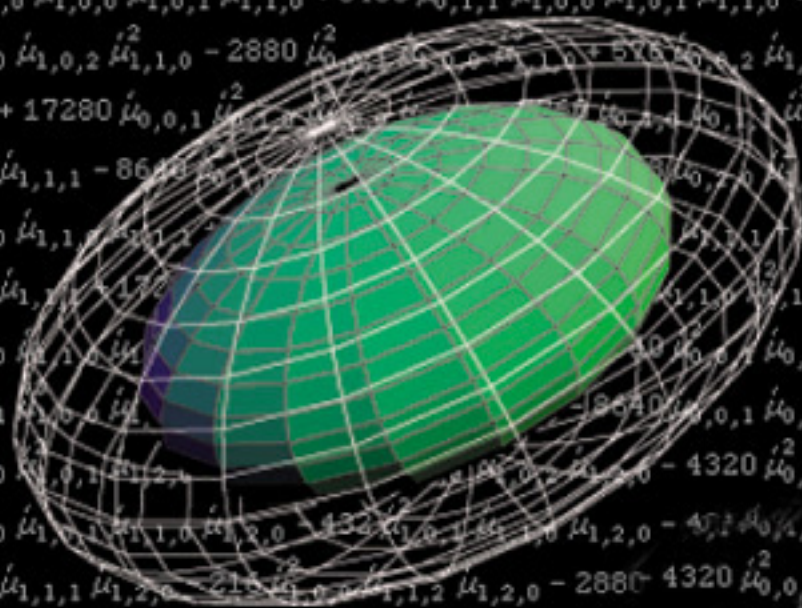


SPRINGER TEXTS IN STATISTICS

# MATHEMATICAL STATISTICS

with  
*Mathematica*<sup>®</sup>



COLIN ROSE  
MURRAY D. SMITH

# Mathematical Statistics with *Mathematica*

## Chapter 5 – Systems of Distributions

5.1	Introduction	149
5.2	The Pearson Family	149
A	Introduction	149
B	Fitting Pearson Densities	151
C	Pearson Types	157
D	Pearson Coefficients in Terms of Moments	159
E	Higher Order Pearson-Style Families	161
5.3	Johnson Transformations	164
A	Introduction	164
B	$S_L$ System (Lognormal)	165
C	$S_U$ System (Unbounded)	168
D	$S_B$ System (Bounded)	173
5.4	Gram–Charlier Expansions	175
A	Definitions and Fitting	175
B	Hermite Polynomials; Gram–Charlier Coefficients	179
5.5	Non-Parametric Kernel Density Estimation	181
5.6	The Method of Moments	183
5.7	Exercises	185

---

**Please reference this 2002 edition as:**

Rose, C. and Smith, M.D. (2002)  
*Mathematical Statistics with Mathematica*, Springer-Verlag, New York.

---

**Latest edition**

For the latest up-to-date edition, please visit: [www.mathStatica.com](http://www.mathStatica.com)

# Chapter 5

## Systems of Distributions

---

### 5.1 Introduction

This chapter discusses three systems of distributions: (i) the Pearson family, §5.2, which defines a density in terms of its slope; (ii) the Johnson system, §5.3, which describes a density in terms of transformations of the standard Normal; and (iii) a Gram–Charlier expansion, §5.4, which represents a density as a series expansion of the standard Normal density.

The Pearson system, in particular, is of interest in its own right because it nests many common distributions such as the Gamma, Normal, Student’s  $t$ , and Beta as special cases. The family of stable distributions is discussed in Chapter 2. Non-parametric kernel density estimation is briefly discussed in §5.5, while the method of moments estimation technique (used throughout the chapter) is covered in §5.6.

---

### 5.2 The Pearson Family

#### 5.2 A Introduction

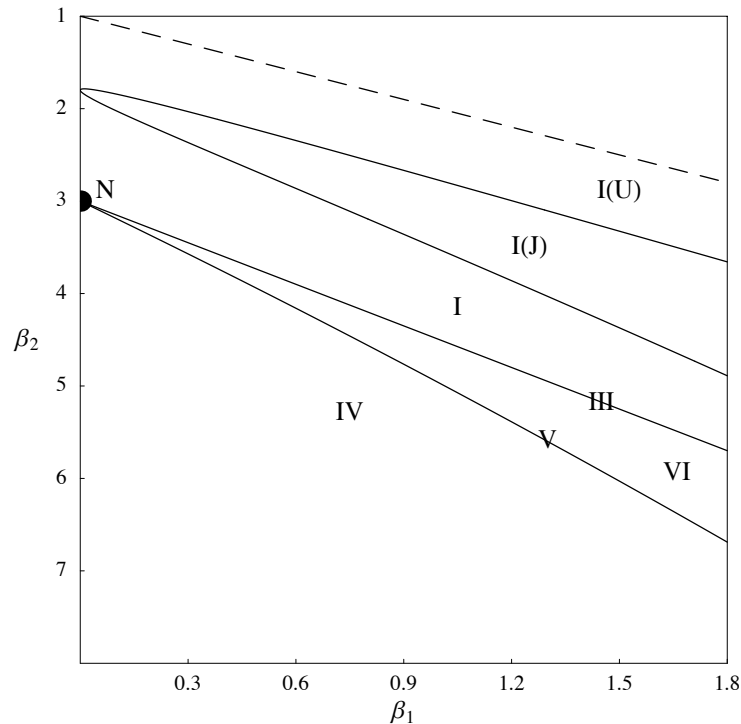
The Pearson system is the family of solutions  $p(x)$  to the differential equation

$$\frac{dp(x)}{dx} = - \frac{a+x}{c_0 + c_1 x + c_2 x^2} p(x) \quad (5.1)$$

that yield well-defined density functions. The shape of the resulting distribution will clearly depend on the Pearson parameters  $(a, c_0, c_1, c_2)$ . As we shall see later, these parameters can be expressed in terms of the first four moments of the distribution (§5.2D). Thus, if we know the first four moments, we can construct a density function that is consistent with those moments. This provides a rather neat way of constructing density functions that approximate a given set of data. Karl Pearson grouped the family into a number of *types* (§5.2 C). These *types* can be classified in terms of  $\beta_1$  and  $\beta_2$  where

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2}. \quad (5.2)$$

The value of  $\sqrt{\beta_1}$  is often used as a measure of *skewness*, while  $\beta_2$  is often used as a measure of *kurtosis*. Figure 1 illustrates this classification system in  $(\beta_1, \beta_2)$  space.




**Fig. 1:** The  $\beta_1, \beta_2$  chart for the Pearson system

The classification consists of several types, as listed in Table 1.

Main types :	<i>Type I</i> including <i>I(U)</i> and <i>I(J)</i> , <i>Type IV</i> and <i>Type VI</i>
Transition types :	<i>Type III</i> (a line), <i>Type V</i> (a line)
Symmetrical types :	If the distribution is symmetrical, then $\mu_3 = 0$ , so $\beta_1 = 0$ . This yields three special cases : <ul style="list-style-type: none"> <li>• The N at (0, 3) denotes the Normal distribution.</li> <li>• <i>Type II</i> (not labelled) occurs when <math>\beta_1 = 0</math> and <math>\beta_2 &lt; 3</math>, and is thus just a special case of <i>Type I</i>.</li> <li>• <i>Type VII</i> occurs when <math>\beta_1 = 0</math> and <math>\beta_2 &gt; 3</math> (a special case of <i>Type IV</i>).</li> </ul>

**Table 1:** Pearson types

The dashed line denotes the upper limit for all distributions. The vertical axis is ‘upside-down’. This has become an established (though rather peculiar) convention which we follow. *Type I*, *I(U)* and *I(J)* all share the same functional form—they are all *Type I*. However, they differ in appearance: *Type I(U)* yields a U-shaped density, while *Type I(J)* yields a J-shaped density.<sup>1</sup> The electronic notebook version of this chapter provides an animated tour of the Pearson system here: 

## 5.2 B Fitting Pearson Densities

This section illustrates how to construct a Pearson distribution that is consistent with a set of data whose first four moments are known. With **mathStatica**, this is a two step process:

- (i) Use `PearsonPlot`  $[\{\mu_2, \mu_3, \mu_4\}]$  to ascertain which Pearson *Type* is consistent with the data.
- (ii) If it is say *Type III*, then `PearsonIII`  $[\mu, \{\mu_2, \mu_3, \mu_4\}, x]$  yields the desired density function  $f(x)$  (and its domain).

The full set of functions is:

PearsonI	PearsonII	PearsonIII	PearsonIV
PearsonV	PearsonVI	PearsonVII	

In the following examples, we categorise data as follows:

- Is it *population* data or *sample* data?
- Is it *raw* data or *grouped* data?

⊕ **Example 1:** Fitting a Pearson Density to *Raw Population* Data

The `marks.dat` data set lists the final marks of all 891 first year students in the Department of Econometrics at the University of Sydney in 1996. It is raw data because it has not been grouped or altered in any way, and may be thought of as population data (as opposed to sample data) because the entire population's results are listed in the data set. To proceed, we first load the data set into *Mathematica*:

```
data = ReadList["marks.dat"];
```

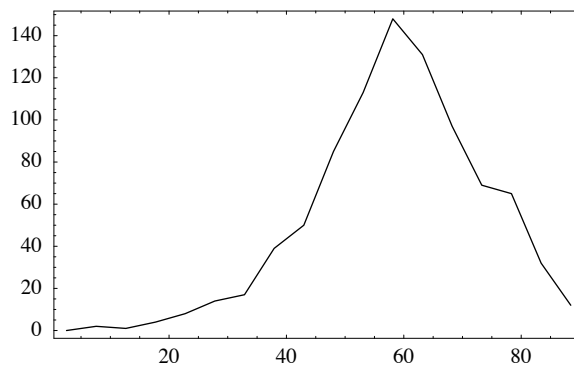
and then find its mean:

```
mean = SampleMean[data] // N
```

```
58.9024
```

We can use the **mathStatica** function `FrequencyPlot` to get an intuitive visual perspective on this data set:

```
FrequencyPlot[data];
```



**Fig. 2:** Frequency polygon of student marks

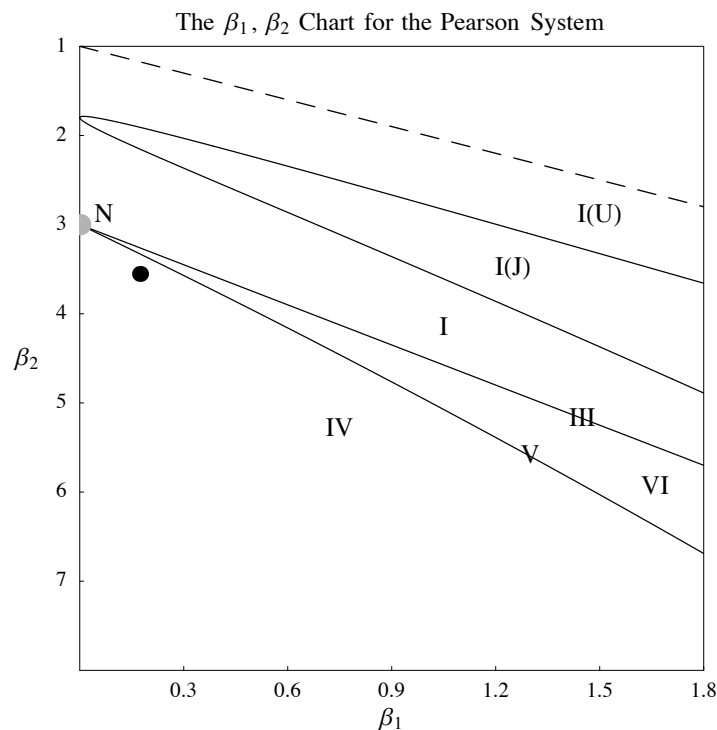
The  $x$ -axis in Fig. 2 represents the range of possible marks from 0 to 100, while the  $y$ -axis plots frequency. Of course, there is nothing absolute about the shape of this plot, because the shape varies with the chosen bandwidth  $c$ . To see this, evaluate `FrequencyPlot[data, {0, 100, c}]` at different values of  $c$ , changing the bandwidth from, say, 4 to 12. Although the shape changes, this empirical pdf nevertheless does give a rough idea of what our Pearson density will look like. Alternatively, see the non-parametric kernel density estimator in §5.5.

Next, we need to find the population central moments  $\mu_2, \mu_3, \mu_4$ . Since we have population data, we can use the `CentralMoment` function in *Mathematica*'s `Statistics`DescriptiveStatistics`` package, which we load as follows:

```
<< Statistics`
 $\mu_{234} = \text{Table}[\text{CentralMoment}[\text{data}, \mathbf{r}], \{\mathbf{r}, 2, 4\}] // \mathbf{N}$ 
```

Step (i): `PearsonPlot[ $\mu_{234}$ ]` calculates  $\beta_1$  and  $\beta_2$  from  $\mu_{234}$ , and then indicates which Pearson *Type* is appropriate for this data set by plotting a large black dot at the point  $(\beta_1, \beta_2)$ :

```
PearsonPlot[ $\mu_{234}$ ];
 $\{\beta_1 \rightarrow 0.173966, \beta_2 \rightarrow 3.55303\}$ 
```



**Fig. 3:** The marks data is of *Type IV*

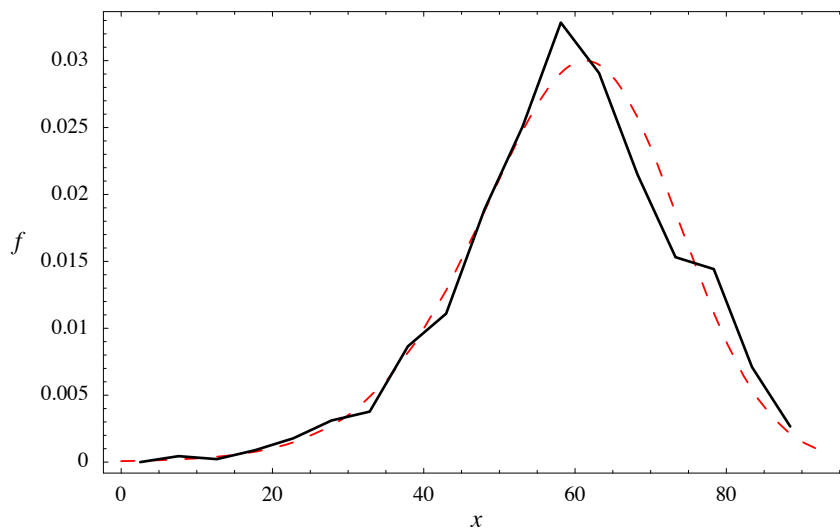
Step (ii): The large black dot is within the *Type IV* zone (the most feared of them all!), so the fitted Pearson density  $f(x)$  and its domain are given by:

$$\{\mathbf{f}, \mathbf{domain}[\mathbf{f}]\} = \mathbf{PearsonIV}[\mathbf{mean}, \mu_{234}, \mathbf{x}]$$

$$\left\{ \frac{1.14587 \times 10^{25} e^{13.4877 \operatorname{ArcTan}[1.55011 - 0.0169455 x]}}{(448.276 - 6.92074 x + 0.0378282 x^2)^{13.2177}}, \{x, -\infty, \infty\} \right\}$$

The `FrequencyPlot` function can now be used to compare the empirical pdf (—) with the fitted Pearson pdf (---):

```
p1 = FrequencyPlot[data, f];
```



**Fig. 4:** The empirical pdf (—) and fitted Pearson pdf (---) for the marks data

⊕ **Example 2:** Fitting a Pearson Density to Raw Sample Data

The file `grain.dat` contains data that measures the yield from 1500 different rows of wheat. The data comes from Andrews and Herzberg (1985) and StatLib. We shall treat it as raw sample data. To proceed, we first load the data set into *Mathematica*:

```
data = ReadList["grain.dat"];
```

and find its sample mean:

```
mean = SampleMean[data] // N
```

```
587.722
```

Because this is sample data, the population central moments  $\mu_2, \mu_3, \mu_4$  are unknown. We shall not use the `CentralMoment` function from *Mathematica's* *Statistics* package to estimate the population central moments, because the `CentralMoment` function is a biased estimator. Instead, we shall use **mathStatistica's** `UnbiasedCentralMoment` function, as discussed in Chapter 7, because it is an unbiased estimator of population central moments (and has many other desirable properties). As it so happens, the bias from using the `CentralMoment` function will be small in this example because the sample size is large, but that may not always be the case. Here, then, is our estimate of the vector  $(\mu_2, \mu_3, \mu_4)$ :

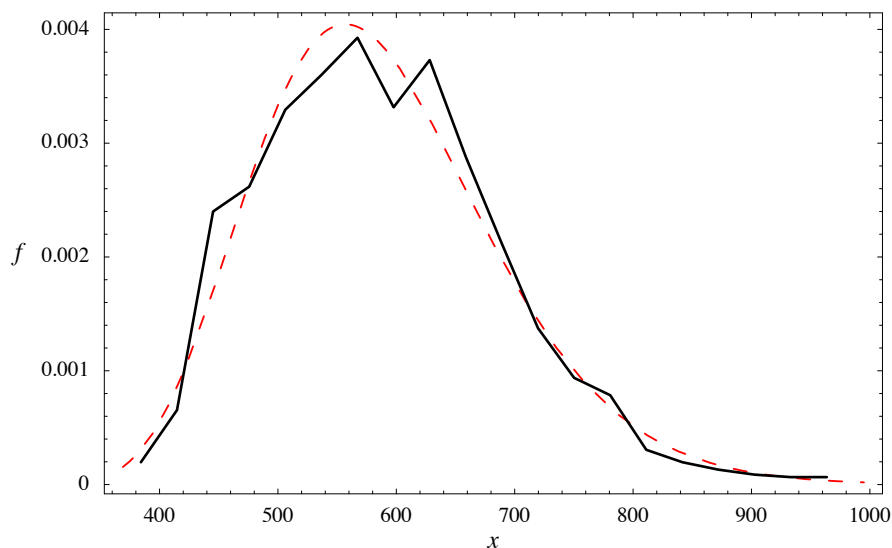
```
 $\hat{\mu}_{234} = \text{Table}[\text{UnbiasedCentralMoment}[\text{data}, \mathbf{r}], \{\mathbf{r}, 2, 4\}]$ 
{9997.97, 576417., 3.39334 × 108}
```

`PearsonPlot` [ $\hat{\mu}_{234}$ ] shows that this is close to *Type III*, so we fit a Pearson density,  $f(x)$ , to *Type III*:

```
{f, domain[f]} = PearsonIII[mean,  $\hat{\mu}_{234}$ , x]
{2.39465 × 10-35 e-0.0324339 x (-7601.05 + 30.832 x)10.0661,
{x, 246.531, ∞}}
```

Once again, the `FrequencyPlot` function compares the empirical pdf (—) with the fitted Pearson pdf (---):

```
FrequencyPlot[data, f];
```



**Fig. 5:** The empirical pdf (—) and fitted Pearson pdf (---) for wheat yield data



⊕ **Example 3:** Fitting a Pearson Density to *Grouped Data*

Table 2 stems from Elderton and Johnson (1969, p. 5):

age X	freq
< 19	34
20–24	145
25–29	156
30–34	145
35–39	123
40–44	103
45–49	86
50–54	71
55–59	55
60–64	37
65–69	21
70–74	13
75–79	7
80–84	3
85–89	1

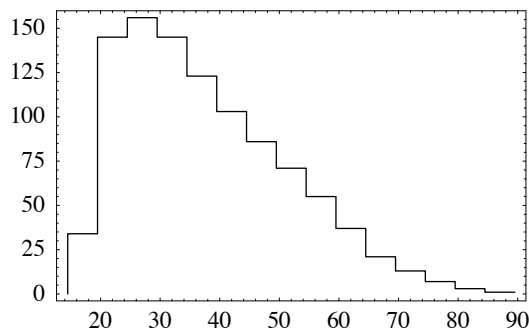
**Table 2:** The number of sick people at different ages (in years)

Here, ages 20–24 includes those aged from  $19\frac{1}{2}$  up to  $24\frac{1}{2}$ , and so on. Let  $X$  denote the mid-point of each class interval of ages (note that these are equally spaced), while  $\text{freq}$  denotes the frequency of each interval. Finally, let  $\tau$  denote the relative frequency. The mid-point of the first class is taken to be 17 to ensure equal bandwidths. Then:

```
x = {17, 22, 27, 32, 37, 42, 47, 52, 57, 62, 67, 72, 77, 82, 87};
freq = {34, 145, 156, 145, 123, 103, 86, 71, 55, 37, 21, 13, 7, 3, 1};
τ = freq / (Plus @@ freq);
```

The **mathStatica** function `FrequencyGroupPlot` provides a ‘histogram’ of this grouped data:

```
FrequencyGroupPlot [{X, freq}];
```



**Fig. 6:** ‘Histogram’ of the number of sick people at different ages

which gives some idea of what the fitted Pearson density should look like.

When data is given in table form, the mean is conventionally taken as  $\sum_{i=1}^k X_i \tau_i$ , where  $X_i$  is the mid-point of each interval, and  $\tau_i$  is the relative frequency of each interval, over the  $k$  class intervals. Thus:

$$\text{mean} = \mathbf{X} \cdot \boldsymbol{\tau} // \mathbf{N}$$

$$37.875$$

A quick and slightly dirty<sup>2</sup> (though widely used) estimator of the  $r^{\text{th}}$  central moment for grouped data is given by:

$$\text{DirtyMu}[\mathbf{r\_}] := (\mathbf{X} - \text{mean})^{\mathbf{r\_}} \cdot \boldsymbol{\tau}$$

Then our estimates of  $(\mu_2, \mu_3, \mu_4)$  are:

$$\hat{\boldsymbol{\mu}}_{234} = \{\text{DirtyMu}[2], \text{DirtyMu}[3], \text{DirtyMu}[4]\}$$

$$\{191.559, 1888.36, 107703.\}$$

which is *Type I*, as `PearsonPlot` [ $\hat{\boldsymbol{\mu}}_{234}$ ] will verify. Then, the fitted Pearson density is:

$$\{\mathbf{f}, \text{domain}[\mathbf{f}]\} = \text{PearsonI}[\text{mean}, \hat{\boldsymbol{\mu}}_{234}, \mathbf{x}]$$

$$\{9.70076 \times 10^{-8} (94.3007 - 1. \mathbf{x})^{2.77976} (-16.8719 + 1. \mathbf{x})^{0.406924}, \{ \mathbf{x}, 16.8719, 94.3007 \}\}$$

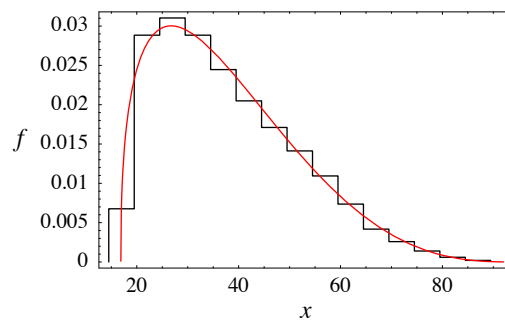
Of course, the density  $f(x)$  should be consistent with the central moments that generated it. Thus, if we calculated the first few central moments of  $f(x)$ , we should obtain  $\{\mu_2 \rightarrow 191.559, \mu_3 \rightarrow 1888.36, \mu_4 \rightarrow 107703\}$ , as above. A quick check verifies these results:

$$\text{Expect}[(\mathbf{x} - \text{mean})^{\{2, 3, 4\}}, \mathbf{f}]$$

$$\{191.559, 1888.36, 107703.\}$$

The `FrequencyGroupPlot` function can now be used to compare the ‘histogram’ with the smooth fitted Pearson pdf:

$$\text{FrequencyGroupPlot}[\{\mathbf{X}, \text{freq}\}, \mathbf{f};$$



**Fig. 7:** The ‘histogram’ and the fitted Pearson pdf (smooth)

## 5.2 C Pearson Types

Recall that the Pearson family is defined as the set of solutions to

$$\frac{dp(x)}{dx} = -\frac{a+x}{c_0 + c_1 x + c_2 x^2} p(x).$$

In *Mathematica*, the solution to this differential equation can be expressed as:

$$\text{Pearson} := \text{DSolve}\left[p'[\mathbf{x}] == -\frac{(\mathbf{a} + \mathbf{x}) p[\mathbf{x}]}{c_0 + c_1 \mathbf{x} + c_2 \mathbf{x}^2}, p[\mathbf{x}], \mathbf{x}\right]$$

Since  $\frac{dp}{dx} = 0$  when  $x = -a$ , the latter defines the mode, while the shape of the density will depend on the roots of the quadratic  $c_0 + c_1 x + c_2 x^2$ . The various Pearson *Types* correspond to the different forms this quadratic may take. We briefly consider the main seven types, in no particular order. Before doing so, we set up `MrClean` to ensure that we start our analysis of each *Type* with a clean slate:

```
MrClean := ClearAll[a, c0, c1, c2, p, x];
```

*Type IV* occurs when  $c_0 + c_1 x + c_2 x^2$  does not have real roots. In *Mathematica*, this is equivalent to finding the solution to the differential equation without making any special assumption at all about the roots. This works because *Mathematica* typically finds the most general solution, and does not assume the roots are real:

```
MrClean; Pearson // Simplify
```

$$\left\{ \left\{ p[x] \rightarrow e^{\frac{(c_1 - 2 a c_2) \text{ArcTan}\left[\frac{c_1 + 2 c_2 x}{\sqrt{-c_1^2 + 4 c_0 c_2}}\right]}{c_2 \sqrt{-c_1^2 + 4 c_0 c_2}}} (c_0 + x (c_1 + c_2 x))^{-\frac{1}{2 c_2}} C[1] \right\} \right\}$$

The domain is  $\{x, -\infty, \infty\}$ . Under *Type IV*, numerical integration is usually required to find the constant of integration  $C[1]$ .

*Type VII* is the special symmetrical case of *Type IV*, and it occurs when  $c_1 = a = 0$ . This nests Student's *t* distribution:

```
% /. {c1 -> 0, a -> 0}
```

$$\left\{ \left\{ p[x] \rightarrow (c_0 + c_2 x^2)^{-\frac{1}{2 c_2}} C[1] \right\} \right\}$$

*Type III* (Gamma distribution) occurs when  $c_2 = 0$ :

```
MrClean; c2 = 0; Pearson // Simplify
```

$$\left\{ \left\{ p[x] \rightarrow e^{-\frac{x}{c_1}} (c_0 + c_1 x)^{\frac{c_0 - a c_1}{c_1^2}} C[1] \right\} \right\}$$

In order for this solution to be a well-defined pdf, we require  $p(x) > 0$ . Thus, if  $c_1 > 0$ , the domain is  $x > -c_0/c_1$ ; if  $c_1 < 0$ , the domain is  $x < -c_0/c_1$ .

Type V occurs when the quadratic  $c_0 + c_1 x + c_2 x^2$  has one real root. This occurs when  $c_1^2 - 4 c_0 c_2 = 0$ . Hence:

$$\text{MrClean; } c_0 = \frac{c_1^2}{4 c_2} ; \text{ Pearson // Simplify}$$

$$\left\{ \left\{ p[x] \rightarrow e^{\frac{-c_1 + 2 a c_2}{c_2 (c_1 + 2 c_2 x)} - 1/c_2} C[1] \right\} \right\}$$

The Normal distribution is obtained when  $c_1 = c_2 = 0$ :

$$\text{MrClean; } c_1 = 0; \quad c_2 = 0; \quad \text{Pearson}$$

$$\left\{ \left\{ p[x] \rightarrow e^{-\frac{a x}{c_0} - \frac{x^2}{2 c_0}} C[1] \right\} \right\}$$

Completing the square allows us to write this as:

$$p[x] = k e^{-\frac{(x+a)^2}{2 c_0}} ; \quad \text{domain}[p[x]] = \{x, -\infty, \infty\};$$

where, in order to be a well-defined density, constant  $k$  must be such that the density integrates to unity; that is, that  $P(X < \infty) = 1$ :

$$\text{Solve[ Prob}[\infty, p[x]] == 1, k]$$

- This further assumes that:  $\{c_0 > 0\}$

$$\left\{ \left\{ k \rightarrow \frac{1}{\sqrt{c_0} \sqrt{2 \pi}} \right\} \right\}$$

The result is thus Normal with mean  $-a$ , and variance  $c_0 > 0$ .

That leaves *Type I*, *Type II* and *Type VI*. These cases occur if  $c_0 + c_1 x + c_2 x^2 = 0$  has two *real* roots,  $r_1$  and  $r_2$ . In particular, *Type I* occurs if  $r_1 < 0 < r_2$  (roots are of *opposite* sign), with domain  $r_1 < x < r_2$ . This nests the Beta distribution. *Type II* is identical to *Type I*, except that we now further assume that  $r_1 = -r_2$ . This yields a symmetrical curve with  $\beta_1 = 0$ . *Type VI* occurs if  $r_1$  and  $r_2$  are the *same* sign; the domain is  $x > r_2$  if  $0 < r_1 < r_2$ , or  $x < r_2$  if  $r_2 < r_1 < 0$ . In the case of *Type VI*, with two real roots of the same sign, one can express  $c_0 + c_1 x + c_2 x^2$  as  $c_2(x - r_1)(x - r_2)$ . The family of solutions is then:

**MrClean;**

$$\text{DSolve} \left[ p'[x] == - \frac{a + x}{c_2 (x - r_1) (x - r_2)} p[x], p[x], x \right] //$$

**Simplify**

$$\left\{ \left\{ p[x] \rightarrow (-r_1 + x)^{\frac{a+r_1}{-c_2 r_1 + c_2 r_2}} (-r_2 + x)^{\frac{a+r_2}{c_2 r_1 - c_2 r_2}} C[1] \right\} \right\}$$

where the constant of integration can now be solved for the relevant domain.

## 5.2 D Pearson Coefficients in Terms of Moments

**ClearAll[a, c0, c1, c2, eqn,  $\mu$ ]**

It is possible to express the Pearson coefficients  $a$ ,  $c_0$ ,  $c_1$  and  $c_2$  in terms of the first four raw moments  $\dot{\mu}_r$  ( $r = 1, 4$ ). To do so, we first multiply both sides of (5.1) by  $x^r$  and integrate over the domain of  $X$ :

$$\int_{-\infty}^{\infty} x^r (c_0 + c_1 x + c_2 x^2) \frac{dp(x)}{dx} dx = - \int_{-\infty}^{\infty} x^r (a + x) p(x) dx. \quad (5.3)$$

If we integrate the left-hand side by parts,

$$\int_{-\infty}^{\infty} f g' dx = f g \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f' g dx \quad \text{with } g' = \frac{dp(x)}{dx}$$

and break the right-hand side into two, then (5.3) becomes

$$\begin{aligned} & \left. x^r (c_0 + c_1 x + c_2 x^2) p(x) \right|_{-\infty}^{\infty} \\ & - \int_{-\infty}^{\infty} \{ r c_0 x^{r-1} + (r+1) c_1 x^r + (r+2) c_2 x^{r+1} \} p(x) dx \\ & = - \int_{-\infty}^{\infty} a x^r p(x) dx - \int_{-\infty}^{\infty} x^{r+1} p(x) dx \end{aligned} \quad (5.4)$$

If we assume that  $x^r p(x) \rightarrow 0$  at the extremum of the domain, then the first expression on the left-hand side vanishes, and after substituting raw moments  $\dot{\mu}$  for integrals, we are left with

$$-r c_0 \dot{\mu}_{r-1} - (r+1) c_1 \dot{\mu}_r - (r+2) c_2 \dot{\mu}_{r+1} = -a \dot{\mu}_r - \dot{\mu}_{r+1}. \quad (5.5)$$

This recurrence relation defines any moment in terms of lower moments. Further, since the density must integrate to unity, we have the boundary condition  $\dot{\mu}_0 = 1$ . In *Mathematica* notation, we write this relation as:

```
eqn[r_] :=
  (-r c0  $\dot{\mu}_{r-1}$  - (r+1) c1  $\dot{\mu}_r$  - (r+2) c2  $\dot{\mu}_{r+1}$  == -a  $\dot{\mu}_r$  -  $\dot{\mu}_{r+1}$ )
  /.  $\dot{\mu}_0 \rightarrow 1$ 
```

We wish to find  $a$ ,  $c_0$ ,  $c_1$  and  $c_2$  in terms of  $\dot{\mu}_r$ . Putting  $r = 0, 1, 2$  and  $3$  yields the required 4 equations (for the 4 unknowns) which we now solve simultaneously to yield the solution:

**z = Solve[Table[eqn[r], {r, 0, 3}], {a, c0, c1, c2}]**  
**// Simplify**

$$a \rightarrow \frac{20 \mu_1^2 \mu_2 \mu_3 - 12 \mu_1^3 \mu_4 - \mu_3 (3 \mu_2^2 + \mu_4) + \mu_1 (-9 \mu_2^3 - 8 \mu_3^2 + 13 \mu_2 \mu_4)}{2 (9 \mu_2^3 + 4 \mu_1^3 \mu_3 - 16 \mu_1 \mu_2 \mu_3 + 6 \mu_3^2 - 5 \mu_2 \mu_4 + \mu_1^2 (-3 \mu_2^2 + 5 \mu_4))}$$

$$c_0 \rightarrow \frac{\mu_1 \mu_3 (\mu_2^2 + \mu_4) + \mu_2 (3 \mu_3^2 - 4 \mu_2 \mu_4) + \mu_1^2 (-4 \mu_3^2 + 3 \mu_2 \mu_4)}{2 (9 \mu_2^3 + 4 \mu_1^3 \mu_3 - 16 \mu_1 \mu_2 \mu_3 + 6 \mu_3^2 - 5 \mu_2 \mu_4 + \mu_1^2 (-3 \mu_2^2 + 5 \mu_4))}$$

$$c_1 \rightarrow \frac{8 \mu_1^2 \mu_2 \mu_3 - 6 \mu_1^3 \mu_4 - \mu_3 (3 \mu_2^2 + \mu_4) + \mu_1 (-3 \mu_2^3 - 2 \mu_3^2 + 7 \mu_2 \mu_4)}{2 (9 \mu_2^3 + 4 \mu_1^3 \mu_3 - 16 \mu_1 \mu_2 \mu_3 + 6 \mu_3^2 - 5 \mu_2 \mu_4 + \mu_1^2 (-3 \mu_2^2 + 5 \mu_4))}$$

$$c_2 \rightarrow \frac{6 \mu_2^3 + 4 \mu_1^3 \mu_3 - 10 \mu_1 \mu_2 \mu_3 + 3 \mu_3^2 - 2 \mu_2 \mu_4 + \mu_1^2 (-3 \mu_2^2 + 2 \mu_4)}{2 (9 \mu_2^3 + 4 \mu_1^3 \mu_3 - 16 \mu_1 \mu_2 \mu_3 + 6 \mu_3^2 - 5 \mu_2 \mu_4 + \mu_1^2 (-3 \mu_2^2 + 5 \mu_4))}$$

If we work *about the mean*, then  $\mu_1 = 0$ , and  $\mu_r = \mu_r$  for  $r \geq 2$ . The formulae then become:

**z /. { $\mu_1 \rightarrow 0$ ,  $\mu \rightarrow \mu$ }**

$$a \rightarrow -\frac{\mu_3 (3 \mu_2^2 + \mu_4)}{2 (9 \mu_2^3 + 6 \mu_3^2 - 5 \mu_2 \mu_4)}$$

$$c_0 \rightarrow \frac{\mu_2 (3 \mu_3^2 - 4 \mu_2 \mu_4)}{2 (9 \mu_2^3 + 6 \mu_3^2 - 5 \mu_2 \mu_4)}$$

$$c_1 \rightarrow -\frac{\mu_3 (3 \mu_2^2 + \mu_4)}{2 (9 \mu_2^3 + 6 \mu_3^2 - 5 \mu_2 \mu_4)}$$

$$c_2 \rightarrow \frac{6 \mu_2^3 + 3 \mu_3^2 - 2 \mu_2 \mu_4}{2 (9 \mu_2^3 + 6 \mu_3^2 - 5 \mu_2 \mu_4)}$$

Note that  $a$  and  $c_1$  are now equal; this only applies when one works *about the mean*. With these definitions, the Pearson *Types* of §5.2 C can now be expressed in terms of the first 4 moments, instead of parameters  $a$ ,  $c_0$ ,  $c_1$  and  $c_2$ . This is, in fact, how the various automated Pearson fitting functions are constructed (§5.2 B).

## 5.2 E Higher Order Pearson-Style Families

Instead of basing the Pearson system upon the quadratic  $c_0 + c_1 x + c_2 x^2$ , one can instead consider using higher order polynomials as the foundation stone. If the moments of the population are known, then this endeavour must unambiguously yield a better fit. If, however, the observed data is a random sample drawn from the population, there is a trade-off: a higher order polynomial implies that higher order moments are required, and the estimates of the latter may be unreliable (have high variance), unless the sample size is ‘large’.

In this section, we consider a Pearson-style system based upon a cubic polynomial. This will be the family of solutions  $p(x)$  to the differential equation

$$\frac{dp(x)}{dx} = - \frac{a + x}{c_0 + c_1 x + c_2 x^2 + c_3 x^3} p(x). \quad (5.6)$$

Adopting the method introduced in §5.2 D once again yields a recurrence relation, but now with one extra term on the left-hand side. Equation (5.5) now becomes

$$-r c_0 \dot{\mu}_{r-1} - (r+1) c_1 \dot{\mu}_r - (r+2) c_2 \dot{\mu}_{r+1} - (r+3) c_3 \dot{\mu}_{r+2} = -a \dot{\mu}_r - \dot{\mu}_{r+1}. \quad (5.7)$$

Given the boundary condition  $\dot{\mu}_0 = 1$ , we enter this recurrence relation into *Mathematica* as:

```
eqn2[r_] :=
  (-r c0 \dot{\mu}_{r-1} - (r+1) c1 \dot{\mu}_r - (r+2) c2 \dot{\mu}_{r+1} - (r+3) c3 \dot{\mu}_{r+2} ==
   -a \dot{\mu}_r - \dot{\mu}_{r+1}) /. \dot{\mu}_0 -> 1
```

Our objective is to find  $a, c_0, c_1, c_2$  and  $c_3$  in terms of  $\dot{\mu}_r$ . Putting  $r = 0, 1, 2, 3, 4$  yields the required 5 equations (for the 5 unknowns) which we now solve simultaneously:

```
Z1 = Solve[Table[eqn2[r], {r, 0, 4}], {a, c0, c1, c2, c3}];
```

The solution is rather long, so we will not print it here. However, if we work *about the mean*, taking  $\dot{\mu}_1 = 0$ , and  $\dot{\mu}_r = \mu_r$  for  $r \geq 2$ , the solution reduces to:

```
Z2 = Z1[[1]] /. {\dot{\mu}_1 -> 0, \dot{\mu} -> \mu} // Simplify;
Z2 // TableForm
```

$$a \rightarrow \frac{-117 \mu_2^3 \mu_3 \mu_4 - 16 \mu_3^3 \mu_4 + 81 \mu_4^3 \mu_5 + 5 \mu_4^2 \mu_5 + \mu_2 \mu_3 (25 \mu_4^2 + 24 \mu_3 \mu_5) + 3 \mu_2^2 (16 \mu_3^3 - 18 \mu_4 \mu_5 + 7 \mu_3 \mu_6) + \mu_3 (-12 \mu_2^2 + 7 \mu_4 \mu_6)}{2(96 \mu_3^4 - 27 \mu_2^4 \mu_4 - 50 \mu_4^3 + 93 \mu_3 \mu_4 \mu_5 + 15 \mu_2^2 (7 \mu_4^2 + 9 \mu_3 \mu_5) + 9 \mu_2^3 (2 \mu_3^2 - 7 \mu_6) - 42 \mu_3^2 \mu_6 + \mu_2 (-272 \mu_3^2 \mu_4 - 36 \mu_2^2 + 35 \mu_4 \mu_6))}$$

$$c_0 \rightarrow \frac{-3 \mu_2^3 (4 \mu_4^2 - 3 \mu_3 \mu_5) + 8 \mu_3^3 (-2 \mu_4^2 + 3 \mu_3 \mu_5) + \mu_2 (40 \mu_4^3 - 77 \mu_3 \mu_4 \mu_5 + 21 \mu_3^2 \mu_6) + \mu_2^2 (3 \mu_3^3 \mu_4 + 36 \mu_2^2 - 28 \mu_4 \mu_6)}{2(-96 \mu_3^4 + 27 \mu_2^4 \mu_4 + 50 \mu_4^3 - 93 \mu_3 \mu_4 \mu_5 - 15 \mu_2^2 (7 \mu_4^2 + 9 \mu_3 \mu_5) + 42 \mu_3^2 \mu_6 + \mu_2^3 (-18 \mu_3^2 + 63 \mu_6) + \mu_2 (272 \mu_3^2 \mu_4 + 36 \mu_2^2 - 35 \mu_4 \mu_6))}$$

$$c_1 \rightarrow \frac{-33 \mu_2^3 \mu_3 \mu_4 - 16 \mu_3^3 \mu_4 + 27 \mu_4^3 \mu_5 + 5 \mu_4^2 \mu_5 + \mu_2 \mu_3 (37 \mu_4^2 + 6 \mu_3 \mu_5) + 3 \mu_2^2 (4 \mu_3^3 - 16 \mu_4 \mu_5 + 7 \mu_3 \mu_6) + \mu_3 (-12 \mu_2^2 + 7 \mu_4 \mu_6)}{2(96 \mu_3^4 - 27 \mu_2^4 \mu_4 - 50 \mu_4^3 + 93 \mu_3 \mu_4 \mu_5 + 15 \mu_2^2 (7 \mu_4^2 + 9 \mu_3 \mu_5) + 9 \mu_2^3 (2 \mu_3^2 - 7 \mu_6) - 42 \mu_3^2 \mu_6 + \mu_2 (-272 \mu_3^2 \mu_4 - 36 \mu_2^2 + 35 \mu_4 \mu_6))}$$

$$c_2 \rightarrow \frac{-48 \mu_3^4 + 18 \mu_2^4 \mu_4 + 20 \mu_4^3 - 39 \mu_3 \mu_4 \mu_5 - 3 \mu_2^2 (22 \mu_4^2 + 23 \mu_3 \mu_5) - 6 \mu_2^3 (2 \mu_3^2 - 7 \mu_6) + 21 \mu_3^2 \mu_6 + \mu_2 (143 \mu_3^2 \mu_4 + 12 \mu_2^2 - 14 \mu_4 \mu_6)}{2(-96 \mu_3^4 + 27 \mu_2^4 \mu_4 + 50 \mu_4^3 - 93 \mu_3 \mu_4 \mu_5 - 15 \mu_2^2 (7 \mu_4^2 + 9 \mu_3 \mu_5) + 42 \mu_3^2 \mu_6 + \mu_2^3 (-18 \mu_3^2 + 63 \mu_6) + \mu_2 (272 \mu_3^2 \mu_4 + 36 \mu_2^2 - 35 \mu_4 \mu_6))}$$

$$c_3 \rightarrow \frac{14 \mu_2^2 \mu_3 \mu_4 - 9 \mu_3^3 \mu_5 + \mu_3 (2 \mu_4^2 - 3 \mu_3 \mu_5) + \mu_2 (-6 \mu_3^3 + 4 \mu_4 \mu_5)}{-96 \mu_3^4 + 27 \mu_2^4 \mu_4 + 50 \mu_4^3 - 93 \mu_3 \mu_4 \mu_5 - 15 \mu_2^2 (7 \mu_4^2 + 9 \mu_3 \mu_5) + 42 \mu_3^2 \mu_6 + \mu_2^3 (-18 \mu_3^2 + 63 \mu_6) + \mu_2 (272 \mu_3^2 \mu_4 + 36 \mu_2^2 - 35 \mu_4 \mu_6)}$$

which is comparatively compact (for a more legible rendition, see the electronic notebook). Whereas the second-order (quadratic) Pearson family can be expressed in terms of the first 4 moments, the third-order (cubic) Pearson-style family requires the first 6 moments. Note that  $a$  and  $c_1$  are no longer equal.

⊕ **Example 4:** Fitting a Third-Order (Cubic) Pearson-Style Density

In this example, we fit a third-order (cubic) Pearson-style density to the data set: `marks.dat`. *Example 1* fitted the standard second-order (quadratic) Pearson distribution to this data set. It will be interesting to see how a third-order Pearson-style distribution compares. First, we load the required data set into *Mathematica*, if this has not already been done:

```
data = ReadList ["marks.dat"];
```

The population central moments  $\mu_2, \mu_3, \mu_4, \mu_5$  and  $\mu_6$  are given by:

```
<< Statistics`
 $\hat{\mu}$  = Table [ $\mu_x \rightarrow$  CentralMoment [data, x] // N, {x, 2, 6}]
{ $\mu_2 \rightarrow 193.875, \mu_3 \rightarrow -1125.94, \mu_4 \rightarrow 133550.,$ 
 $\mu_5 \rightarrow -2.68578 \times 10^6, \mu_6 \rightarrow 1.77172 \times 10^8$ }
```

In the quadratic system, this data was of *Type IV* (the most general form). Consequently, in the cubic system, we will once again try the most general solution (*i.e.* without making any assumptions about the roots of the cubic polynomial). The solution then is:

$$\mathbf{DSolve} \left[ \mathbf{p}'[\mathbf{x}] == - \frac{\mathbf{a} + \mathbf{x}}{\mathbf{c0} + \mathbf{c1} \mathbf{x} + \mathbf{c2} \mathbf{x}^2 + \mathbf{c3} \mathbf{x}^3} \mathbf{p}[\mathbf{x}], \mathbf{p}[\mathbf{x}], \mathbf{x} \right]$$

$$\left\{ \left\{ \mathbf{p}[\mathbf{x}] \rightarrow \mathbf{e}^{-\text{RootSum}[\mathbf{c0} + \mathbf{c1} \#1 + \mathbf{c2} \#1^2 + \mathbf{c3} \#1^3 \&, \frac{\mathbf{a} \text{Log}[\mathbf{x} - \#1] + \text{Log}[\mathbf{x} - \#1] \#1}{\mathbf{c1} + 2 \mathbf{c2} \#1 + 3 \mathbf{c3} \#1^2} \&]} \mathbf{C}[1] \right\} \right\}$$

*Mathematica* provides the solution in terms of a `RootSum` object. If we now replace the Pearson coefficients  $\{a, c_0, c_1, c_2, c_3\}$  with central moments  $\{\mu_2, \mu_3, \mu_4, \mu_5, \mu_6\}$  via `Z2` derived above, and then replace the latter with the empirical  $\hat{\mu}$ , we obtain:

```
sol = e-RootSum [c0+c1 #1+c2 #12+c3 #13 &,  $\frac{a \text{Log}[\mathbf{x} - \#1] + \text{Log}[\mathbf{x} - \#1] \#1}{\mathbf{c1} + 2 \mathbf{c2} \#1 + 3 \mathbf{c3} \#1^2} \&]$  &] / .
Z2 /.  $\hat{\mu}$  // Simplify
((-31.6478 - 52.712 i) + x)-9.86369+6.66825 i
((-31.6478 + 52.712 i) + x)-9.86369-6.66825 i (556.021 + x)19.7274
```

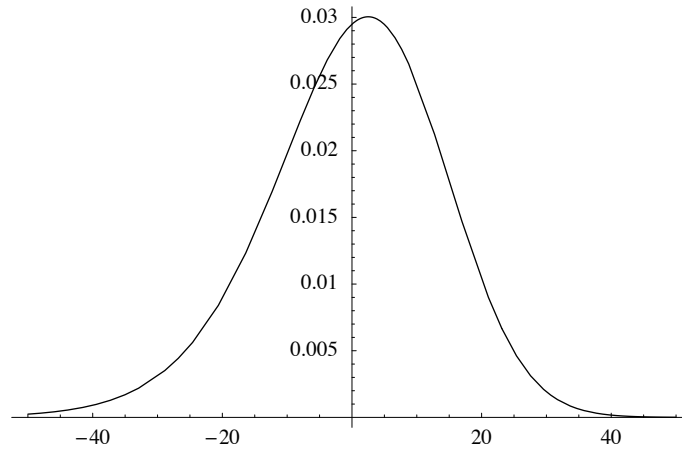
while the constant of integration over, say,  $\{x, -100, 100\}$  is:

```
cn = NIntegrate [sol, {x, -100, 100}]
4.22732  $\times 10^{32}$ 
```



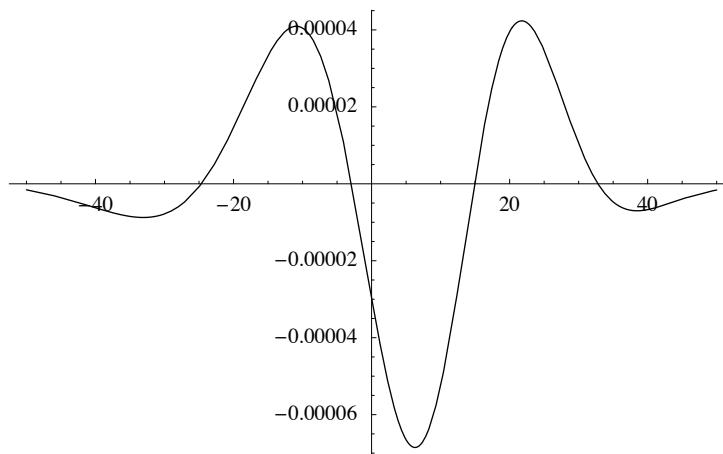
A quick plot illustrates:

```
Plot[sol / cn, {x, -50, 50}];
```



**Fig. 8:** Cubic Pearson fit for the marks data set

This looks identical to the plot of  $f$  derived in *Example 1*, except the origin is now at zero, rather than at the mean. If  $f$  from *Example 1* is derived with zero mean, one can then `Plot[f-sol/cn, {x, -50, 50}]` to see the difference between the two solutions. Doing so yields Fig. 9.



**Fig. 9:** The difference between the quadratic and cubic Pearson fit

The difference between the plots is remarkably small (note the scale on the vertical axis). This outcome is rather reassuring for those who prefer to use the much simpler quadratic Pearson system. ■

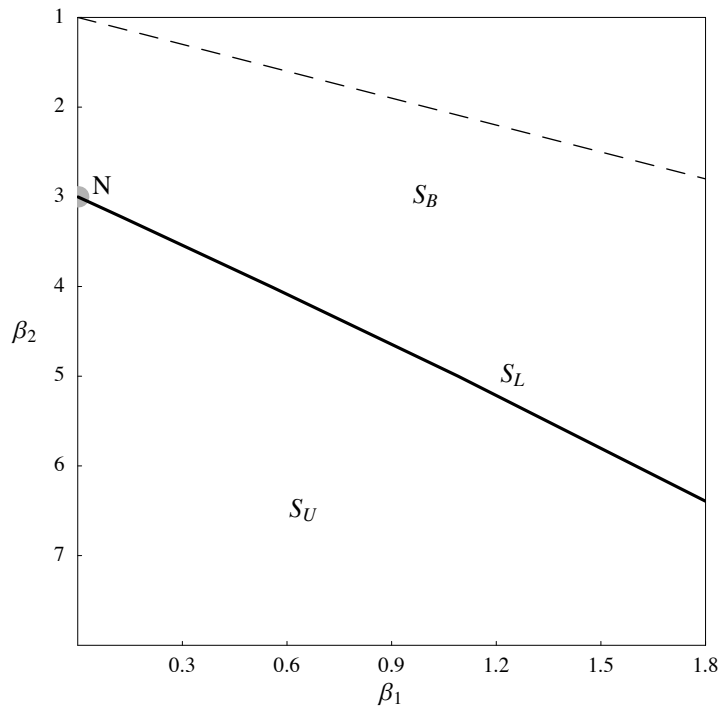
## 5.3 Johnson Transformations

### 5.3 A Introduction

Recall that the Pearson family provides a unique distribution for every possible  $(\beta_1, \beta_2)$  combination. The Johnson family provides the same feature, and does so by using a set of three transformations of the standard Normal. In particular, if  $Z \sim N(0, 1)$  with density  $\phi(z)$ , and  $Y$  is a transform of  $Z$ , then the Johnson family is given by:

- (1)  $S_L$  (Lognormal)  $Y = \exp\left(\frac{Z-\gamma}{\delta}\right) \Leftrightarrow Z = \gamma + \delta \log(Y) \quad (0 < y < \infty)$
- (2)  $S_U$  (Unbounded)  $Y = \sinh\left(\frac{Z-\gamma}{\delta}\right) \Leftrightarrow Z = \gamma + \delta \sinh^{-1}(Y) \quad (-\infty < y < \infty)$
- (3)  $S_B$  (Bounded)  $Y = \frac{1}{1 + \exp\left(-\frac{Z-\gamma}{\delta}\right)} \Leftrightarrow Z = \gamma + \delta \log\left(\frac{Y}{1-Y}\right) \quad (0 < y < 1)$

Applying a second transform  $X = \xi + \lambda Y$  (or equivalently  $Y = \frac{X-\xi}{\lambda}$ ) expands the system from two parameters  $(\gamma, \delta)$  to the full set of four  $(\gamma, \delta, \xi, \lambda)$ , where  $\delta$  and  $\lambda$  are taken to be positive. Since  $X = \xi + \lambda Y$ , the shape of the distribution of  $X$  will be the same as that of  $Y$ . Hence, the parameters may be interpreted as follows:  $\gamma$  and  $\delta$  determine the shape of the distribution of  $X$ ;  $\lambda$  is a scale factor; and  $\xi$  is a location factor. Figure 10 illustrates the classification system in  $(\beta_1, \beta_2)$  space.



**Fig. 10:** The  $\beta_1, \beta_2$  chart for the Johnson system

Several points are of note:

- (i) The classification consists of two *main* types, namely  $S_U$  and  $S_B$ . These are separated by a *transition* type, the  $S_L$  line, which corresponds to the family of Lognormal distributions. The N at  $(\beta_1, \beta_2) = (0, 3)$  once again denotes the Normal distribution, which may be thought of as a limiting form of the three systems as  $\delta \rightarrow \infty$ .
- (ii) The  $S_U$  system is termed *unbounded* because the domain here is  $\{y: y \in \mathbb{R}\}$ . The  $S_B$  system is termed *bounded* because the domain for this system is  $\{y: 0 < y < 1\}$ .
- (iii) The dashed line represents the bound on all distributions, and is given by  $\beta_2 - \beta_1 = 1$ .

Whereas the Pearson system can be easily ‘automated’ for fitting purposes, the Johnson system requires some hands-on fine tuning. We consider each system in turn:  $S_L$  (§5.3 B);  $S_U$  (§5.3 C); and  $S_B$  (§5.3 D).

### 5.3 B $S_L$ System (Lognormal)

Let  $Z \sim N(0, 1)$  with density  $\phi(z)$ :

$$\phi = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}; \quad \text{domain}[\phi] = \{z, -\infty, \infty\} \ \&\& \ \{\gamma \in \text{Reals}, \delta > 0\};$$

The  $S_L$  system is defined by the transformation  $Y = \exp\left(\frac{Z-\gamma}{\delta}\right)$ . Then, the density of  $Y$ , say  $g(y)$ , is:

$$g = \text{Transform}\left[y == e^{\frac{z-\gamma}{\delta}}, \phi\right]$$

$$\text{domain}[g] = \text{TransformExtremum}\left[y == e^{\frac{z-\gamma}{\delta}}, \phi\right]$$

$$\frac{e^{-\frac{1}{2}(\gamma + \delta \text{Log}[y])^2} \delta}{\sqrt{2\pi} y}$$

$$\{y, 0, \infty\} \ \&\& \ \{\gamma \in \text{Reals}, \delta > 0\}$$

The Lognormal density is positively skewed, though as  $\delta$  increases, the curve tends to symmetry. In Fig. 11, the density on the far left corresponds to a ‘small’  $\delta$ , while each successive density to the right corresponds to a doubling of  $\delta$ .

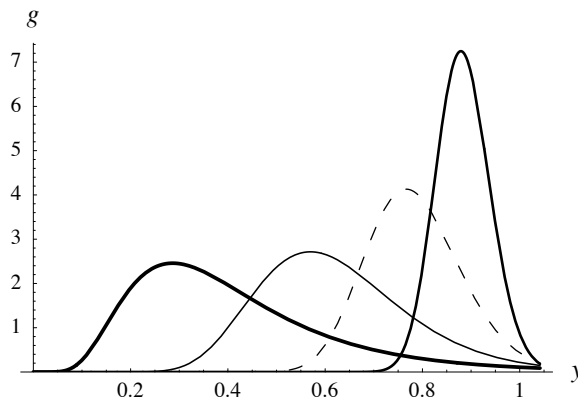


Fig. 11: The Lognormal pdf  $g(y)$  when  $\gamma = 2$ , and  $\delta = 2, 4, 8$  and  $16$

Since  $Y = \exp\left(\frac{Z-\gamma}{\delta}\right)$ , and  $Z$  has density  $\phi(z)$ , the  $r^{\text{th}}$  raw moment  $E[Y^r]$  can be expressed as:

$$\Omega = \mathbf{Expect} \left[ e^{\frac{(z-\gamma)r}{\delta}}, \phi \right]$$

$$e^{\frac{r(z-\gamma)\delta}{2\sigma^2}}$$

Thus, the first 4 raw moments (rm) are:

$$\mathbf{rm} = \mathbf{Table} \left[ \mu_r \rightarrow \Omega, \{\mathbf{r}, 4\} \right]$$

$$\left\{ \mu_1 \rightarrow e^{\frac{1-2\gamma\delta}{2\sigma^2}}, \mu_2 \rightarrow e^{\frac{2-2\gamma\delta}{\sigma^2}}, \mu_3 \rightarrow e^{\frac{3(3-2\gamma\delta)}{2\sigma^2}}, \mu_4 \rightarrow e^{\frac{2(4-2\gamma\delta)}{\sigma^2}} \right\}$$

This can be expressed in terms of central moments (cm), as follows:

$$\mathbf{cm} = \mathbf{Table} \left[ \mathbf{CentralToRaw}[\mathbf{r}] /. \mathbf{rm} // \mathbf{Simplify}, \{\mathbf{r}, 2, 4\} \right];$$

**cm // TableForm**

$$\mu_2 \rightarrow e^{\frac{1-2\gamma\delta}{\sigma^2}} \left( -1 + e^{\frac{1}{\sigma^2}} \right)$$

$$\mu_3 \rightarrow e^{\frac{3-6\gamma\delta}{2\sigma^2}} \left( -1 + e^{\frac{1}{\sigma^2}} \right)^2 \left( 2 + e^{\frac{1}{\sigma^2}} \right)$$

$$\mu_4 \rightarrow e^{\frac{2-4\gamma\delta}{\sigma^2}} \left( -1 + e^{\frac{1}{\sigma^2}} \right)^2 \left( -3 + 3 e^{\frac{2}{\sigma^2}} + 2 e^{\frac{3}{\sigma^2}} + e^{\frac{4}{\sigma^2}} \right)$$

Then  $\beta_1$  and  $\beta_2$  can be expressed as:

$$\beta_1 = \frac{\mu_3}{\mu_2^2} /. \mathbf{cm} // \mathbf{Simplify}$$

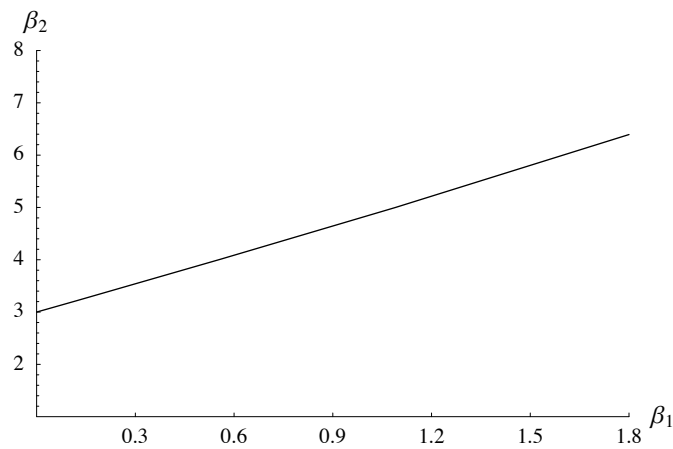
$$\left( -1 + e^{\frac{1}{\sigma^2}} \right) \left( 2 + e^{\frac{1}{\sigma^2}} \right)^2$$

and

$$\beta_2 = \frac{\mu_4}{\mu_2^2} /. \mathbf{cm} // \mathbf{Simplify}$$

$$-3 + 3 e^{\frac{2}{\sigma^2}} + 2 e^{\frac{3}{\sigma^2}} + e^{\frac{4}{\sigma^2}}$$

These equations define the Lognormal curve parametrically in  $(\beta_1, \beta_2)$  space, as  $\delta$  increases from 0 to  $\infty$ , as Fig.12 illustrates. In *Mathematica*, one can use `ParametricPlot` to derive this curve.



**Fig. 12:** The Lognormal curve in  $(\beta_1, \beta_2)$  space

This is identical to the  $S_L$  curve shown in Fig. 10 (The Johnson Plot), except that the vertical axis is not inverted here. Despite appearances, the curve in Fig. 12 is not linear; this is easy to verify with a ruler. In the limit, as  $\delta \rightarrow \infty$ ,  $\beta_1$  and  $\beta_2$  tend to 0 and 3, respectively:

```
Limit[{beta_1, beta_2}, delta -> infinity]
{0, 3}
```

so that the Normal distribution is obtained as a limit case of the Lognormal.

Given an empirical value for  $\beta_1$  (or  $\beta_2$ ), we can now ‘solve’ for  $\delta$ . This is particularly easy since  $\gamma$  is not required. For instance, if  $\hat{\beta}_1 = 0.829$ :

```
Solve[beta_1 == 0.829, delta]
```

```
- Solve::ifun : Inverse functions are being
  used by Solve, so some solutions may not be found.

{ {delta -> -3.46241} ,
  {delta -> -0.457213 - 0.354349 i} ,
  {delta -> -0.457213 + 0.354349 i} ,
  {delta -> 0.457213 - 0.354349 i} ,
  {delta -> 0.457213 + 0.354349 i} ,
  {delta -> 3.46241} }
```

Since we require  $\delta$  to be both real and positive, only the last of these solutions is feasible. One can now find  $\gamma$  by comparing  $\mu_2$  (derived above) with its empirical estimate  $\hat{\mu}_2$ .

### 5.3 C $S_U$ System (Unbounded)

Once again, let  $Z \sim N(0, 1)$  with density  $\phi(z)$ :

$$\phi = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}; \quad \text{domain}[\phi] = \{z, -\infty, \infty\} \ \&\& \ \{\gamma \in \text{Reals}, \delta > 0\};$$

The  $S_U$  system is defined by the transformation  $Y = \sinh\left(\frac{Z-\gamma}{\delta}\right)$ . Hence, the density of  $Y$ , say  $g(y)$ , is:

$$\begin{aligned} g &= \text{Transform}[y == \text{Sinh}\left[\frac{z-\gamma}{\delta}\right], \phi] \\ \text{domain}[g] &= \text{TransformExtremum}[y == \text{Sinh}\left[\frac{z-\gamma}{\delta}\right], \phi] \\ &= \frac{e^{-\frac{1}{2}(\gamma + \delta \text{ArcSinh}[y])^2} \delta}{\sqrt{2\pi} \sqrt{1+y^2}} \\ &\{y, -\infty, \infty\} \ \&\& \ \{\gamma \in \text{Reals}, \delta > 0\} \end{aligned}$$

Figure 13 indicates shapes that are typical in the  $S_U$  family.

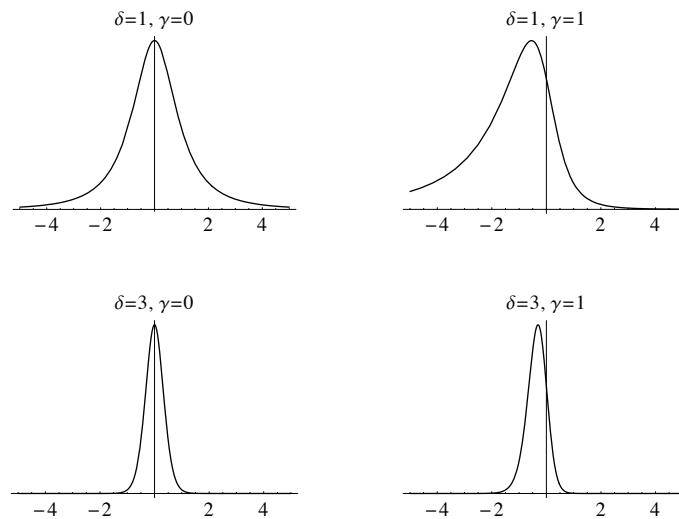


Fig. 13: Typical pdf shapes in the  $S_U$  family

Since  $Y = \sinh\left(\frac{Z-\gamma}{\delta}\right)$ , and  $Z$  has density  $\phi(z)$ , the  $r^{\text{th}}$  moment  $E[Y^r]$  can be expressed:

$$\Omega := \text{Expect}\left[\text{Sinh}\left[\frac{z-\gamma}{\delta}\right]^r, \phi\right] // \text{ExpToTrig} // \text{FullSimplify}$$

This time, *Mathematica* cannot find the solution as a function of  $r$ , which is why we use a delayed evaluation ( $:=$ ) instead of an immediate evaluation ( $=$ ).

The first 4 raw moments (rm) are now given by:

$$\mathbf{rm} = \mathbf{Table}[\mu_r \rightarrow \Omega, \{\mathbf{r}, 4\}]; \quad \mathbf{rm} // \mathbf{TableForm}$$

$$\begin{aligned} \mu_1 &\rightarrow -e^{\frac{1}{2\delta^2}} \sinh\left[\frac{\gamma}{\delta}\right] \\ \mu_2 &\rightarrow \frac{1}{2} \left(-1 + e^{\frac{2}{\delta^2}} \cosh\left[\frac{2\gamma}{\delta}\right]\right) \\ \mu_3 &\rightarrow -\frac{1}{4} e^{\frac{1}{2\delta^2}} \left(-3 \sinh\left[\frac{\gamma}{\delta}\right] + e^{\frac{4}{\delta^2}} \sinh\left[\frac{3\gamma}{\delta}\right]\right) \\ \mu_4 &\rightarrow \frac{1}{8} \left(3 - 4 e^{\frac{2}{\delta^2}} \cosh\left[\frac{2\gamma}{\delta}\right] + e^{\frac{8}{\delta^2}} \cosh\left[\frac{4\gamma}{\delta}\right]\right) \end{aligned}$$

This can be expressed in terms of central moments (cm), as follows:<sup>3</sup>

$$\mathbf{cm} = \mathbf{Table}[\mathbf{CentralToRaw}[\mathbf{r}] /. \mathbf{rm} // \mathbf{FullSimplify}, \{\mathbf{r}, 2, 4\}]$$

$$\begin{aligned} \{\mu_2 &\rightarrow \frac{1}{2} \left(-1 + e^{\frac{1}{\delta^2}}\right) \left(1 + e^{\frac{1}{\delta^2}} \cosh\left[\frac{2\gamma}{\delta}\right]\right), \\ \mu_3 &\rightarrow -\frac{1}{4} e^{\frac{1}{2\delta^2}} \left(-1 + e^{\frac{1}{\delta^2}}\right)^2 \left(3 \sinh\left[\frac{\gamma}{\delta}\right] + e^{\frac{1}{\delta^2}} \left(2 + e^{\frac{1}{\delta^2}}\right) \sinh\left[\frac{3\gamma}{\delta}\right]\right), \\ \mu_4 &\rightarrow \frac{1}{8} \left(3 + e^{\frac{2}{\delta^2}} \left(e^{\frac{6}{\delta^2}} \cosh\left[\frac{4\gamma}{\delta}\right] + 4 \cosh\left[\frac{2\gamma}{\delta}\right] \left(-1 + 6 e^{\frac{1}{\delta^2}} \sinh\left[\frac{\gamma}{\delta}\right]^2\right) - \right. \right. \\ &\quad \left. \left. 8 \sinh\left[\frac{\gamma}{\delta}\right] \left(3 \sinh\left[\frac{\gamma}{\delta}\right]^3 + e^{\frac{3}{\delta^2}} \sinh\left[\frac{3\gamma}{\delta}\right]\right)\right)\right) \} \end{aligned}$$

Then  $\beta_1$  and  $\beta_2$  can be expressed as:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} /. \mathbf{cm} // \mathbf{Simplify}$$

$$\frac{e^{\frac{1}{\delta^2}} \left(-1 + e^{\frac{1}{\delta^2}}\right) \left(3 \sinh\left[\frac{\gamma}{\delta}\right] + e^{\frac{1}{\delta^2}} \left(2 + e^{\frac{1}{\delta^2}}\right) \sinh\left[\frac{3\gamma}{\delta}\right]\right)^2}{2 \left(1 + e^{\frac{1}{\delta^2}} \cosh\left[\frac{2\gamma}{\delta}\right]\right)^3}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} /. \mathbf{cm} // \mathbf{Simplify}$$

$$\frac{3 + e^{\frac{2}{\delta^2}} \left(e^{\frac{6}{\delta^2}} \cosh\left[\frac{4\gamma}{\delta}\right] + 4 \cosh\left[\frac{2\gamma}{\delta}\right] \left(-1 + 6 e^{\frac{1}{\delta^2}} \sinh\left[\frac{\gamma}{\delta}\right]^2\right) - 8 \sinh\left[\frac{\gamma}{\delta}\right] \left(3 \sinh\left[\frac{\gamma}{\delta}\right]^3 + e^{\frac{3}{\delta^2}} \sinh\left[\frac{3\gamma}{\delta}\right]\right)\right)}{2 \left(-1 + e^{\frac{1}{\delta^2}}\right)^2 \left(1 + e^{\frac{1}{\delta^2}} \cosh\left[\frac{2\gamma}{\delta}\right]\right)^2}$$

#### o *Fitting the $S_U$ System*

To fit the  $S_U$  system, we adopt the following steps:

- (i) Given values for  $(\beta_1, \beta_2)$ , solve for  $(\delta, \gamma)$ , noting that  $\delta > 0$ , and that the sign of  $\gamma$  is opposite to that of  $\mu_3$ .
- (ii) This gives us  $g(y | \gamma, \delta)$ . Given the transform  $X = \xi + \lambda Y$ , solve for  $\xi$ , and  $\lambda > 0$ .

⊕ **Example 5:** Fit a Johnson Density to the `marks.dat` Population Data Set

First, load the data set, if this has not already been done:

```
data = ReadList["marks.dat"];
```

The mean of this data set is:

```
mean = SampleMean[data] // N
```

```
58.9024
```

Empirical values for  $\mu_2$ ,  $\mu_3$  and  $\mu_4$  are once again given by:

```
<< Statistics`
```

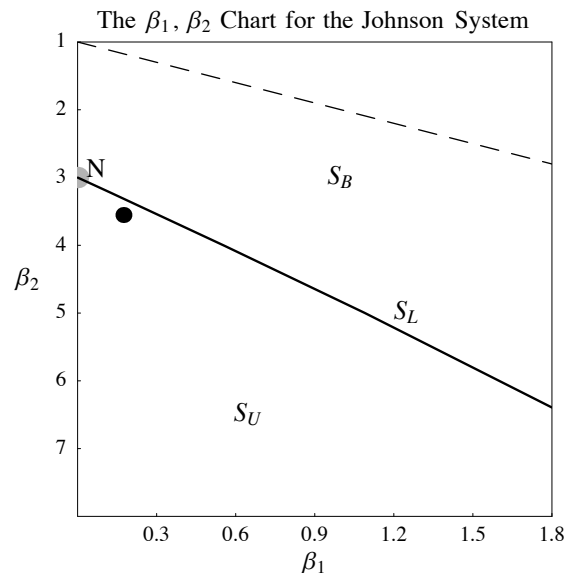
```
 $\mu_{234} = Table[CentralMoment[data, r], {r, 2, 4}] // N$ 
```

```
{193.875, -1125.94, 133550.}
```

If we were working with sample data, we would replace the `CentralMoment` function with `UnbiasedCentralMoment` (just cut and paste). Just as `PearsonPlot` was used in *Example 1* to indicate the appropriate *Pearson Type*, we now use `JohnsonPlot` to indicate which of the Johnson systems is suitable for this data set:

```
JohnsonPlot[ $\mu_{234}$ ];
```

```
{ $\beta_1 \rightarrow 0.173966$ ,  $\beta_2 \rightarrow 3.55303$ }
```



**Fig. 14:** The marks data lies in the  $S_U$  system

The black dot, depicting  $(\beta_1, \beta_2)$  for this data set, lies in the  $S_U$  system. We derived  $\beta_1$  and  $\beta_2$  in terms of  $\delta$  and  $\gamma$  above. Thus, given values  $\{\beta_1 \rightarrow 0.173966, \beta_2 \rightarrow 3.55303\}$ , it is



now possible to ‘solve’ for  $(\delta, \gamma)$ . The `FindRoot` function simultaneously solves the two equations for  $\delta$  and  $\gamma$ :

```
sol = FindRoot [
      {  $\beta_1 == 0.17396604431160143`$ ,
         $\beta_2 == 3.5530347934625883`$  }, { $\delta, 2$ }, { $\gamma, 2$ }]
{ $\delta \rightarrow 3.74767$ ,  $\gamma \rightarrow 2.0016$ }
```

Note that `FindRoot` is a numerical technique that returns the first solution it finds, so different starting points may yield different solutions. In evaluating the solution, it helps to note that  $\delta$  should be positive, while  $\gamma$  should be opposite in sign to  $\mu_3$ . Johnson (1949, p. 164) and Johnson *et al.* (1994, p. 36) provide a diagram known as an *abac* that provides a rough estimate of  $\gamma$  and  $\delta$ , given values for  $\beta_1$  and  $\beta_2$ . These rough estimates make an excellent starting point for the `FindRoot` function. In a similar vein, see Bowman and Shenton (1980).

The full 4-parameter  $(\gamma, \delta, \xi, \lambda)$  Johnson  $S_U$  system is obtained by applying the further transformation  $X = \xi + \lambda Y$  or equivalently  $Y = \frac{X-\xi}{\lambda}$ . Since we are adding two new parameters, we shall add some assumptions about them:

```
domain[g] = domain[g] && { $\xi \in \text{Reals}$ ,  $\lambda > 0$ };
```

Then the density of  $X = \xi + \lambda Y$ , say  $f(x)$ , is:

```
f = Transform[x ==  $\xi + \lambda y$ , g]
domain[f] = TransformExtremum[x ==  $\xi + \lambda y$ , g]


$$\frac{e^{-\frac{1}{2}(\gamma + \delta \text{ArcSinh}[\frac{x-\xi}{\lambda}])^2} \delta}{\sqrt{2\pi} \lambda \sqrt{1 + \frac{(x-\xi)^2}{\lambda^2}}}$$


{x,  $-\infty$ ,  $\infty$ } && { $\gamma \in \text{Reals}$ ,  $\delta > 0$ ,  $\xi \in \text{Reals}$ ,  $\lambda > 0$ }
```

where  $\gamma$  and  $\delta$  have already been found. Since  $X = \xi + \lambda Y$ ,  $\text{Var}(X) = \lambda^2 \text{Var}(Y)$ . Here,  $\text{Var}(Y)$  was found above as  $\mu_2(\gamma, \delta)$  (part of `cm`), while  $\text{Var}(X)$  is taken to be the empirical variance 193.875 of the data set. Thus, at the fitted values, the equation  $\text{Var}(X) = \lambda^2 \text{Var}(Y)$  becomes:

```
193.875 ==  $\lambda^2 \mu_2$  /. cm /. sol
193.875 == 0.101355  $\lambda^2$ 
```

Solving for  $\lambda$  yields:

```
 $\hat{\lambda} = \text{Solve}[\%, \lambda]$ 
{{ $\lambda \rightarrow -43.7359$ }, { $\lambda \rightarrow 43.7359$ }}
```

Since we require  $\lambda > 0$ , the second solution is the desired one. That leaves  $\xi$  ...

Since  $X = \xi + \lambda Y$ ,  $E[X] = \xi + \lambda E[Y]$ . Here,  $E[Y]$  was found above as  $\hat{\mu}_1(\gamma, \delta)$  (part of `rm`), while  $E[X]$  is taken to be the empirical mean of the data set. Thus, at the fitted values,  $E[X] = \xi + \lambda E[Y]$  becomes:

$$\text{mean} == \xi + \lambda \hat{\mu}_1 /. \text{rm} /. \text{sol} /. \hat{\lambda}[[2]]$$

$$58.9024 == -25.3729 + \xi$$

Solving for  $\xi$  yields:

$$\hat{\xi} = \text{Solve}[\%, \xi]$$

$$\{\{\xi \rightarrow 84.2752\}\}$$

The desired fitted density  $f(x)$  is thus:

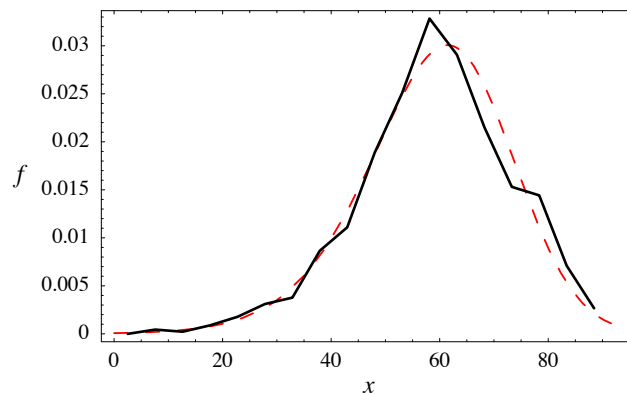
$$\mathbf{f} = \mathbf{f} /. \text{sol} /. \hat{\lambda}[[2]] /. \hat{\xi}[[1]]$$

$$\frac{0.0341848 e^{-\frac{1}{2} (2.0016 + 3.74767 \text{ArcSinh}[0.0228645 (-84.2752 + x)])^2}}{\sqrt{1 + 0.000522787 (-84.2752 + x)^2}}$$

which has an unbounded domain, like all  $S_U$  distributions.

As in *Example 1*, the **mathStatica** function `FrequencyPlot` allows one to compare the fitted density with the empirical pdf of the data:

```
p2 = FrequencyPlot[data, f];
```



**Fig. 15:** The empirical pdf (—) and the fitted Johnson  $S_U$  pdf (---)

This Johnson  $S_U$  fitted density appears almost identical to the `PearsonIV` fit derived in *Example 1*. The final diagram in *Example 1* was labelled `p1`. If `p1` is still in memory, the command `Show[p1/.Hue[___]→Hue[.4], p2]` shows both plots together, but now with the fitted Pearson curve in green rather than red, enabling a visual comparison (note that `Hue[___]` contains two `_` characters). The curves are so similar that only a tiny tinge of green would be visible on screen. ■

### 5.3 D $S_B$ System (Bounded)

Once again, let  $Z \sim N(0, 1)$  with density  $\phi(z)$ :

$$\phi = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}; \quad \text{domain}[\phi] = \{z, -\infty, \infty\} \ \&\& \ \{\gamma \in \text{Reals}, \delta > 0\};$$

The  $S_B$  (bounded) system is defined by the transformation  $Y = (1 + \exp(-\frac{Z-\gamma}{\delta}))^{-1}$ . Then, the density of  $Y$ , say  $g(y)$ , is:

$$\begin{aligned} \mathbf{g} &= \text{Transform}[\mathbf{y} == (1 + e^{-\frac{z-\gamma}{\delta}})^{-1}, \phi] \\ \text{domain}[\mathbf{g}] &= \text{TransformExtremum}[\mathbf{y} == (1 + e^{-\frac{z-\gamma}{\delta}})^{-1}, \phi] \\ &= \frac{e^{-\frac{1}{2}(\gamma - \delta \text{Log}[-1 + \frac{1}{y}])^2} \delta}{\sqrt{2\pi} (y - y^2)} \\ &\{y, 0, 1\} \ \&\& \ \{\gamma \in \text{Reals}, \delta > 0\} \end{aligned}$$

The full 4-parameter  $(\gamma, \delta, \xi, \lambda)$  Johnson  $S_B$  system is obtained by applying the further transformation  $X = \xi + \lambda Y$  or equivalently  $Y = \frac{X-\xi}{\lambda}$ . Since we are adding two new parameters, we shall add some assumptions about them:

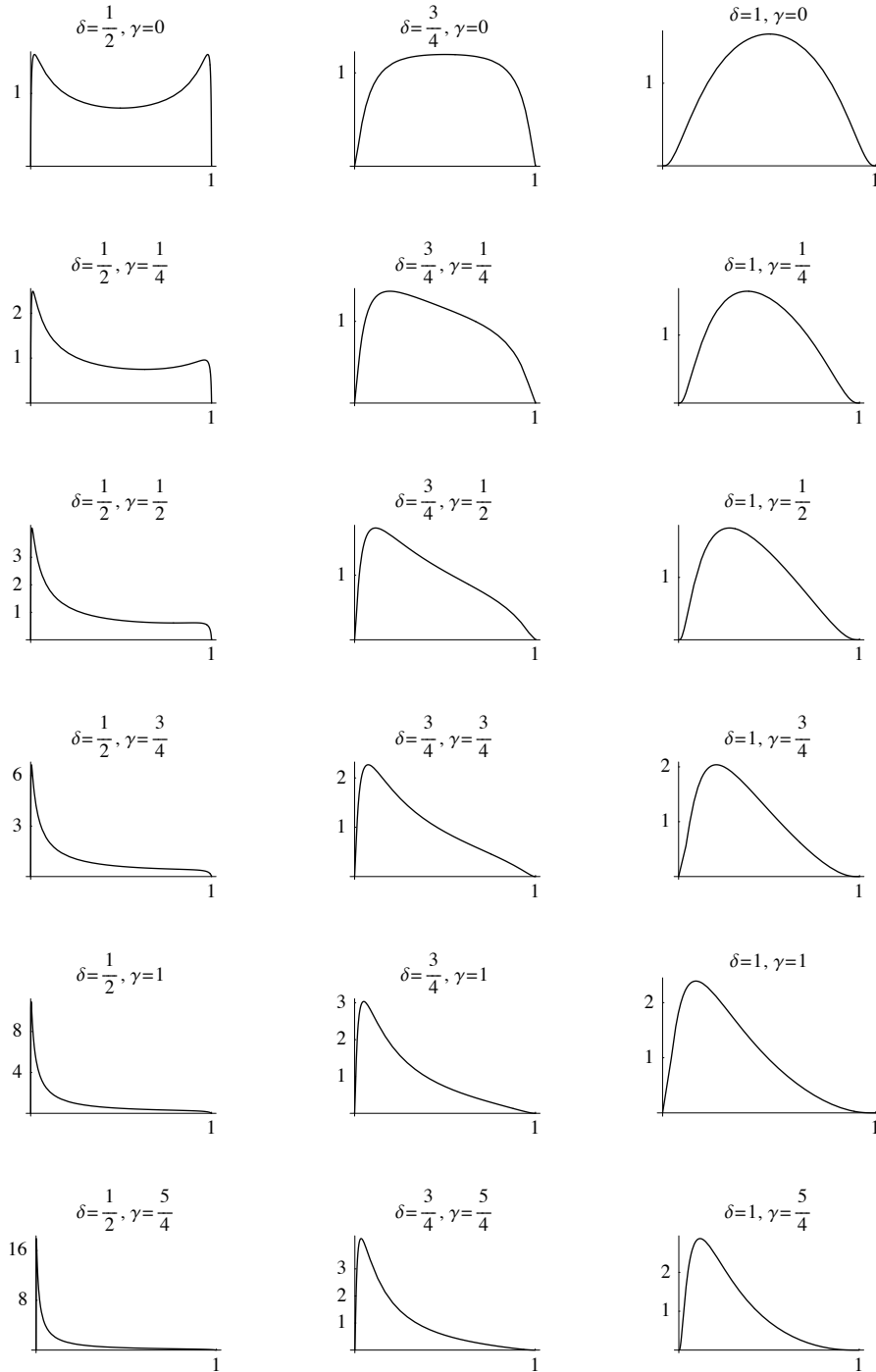
$$\text{domain}[\mathbf{g}] = \text{domain}[\mathbf{g}] \ \&\& \ \{\xi \in \text{Reals}, \lambda > 0\};$$

Then the density of  $X$ , say  $f(x)$ , is:

$$\begin{aligned} \mathbf{f} &= \text{Transform}[\{\mathbf{x} == \xi + \lambda \mathbf{y}\}, \mathbf{g}] \\ \text{domain}[\mathbf{f}] &= \text{TransformExtremum}[\{\mathbf{x} == \xi + \lambda \mathbf{y}\}, \mathbf{g}] \\ &= \frac{e^{-\frac{1}{2}(\gamma - \delta \text{Log}[-1 + \frac{\lambda}{x-\xi}])^2} \delta \lambda}{\sqrt{2\pi} (x - \xi) (-x + \lambda + \xi)} \\ &\{x, \xi, \lambda + \xi\} \ \&\& \ \{\gamma \in \text{Reals}, \delta > 0, \xi \in \text{Reals}, \lambda > 0\} \end{aligned}$$

Figure 16 shows some plots from the  $S_B$   $(\gamma, \delta)$  family.

The moments of the  $S_B$  system are extremely complicated. Johnson (1949) obtained a solution for  $\mu_1$ , though this does not have a closed form; nor can it be implemented usefully in *Mathematica*. As such, the method of moments is not generally used for fitting  $S_B$  systems. Instead, a method of percentile points is used, which equates percentile points of the observed and fitted curves. This approach is not an exact methodology, and we refer the interested reader to Johnson (1949) or Elderton and Johnson (1969, p.131). Alternatively, one can always use the automated Pearson fitting functions as a substitute, which is inevitably a much simpler strategy.



**Fig. 16:** Some pdf shapes in the  $S_B$  family

## 5.4 Gram–Charlier Expansions

### 5.4 A Definitions and Fitting

Let  $\phi(z)$  denote a standard Normal density:

$$\phi = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}; \quad \text{domain}[\phi] = \{z, -\infty, \infty\};$$

and let  $\psi(z)$  denote an arbitrary pdf that has been standardised so that its mean is 0 and variance is 1. If  $\psi(z)$  can be expanded as a series of derivatives of  $\phi(z)$ , then

$$\psi(z) = \sum_{j=0}^{\infty} c_j (-1)^j \frac{d^j \phi(z)}{d z^j}. \quad (5.8)$$

This assumes the expansion is convergent—Stuart and Ord (1994, Section 6.22) provide conditions in this regard. Further, let  $H_j(z) = \frac{(-1)^j}{\phi(z)} \frac{d^j \phi(z)}{d z^j}$ ;  $H_j(z)$  is known as a Hermite polynomial and has a number of interesting properties (see §5.4 B). Then (5.8) may be written as

$$\psi(z) = \phi(z) \sum_{j=0}^{\infty} c_j H_j(z). \quad (5.9)$$

Then, for sufficiently large  $t$ ,  $\psi(z) \approx \phi(z) \sum_{j=0}^t c_j H_j(z)$ . In *Mathematica*, we explicitly model this as a function of  $t$ :

$$\psi[t\_ ] := \phi \sum_{j=0}^t c[j] H[j]$$

This has two components: (i)  $H_j(z)$  and (ii)  $c_j$ .

(i) The Hermite polynomial  $H_j(z)$  is defined by:<sup>4</sup>

$$H[j\_ ] := \frac{(-1)^j}{\phi} \partial_{\{z, j\}} \phi \quad // \text{Expand}$$

Then the first few Hermite polynomials are:

```
Table[H_j -> H[j], {j, 0, 10}]
// TableForm // TraditionalForm
```

$$H_0 \rightarrow 1$$

$$H_1 \rightarrow z$$

$$H_2 \rightarrow z^2 - 1$$

$$H_3 \rightarrow z^3 - 3z$$

$$H_4 \rightarrow z^4 - 6z^2 + 3$$

$$H_5 \rightarrow z^5 - 10z^3 + 15z$$

$$H_6 \rightarrow z^6 - 15z^4 + 45z^2 - 15$$

$$H_7 \rightarrow z^7 - 21z^5 + 105z^3 - 105z$$

$$H_8 \rightarrow z^8 - 28z^6 + 210z^4 - 420z^2 + 105$$

$$H_9 \rightarrow z^9 - 36z^7 + 378z^5 - 1260z^3 + 945z$$

$$H_{10} \rightarrow z^{10} - 45z^8 + 630z^6 - 3150z^4 + 4725z^2 - 945$$

- (ii) The  $c_j$  terms are formally derived in §5.4 B where it is shown that  $c_j$  is a function of the first  $j$  moments of  $\psi(z)$ . Since we are basing the expansion on  $\phi(z)$  (a standardised Normal),  $c_j$  is given here in terms of standardised moments (*i.e.* assuming  $\mu_1 = \mu_1 = 0$ ,  $\mu_2 = 1$ ). The solution takes a similar functional form to  $H_j(x)$ , which we can exploit in *Mathematica* through pattern matching:

$$\mathbf{c}[j\_]:= \frac{\mathbf{H}[j]}{j!} /. \mathbf{z}^{i\_} \rightarrow \mu_i /. \{\mu_1 \rightarrow 0, \mu_2 \rightarrow 1\}$$

The first few  $c_j$  terms are given by:

**Table[c<sub>j</sub> → c[j], {j, 0, 10}] // TableForm**

$$c_0 \rightarrow 1$$

$$c_1 \rightarrow 0$$

$$c_2 \rightarrow 0$$

$$c_3 \rightarrow \frac{\mu_3}{6}$$

$$c_4 \rightarrow \frac{1}{24} (-3 + \mu_4)$$

$$c_5 \rightarrow \frac{1}{120} (-10 \mu_3 + \mu_5)$$

$$c_6 \rightarrow \frac{1}{720} (30 - 15 \mu_4 + \mu_6)$$

$$c_7 \rightarrow \frac{105 \mu_3 - 21 \mu_5 + \mu_7}{5040}$$

$$c_8 \rightarrow \frac{-315 + 210 \mu_4 - 28 \mu_6 + \mu_8}{40320}$$

$$c_9 \rightarrow \frac{-1260 \mu_3 + 378 \mu_5 - 36 \mu_7 + \mu_9}{362880}$$

$$c_{10} \rightarrow \frac{3780 - 3150 \mu_4 + 630 \mu_6 - 45 \mu_8 + \mu_{10}}{3628800}$$

We can now evaluate the *Mathematica* function  $\psi[t]$  for arbitrarily large  $t$ , as a function of the first  $t$  (standardised) moments of  $\psi(z)$ . Here is an example with  $t = 7$ :

$\psi[7]$

$$\frac{1}{\sqrt{2\pi}} \left( e^{-\frac{z^2}{2}} \left( 1 + \frac{1}{6} (-3z + z^3) \mu_3 + \frac{1}{24} (3 - 6z^2 + z^4) (-3 + \mu_4) + \frac{1}{120} (15z - 10z^3 + z^5) (-10\mu_3 + \mu_5) + \frac{1}{720} (-15 + 45z^2 - 15z^4 + z^6) (30 - 15\mu_4 + \mu_6) + \frac{(-105z + 105z^3 - 21z^5 + z^7) (105\mu_3 - 21\mu_5 + \mu_7)}{5040} \right) \right)$$

⊕ **Example 6:** Fit a Gram–Charlier Density to the marks.dat Population Data

First, load the data if this has not already been done:

```
data = ReadList ["marks.dat"];
```

Once again, its mean is:

```
mean = SampleMean[data] // N
```

```
58.9024
```

Evaluating the first 6 central moments (cm) yields:

```
<< Statistics`
```

```
cm = Table[CentralMoment[data, r] // N, {r, 1, 6}]
```

```
{0., 193.875, -1125.94,  
133550., -2.68578 × 106, 1.77172 × 108}
```

(Once again, if we were working with sample data, we would replace the CentralMoment function with UnbiasedCentralMoment in the line above.) To obtain standardised moments, note that  $\mu_i^{\text{standardised}} = \mu_i / \mu_2^{i/2}$ . Then, empirical values for the first 6 standardised moments (sm) are:

```
sm = Table[ $\mu_i \rightarrow \frac{\text{cm}[[i]]}{\text{cm}[[2]]^{i/2}}$ , {i, 1, 6}]
```

```
{ $\mu_1 \rightarrow 0.$ ,  $\mu_2 \rightarrow 1.$ ,  $\mu_3 \rightarrow -0.417092$ ,  
 $\mu_4 \rightarrow 3.55303$ ,  $\mu_5 \rightarrow -5.13177$ ,  $\mu_6 \rightarrow 24.3125$ }
```

Evaluating  $\psi[6]$  at these values yields:

```

       $\psi_6 = \psi[6] /. sm // Simplify$ 
domain[ $\psi_6$ ] = {z, -∞, ∞};

0.000563511 e- $\frac{z^2}{2}$  (-5.24309 + z) (-3.14529 + z)
(8.28339 - 1.45564 z + z2) (5.43111 + 4.17537 z + z2)

```

The above gives the density in standardised units. To find the density in original units, say  $f(x)$ , transform from  $Z = \frac{X-\mu}{\sigma}$  to  $X = \mu + \sigma Z$ :

```

eqn = {x == mean +  $\sqrt{cm[[2]]}$  z};

f = Transform[eqn,  $\psi_6$ ]
domain[f] = TransformExtremum[eqn,  $\psi_6$ ]

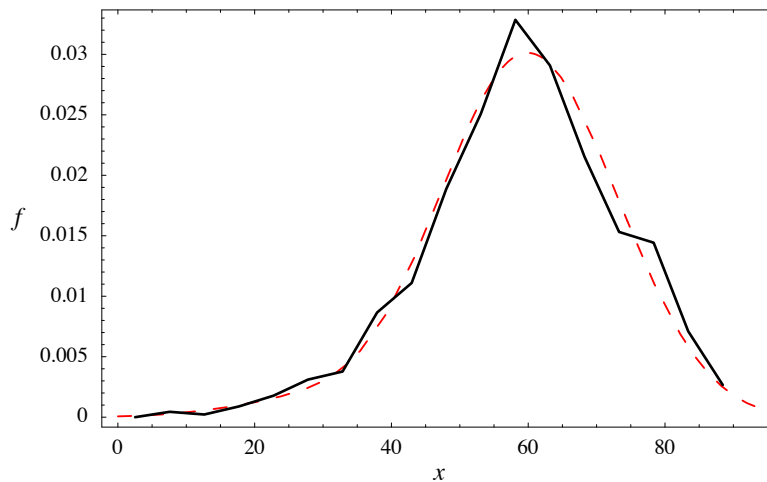
5.55363 × 10-12 e-0.00257898 (-58.9024+x)2
(-131.907 + x) (-102.697 + x)
(6269.27 - 138.073 x + x2) (1098.01 - 59.6673 x + x2)

{x, -∞, ∞}

```

Once again, `FrequencyPlot` allows one to compare the empirical pdf with the fitted density:

```
p3 = FrequencyPlot[data, f];
```



**Fig. 17:** The empirical pdf (—) and the fitted Gram–Charlier pdf (---)

This fitted Gram–Charlier density is actually very similar to the previous Johnson and `PearsonIV` results. The final Pearson fit was labelled `p1`. If it is still in memory, the command `Show[p1 /. Hue[___] → Hue[.4], p3]` shows both plots together, but now with the fitted Pearson curve in green rather than red, enabling a visual comparison (note that `Hue[___]` contains two `_` characters). On screen, the difference is apparent, but very slight. ■



*Some Advantages and Disadvantages of Gram–Charlier Expansions*

By construction, Pearson densities must be unimodal; this follows from equation (5.1), since  $d p / d x = 0$  at  $x = -a$ . Given bimodal data, Pearson densities may yield a very poor fit. In the Johnson family, both the  $S_L$  and  $S_U$  systems are unimodal. Although the  $S_B$  system can produce bimodal densities under certain conditions, the latter is not pleasant to work with. By contrast, Gram–Charlier expansions can produce mildly multimodal densities. On the downside, however, Gram–Charlier expansions have an undesirable tendency to sometimes produce small negative frequencies, particularly in the tails. In an ideal world, these negatives frequencies could be avoided by taking higher order expansions. This in turn requires higher order moments, which in turn have high variance and may be unreliable unless the sample size is sufficiently large. Finally, from a practical viewpoint, Gram–Charlier expansions are often ‘unstable’ in the sense that adding an extra ( $t + 1^{\text{th}}$ ) term may actually yield a worse fit, so some care is required in choosing an appropriate value for  $t$ .

**5.4 B Hermite Polynomials; Gram–Charlier Coefficients**

Let  $j$  denote the degree of the polynomial  $P_j(z)$ . Then, the family of polynomials  $P_j(z)$ ,  $j = 0, 1, 2, \dots$ , is said to be *orthogonal* to the weight function  $w(z)$  if

$$\int_{-\infty}^{\infty} P_i(z) P_j(z) w(z) dz = 0 \quad \text{for } i \neq j. \tag{5.10}$$

*Hermite polynomials* are orthogonal to the weight function  $w(z) = e^{-z^2/2}$ . They are defined by

$$H_j(z) = \frac{(-1)^j}{w(z)} \frac{d^j w(z)}{dz^j} = \frac{(-1)^j}{\phi(z)} \frac{d^j \phi(z)}{dz^j} \tag{5.11}$$

and have the property that

$$\int_{-\infty}^{\infty} H_i(z) H_j(z) \phi(z) dz = \begin{cases} 0 & \text{if } i \neq j \\ j! & \text{if } i = j \end{cases} \tag{5.12}$$

To illustrate the point, compare: (Note: H[ j ] and  $\phi$  were inputted in §5.4 A)

$$\int_{-\infty}^{\infty} \mathbf{H}[2] \mathbf{H}[3] \phi dz$$

0

with

$$\int_{-\infty}^{\infty} \mathbf{H}[3] \mathbf{H}[3] \phi dz$$

6

Multiplying both sides of (5.9) by  $H_i(z)$  yields

$$H_i(z) \psi(z) = \sum_{j=0}^{\infty} c_j H_i(z) H_j(z) \phi(z). \quad (5.13)$$

Integrating both sides yields, by the orthogonal property (5.12),

$$\int_{-\infty}^{\infty} H_i(z) \psi(z) dz = c_i i! \quad (5.14)$$

Thus,

$$c_i = \frac{1}{i!} E[H_i(z)] \quad (5.15)$$

where the expectation is carried out with respect to  $\psi(z)$ . We already know the form of the Hermite polynomials. For instance,  $H_6(z)$  is:

**H [6]**

$$-15 + 45 z^2 - 15 z^4 + z^6$$

It immediately follows that  $E[H_6(z)] = (-15 + 45 \acute{\mu}_2 - 15 \acute{\mu}_4 + \acute{\mu}_6)$  where  $\acute{\mu}_i$  denotes the  $i^{\text{th}}$  raw moment of  $\psi(z)$ . In *Mathematica*, this conversion from  $z^i$  to  $\acute{\mu}_i$  can be neatly achieved through pattern matching:

**H [6] /. z<sup>i</sup> -> \acute{\mu}\_i**

$$-15 + 45 \acute{\mu}_2 - 15 \acute{\mu}_4 + \acute{\mu}_6$$

Finally, since we have assumed that  $\psi(z)$  is a standardised density, replace  $\acute{\mu}$  with  $\mu$ , and let  $\mu_1 = 0$  and  $\mu_2 = 1$ . Then  $c_6$  reduces to  $(30 - 15 \mu_4 + \mu_6)/6!$ . These substitutions accord with the definition of the  $c[j]$  function in §5.4 A, and so  $c[6]$  yields:

**c [6]**

$$\frac{1}{720} (30 - 15 \mu_4 + \mu_6)$$

Finally, the nomenclature ‘Gram–Charlier Expansion of *Type A*’ suggests other types of expansions also exist. Indeed, just as *Type A* uses the standard Normal  $\phi(z)$  as a generating function, Charlier’s ‘*Type B*’ uses the Poisson weight function  $e^{-\lambda} \lambda^x / x!$  as its generating function, defined for  $x = 0, 1, 2, \dots$ . This has the potential to perform better than the standard Normal when approximating skew densities. However, it assumes a discrete ordinate system and perhaps for this reason is rarely used.

## 5.5 Non-Parametric Kernel Density Estimation

Kernel density estimation does not typically belong in a chapter on *Systems of Distributions*. However, just as a Pearson curve gives an impression of the distribution of the underlying population, so too does kernel density estimation, which helps explain why it is included here.

One of the virtues of working with families of distributions, rather than a specific distribution, is that it reduces the chance of making the wrong parametric assumption about the distribution's correct form. Instead of assuming a particular functional form, one assumes a particular family, which is more general. If our assumption is correct, then our estimates should be efficient. However, assumptions do not always hold, and by locking our analysis into an incorrect assumptional framework, we can end up doing rather poorly. As such, it is usually wise to conduct a preliminary investigation of the data based upon minimal assumptions. Smoothing methods serve to do this, as density smoothness is all that is imposed. The so-called *kernel density estimator* is

$$\hat{f}(y) = \frac{1}{nc} \sum_{i=1}^n K\left(\frac{y-Y_i}{c}\right) \quad (5.16)$$

where  $(Y_1, \dots, Y_n)$  is a random sample of size  $n$  collected on a random variable  $Y$ . The function  $K$  is known as the kernel and is specified by the analyst; it is often chosen to be a density function with zero mean and finite variance. Parameter  $c > 0$  is known as the *bandwidth* and it too is specified by the analyst; small values of  $c$  produce a rough estimate, while large values produce a very smooth estimate. For further details on kernel density estimation, see Silverman (1986) and Simonoff (1996); Stine (1996) gives an implementation under *Mathematica* Version 2.2.

### ⊕ *Example 7: Non-Parametric Kernel Density Estimation*

In practice, the kernel density estimate is presented in the form of a plot, and this is exactly the output produced by the **mathStatica** function `NPKDEPlot` (non-parametric kernel density estimator). To illustrate its use, we apply it to Parzen's (1979) yearly 'Snowfall in Buffalo' data (63 data points collected from 1910 to 1972, and measured in inches):

```
data = ReadList ["snowfall.dat"];
```

Two steps are required:

- (i) Specify the kernel  $K$
- (ii) Choose the bandwidth  $c$

We can then use `NPKDEPlot` to plot the kernel density estimate.

Step (i): In this example, we select  $K$  to be of form

$$K(u) = \frac{(2r+1)!!}{r! 2^{r+1}} (1-u^2)^r, \quad -1 \leq u \leq 1 \quad (5.17)$$

where  $r = 1, 2, 3, \dots$  denotes the weight of the kernel, and  $!!$  is the double factorial function. The  $r = 1$  case yields the Epanechnikov kernel (`ep`):

$$\mathbf{ep} = \frac{3}{4} (1 - u^2); \quad \mathbf{domain[ep]} = \{u, -1, 1\};$$

Other common choices for  $K$  include the bi-weight kernel ( $r = 2$ ), the tri-weight kernel ( $r = 3$ ), and the Gaussian kernel  $(2\pi)^{-1/2} \exp(-u^2/2)$  which is defined everywhere on the real line.

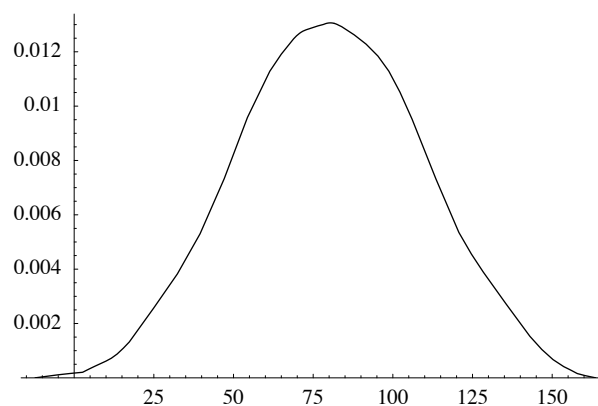
Step (ii): Next, we select the bandwidth  $c$ . This is most important, and experimenting with different values of  $c$  is advisable. A number of methods exist to automate bandwidth choice; `mathStatica` implements both the Silverman (1986) approach (default) and the more sophisticated (but much slower) Sheather and Jones (1991) method. They can be used as stand-alone bandwidth selectors, or, better still, as a starting point for experimentation. For the snowfall data set, the Sheather–Jones optimal bandwidth (using the Epanechnikov kernel) is:

```
c = Bandwidth[data, ep, Method → SheatherJones]
```

```
37.2621
```

Since  $K$  and  $c$  have now been specified, we can plot the smoothed *non-parametric kernel density estimate* using the `NPKDEPlot[data, K, c]` function:

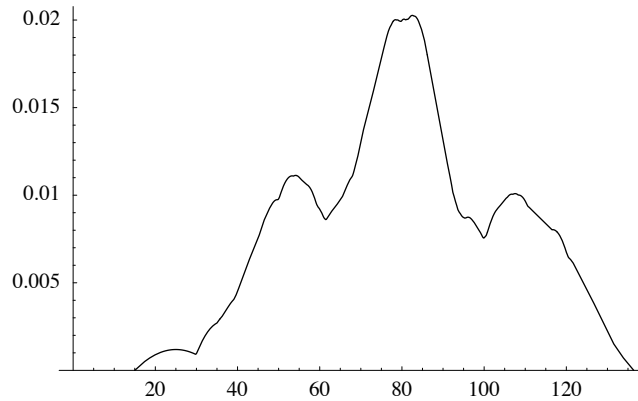
```
NPKDEPlot[data, ep, c];
```




**Fig. 18:** Plot of the non-parametric kernel density estimate, snowfall data ( $c = 37.26$ )

This estimate has produced a distinct mode for snowfall of around 80 inches. Suppose we keep the same kernel, but choose a smaller bandwidth with  $c = 10$ :

```
NPKDEPlot[data, ep, 10];
```



**Fig. 19:** Plot of the non-parametric kernel density estimate, snowfall data ( $c = 10$ ) 

Our new estimate exposes two lesser modes on either side of the 80-inch mode, at around 53 inches and 108 inches. A comparison of the two estimates suggests that the Sheather–Jones bandwidth is too large for this data set and has over-smoothed. This observation is in line with Parzen (1979, p.114) who reports that a trimodal shape for this data is “the more likely answer”. This serves to highlight the importance of the experimentation process. Clicking the ‘View Animation’ button in the electronic notebook brings up an animation in which the bandwidth  $c$  varies from 4 to 25 in step sizes of  $1/4$ . This provides a rather neat way to visualise how the shape of the estimate changes with  $c$ .

## 5.6 The Method of Moments

The *method of moments* is employed throughout this chapter to estimate unknown parameters. This technique essentially equates sample moments with population moments. The latter are generally functions of unknown parameters, and are then solved for those parameters.

To be specific, suppose the random variable  $Y$  has density  $f(y; \theta)$ , where  $\theta$  is a  $(k \times 1)$  vector containing all unknown parameters. Now construct the first  $r$  raw moments of  $Y$ . That is, construct  $\mu_i = E[Y^i]$  for  $i = 1, \dots, r$  and  $r \geq k$  (in all our examples, it suffices to set  $r = k$ ). Generally, each moment will depend (often non-linearly) upon the parameters, so  $\mu_i = \mu_i(\theta)$ . Now let  $(Y_1, \dots, Y_n)$  denote a random sample of size  $n$  collected on  $Y$ . We then construct the sample raw moments  $m_i = \frac{1}{n} \sum_{j=1}^n Y_j^i$  for each  $i$ . The method of moments estimator, denoted by  $\hat{\theta}$ , solves the set of  $k$  equations  $\mu_i(\hat{\theta}) = m_i$  for  $\hat{\theta}$ . The estimator is defined by equating the population moment with the sample moment, even though population moments and sample moments are generally not equal; that is,  $\mu_i(\theta) \neq m_i$ . This immediately questions the validity of the method of moments estimator. While not pursuing the answer in any detail here, we shall merely assert that the estimator may be justified using asymptotic arguments; for further discussion, see Mittelhammer (1996). Asymptotic theory is considered in detail in Chapter 8.

⊕ **Example 8:** The Bernoulli Distribution

Let  $Y \sim \text{Bernoulli}(\theta)$ , where  $\theta = P(Y = 1)$ , with pmf  $g(y)$ :

$$\mathbf{g} = \theta^y (1 - \theta)^{1 - y};$$

$$\mathbf{domain}[\mathbf{g}] = \{\mathbf{y}, 0, 1\} \ \&\& \ \{0 < \theta < 1\} \ \&\& \ \{\mathbf{Discrete}\};$$

The population mean of  $Y$  is easily derived as:

$$\hat{\mu}_1 = \mathbf{Expect}[\mathbf{y}, \mathbf{g}]$$

$\theta$

For a random sample of size  $n$ , the method of moments estimator is defined as the solution to  $\hat{\mu}_1(\hat{\theta}) = \hat{m}_1$ , which needs no further effort in this case:  $\hat{\theta} = \hat{m}_1$ . ■

⊕ **Example 9:** The Gamma Distribution

Let  $Y \sim \text{Gamma}(a, b)$  denote the Gamma distribution with parameter  $\theta = \begin{pmatrix} a \\ b \end{pmatrix}$  and pdf  $f(y)$ :

$$\mathbf{f} = \frac{\mathbf{y}^{a-1} e^{-y/b}}{\Gamma[\mathbf{a}] \mathbf{b}^a}; \quad \mathbf{domain}[\mathbf{f}] = \{\mathbf{y}, 0, \infty\} \ \&\& \ \{\mathbf{a} > 0, \mathbf{b} > 0\};$$

To estimate  $\theta$  using the method of moments, we require the first two population raw moments:

$$\hat{\mu}_1 = \mathbf{Expect}[\mathbf{y}, \mathbf{f}]$$

$$\hat{\mu}_2 = \mathbf{Expect}[\mathbf{y}^2, \mathbf{f}]$$

$a b$

$a (1 + a) b^2$

Then, the method of moments estimator of parameters  $a$  and  $b$  is obtained via:

$$\mathbf{Solve}[\{\hat{\mu}_1 == \hat{m}_1, \hat{\mu}_2 == \hat{m}_2\}, \{\mathbf{a}, \mathbf{b}\}]$$

$$\left\{ \left\{ \mathbf{a} \rightarrow -\frac{\hat{m}_1^2}{\hat{m}_1^2 - \hat{m}_2}, \mathbf{b} \rightarrow \frac{-\hat{m}_1 + \hat{m}_2}{\hat{m}_1} \right\} \right\}$$

*Mathematica* gives the solution as a replacement rule for  $a$  and  $b$ . Note that the symbols  $\hat{\mu}_1$  and  $\hat{\mu}_2$  are ‘reserved’ for use by **mathStatistica**’s moment converter functions. To avoid any confusion, it is best to `Unset` them:

$$\hat{\mu}_1 = .; \hat{\mu}_2 = .;$$

... prior to leaving this section. ■

## 5.7 Exercises

- Identify where each of the following distributions will be found on a Pearson diagram:
  - Exponential( $\lambda$ )
  - standard Logistic
  - Azzalini's skew-Normal distribution with  $\lambda > 0$  (see Chapter 2, Exercise 2).
- The data "stock.dat" provides monthly US stock market returns from 1834 to 1925, yielding a sample of 1104 observations. The data is the same as that used in Pagan and Ullah (1999, Section 2.10).<sup>5</sup>
  - Fit a Pearson density to this data.
  - Estimate the density of stock market returns using a non-parametric kernel density estimator, with a Gaussian kernel.
  - Compare the Pearson fit to the kernel density estimate.

To load the data, use: `ReadList["stock.dat"]`.

- Derive the equation describing the *Type III* and *Type V* lines in the Pearson diagram. [Hint: use the recurrence relation (5.5) to solve the moments  $(\mu_1, \mu_2, \mu_3, \mu_4)$  as a function of the Pearson coefficients  $(a, c_0, c_1, c_2)$ . Hence, find  $\beta_1$  and  $\beta_2$  in terms of  $(a, c_0, c_1, c_2)$ . Then impose the parameter assumptions that define *Type III* and *Type V*, and find the relation between  $\beta_1$  and  $\beta_2$ .]\*
- Exercise 3 derived the formulae describing the *Type III* and *Type V* lines, respectively, as:

$$\textit{Type III:} \quad \beta_2 = \frac{3}{2} \beta_1 + 3$$

$$\textit{Type V:} \quad \beta_2 = \frac{3(-16-13\beta_1-2(4+\beta_1)^{3/2})}{\beta_1-32}$$

Use these results to show that a Gamma distribution defines the *Type III* line in a Pearson diagram, and that an Inverse Gamma distribution defines the *Type V* line.

- Let random variable  $X \sim \text{Beta}(a, 1)$  with density  $f(x) = ax^{a-1}$ , for  $0 < x < 1$ ; this is also known as a Power Function distribution. Show that this distribution defines the *Type I(J)* line(s) on a Pearson diagram, as parameter  $a$  varies.
- Let random variable  $X$  have a standard Extreme Value distribution. Find  $\mu$  and  $\{\mu_2, \mu_3, \mu_4\}$ . Fit a Pearson density to these moments. Compare the true pdf (Extreme Value) with the Pearson fit.
- Recall that the Johnson family is based on transformations of  $Z \sim N(0, 1)$ . In similar vein, a Johnson-style family can be constructed using transformations of  $Z \sim \text{Logistic}$  (Tadikamalla and Johnson (1982)). Thus, if  $Z \sim \text{Logistic}$ , find the pdf of  $Y = \sinh\left(\frac{Z-\gamma}{\delta}\right)$ ,  $\gamma \in \mathbb{R}$ ,  $\delta > 0$ . Plot the pdf when  $\gamma = 0$  and  $\delta = 1, 2$  and  $3$ . Find the first 4 raw moments of random variable  $Y$ .
- Construct a non-parametric kernel density estimator plot of the "sd.dat" data set (which measures the diagonal length of 100 forged Swiss bank notes and 100 real Swiss bank notes) using a Logistic kernel and the Silverman optimal bandwidth.