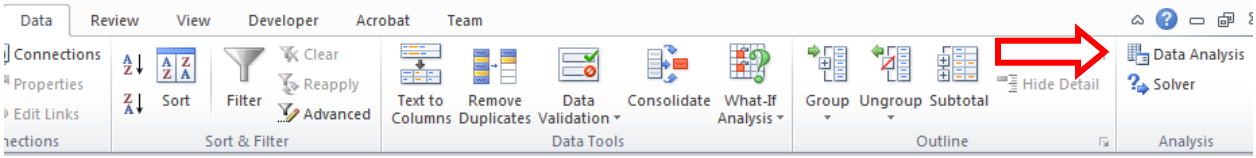
A. Using Excel

Step 1 – collect the data

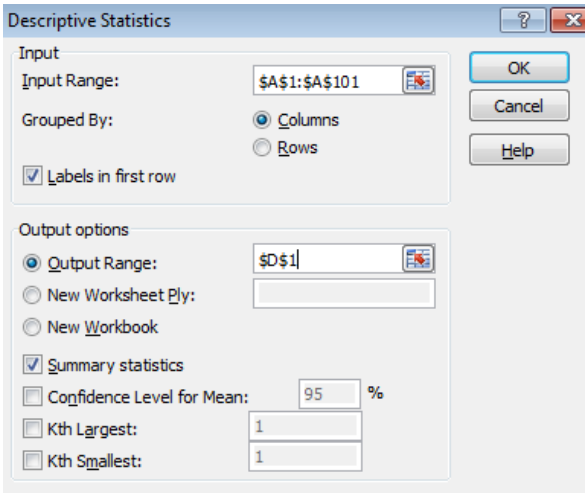
For this exercise we will assume that we have collected the service times for 100 customers. Open the file **Training data.xlsx**.

Step 2 – descriptive statistics

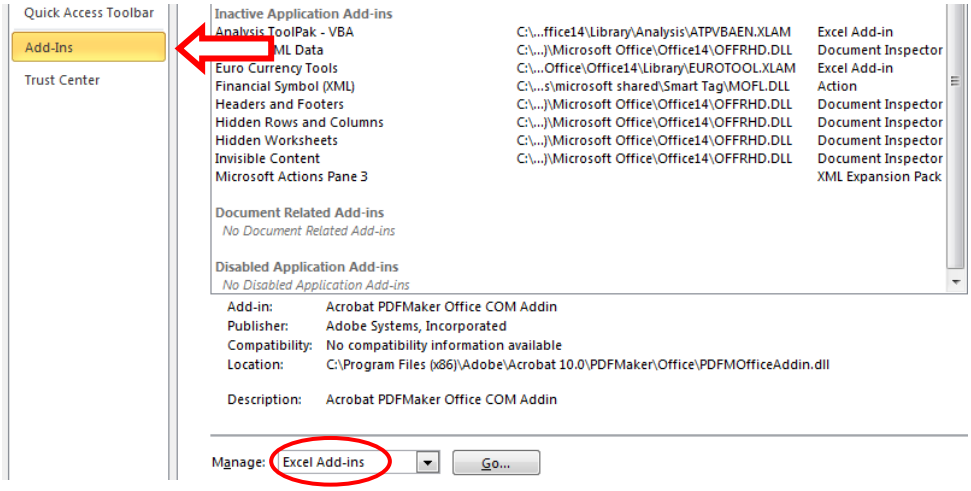
Use the **Descriptive Statistics** tool in Excel (**Data** > **Data Analysis**) to obtain the basic statistics about the data.



This should show the following dialog. Please make sure you have filled the parameters correctly and click **Ok**.



Note: if the **Data analysis** option is not available under tab **Data**, just select the **File** > **Options** > **Add-Ins**. Make sure **Excel Add-ins** is selected in the list box **Manage** and click button **Go…**(see the figure below).

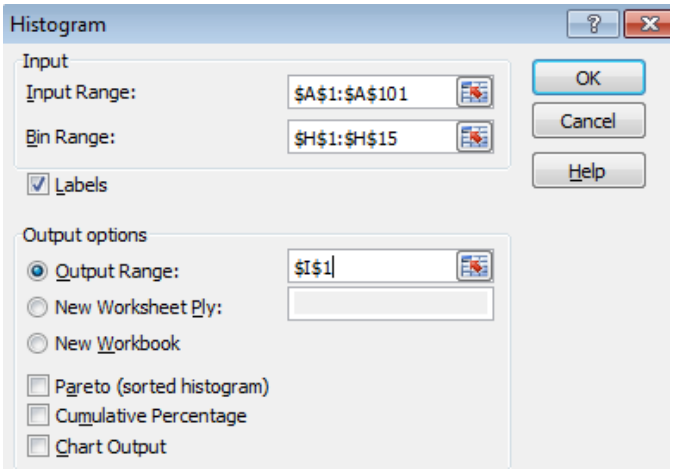


In the dialogue box, check the box for **Analysis ToolPak** and click button **OK**. The **Data Analysis** should now be available under tab **Data**.

Step 3 – histogram values

In order to draw the histogram, we need to generate the number of values in suitable intervals (choose the bin values based on the max and min values from the descriptive statistics – for this data I suggest using 1, 2, 3 …. See column H in the figure on the left below). The bin values (column H) are the maximum values in each interval (i.e. the frequencies are the number of data values between the previous bin value and the current bin value). Once you have prepared the bin data, use the **Histogram** tool from the **Data** > **Data Analysis** to activate the dialog shown on the figure on the right below. Enter the required parameters as shown.

G	H
Min	*Max (bin)*
0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
9	10
10	11
11	12
12	13
13	14



You need to reformat the result to make it looks like the figure below. Add the extra columns and enter the interval (column I, select cell type as text) and the mid-point (column J), since these will be useful in later steps.

	A	B	C	D	E	F	G	H	I	J	K
1	Data			*Data*			*Min*	*Max (bin)*	*Interval*	*Mid-point*	*Frequency*
2	6.5						0	1	0-1	0.5	4
3	8.8			Mean	4.842		1	2	1-2	1.5	10
4	4.0			Standard Error	0.265197		2	3	2-3	2.5	16
5	1.5			Median	4.4		3	4	3-4	3.5	12
6	2.4			Mode	6.5		4	5	4-5	4.5	14
7	5.3			Standard Deviat	2.651974		5	6	5-6	5.5	16
8	4.2			Sample Varianc	7.032966		6	7	6-7	6.5	10
9	2.4			Kurtosis	0.823057		7	8	7-8	7.5	6
10	3.8			Skewness	0.80609		8	9	8-9	8.5	5
11	1.3			Range	13.9		9	10	9-10	9.5	3
12	4.3			Minimum	0.1		10	11	10-11	10.5	2
13	4.0			Maximum	14		11	12	11-12	11.5	0
14	5.8			Sum	484.2		12	13	12-13	12.5	1
15	6.5			Count	100		13	14	13-14	13.5	1
16	5.9							More			0
17	4.4							Total			100
18	4.1										
19	5.8										

Step 4 – draw a histogram

To draw the histogram, you need to use the column chart option (**Insert** > **Column**, select the first 2D column). If the chart is not shown automatically, you need to select the blank frame and select **Design** > **Select Data**. Select the frequency values as the data for the chart by clicking on **add**, set the series name to $K$1 and the series values to $K$2:$K$15, click **Ok**. Next, you need to select the interval values as the x-axis (or horizontal axis) labels for the data by clicking the **edit** button on the right. Set the axis label range to $I$2:$I$15.

What are the general characteristics of the data? Bear in mind that we only have sample of 100 values and so we would expect some irregularities. Which standard distributions might apply (also think about what you would expect for a service time distribution)?

The distribution is positively skewed as would be expected for this type of data. The Erlang (Gamma) and the Lognormal are likely distributions.

Step 5 – calculating the parameters for the distributions

Erlang (Gamma)

The gamma distribution has the parameters α and β. The maximum likelihood estimators are complicated to calculate. Instead we will use the fact that mean = $\alpha\beta$ and variance = $\alpha\beta^2$. Hence we will estimate the parameter values by:

β = variance / mean
α = mean / β

Calculate these values in the spreadsheet using the mean and variance (= standard deviation ^{2}) from the data. You should get $\alpha = 3.33$ and $\beta = 1.45$.

The Erlang is a special case when α is an integer (this is usually denoted K). Using the Erlang often keeps things simpler and is sufficiently accurate. Take K as the nearest integer: hence K = 3. Calculate β using this value of K (mean = $\alpha\beta$ and so β = mean / K = 1.61).

Lognormal

Create a new column next to the data and put in the ln of each data value (Excel function LN). Apply the descriptive statistics to this data. The mean and standard deviation of the ln values are the required parameters for the lognormal.

The spreadsheet should now look like (the parameters we have calculated are in bold):

	A	B	C	D	E	F	G	H	I	J	K	L
1	Data	ln(data)		*Data*			*Min*	*Max (bin)*	*Interval*	*Mid-point*	*Frequency*	
2	6.5	1.9					0	1	0-1	0.5	4	
3	8.8	2.2		Mean	4.842		1	2	1-2	1.5	10	
4	4.0	1.4		Standard Error	0.265197		2	3	2-3	2.5	16	
5	1.5	0.4		Median	4.4		3	4	3-4	3.5	12	
6	2.4	0.9		Mode	6.5		4	5	4-5	4.5	14	
7	5.3	1.7		Standard Deviat	2.651974		5	6	5-6	5.5	16	
8	4.2	1.4		Sample Varianc	7.032966		6	7	6-7	6.5	10	
9	2.4	0.9		Kurtosis	0.823057		7	8	7-8	7.5	6	
10	3.8	1.3		Skewness	0.80609		8	9	8-9	8.5	5	
11	1.3	0.3		Range	13.9		9	10	9-10	9.5	3	
12	4.3	1.5		Minimum	0.1		10	11	10-11	10.5	2	
13	4.0	1.4		Maximum	14		11	12	11-12	11.5	0	
14	5.8	1.8		Sum	484.2		12	13	12-13	12.5	1	
15	6.5	1.9		Count	100		13	14	13-14	13.5	1	
16	5.9	1.8						More			0	
17	4.4	1.5						Total			100	
18	4.1	1.4		*ln(data)*								
19	5.8	1.8									Gamma	Erlang
20	4.2	1.4		Mean	**1.389918**					Alpha (K) =	3.334	**3.000**
21	4.4	1.5		Standard Error	0.07099					Beta =	1.452	**1.614**
22	0.7	-0.4		Median	1.481605							
23	5.8	1.8		Mode	1.871802							
24	7.1	2.0		Standard Deviat	**0.709899**							
25	4.1	1.4		Sample Varianc	0.503956							
26	6.4	1.9		Kurtosis	6.528274							
27	8.9	2.2		Skewness	-1.75797							
28	2.1	0.7		Range	4.941642							
29	6.3	1.8		Minimum	-2.30259							
30	2.9	1.1		Maximum	2.639057							
31	14.0	2.6		Sum	138.9918							
32	8.9	2.2		Count	100							
33	6.6	1.9										

Step 6 – Chi-Square Test

Preparing the data for chi-square test

To calculate the expected number of values in each of your histogram bins for the lognormal and Erlang distributions, we will use the `GAMMADIST` and `LOGNORMDIST` functions. These give the probabilities of the distributions (see the Excel help for more information).

To obtain the expected number of values in the interval 0-1:

- Enter `=100*(GAMMADIST(H2,$L$20,$L$21,TRUE))` to cell L2
- Enter `=100*(LOGNORMDIST(H2,$E$20,$E$24))` to cell M2

To obtain the expected number of values in the interval 1-2:

- Enter the following formula to cell L3
  `=100*(GAMMADIST(H3,$L$20,$L$21,TRUE)-GAMMADIST(G3,$L$20,$L$21,TRUE))`
- Enter the following formula to cell M3
  `=100*(LOGNORMDIST(H3,$E$20,$E$24)-LOGNORMDIST(G3,$E$20,$E$24))`

H3 and G3 contain the maximum and minimum values in the interval (i.e. 2 and 1). L20 and L21 contain the K and β parameters. TRUE means that `GAMMADIST` gives the cumulative probability. Hence the first `GAMMADIST` value gives the probability of a value < 2 and the second gives the probability of a value < 1. The value inside the brackets therefore gives the probability of selecting a value from the Gamma distribution of between 1 and 2. Multiplying by 100 gives the expected number of values in the interval out of a total sample of 100.
[Note: In Excel if you type `=GAMMADIST()` in the formula bar and click the equals sign in the formula bar it will prompt you for the required parameters].

The lognormal formula works in the same way (the `LOGNORMDIST` function automatically gives the cumulative probability and so doesn't need the TRUE parameter). E20 and E24 contain the mean and standard deviation of the LN data values.

Copy these formulae down for the other intervals.

Carrying out Chi-square test

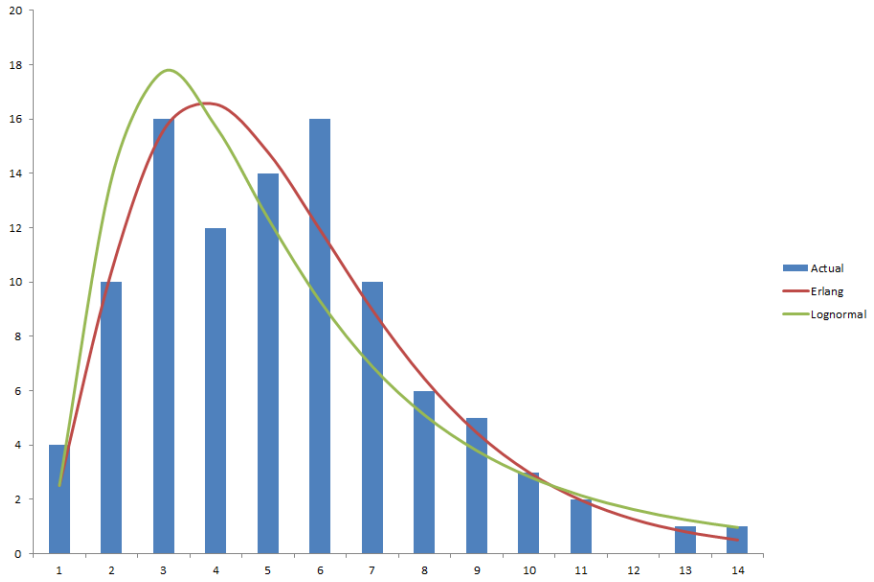
The chi-square test compares the actual and expected values and so we can use the values from step 6. As discussed in the lectures this is only a guide.

As you hopefully remember from the introductory statistics course, we need to combine any intervals with expected values of less than 5 values. This is because the chi-square statistic for each interval divides by the expected value and so categories with a small expected value can dominate the test too much.

Decide on your intervals and create columns with the actual and expected frequencies in each interval. Then use the `CHITEST` Excel function to calculate the p value (see the Excel help for more information). We have a statistically significant difference if the p value is less than our significance level (i.e. for a 5% significance level, if p < 0.05). Neither distribution should give this. In general, the larger the p value, the better the fit. Part of your spreadsheet should look like the following figure.

	G	H	I	J	K	L	M	N	O	P	Q	R
1	*Min*	*Max (bin)*	*Interval*	*Mid-point*	*Actual*	Erlang	Lognormal		Interval	Actual	Expected	Expected
2	0	1	0-1	0.5	4	2.509	2.512					
3	1	2	1-2	1.5	10	10.402	13.805		0-2	14	12.911	16.317
4	2	3	2-3	2.5	16	15.604	17.760		2-3	16	15.604	17.760
5	3	4	3-4	3.5	12	16.546	15.719		3-4	12	16.546	15.719
6	4	5	4-5	4.5	14	14.777	12.346		4-5	14	14.777	12.346
7	5	6	5-6	5.5	16	11.913	9.290		5-6	16	11.913	9.290
8	6	7	6-7	6.5	10	8.974	6.892		6-7	10	8.974	6.892
9	7	8	7-8	7.5	6	6.440	5.106		7-8	6	6.440	5.106
10	8	9	8-9	8.5	5	4.457	3.798		over 8	12	12.835	16.570
11	9	10	9-10	9.5	3	3.000	2.844		Total	100	100	100
12	10	11	10-11	10.5	2	1.974	2.147					
13	11	12	11-12	11.5	0	1.275	1.634		**Chi test p value =**		0.885	0.234
14	12	13	12-13	12.5	1	0.811	1.254					
15	13	14	13-14	13.5	1	0.510	0.970		The chi-square test at 5% significance level			
16		More			0	0.809	3.924		rejects the distribution if the p value < 0.05.			
17		Total			100	100	100					
18												
19					Gamma	Erlang						
20				Alpha (K) =	3.334	3.000						
21				Beta =	1.452	1.614						
22												

You can plot the data and the two distributions as shown in the following figure.



Which distribution do you think gives the best fit?
In your opinion, is the fit good enough?