PW #4 Distributions. Normal distribution.

4.1. Distributions

The word histogram is derived from Greek: histos ‘anything set upright’; gramma ‘drawing, record, writing’. In statistics, a histogram is a graphical display of tabulated frequency. The histogram differs from a bar chart in that it is the area of the bar that denotes the value, not the height. The bars must be adjacent. Histograms allow you to explore your data by displaying the distribution of a continuous variable (percentage of sample) against categories of the value. You can obtain the shape of the distribution and whether the data are distributed symmetrically.

The world of statistics includes dozens of different distributions for categorical and numerical data; the most common ones have their own names. One of the most well-known distributions is called the **normal distribution (Gaussian distribution),** also known as the **bell-shaped curve**. The normal distribution is based on numerical data that is continuous; its possible values lie on the entire real number line. Its overall shape, when the data are organized in graph form, is a symmetric bell-shape. In other words, most (around 68%) of the data are centered around the mean (giving you the middle part of the bell), and as you move farther out on either side of the mean, you find fewer and fewer values (representing the downward sloping sides on either side of the bell).

Due to symmetry, the **mean** and the **median** lie at the **same point**, directly in the **center of the normal distribution**. The standard deviation is measured by the distance from the mean to the inflection point (where the curvature of the bell changes from concave up to concave down).

Inspecting a histogram is one of the most popular ways to understand what your data "looks like", particularly to see if it is **normally distributed**. When inspecting a histogram for normality you are looking for the classic **"bell curve"** shape that is exhibited by a normal distribution. If your data is approximately normally distributed, it should have a shape very similar to this "bell curve" shape.

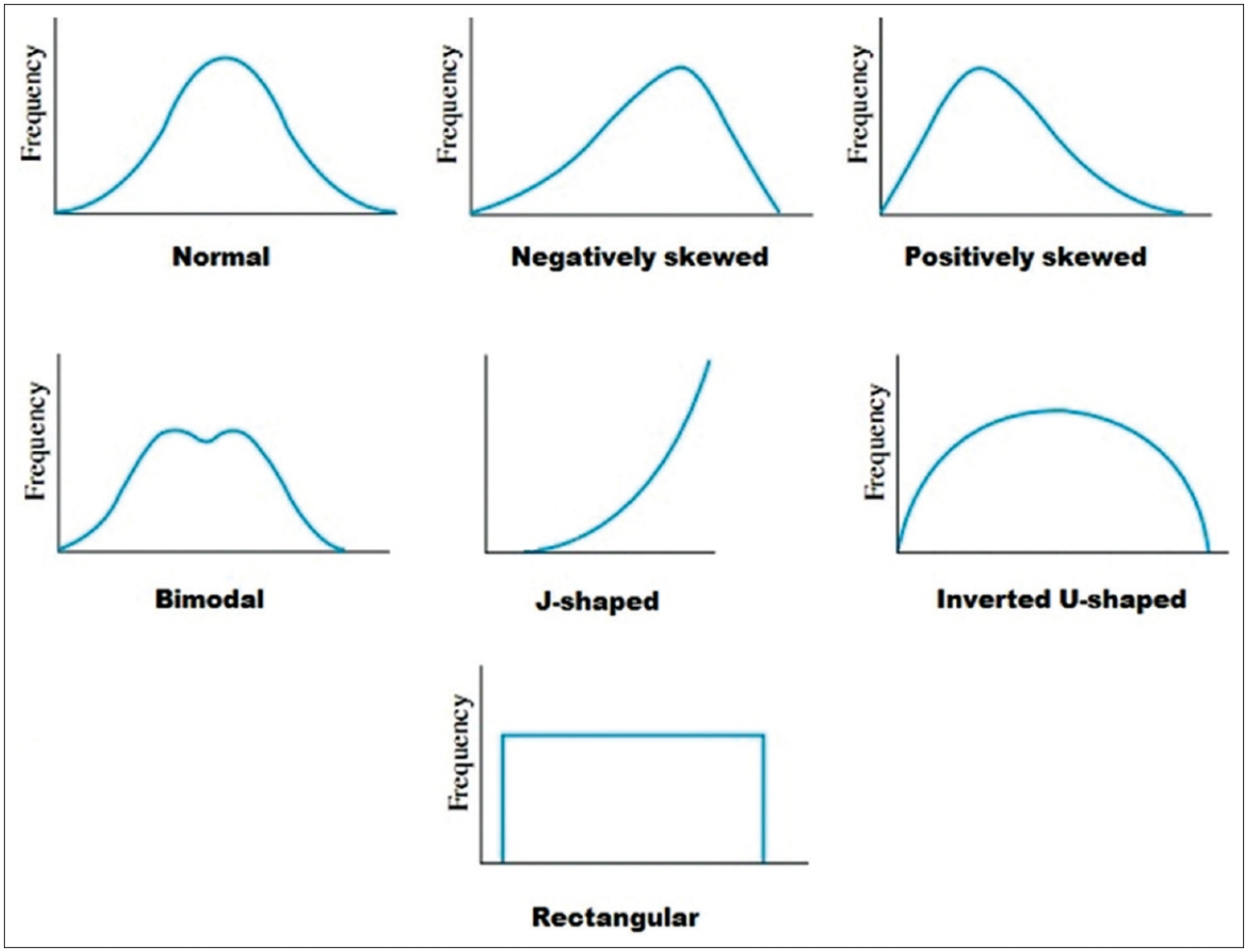
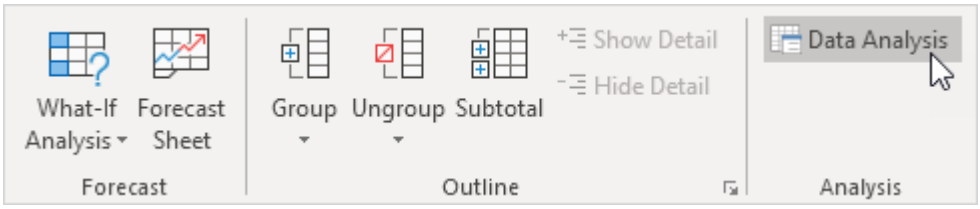


Figure 1. Examples of distribution shapes

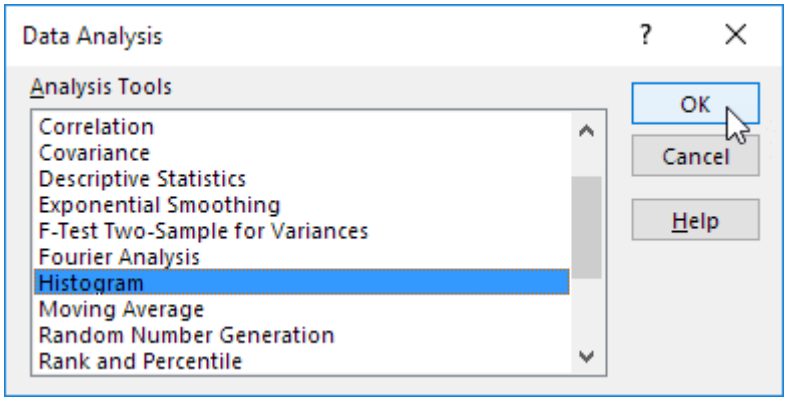
4.2. MS Excel functions for normal distribution

1. Enter MS Excel and open the file
2. On the **Data tab**, in the Analysis group, click **Data Analysis**

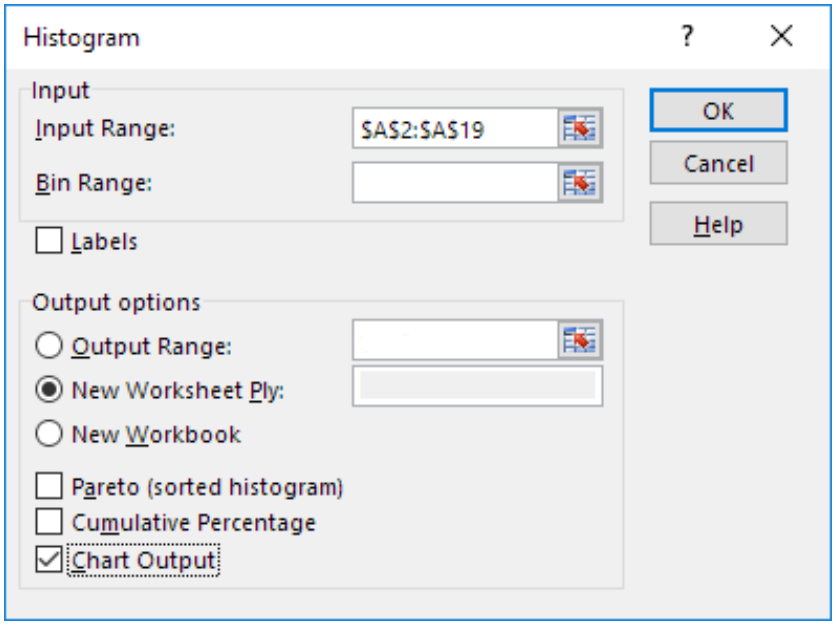


Note: can't find the Data Analysis button? Click here to load the Analysis ToolPak add-in.

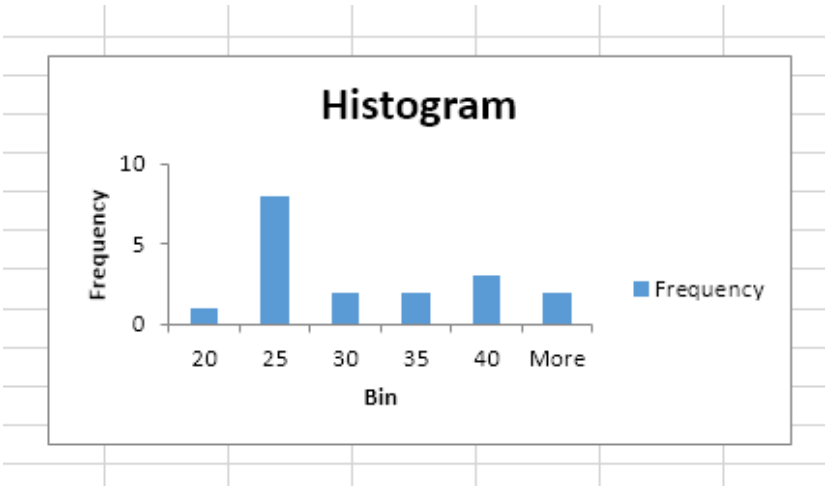
3. Select **Histogram** and click **OK**



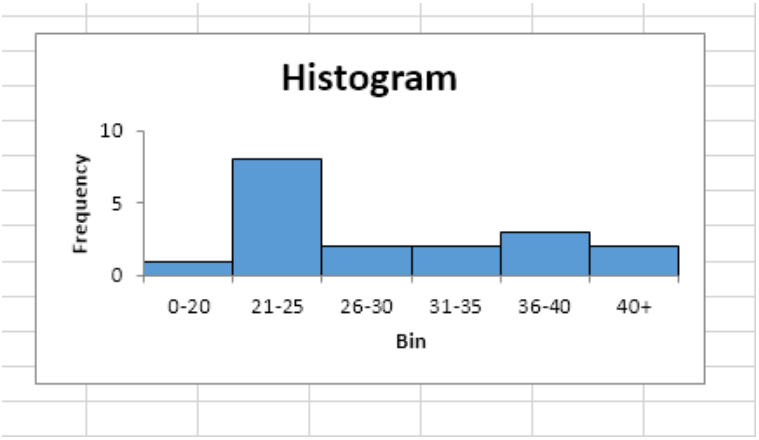
4. Select the range of the numeric variable. (if you include the first row then check **Labels**)
5. Click the Output Range option button, click in the Output Range box and select **New Worksheet**
6. Check **Chart Output** and hit **OK**



7. Click the legend on the right side and press Delete.
8. Properly label your bins.
9. To remove the space between the bars, right click a bar, click Format Data Series and change the Gap Width to 0%.



10. To add borders, right click a bar, click Format Data Series, click the Fill & Line icon, click Border and select a color.



4.3. Statistical inference

Because the population, in most instances, is not feasible to be measured directly, its characteristics can only be estimated using statistics calculated at the level of a sample extracted from the population to be analyzed.

The process of estimating the characteristics of a population and generalizing the observations from the sample level to the population level is called statistical inference.

The first type of statistical inference is the direct statistical inference. In this case, the estimation of the characteristics of a population will be done through reliable intervals. The interpretation of these obtained intervals is that "we estimate that the population parameters will be found within the range, an assertion for which we have a confidence level of x%".

Based on the properties of the normal distribution and assuming that the distribution of values within the studied sample is the correspondence of the distribution of values in the population, we can estimate:

- The range of values within which we will find a percentage of the population, centered on the mean.
- The range of values within which we will find the mean population, with the desired degree of confidence

Remember: *When choosing the degree of confidence used for the estimation of the mean, it must be taken into account that the choice of a higher degree of confidence leads to the decrease of the precision with which the estimation is made, and vice versa. For example, we can imagine the degree of confidence for estimating being a target. The larger the target, the more likely it is to be reached, including by a precision target less good, it grows. Therefore, confidence in the proposed range increases as the range size increases.*

4.4. The 68-95-99.7 rule

Taking into account the properties of the Gaussian distribution, for example we will use estimation ranges of 68%, 95% and 99.7%. However, it should be noted that estimation is possible for any confidence level chosen, using the z-score, characteristic of the normal distribution, z-score which has the equivalent for any confidence level of the estimate.

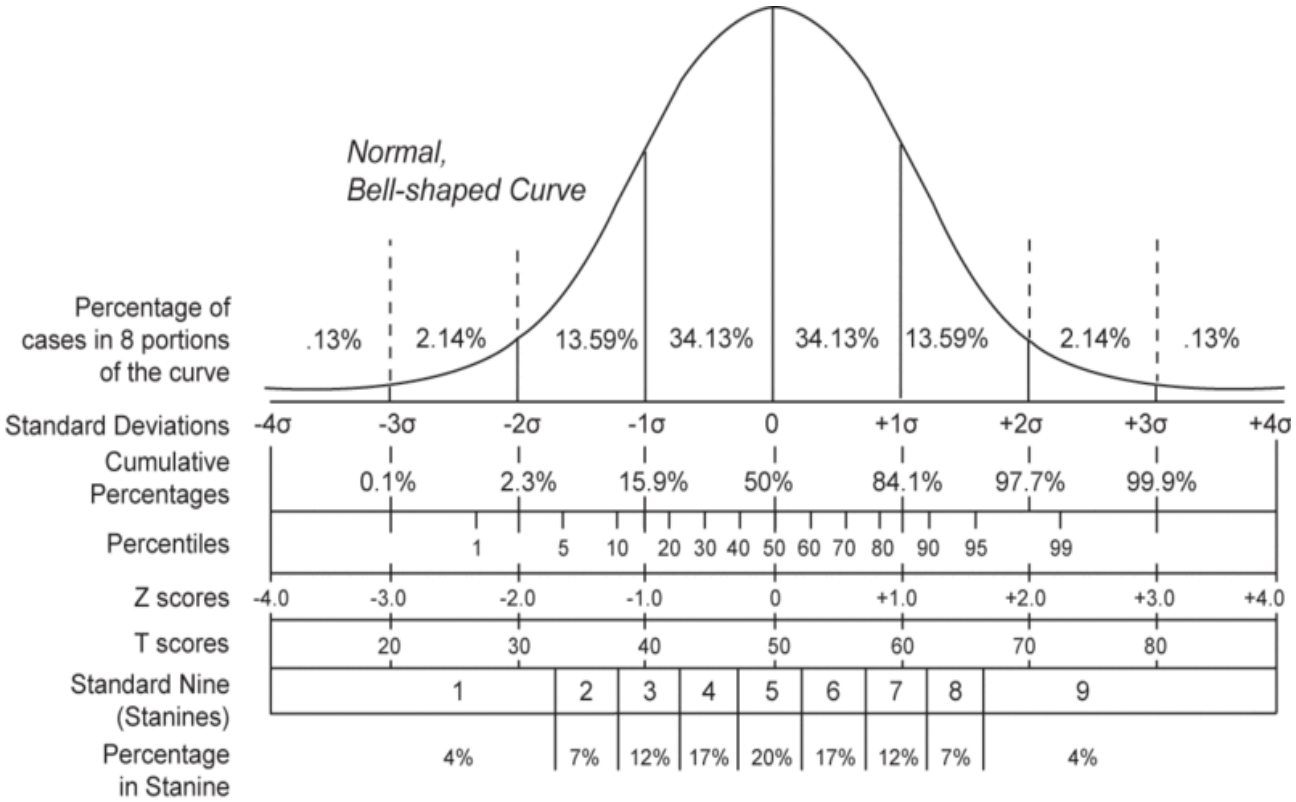


Figure 1. Normal distribution, Z & T scores

It is known that, in a normal distribution of values, in the following ranges of values will be found:

- Between [mean – standard deviation; mean + standard deviation] will be found 68% of the constituents of the population
- Between [mean – 2*standard deviation; mean + 2*standard deviation] will be found 95% of the constituents of the population
- Between [mean – 3*standard deviation; mean + 3*standard deviation] will be found 69.7% of the constituents of the population

In statistics, the 68–95–99.7 rule, also known as the empirical rule, is a shorthand used to remember the percentage of values that lie within a band around the mean in a normal distribution with a width of two, four and six standard deviations (S), respectively; more accurately, 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations of the mean, respectively.

In the empirical sciences the so-called three-sigma rule of thumb expresses a conventional heuristic that nearly all values are taken to lie within three standard deviations of the mean, and thus it is empirically useful to treat 99.7% probability as near certainty.[1] The usefulness of this heuristic depends significantly on the question under consideration. In the social sciences, a result may be considered "significant" if its confidence level is of the order of a two-sigma effect (95%), while in particle physics, there is a convention of a five-sigma effect (99.99994% confidence) being required to qualify as a discovery.

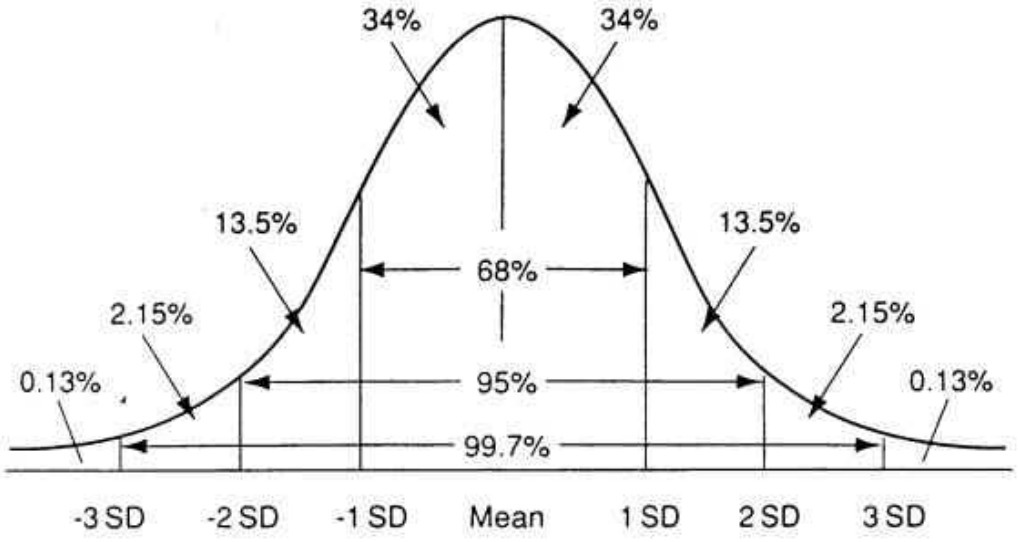


Figure 2.Normal distribution and the three-sigma rule of thumb

Example exercise

1. Using the properties of the normal distribution, estimate in what range of weights, centered on the average, we will find **68% of the components** of the population of medical students.

Solution

We apply the formula previously described at **The 68-95-99.7 rule.**
In order to find the interval in which lye 68% of population, we will use the below formula:

[mean – standard deviation; mean + standard deviation]

[81.21 – 15.13 ; 81.21 + 15.13] → [66.08 ; 96.34]

We can conclude that 68% of the population of medical students have a weight between 66.08 and 96.34kg.