Descriptive Statistics

Consider sample of 50 firms who were asked how many industrial disputes thay had had in the past year present results as:

No. disputes	Frequency	RF	CF	Degrees	Fx$_i$
0	14	0.28	14	100.8	0
1	8	0.16	22	57.6	8
2	12	0.24	34	86.4	24
3	6	0.12	40	43.2	18
4	3	0.06	43	21.6	12
5	4	0.08	47	28.8	20
6	2	0.04	49	14.4	12
7	1	0.02	50	7.2	7
	50	1.00	100	360	

RF= Relative frequency=$\frac{F_i}{N}$

CF=cumulative frequency=$> \frac{F_i}{N}$

Degrees = $\frac{F_i}{N} \times 360^{\text{E}}$ for pie chart

¾ Can represent in a number of ways: Excel

- Frequency
- Relative frequency
- Cumulative frequency
- Pie chart
- Histogram: bar chart where area under curve represents frequency

¾ here dealing with a discrete distribution, number of disputes 0,1,2 etc...

¾ but say we wanted to look at the amount of time lost during a dispute

¾ Then have a continuous variable and likely that each firm will have a different number of hours even if they had the same number of disputes

¾ But can group data into a range:

Hours	freq	Alternative	Alternative
0 - 2	20	0.0 - 2.99	0 - 2
2 - 5	9	3.0 - 4.99	3-5
5	10	4.0 -	6 -
lower -upper			
Total	39		

¾ Call these class intervals an how we choose them is something of an art

- might have some theoretical reason eg dispute of 8 hours is defined as a strike and if greater than 48 hours a serious strike etc
- Do not need to have equal class intervals. Often the last one will be open ended.

¾ Once data grouped inthis way can carry on as before

- often will centre on middle of class interval in drawing bar charts. Mid point of class interval =$\frac{lower+upper}{2}$
- Histogram more complex as column width not equal to 1 as in first example so have to adjust the height of the bars to reflect the different column widths.

- Consider the first 4 classes in the first example

No.	F	Write as	C	F	Then	F/C	
0	14		0-1	15		14/1	14
1	8		1-2	8		8/1	8
2	12		2-4	18		18/2	9
3	6						

¾ - We get frequency density by dividing the frequency by the interval.
  - Get a relative frequency histogram if make total area under the curve 1

¾ Note as reduce the intervals in bar chart move towards a curve: continuous

¾ There are a number of special types of frequency curves which are graphical represenatations of theoretical distributions, such as:
  - Rectangular distribution
  - Normal distribution
  - Skewed distribution
  - Bimodal distribution

¾ Useful to be able to summarise the data with summary statistics that describe characteristics of the distributions. might want to compare a number of them or make some evaluation.

¾ There are two types:
  - Measures of central tendency -where is most of the weight; what value is the distribution centred around
  - Measures of dispersion -how much is the distribution spread

¾ Measures of Central tendency:
  - Mode: most frequent value -might be more than one
  - Median: an order statistic, asks what is the middle value, the value below which half of the sample lies inerquartile range etc
  - Mean (arithmetic mean): what usually refer to as average

$$\overline{X} = \frac{> X_i}{N} \text{ or in case of discrete} = \frac{> F_i X_i}{N}$$

  - Other statistics are are range, percentiles, interquartile range etc...

¾ All are useful and have own advantages:
  - Mode: often enough just to know this, but may be more than one
  - Median: useful in categorical especially if open ended when it is better than the arithmetic mean
  - Mean ($\overline{X}$) is he most informative, useful and widely used but
    - * Care is needed: when using categorical data.
    - * Means of different data sets may be averaged to give new oevrall mean without reference to underlying data
    - * If are to compare means data sets must be comparable
    - * Problems of outliers
    - * Are different types -harmonic and geometric: need to take care

¾ Measures of Dispersion (spread)
  - Range: simply largest minus the smallest value
  - Mean/average deviation: looks at the absolute value of the deviations from the mean of the distribution
  - Variance (a^2): similar to previous, but weights the deviations by squaring them, so that the further away the more the weight. If small then the duistibution is thin, if large then the distribution is flat
  - Standard deviation: square root of the variance (a) -of use in statistical theory

¾ Dont have to work about the mean but invariably do

¾ Problem in making comparisions:
  - Ok if the same sample of firms involved but if samples differ cant compare
  - So standardise the standard deviation using the mean to geive the coefficient of variation:

$$CV = \frac{a}{\overline{X}}$$

¾ Also measure of skew: show how assymetric distribution is:
  - Normal distribution is symmetric (mode=median=mode)
  - Can have positive (more weight to left: mode<median<mean) and negative (more weight to the right:

mean<median<mode) skewed distributions

- Are some statistics that are used to summarise degree of skewness: Pearson coefficients:

$$\frac{mean\ ?\ \mathrm{mod}e}{a} \text{ and } \frac{3\text{Ý}mean\ ?\ median\text{Þ}}{a}$$

- Also kurtosis: degree of peakedness of the distribution -usually taken relative to the normal distribution