### Distribution Fitting

### For any statistical analysis that requires probability calculations, it is necessary to define first which distribution fits the data best.

### For example, if we want to determine the process capability of a process, we need to define first which distribution fits the best to the process outputs data so that we can determine the probability of producing materials beyond the specifications.

### It is important to identify the distribution that accurately reflects our data. If we select the wrong distribution, our calculations will be wrong.

### What is distribution fitting?

Distribution fitting is the process used to select a statistical distribution that best fits the data.  Examples of statistical distributions include the normal, Poisson, binomial, exponential, gamma, Weibull distributions, etc.

### Not everything fits a normal distribution

Life would have been great if we could just assume that our data were normally distributed. It is the most frequently used distribution because many things are normally distributed, but not everything.  The normal distribution is defined by the average and standard deviation. It is symmetrical about the average.

Distributions that are skewed to the left or to the right cannot be modelled as normal distributions. Other distributions are bounded and cannot be modelled as normal distributions. And there are symmetrical distributions that may fit our data better than a normal distribution does. Picking the wrong distribution gives us inaccurate results.

**Identifying the best distribution**

* Multiple distributions are usually tested against the data to determine which one fits the data the best. We cannot just look at the shape of the distribution and assume it is a good fit to our data.
* There are visual techniques that help us first to identify which distributions are likely to be appropriate. The most important one is examining a histogram with the distribution overlaid and comparing the empirical model to the theoretical model.
* Statistical techniques are then used to ***estimate the parameters*** of the various distributions. Once this estimation process is complete, we use ***goodness of fit techniques*** to help determine which distribution fits our data best.

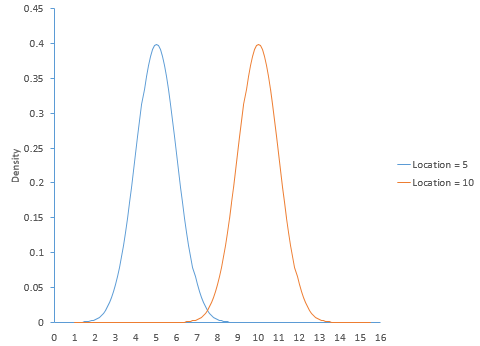
**Parameters of a Distribution**

The parameters of a distribution define the distribution. Distribution fitting involves estimating the parameters that define the various distributions.

* There are four parameters primarily used in distribution fitting.  These four parameters are:
* Location parameter
* Scale parameter
* Shape parameter
* Threshold parameter
* Not all parameters exist for each distribution.  For example, the normal distribution has only two parameters: location (the average) and scale (the standard deviation). These two parameters completely define the normal distribution.

**Location parameter**

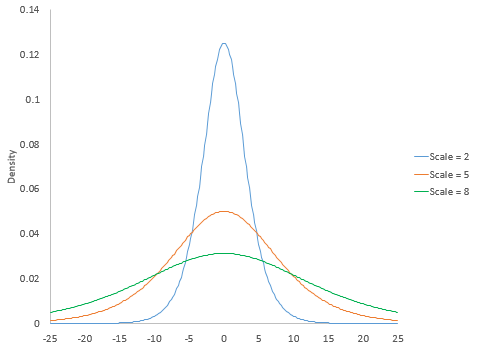
* The location parameter of a distribution indicates ***where the distribution lies along the X-axis*** (the horizontal axis).
* Fig. 1 shows two ***normal distributions*** with two different location parameters: 5 and 10.
* Both have the same standard deviation (or scale in parameter terms).
* The normal distribution does not have a shape parameter. The shape of normal distribution is defined by the values of location and scale parameters.



**Fig.1**: Normal distribution with different locations

**Scale parameter**

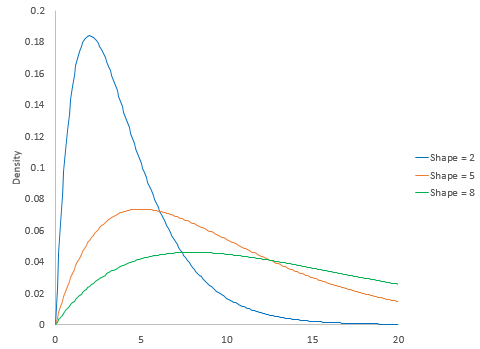
* The scale parameter of a distribution ***determines how much spread there is*** in the distribution.
* The **larger** the value of scale parameter, the **more spread** there is in the distribution.  The **smaller** the value of scale parameter, the **less spread** there is in the distribution.
* Fig. 2 shows the ***logistic distribution*** with three different scale parameters: 2, 5, and 8. The location for all three curves is 0.
* The logistic distributions do not have a shape parameter. The shape of logistic distribution is defined by the values of location and scale parameters.



**Fig. 2:** Logistic distribution with different scale parameters

**Shape parameter**

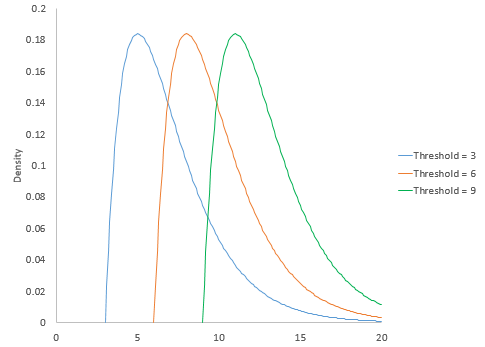
* Some distributions (e.g. Gamma, Weibull distribution) have shape parameters.
* The shape parameter of a distribution ***allows the distribution to take different shapes***.
* *The* ***larger the value*** *of shape parameter, the more* ***the distribution tends to be skewed to the left****.*
* *The* ***smaller the value*** *shape parameter,* ***the more the distribution tends to be skewed to the right****.*
* Fig. 3 shows the gamma distribution with three different shape parameters: 2, 5, and 8.
* The scale parameter for the gamma distribution in Fig. 3 is 2.  The gamma distribution does not have a location parameter.



**Fig. 3:** Gamma distribution with different shape parameters

**Threshold parameter**

* The threshold parameter of a distribution defines the minimum value of the distribution along the *X*-axis.
* The distribution cannot have any values below this threshold.
* Fig. 4 is the gamma distribution with three different threshold values: 3, 6 and 9.
* Both the scale and shape parameters of the shown gamma distribution are 2.



**Fig. 4:** Gamma distribution with different threshold values

**Parameter estimation**

A number of statistical techniques can be used to estimate the parameters for a distribution.  The most popularly used techniques for estimation of the parameters is the maximum likelihood estimation (MLE) technique. In this process, parameters are chosen such that likelihood function (L) is maximized or negative of likelihood function (L) is minimized.

Once this estimation is complete, we use goodness of fit techniques to help determine which distribution fits our data best.

**Testing goodness of fit**

* The goodness-of-fit test is a statistical hypothesis test to see how well sample data fit a distribution from an assumed/stated population.
* Put differently, this test shows if our sample data represents the data we would expect to find in the actual population or if it is somehow different.
* Goodness-of-fit establishes the discrepancy between the observed values and those that would be expected in the assumed/stated population.
* The most important method for determining goodness-of-fit is the ***frequency chi-square test***.

**Frequency chi-square () test**

* Suppose a population consists of mutually exclusive classes, the proportion of members falling in the class being (). This classification may be with respect to either an attribute or a variable (in case of a continuous variable, the classification is achieved by dividing the whole range of the variable into arbitrarily defined intervals.)
* Then, .
* Let a sample of size is drawn from this population and the number of members belonging to the class is (), and .
* It can be shown that
* Then, (Approximately), if is sufficiently large.
* The approximate normal deviates are, however, subject to the linear constraint . Thus,

(Approximately)

i.e. (1)

* In statistical literature, is referred to as a Pearsonian or a Frequency .
* The Pearsonian or a Frequency is an approximation. The condition for validity of the approximation is that the theoretical frequencies should be sufficiently large.
* Both practical and theoretical investigation show that the approximation is usually good if for each , provided the number of classes is also greater than or equal to 5.
* If the number of classes is smaller than 5, it is advisable to have each of the expected frequencies somewhat greater than 5.
* When it is found that for some class is less than 5, one should amalgamate this class with one or more of the adjacent classes so as to make the theoretical frequency in the combined class greater than or equal to 5.
* The number of degrees of freedom (*df*) will then be calculated taking into account the number of class after amalgamation, i.e.

On the other hand, in many situations **the values are not specified by the hypothesis but known that are dependent on some unknown population parameter(s). For instance, when the hypothesis says that the population distribution is of the Poisson or of the normal type, without specifying the value of or those of and .**

* Let us suppose that the hypothetical population depends on unknown parameters (), which may be estimated from the given sample itself. If the estimates of the population proportions in the ith class is denoted by , then

(2)

will still be approximately a if are large enough.

* However, the degrees of freedom will get reduced for estimation of each parameter. This is because estimation of each parameter imposes a constraint on the approximate standard normal deviates, .
* The number of degrees of freedom of the statistic will, therefore, be

**Problem 1**

In the course of an experiment on the breeding of peas, a botanist obtained 556 peas, of which 315 were round and yellow, 108 were round and green, 101 were angular and yellow, and 32 were angular and green. According to genetic theory, such peas should be obtained in the ratios 9:3:3:1. Are the experimental results compatible with the theory?

**Ans:** If we denote by , , and the proportions of peas in the four classes in the whole population of peas that may be obtained in experiments of this type, then the null hypothesis to be tested is

, , ,

A test for the hypothesis is then carried out by the statistic , which is distributed as a with .

The calculations of this for the given sample are shown below in the following table:

**Table 1:** Calculation of values

|  |  |  |  |
| --- | --- | --- | --- |
| Class | Obs. Freq.  () | Exp. Freq.  () |  |
| Round & Yellow | 315 | 312.75 | 0.016 |
| Round & Green | 108 | 104.25 | 0.135 |
| Angular & Yellow | 101 | 104.25 | 0.101 |
| Angular & Green | 32 | 34.75 | 0.218 |
| Total | | | 0.470 |

Hence, (calculated) = 0.470 and

(tabulated) = 7.815

This implies that the calculated value is insignificant at 5% level. Thus, the experimental results seem compatible with the genetic theory.

**Problem 2:** The heights (in cm) of 177 Indian adult males are taken. The frequency distribution of these measurement data are presented below:

|  |  |
| --- | --- |
| Class interval of height (cm) | Frequency |
| 144.55-149.55 | 1 |
| 149.55-154.55 | 3 |
| 154.55-159.55 | 24 |
| 159.55-164.55 | 58 |
| 164.55-169.55 | 60 |
| 169.55-174.55 | 27 |
| 174.55-179.55 | 2 |
| 179.55-184.55 | 2 |

Test if it would be appropriate to assume that the population distribution of height follows normal distribution.

**Ans**: Let us first fit a normal distribution to the given frequency distribution of height of Indian adult males. For the distribution of height of Indian adult males, the mean and standard deviation are found to be

and

The calculations for the fitted normal distributions are shown in the following table

**Table 1**: Fitting a normal distribution to the height distribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class interval  () |  |  |  | Exp. Freq  [] | Obs. Freq. |
| - 144.55 | -3.689 | 0.0001126 | 0.0001126\* | 0.20 | 0 |
| 144.55-149.55 | -2.775 | 0.0027604 | 0.0026478 | 0.469 | 1 |
| 149.55-154.55 | -1.861 | 0.0313727 | 0.0286123 | 5.064 | 3 |
| 154.55-159.55 | -0.947 | 0.1718219 | 0.1404492 | 24.860 | 24 |
| 159.55-164.55 | -0.034 | 0.4864387 | 0.3146168 | 55.687 | 58 |
| 164.55-169.55 | 0.880 | 0.8105703 | 0.3241316 | 57.371 | 60 |
| 169.55-174.55 | 1.794 | 0.9635916 | 0.1530213 | 27.085 | 27 |
| 174.55-179.55 | 2.708 | 0.9966152 | 0.0330236 | 5.845 | 2 |
| 179.55-185.55 | 3.621 | 0.9998533 | 0.0032381 | 0.573 | 2 |
| 184.55 - |  | 1 | 0.0001467\*\* | 0.026 | 0 |
| Total | | | | 177.000 | 177 |

\*It is the probability

\*\*It is the probability

We now may test for the goodness of fit of the normal distribution to the population of the height by means of statistic.

The computation of the is shown in the table below. It may be noted that here the first three and the last three class-intervals of the variable have been amalgamated to ensure that expected frequency in each cell is at least 5.

**Table 2:** Calculation of values

|  |  |  |  |
| --- | --- | --- | --- |
| Class interval  () | Obs. Freq.  () | Exp. Freq.  () |  |
| – 154.55 | 4 | 5.553 | 0.434 |
| 154.55 – 159.55 | 24 | 24.860 | 0.030 |
| 159.55 – 164.55 | 58 | 55.687 | 0.096 |
| 164.55 – 169.55 | 60 | 57.371 | 0.120 |
| 169.55 – 174.55 | 27 | 27.085 | 0.000 |
| 14.55 – | 4 | 6.444 | 0.927 |
| Total | 177 | 177.000 | 1.607 |

From the above table, (calculated) = 1.607

Remembering that in fitting the normal distribution, two parameters had to be estimated from the sample, we find that the statistic now has . The relevant tabulated values are = 7.815 and = 11.345.

Since the calculated value is much smaller than the tabulated values under the hypothesis of normality, the normal distribution seems to have given a very good fit.

**Exercise 1:** The following table gives the frequency distribution of number of weed seeds per packet for 196 one-lb packets of a variety of pulses.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. of Weed seeds | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Frequency | 7 | 33 | 54 | 37 | 34 | 16 | 8 | 5 | 1 | 1 |

Fit a Poisson distribution to these data and then test goodness of fit of the Poisson distribution.

**Exercise 2:** A six-faced die was thrown 300 times, and the number of points obtained at each throw was recorded. In this way, the following frequency distribution was formed. Use these data to test whether the die was unbiased.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. of points per throw | 1 | 2 | 3 | 4 | 5 | 6 |
| Frequency | 31 | 52 | 46 | 40 | 54 | 77 |

**Exercise 3**: Suppose we have sample of 100 data points shown in the following table. Fit an exponential distribution to these data and test the goodness of fit.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2.463 | 4.047 | 0.782 | 2.897 | 3.354 | 1.249 | 7.047 | 2.581 | 3.953 | 1.224 |
| 1.649 | 5.353 | 4.236 | 2.195 | 1.635 | 1.995 | 4.996 | 0.618 | 8.056 | 2.469 |
| 4.378 | 1.840 | 2.458 | 2.304 | 3.545 | 4.797 | 8.091 | 5.821 | 4.320 | 3.605 |
| 2.707 | 2.603 | 0.383 | 0.825 | 3.821 | 3.436 | 1.341 | 0.574 | 1.141 | 2.546 |
| 2.289 | 2.640 | 4.804 | 0.895 | 4.304 | 4.699 | 0.329 | 1.259 | 2.322 | 2.208 |
| 0.920 | 3.100 | 5.409 | 1.433 | 1.453 | 2.134 | 2.864 | 5.030 | 1.234 | 4.252 |
| 1.082 | 2.950 | 3.801 | 5.114 | 2.563 | 3.271 | 1.152 | 3.104 | 1.366 | 2.939 |
| 0.829 | 3.257 | 0.466 | 0.652 | 3.544 | 2.810 | 4.415 | 2.019 | 6.170 | 4.770 |
| 0.839 | 4.315 | 1.070 | 0.983 | 3.529 | 6.265 | 3.872 | 3.268 | 0.465 | 6.355 |
| 2.661 | 1.755 | 2.122 | 2.064 | 3.136 | 5.720 | 6.757 | 1.353 | 3.339 | 4.384 |

**Probability-Probability Plot (P-P plot)**

* In statistics, a **P–P plot** (probability–probability plot  or percent–percent plot) is used to visually compare data coming from different datasets (distributions). In other words, a P-P plot is used for assessing how closely two [data sets](https://en.wikipedia.org/wiki/Data_set) agree. The possible scenarios involve comparing:
* two empirical sets
* one empirical and one theoretical set
* two theoretical sets
* The most common use for probability plots is the middle one, when we compare observed (empirical) data to data coming from a specified theoretical distribution like normal, Poisson, exponential etc.

**The Procedure**

* A P–P plot plots two [cumulative distribution functions](https://en.wikipedia.org/wiki/Cumulative_distribution_function) (cdfs) against each other. Given two probability distributions, with cdfs "*F*" and "*G*", it plots {\displaystyle (F(z),G(z))} as *z* ranges from {\displaystyle -\infty } to {\displaystyle \infty .}.
* As a cdf has range [0,1], the domain of this parametric graph is [{\displaystyle (-\infty ,\infty )}{\displaystyle -\infty } {\displaystyle \infty .}],  and the range is the unit square .{\displaystyle [0,1]\times [0,1].}
* Thus for input z, the output is the pair of numbers giving what *percentage* of  and what percentageof   fall at or below *z.*
* The comparison line is the 45° line from (0, 0) to (1, 1).
* The distributions are equal if and only if the plot falls on this line. Any deviation indicates a difference between the distributions.

**Examples:**

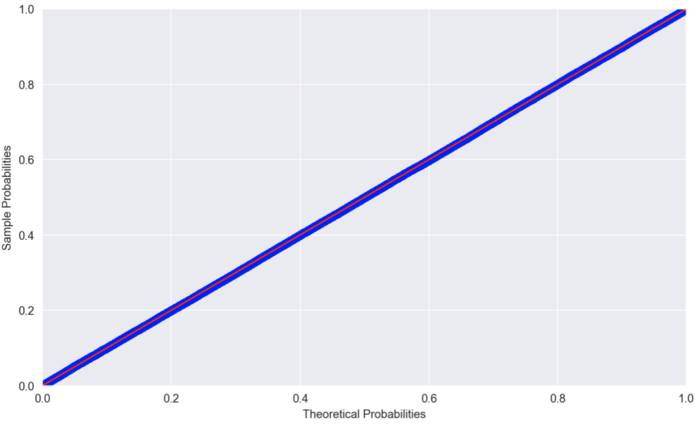


Fig. 1: A P-P plot comparing random numbers drawn from *N*(0, 1) to Standard Normal

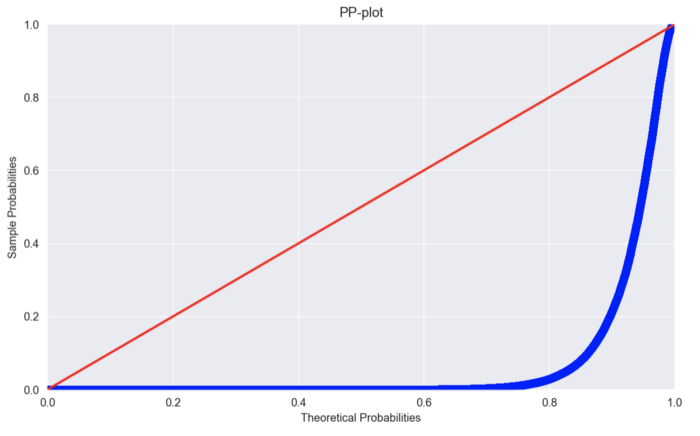


Fig. 2: A P-P plot comparing random variables drawn from *N*(1, 2.5) vs. *N*(5, 1)

**Some key information on P-P plots**

* The point on the plot indicates what percentage of data lies at or below *z* in both *f* and *g*(as per definition of the CDF)*.*
* To compare the distributions we check if the points lie on a 45-degree line (*x=y*). In case they deviate, the distributions differ.
* P-P plots are well suited to compare regions of high probability density (centre of distribution) because in these regions the empirical and theoretical CDFs change more rapidly than in regions of low probability density.
* P-P plots require fully specified distributions, so if we are using normal as the theoretical distribution we should specify the location and scale parameters.
* P-P plots can be used to visually evaluate the skewness of a distribution.
* P-P plots are most useful when comparing probability distributions that have a nearby or equal location.

**Plot (Q-Q plot)**

* The quantile-quantile (q-q) plot is a graphical technique for determining i) if two data sets come from populations with a common distribution, or ii) to compare the distribution of a sample to a theoretical distribution.
* By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
* A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
* A 45-degree reference line is also plotted.
* If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
* The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

**Examples**

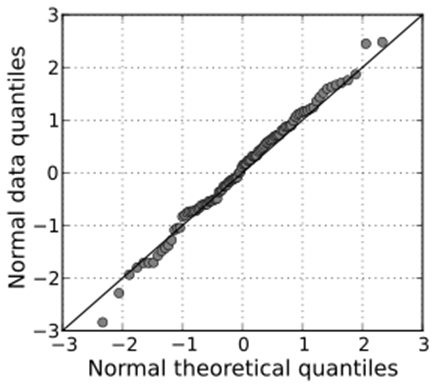
****

Fig.3: Q-Q plot of randomly generated independent standard normal data & standard normal population

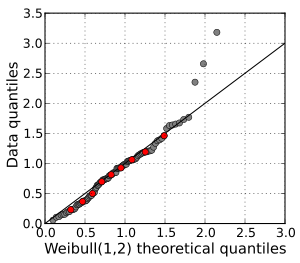
[](https://en.wikipedia.org/wiki/File:Weibull_qq.svg)

Fig. 4.: Q-Q plot of a sample data vs a Weibull distribution. The deciles of the distributions are shown in red.

**Interpretation of Q-Q plot**

* The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right.
* If the two distributions being compared are identical, the Q–Q plot follows the 45° line *y* = *x*.
* If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line *y* = *x*.
* If the general trend of the Q–Q plot is flatter than the line *y* = *x*, the distribution plotted on the horizontal axis is more [dispersed](https://en.wikipedia.org/wiki/Statistical_dispersion) than the distribution plotted on the vertical axis.
* Conversely, if the general trend of the Q–Q plot is steeper than the line *y* = *x*, the distribution plotted on the vertical axis is more [dispersed](https://en.wikipedia.org/wiki/Statistical_dispersion) than the distribution plotted on the horizontal axis.
* Q–Q plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

**Difference between P-P plot and Q-Q plot**

* A P-P plot compares the empirical cumulative distribution function of a data set with a specified theoretical cumulative distribution function F(·).
* A Q-Q plot compares the quantiles of a data distribution with the quantiles of a standardized theoretical distribution from a specified family of distributions.