# SUMMER RESEARCH PROJECT

16-May-2017 to 13-July-2017

---


**Project Supervisor : Prof. Asok Kumar Nanda**

**Project Intern : Anshuman Chakravarty**

# System of Frequency Curves

**Anshuman Chakravarty**
Summer Research Intern
Indian Institute of Science Education and Research Kolkata – 741246, India
Email – achakravarty97@iitkgp.ac.in

## Abstract

In this research project, we have tried to fit a known probability distribution to a particular dataset based on Pearson's Distributions. The sample data can be in different forms, for example, they can be a series of isolated points (single data points) or binned data represented by upper and lower bounds. A conventional way to solve the problem of data fitting is by the least square approximation method. But here, we will use Pearson's method and develop the theory, ground up, from a general differential equation representing a distribution. We will see how to correlate our observations from statistical experiment to a well formed equation or curve using method of moments, learn about the about the inferences we can make using the sample data and classify the observed data into certain types which would in turn tell us the curve which would be a reasonably good fit for the above mentioned dataset.

## Introduction

Data fitting is extremely necessary when trying to study the nature of sample observations. Often we need to predict values of a dependent variable, by interpolation or extrapolation of data points. In such situations, we need to look for a smooth curve which best captures the trends shown by the statistical data. Data Fitting has its uses in a variety of fields including calorimetry, thermodynamics, econometric analysis, hypothesis testing and so on. In the sections that follow, we address this problem of data fitting using a method proposed by Karl Pearson in 1895 and subsequently extended by him in 1901 and 1916 in a series of articles on biostatistics. Pearson has helped in developing Mathematical Statistics to a great extent. Some of his contributions include the correlation and regression coefficients, chi squared test, method of moments and the method of principal component analysis. In fact, the first introduction of histogram is usually credited to Pearson.

## Method of Moments

It is fairly straightforward to see that when we have a graduating curve or formula with $n$ constants and sample observations with $n$ data points as well, we can easily find the constants by solving $n$ equations. But in general, we will have more data points in sample observations than constants in the graduating curve. In such cases the graduating curve will never exactly reproduce any of the sample observations but will gives us a smooth curve running evenly through the roughness of the statistic.

Suppose we have already found a graduating curve to a sample observation, for example

Table 1.

| $x_i$ | $y_i$ | Value from Graduating Curve |
|---|---|---|
| 1 | $y_1$ | $a_1$ |
| 2 | $y_2$ | $a_2$ |
| 3 | $y_3$ | $a_3$ |
| . | . | . |
| . | . | . |
| . | . | . |
| $n$ | $y_n$ | $a_n$ |

Then, the sum of the values, $a_1, a_2, a_3, \ldots$ , i.e. the calculated frequencies from the graduating curve, should be equal to the sum of $y_1, y_2, y_3, \ldots$, i.e. the frequencies from sample observations.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a_i \tag{1}$$

Continuing with this line of thought, we can say that

$$\sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n} c_i a_i$$

where $c_i$ is a numerical coefficient. In the above table, we can see that $x_i = i$, hence

$$mean\ from\ observed\ data = \frac{\sum_{i=1}^{n} i\, y_i}{\sum_{i=1}^{n} y_i}$$

$$mean\ from\ calculated\ data = \frac{\sum_{i=1}^{n} i\, a_i}{\sum_{i=1}^{n} a_i}$$

By equating the two means and using eq.$(1)$, we get

$$\sum_{i=1}^{n} i\, y_i = \sum_{i=1}^{n} i\, a_i$$

Therefore, a reasonable choice for $\boldsymbol{c_i}$ is $\boldsymbol{i}$. To generalize this theory and generate more such equations, we can substitute, $c_i = i^t$, with $t = 0,1,2,3, \ldots$

*Defining moment about a vertical*

The $n^{th}$ moment is defined as the product of the frequency $(y_i)$ and the $n^{th}$ power of the distance of the frequency from the vertical about which the moment is being measured.

For example, instead of $x_i = i$, as in Table 1, if we have $x_i = x_i$, in general, then the $n^{th}$ moment of the whole distribution about a vertical through $x_r$ is equal to

$$n^{th}\ moment\ about\ x_r = \sum_{i=1}^{n} y_i\ (x_i - x_r)^n$$

*Moments for Binned data*

When $x_i$ are not isolated points but areas, we assume the frequencies ($y_i$) to be concentrated in the middle of the bases.

This assumption results in an overestimation or underestimation of some of the calculated moments. Hence, approximate corrections to the estimates of moments are necessary. These changes are known as *Sheppard's Corrections*. If $c$ is the class interval or bin width, $m_k$ is the measured $k^{th}$ moment and $\mu_k$ the corresponding corrected moment, then

$$\mu_1 = m_1$$

$$\mu_2 = m_2 - \frac{c^2}{12}$$

$$\mu_3 = m_3$$

$$\mu_4 = m_4 - \frac{m_2 c^2}{2} + \frac{7c^4}{240}$$

*Unit of Grouping*

It is the difference between successive mid-points of the bins or areas. Alternatively, it can be defined as the width of each bin or height of each class

*Moments about Centroid Vertical*

A vertical through the mean of sample observations is known as the Centroid Vertical. The moments calculated about the Centroid vertical are the most useful as they simplify our work when calculating constants or finding parameters for the different types of Pearson's Curves.
An important observation to be made here is that the distance of mean from any data point of the sample is equal to the $1^{st}$ moment of the whole distribution about a vertical through that point. Hence, the $\mathbf{1^{st}}$ *moment* of the whole distribution about the vertical through mean (Centroid Vertical) is $0$.

*Calculation of Moments from Curves*

In the preceding sections we've seen how to calculate moments from observed data. In this section we'll know how to find the moments from equations of curves. to different granules.
We have previously seen that the $n^{th}$ moment of an ordinate $y_x$ is $y_x x^n$. Therefore, the $n^{th}$ moment of the whole distribution from $x = h$ to $x = k$ is

$$\int_h^k y_x x^n\, dx\ ,$$

where the mean of the distribution is $\frac{\int_h^k y_x x dx}{\int_h^k y_x dx}$

**What is a Frequency Curve?**

A curve or a system of curves which describe frequency distributions by removing the roughness of the ungraduated observations is known as a frequency curve. While forming a differential equation which best describes a frequency curve in general, we make some assumptions:-

- They start at zero.
- They rise to a point of maxima at some rate and again fall at possibly a different rate.
- They have high contact at both ends.
- This approach by Pearson is based on unimodal distributions only.

Considering these assumptions, we can form an equation like this,

$$\frac{dy}{dx} = \frac{y(x+a)}{F(x)}$$

On expanding $F(x)$ using Maclauren's Formula we get

$$\frac{dy}{dx} = \frac{y(x+a)}{b_0 + b_1 x + b_2 x^2 + \cdots}$$

It can easily be shown that at $x = -a$, we have a point of maxima, which implies that, it is where the *mode* lies. And at both ends of the curve, where $y = 0$, we have points of high contact.

Another observation here is that $b_0$cannot be zero. Because if $b_0$is equal to $0$, then for $x = 0$, we observe that the slope will become infinite, indicating a point of discontinuity in the curve.

**Developing an intuition for the differential equation from a Hypergeometric Series**

Suppose we have $p$ black balls and $q$ white balls in a bag. And, $p + q = n$ is the total number of balls. Now we pick $r$ balls at random from this bag.

$$P(all\ the\ r\ balls\ are\ black) = \frac{p(p-1)(p-2)\dots(p-r+1)}{n(n-1)(n-2)\dots(n-r+1)}$$

$$P(1\ ball\ is\ white) = P(\ r-1\ balls\ are\ black) = \frac{p(p-1)(p-2)\dots(p-r+2)rq}{n(n-1)(n-2)\dots(n-r+1)}$$

$P(x-1\ balls\ are\ white) = y_x$(say)

$$= \frac{p(p-1)(p-2)\dots(p-r+x)q(q-1)(q-2)\dots(q-x+2)}{n(n-1)(n-2)\dots(n-r+1)} \frac{r(r-1)(r-2)\dots(r-x+2)}{(x-1)!}$$

$$= \frac{p(p-1)(p-2)\dots(p-r+1)}{n(n-1)(n-2)\dots(n-r+1)} \frac{q(q-1)(q-2)\dots(q-x+2)}{(p-r+x-1)(p-r+x-2)\dots(p-r+1)} \frac{r(r-1)(r-2)\dots(r-x+2)}{(x-1)!}$$

$$Let\ k = \frac{p(p-1)(p-2)\dots(p-r+1)}{n(n-1)(n-2)\dots(n-r+1)}$$

$$Then, y_x = k \frac{q(q-1)(q-2)\dots(q-x+2)}{(p-r+x-1)(p-r+x-2)\dots(p-r+1)} \frac{r(r-1)(r-2)\dots(r-x+2)}{(x-1)!}$$

$$And, y_{x+1} = k \frac{q(q-1)(q-2)\dots(q-x+1)}{(p-r+x)(p-r+x-2)\dots(p-r+1)} \frac{r(r-1)(r-2)\dots(r-x+1)}{x!}$$

$$\Delta y_x = y_{x+1} - y_x = y_x\left(\frac{(r-x+1)(q-x+1)}{x(p-r+x)} - 1\right)$$

$$y_{x+\frac{1}{2}} = \frac{y_{x+1} + y_x}{2} = \frac{y_x}{2}\left(\frac{(r-x+1)(q-x+1)}{x(p-r+x)} + 1\right)$$

$$\frac{\Delta y_x}{y_{x+\frac{1}{2}}} = \frac{1}{y}\frac{dy}{dx} = \frac{2\big((r+1)(q+1) - (n+2)x\big)}{2x^2 - x(2(r+1) + q - p) + (q+1)(r+1)}$$

Thus, we can see that we have reached a point in this exercise where the form of the equation derived is

$$\frac{1}{y}\frac{dy}{dx} = \frac{(x+a)}{b_0 + b_1 x + b_2 x^2}$$

**Determiners of the shape of a curve**

In the sections that follow, we deal with the constants up to $b_2$ and refrain from using further terms like $b_3, b_4$ .... This is because using more terms would mean using higher moments which are calculated from sample data. And as we know higher moments become untrustworthy because of the roughness present in observed data we don't want to use more terms in $F(x)$**.**

To determine the constants $a, b_0, b_1, b_2$ we would require the first four moments only. The first four moments, representative of the mean, variance, measure of skewness and kurtosis of a curve, help us describe the shape of any given frequency curve

Table 2.
$\mu_k$ *denotes the kth Central Moment &* $f(x)$ *is the density function*

| **Measure** | **Definition** | **Type of moment** | **Expression** |
|---|---|---|---|
| Mean($\mu$) | Expected value, i.e. central tendency of the distribution | Raw Moment | $\int_{-\infty}^{+\infty} xf(x)dx$ |
| Variance($\sigma^2$) | Expectation of the squared deviation of a random variable from its mean | Central Moment | $\int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \mu_2$ |
| Measure of Skewness | A measure of the asymmetry of the frequency distribution | Standardised Moment | $E\left(\frac{(x-\mu)^3}{\sigma^3}\right) = \frac{\mu_3}{\mu_2^{3/2}}$ |
| Measure of Kurtosis | A measure of the tailedness of the frequency distribution | Standardised Moment | $E\left(\frac{(x-\mu)^4}{\sigma^4}\right) = \frac{\mu_4}{\mu_2^2}$ |

**Deriving the values of the constants**

So, now we have the differential equation

$$\frac{dy}{dx} = \frac{y(x+a)}{b_0 + b_1 x + b_2 x^2}$$

To get the constants $a, b_0, b_1, b_2$ in terms of the moments we follow these steps

$$\frac{dy}{dx}(b_0 + b_1 x + b_2 x^2) = y(x+a)$$

Multiplying both sides by $x^n$ and integrating,

$$\int x^n \frac{dy}{dx}(b_0 + b_1x + b_2x^2)\,dx = \int x^n y(x+a)dx$$

$$x^n(b_0 + b_1x + b_2x^2)y - \int \{nb_0x^n + (n+1)b_1x^n + (n+2)b_2x^{n+1}\}ydx = \int x^{n+1}ydx + \int ax^nydx$$

Due to high contact at the ends of the curve, the expression, $x^n(b_0 + b_1x + b_2x^2)y$ , vanishes when limits are applied.

We know that, $n^{th}\ moment\ of\ a\ curve = \int yx^n\,dx$. Using this

$$-nb_0\mu_{n-1} - (n+1)b_1\mu_n - (n+2)b_2\mu_{n+1} = \mu_{n+1} + a\mu_n$$

where $\mu_n$ denotes the $n^{th}\ moment\ of\ the\ curve$ about mean. We choose moment about mean to make our calculations easier as $\mu_1 = 0$. Now, if we put $n = 0,1,2,3$ we get four equations to solve for the four unknowns.

On solving

$$\left.\begin{aligned} a\mu_0 + b_1\mu_0 + 2b_2 &= -\mu_1 \\ a\mu_1 + b_0\mu_0 + 2b_1\mu_1 + 3b_2\mu_2 &= -\mu_2 \\ a\mu_2 + 2b_0\mu_1 + 3b_1\mu_2 + 4b_2\mu_3 &= -\mu_3 \\ a\mu_3 + 3b_0\mu_2 + 4b_1\mu_3 + 5b_2\mu_4 &= -\mu_4 \end{aligned}\right\}$$

We get

$a = \frac{\mu_3(\mu_4+3\mu_2^2)}{10\mu_2\mu_4-18\mu_2^3-12\mu_3^2}$ , $b_0 = \frac{\mu_2(4\mu_2\mu_4-3\mu_3^2)}{10\mu_2\mu_4-18\mu_2^3-12\mu_3^2}$ , $b_1 = \frac{\mu_3(\mu_4+3\mu_2^2)}{10\mu_2\mu_4-18\mu_2^3-12\mu_3^2}$ , $b_1 = \frac{2\mu_2\mu_4-3\mu_3^2-6\mu_2}{10\mu_2\mu_4-18\mu_2^3-12\mu_3^2}$

We have already stated that the point $x = a$ is where the mode of the distribution lies. Now, as we have shifted our origin to mean (while calculating moments), $\boldsymbol{a}$ is nothing but the distance between the mean and the mode.

We define two parameters $\beta_1\ and\ \beta_2$ as

$$\beta_1{}^2 = \frac{\mu_3{}^2}{\mu_2^3},\ i.e\ (Measure\ of\ Skewness)^2$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2},,\ i.e\ Measure\ of\ Kurtosis$$

**Kappa Criterion**

To solve the equation, $\frac{dy}{dx} = \frac{y(x+a)}{b_0+b_1x+b_2x^2}$ , by integration, we need to know the properties of the roots of the equation

$$b_0 + b_1x + b_2x^2 = 0$$

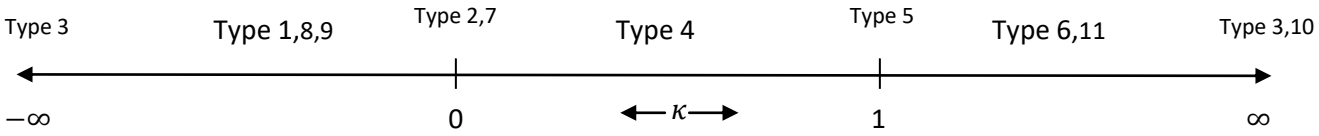
Let us define a parameter, $\kappa = \frac{b_1^2}{4b_0b_2} = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)}$

- $\kappa < 0$ implies that the roots are real and of different signs.
- $0 < \kappa < 1$ implies that the roots are not real, but complex conjugates.
- $\kappa > 1$ implies that the roots are real and of same signs.

The cases mentioned above give us the 3 main types of Pearson's Curves, namely *Type 1, Type 4* and *Type 6* respectively.

- In limiting cases, for example, when $\kappa$ changes from $< 0$ to $> 0$ , it has to go through $0$ , cases such as these give us the *Transition Types*.

For a better understanding, we represent the types on a number line as




**Solving the Ordinary Differential Equation**

***Pearson's Type 1:***

$\kappa < 0$, which implies the roots are real and of opposite sign. Let the roots be $-A_1 \; and \; A_2$ , where both $A_1 \; and \; A_2$ are positive. Hence, we get

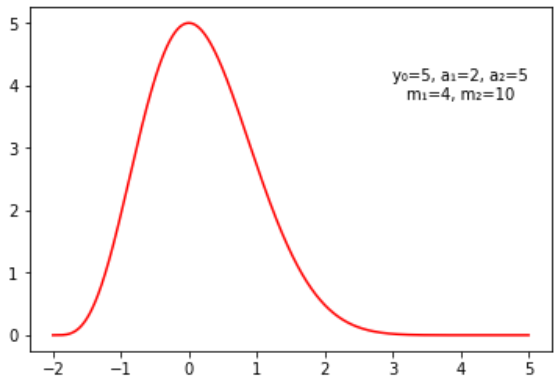
$$\int \frac{dy}{y} = \frac{1}{b_2} \int \frac{(x+a)dx}{(x+A_1)(x-A_2)}$$

$$y = c(x+A_1)^{\frac{a+A_2}{b_2(A_1+A_2)}}(A_2-x)^{\frac{A_1-a}{b_2(A_1+A_2)}} \quad \text{where} - A_1 < x < A_2 \text{ and origin at mean}$$

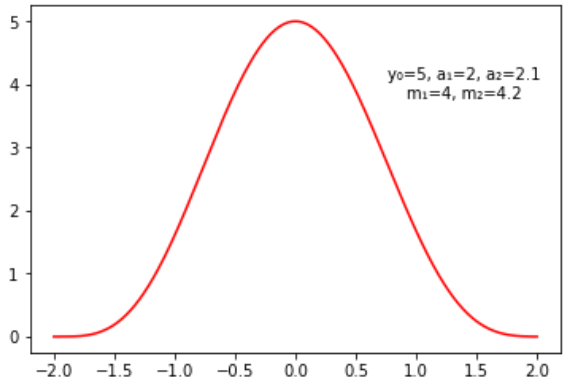
Now we shift the origin at mode by

$$x = x + a, \; a_1 = A_1 - a, \; a_2 = A_2 + a, and \; \frac{m_1}{a_1} = \frac{m_2}{a_2}$$

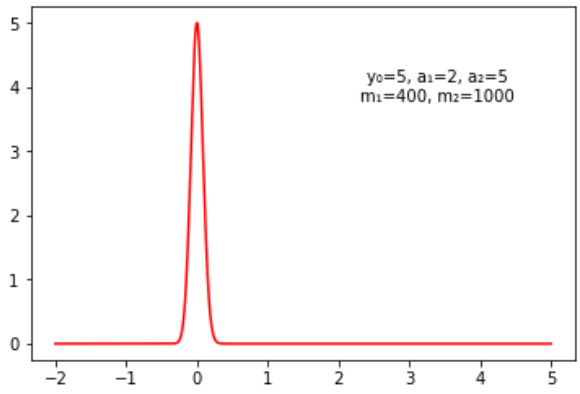
$$y = y_0 \left(1 + \frac{x}{a_1}\right)^{m_1} \left(1 - \frac{x}{a_2}\right)^{m_2} \text{ with limited range } (-a_1, \; a_2)$$



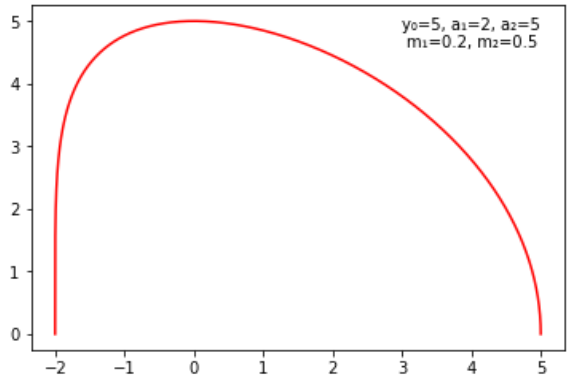

constants $m_1, m_2 > 0, m_1 \neq m_2$



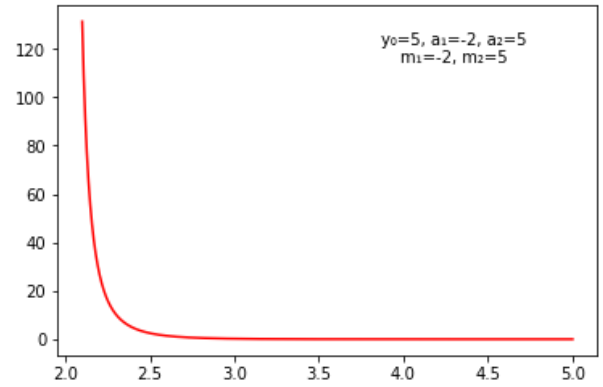

constants $m_1, m_2 > 0, m_1 \approx m_2$



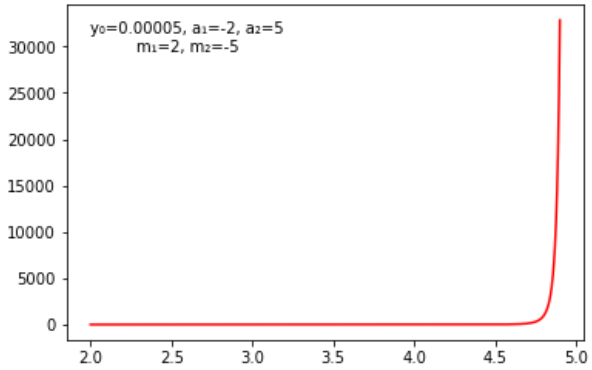

constants $m_1, m_2 > 0, m_1, m_2$ *are large*




constants $m_1, m_2 > 0, m_1, m_2$ *are very small*




constants $m_1 < 0, m_2 > 0$




constants $m_1 > 0, m_2 < 0$

***Pearson's Type 4:***

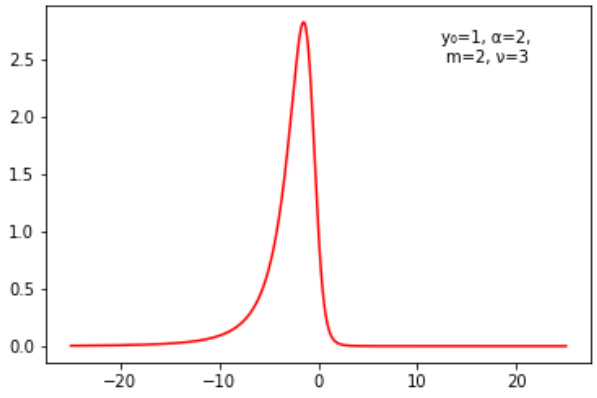
$0 < \kappa < 1$, which implies the roots are not real and are complex conjugates.

$$\int \frac{dy}{y} = \frac{1}{b_2} \int \frac{(x+a)dx}{\left(x + \frac{b_1}{2b_2}\right)^2 + \left(\frac{b_0}{b_2} - \frac{b_1^2}{4b_2^2}\right)},$$

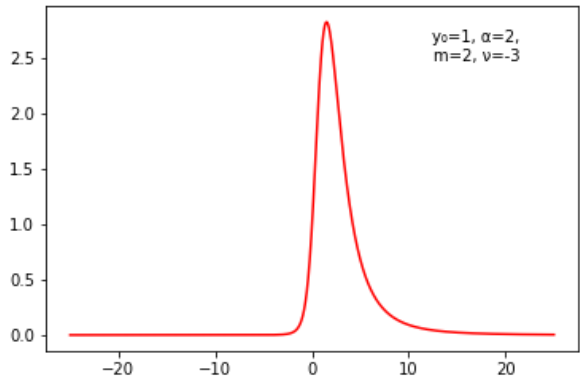
$$let\ \alpha^2 = \left(\frac{b_0}{b_2} - \frac{b_1^2}{4b_2^2}\right), \text{then}$$

$$y = y_0 \left(1 + \frac{x^2}{\alpha^2}\right)^{-m} e^{-\nu\, tan^{-1}(x/\alpha)} \text{ with unlimited range } (-\infty, \infty)$$

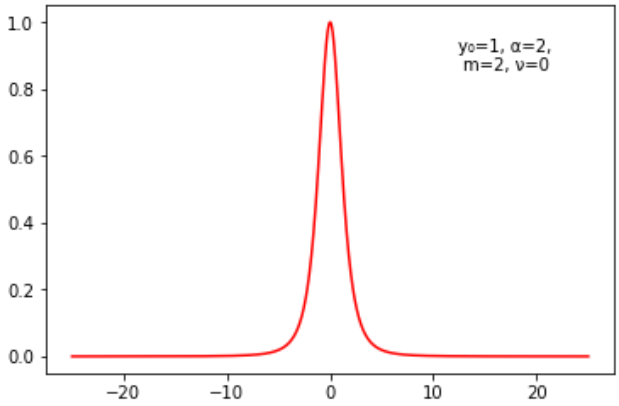
where $m = {}^{-1}/_{2b_2}$, $\nu = {}^{-c}/_{\alpha b_2}$




constant $\nu > 0$




constant $\nu < 0$




constant $\nu = 0$

***Pearson's Type 6:***

$\kappa > 1$, which implies the roots are real and of same sign. Let the roots be $-A_1\ and\ -A_2$, where both $A_1\ and\ A_2$ are positive. Hence, we get

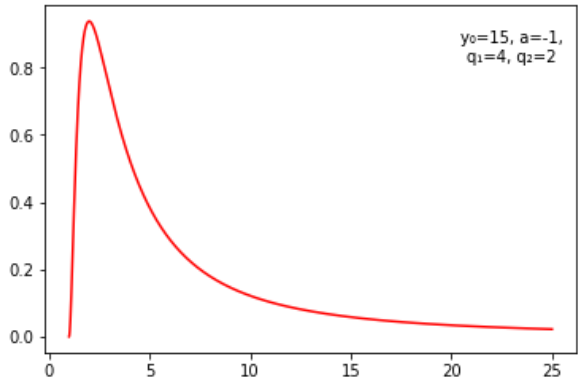
$$\int \frac{dy}{y} = \frac{1}{b_2}\int \frac{(x+a)dx}{(x+A_1)(x+A_2)}$$

$$y = c(x+A_1)^{\frac{a+A_2}{b_2(A_1+A_2)}}(x+A_2)^{\frac{A_1+a}{b_2(A_1+A_2)}} \quad \text{where} - A_1 < x < -A_2 \text{ and origin at mean}$$

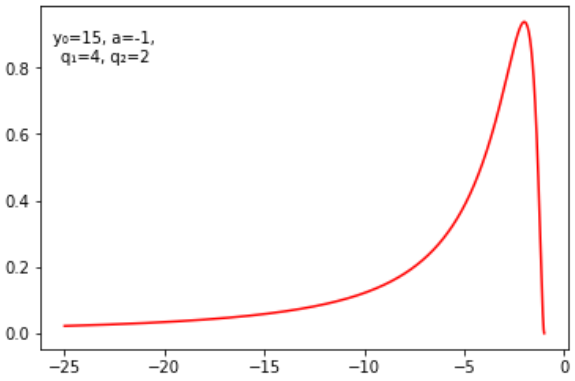
Now we shift the origin such that we can write $x + A_1\ as\ x$, then we have

$$y = y_0(x-a)^{q_2}x^{-q_1} \quad \text{with origin at a distance } a \text{ before start of curve}$$

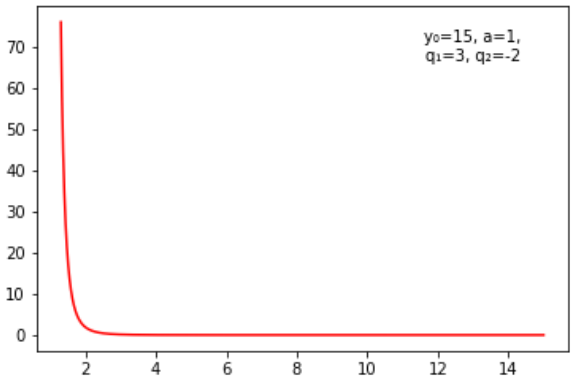
and $q_1 > q_2$, unlimited range in one direction $(a, \infty)\ or\ (-\infty, a)$ depending on the value of $a$




conditions $\mu_3 > 0, a > 0, (q_1, q_2) > 0$




conditions $\mu_3 < 0, a < 0, (q_1, q_2) > 0$

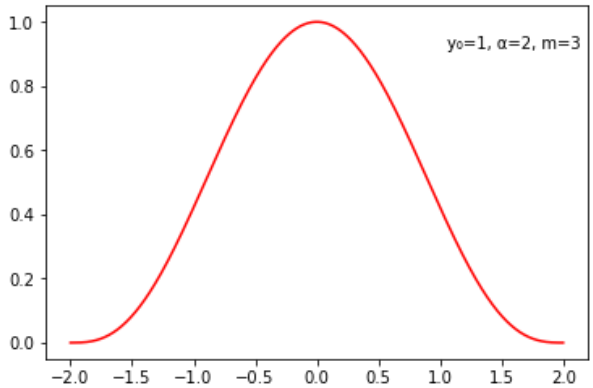



conditions $(\mu_3, a, q_1) > 0, q_2 < 0$

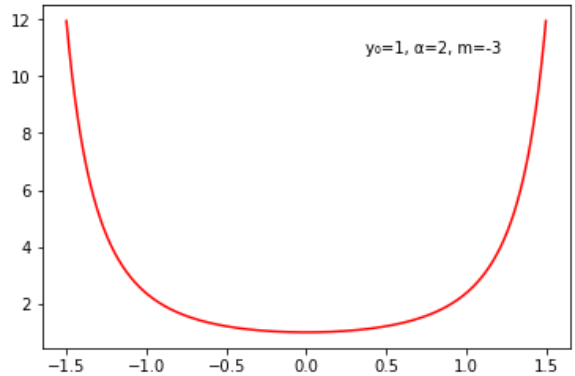
***Pearson's Type 2:***

When $\kappa = 0, \beta_1 = 0, \beta_2 < 3$, we get

$$y = y_0 \left(1 - \frac{x^2}{\alpha^2}\right)^m \text{ where origin is at mode, with limited range } (-a, a)$$




constant $m > 0$

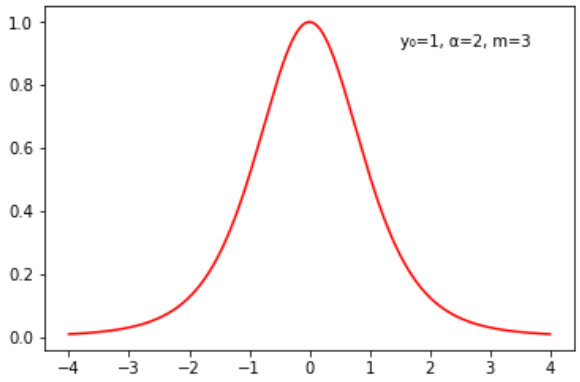



constant $m < 0$

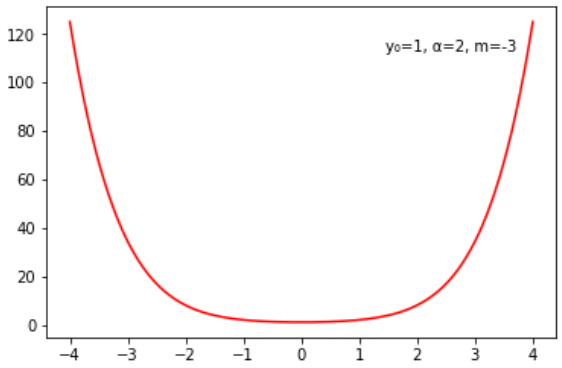
***Pearson's Type 7:***

When $\kappa = 0, \beta_1 = 0, \beta_2 > 3$, we get

$$y = y_0\left(1 + \frac{x^2}{\alpha^2}\right)^{-m}$$ where origin is at mode, with unlimited range $(-\infty, \infty)$




constant $m > 0$

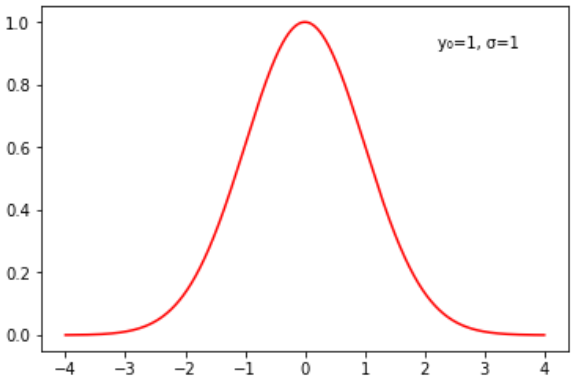



constant $m < 0$

***Normal Curve:***

When $\kappa = 0, \beta_1 = 0, \beta_2 = 3,$ we get

$$y = y_0 e^{-x^2/2\sigma^2}$$ where origin is at mode, with unlimited range $(-\infty, \infty)$

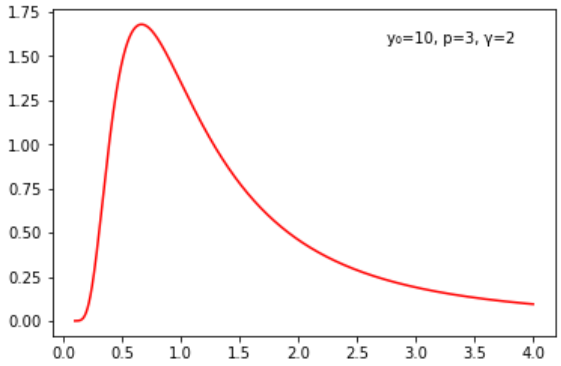



***Pearson's Type 5:***

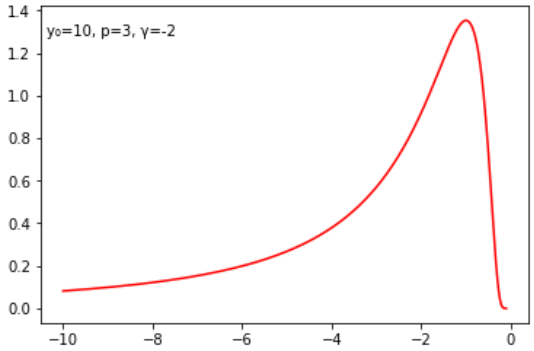
When $\kappa = 1,$ we get

$$y = y_0 x^{-p} e^{-\gamma/x}$$ where origin is at the start of the curve

and $p > 1$, unlimited range in one direction $(0, \infty)$ $or$ $(-\infty, 0)$ depending on the value of $\gamma$




constant $\gamma > 0$



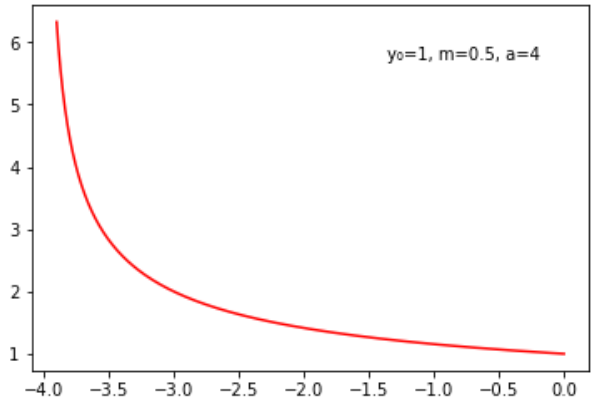

constant $\gamma < 0$

***Pearson's Type 8:***

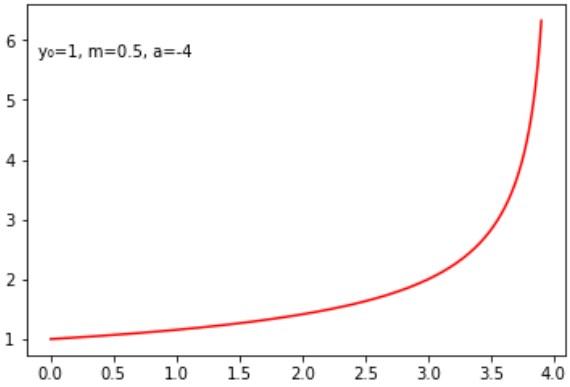
When $\kappa = 1, 5\beta_2 - 6\beta_1 - 9 < 0$ we get

$$y = y_0\left(1 + \frac{x}{a}\right)^{-m}$$ where origin is at the end of the curve

and $0 < m < 1$, limited range from $(-a, 0)$ $or$ $(0, -a)$ depending on the sign of $a$




constant $a > 0$



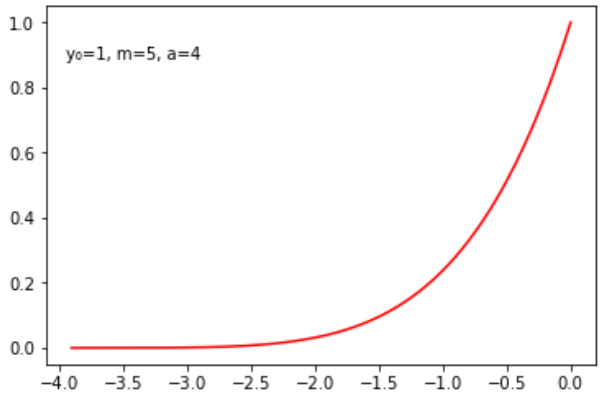

constant $a < 0$

***Pearson's Type 9:***

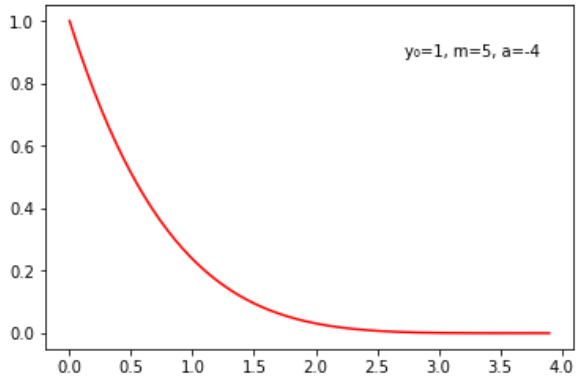
When $\kappa = 1,\ 5\beta_2 - 6\beta_1 - 9 > 0\ and\ 2\beta_2 - 3\beta_1 - 6 < 0$ we get

$$y = y_0\left(1 + \frac{x}{a}\right)^{m}$$ where origin is at the end of the curve

and $0 < m$, limited range from $(-a, 0)$ $or$ $(0, -a)$ depending on the sign of $a$




constant $a > 0$



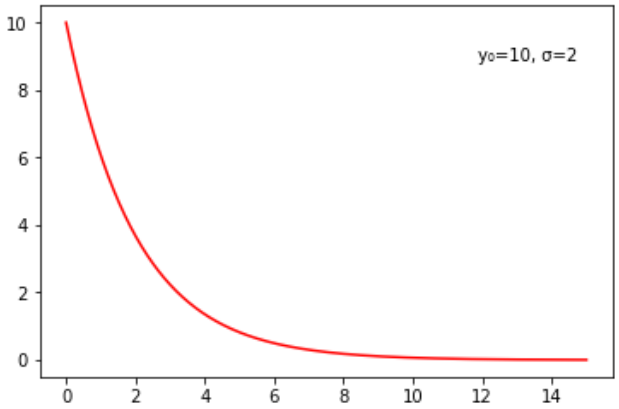

constant $a < 0$

***Pearson's Type 10:***

When $\beta_2 = 9\ and\ \beta_1 = 4$, we get

$$y = y_0 e^{-x/\sigma}$$ where origin is at the start of the curve

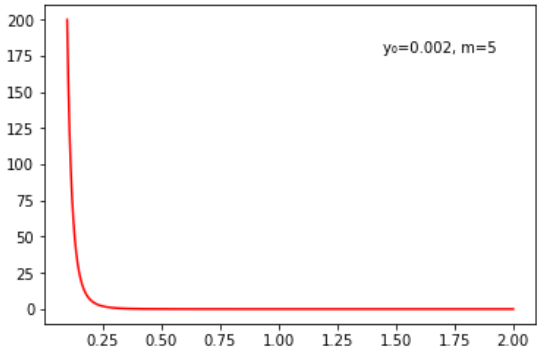
and with unlimited range in one direction $(0, \infty)$




***Pearson's Type 11:***

When $\kappa > 1\ and\ 2\beta_2 - 3\beta_1 - 6 > 0$ we get

$$y = y_0 x^{-m}$$ where origin is at a distance $'b'$ before start of the curve




constant $m > 0$

**Example Data set**

*Data Set Description*: Here we take an example which gives us a measure of the exposed risk of sickness of people in general, belonging to different age groups. In the table presented below we have taken the central age of each age group in the data set.

Table 3.
$'5'$ *is the Unit of Grouping in this case*

| **Central Age of each Group ($x_i$)** | **Exposed Risk of sickness ($f_i$)** |
|---|---|
| 17 | 34 |
| 22 | 145 |
| 27 | 156 |
| 32 | 145 |
| 37 | 123 |
| 42 | 103 |
| 47 | 86 |
| 52 | 71 |
| 57 | 55 |
| 62 | 37 |
| 67 | 21 |
| 72 | 13 |
| 77 | 7 |
| 82 | 3 |
| 87 | 1 |

**Algorithm for determining the Type of Frequency Curve:**

We break down the process of finding which Type of Curve it is into a series of steps,

*Step 1* **:** Divide every $x_i$ with the Unit of Grouping, which is **5** in this case. We basically divide by the class height so that $x_{i+1} - x_i = 1$

*Step 2* **:** Calculate mean of distribution with the help of the formula $\frac{\sum f_i x_i}{N}$ , where $N = \sum f_i$. Also calculate the mode of the distribution. In this case

$\text{Mean} = 7.575$
$\text{Mode} = 5.4$

*Step 3* **:** Calculate the first four moments about mean. In this case

$\mu_1 = 1.78346226676 \text{ X } 10^{-15} \approx 0$
$\mu_2 = 7.662375$
$\mu_3 = 15.10689375$
$\mu_4 = 172.325275078$

*Step 4* **:** Calculate $\beta_1$ and $\beta_2$. In this case

$\beta_1 = \ 0.507294485379$
$\beta_2 = \ 2.93509508361$

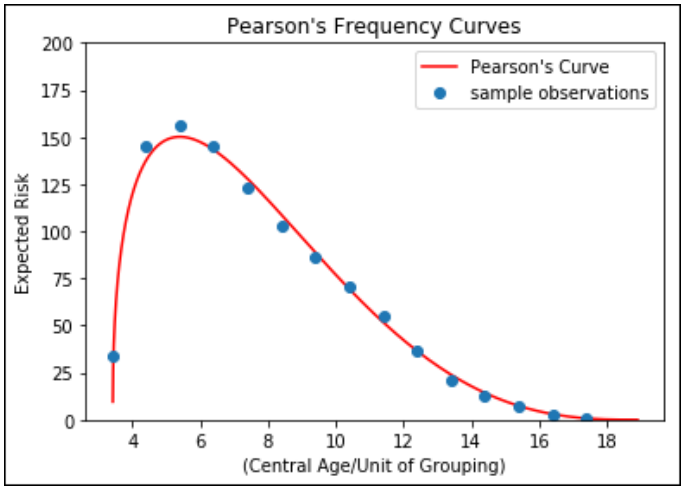
*Step 5* **:** Using $\beta_1$ and $\beta_2$ calculate $\kappa(kappa\ criterion)$. In this case

$\kappa$ = -0.264690949076

*Step 6* **:** Depending on the value of $\kappa$ determine the Type of frequency curve which would be a reasonably good fit for the sample observations. In this case we use **Type 1.**

*Step 7* **:** Once the Type is known, we can find the constants associated with the curve in terms of $\beta_1$, $\beta_2$ and the first four moments.

*Step 8* **:** Shift the curve according to where the origin was shifted while deriving the specific Pearson's Type.




**Results**

As per the calculations made in *Step 7,* the curve we get for the above stated statistical experience is,

$$y = 150.12\left(1+\frac{x}{1.977}\right)^{0.407}\left(1-\frac{x}{13.508}\right)^{2.779} \text{ where origin is at mode}$$

We shift, $x = x - mode,$ and hence get the curve aligned with the sample data points. Now, we can plug in several values of $x$ in this equation to get the corresponding *Expected Risk of Sickness* depending on our needs.

**Goodness of Fit**

A confusion may arise in many cases when the value of $\kappa$ is very close to zero or one. For example, if $\kappa = 0.012$, we might say that a curve of Type 4 would be of best fit, but Type 2 or 7 or the Normal Curve might also give better or satisfactory results. One way to get an idea of which one to use, can be made through the Goodness of Fit test. Here, we will be discussing about Pearson's Chi Squared Test.

In this test, we divide the whole range of the curve into $n$ divisions. This number $n$ is decided based on the Struges' rule. This rule gives us an approximate formula for calculating number of classes

$$n = 1 + 3.322\, log_{10}N$$

where $n$ is the number of classes and $N$ is the total number of observations in the data.

As a result we get $n$ intervals. Next, we calculate the observed and expected frequencies in each of these intervals.

*Observed Frequency in the $i^{th}$ interval*

$$O_i = \sum frequencies\ corresponding\ to\ the\ x\ coordinates\ which\ belong\ to\ that\ interval$$

*Expected Frequency in the $i^{th}$ interval*

$$E_i = \int_{start\ of\ interval}^{end\ of\ interval} y\ dx$$

Now we calculate the test statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where $\chi^2$ asymptotically approaches a $\boldsymbol{\chi^2}$ (chi squared) distribution. Here, the number of degrees of freedom is equal to

$$degrees\ of\ freedom = n - 1 - s$$

where $s = \text{the number of constants estimated in the equation of the frequency curve}$

Now we can get the $p - value$ by comparing the statistic to the $\boldsymbol{\chi^2}$ (chi squared) distribution. When comparing two curves, generally, the one which gives a higher $p - value$, is assumed to be a better fit. Even after this test there might remain some ambiguity if both the $p - values$ are very close to each other.

**Future Projections**

A comparison between the estimated values from Pearson's curves and those from Least Square Approximation can be studied. The error in each case can be measured and tabulated. Also, the whole exercise performed here was pertaining to the problem of curve fitting only in case of *Continuous Distributions.* No efforts were made to study the best possible fit if the known distribution were to be Discrete. This project encourages and motivates us to explore the same scenario of Data Fitting in case of Discrete Distributions.

**Summary**

Thus, with the help of Pearson's System of Frequency Curves we have been able to solve the popular problem of Data Fitting. Here we used a very general Ordinary Differential Equation, keeping in mind the assumptions made on frequency curves, and divided the whole domain of frequency curves into various *Types*. Based on this model, we defined some new parameters, e.g.,$\beta_1$, $\beta_2$ and $\kappa\ (Kappa\ Criterion)$. Finally we decide on a *Type* after recording the value of $\kappa$. Once we know the *Type*, we can find the constants associated with that type of curve and hence find the equation of the frequency curve.

**References**

Elderton, W. P., and N. L. Johnson: "Systems of Frequency Curves". Cambridge University Press, London 1969. 216 S., 17 Abb., 61 Tab., Preis 60 s. net