A small note on the statistical method of moments for fitting a probability model to data

by Nasser Abbasi, Nov 16, 2007
Mathematics 502 probability and statistics, CSUF, Fall 2007

The problem to solve : Given some data, we seek to fit a probability law to the data. In other words, we want to determine the best probability distribution function by which the given data could have been generated according to.

We call the given data the population data. The idea of this method is as follows: Assume that the data was generated according to some distribution, say Normal or Gamma or Poisson, etc... For each one of these Choice we need to determine the relevant distribution parameters to be able to fully specify the pdf.

For example, if we want to fit the population data to the normal distribution, then we need to determine the mean and variance of the data $\{\mu, \sigma^2\}$ since the normal pdf is fully specified by these 2 parameters $f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}$

If we want to fit the population to the Gamma distribution, then we need to determine the parameters $\{\alpha,\lambda\}$ since the Gamma distribution is is fully specified by these 2 parameters $f(x) = \frac{e^{-\frac{x}{\lambda}} x^{-1+\alpha} \lambda^{-\alpha}}{\text{Gamma}[\alpha]}$.

If we have to determine 2 parameters (as in the above 2 cases) then we need 2 equations. But if we wanted to fit the data to Poisson distribution, then we only need one equation since the Poisson pdf is defined in terms on one parameter λ as in $f(x)=\frac{e^{-\lambda}\lambda^x}{x!}$

Let us assume there are n parameters to be determined (i.e. we want to fit the data to some distribution which is defined using n parameters). We call these $\theta_1, \theta_2, \ldots, \theta_n$, so for the case of fitting to a normal distribution $n = 2$, $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

We start by writing down the n probability moments, called $M_1, M_2, \ldots, M_n$ for the selected pdf we want to fit the data to. These are known analytical expressions for the selected pdf and can be looked up or derived from the assumed pdf.

The n^{th} moment is defined as $E(X^n)$. This will give us n equations expressed as functions of the θ_i,

Next we calculate the moments from the data itself and set these to be equal to the moments for the pdf and solve for the θ_i.

An example will help. Suppose to want to fit the data to a normal distribution, then we know that the first moment is given by $M_1 = E(X^1) = \mu$ and that the second moment is given by $M_2 = E(X^2) = \sigma^2 + \mu^2$.
So now we have 2 equations in 2 unknowns

$$\begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma^2 + \mu^2 \end{pmatrix} \tag{1}$$

It is easier to re - write the above as follows

$$\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} M_1 \\ M_2 - M^2{}_1 \end{pmatrix} \tag{2}$$

Now we determine an estimate for M_1 and M_2 from the data, or the sample, and substitute in the above and solve for μ and σ^2

$$\begin{aligned} M_1 &= \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \\ M_2 &= \text{Var}(X) + E(X)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 + \bar{X}^2 \end{aligned} \tag{3}$$

Now using (3) we solve (2) μ and σ^2. These will be estimated values. Hence written as $\hat{\mu}$ and $\hat{\sigma}$

Hence the solution from above gives an estimate of the pdf parameters from the data itself. We can now plot this selected pdf using the calculated parameters on top of the histogram of the data and see how good the fit is. If the fit is not good, we can try to fit the data to a different distribution.

This is another example, suppose we have data we want to fit to a Gamma distribution, hence we know that for a Gamma distribution $M_1 = E(X) = \frac{\alpha}{\lambda}$ and that $M_2 = E(X^2) = \frac{\alpha(\alpha+1)}{\lambda^2}$ hence we have

$$\begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\lambda} \\ \frac{\alpha(\alpha+1)}{\lambda^2} \end{pmatrix} \tag{4}$$

It is easier to re - write the above as follows

$$\begin{pmatrix} \alpha \\ \lambda \end{pmatrix} = \begin{pmatrix} \frac{M^2{}_1}{M_2 - M^2{}_1} \\ \frac{M_1}{M_2 - M^2{}_1} \end{pmatrix} \tag{5}$$

Now using (5) we solve for α and λ using the calculated values for M_1 and M_2 from the data as shown in (3).

Numerical example

In these examples I will first generate random data (the population) from known distributions then take a small random sample from the data (with replacement), then use the method of moments above to estimate the parameters of the population (which is of course known in this case) and fit the found parameters on the population histogram to see how good the fit it.

Example 1, fitting to normal

Using real data

This data is the annual precipitation in Seattle (I think) for the years 1863 to 1999, it was downloaded from http://www.seattlecentral.edu/qelp/sets/049/049.html.
First load the data, and do histogram on it, then try to fit a normal distribution on it and see how good the fit is.

Load the data

In[3]:=
```
data = ReadList[
   "E:\\nabbasi\\data\\nabbasi_web_Page\\my_notes\\math_502_material\\method_of_moments
     \\data.txt", {Number, Number}];
```

In[4]:=
```
sizeOfPopulation = Length[data]
```

Out[4]=
```
137
```

Display few lines of data

In[5]:=
```
TableForm[data[[1 ;; 10, All]],
 TableHeadings → {None, {"year", "annual rain\nin inches"}}]
```

Out[5]//TableForm=

year	annual rain in inches
1863	46.31
1864	38.42
1865	49.65
1866	41.51
1867	49.94
1868	48.43
1869	45.41
1870	48.62
1871	48.84
1872	43.9

Decide on numbers of bins, and make histogram

In[6]:=
```
nBins = 25;
gz = nmaMakeDensityHistogram[data[[All, 2]], nBins];
gpz = GeneralizedBarChart[gz, BarStyle → White, ImageSize → 450];
```

Calculate first and second moments of data

In[9]:=
```
(*sampleSize=30;*)
sample = data[[All, 2]];
m1 = Mean[sample];
m2 = Variance[sample] + Mean[sample]^2;
```

Estimate data parameters. Solve the method of moments equations (this solves equations (2) above)

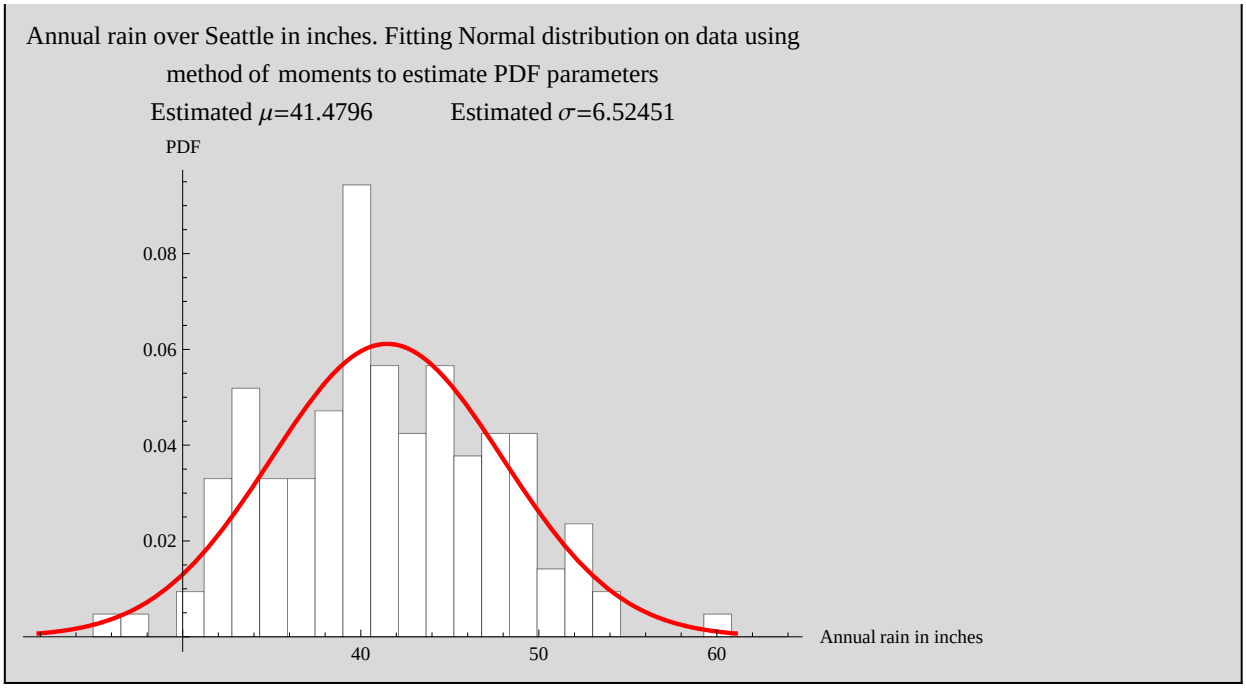
In[12]:=
```
eq1 = μEstimate == m1;
eq2 = varEstimate == m2 - m1^2;
sol = First@Solve[{eq1, eq2}, {μEstimate, varEstimate}]

μ = μEstimate /. sol;
σ = √varEstimate /. sol ;
```

Out[14]=
```
{μEstimate → 41.4796, varEstimate → 42.5692}
```

Plot the fitted PDF using the above estimated parameters

In[17]:=
```
p1 = Plot[PDF[NormalDistribution[μ, σ], x], {x, μ - 3 σ, μ + 3 σ}, PlotStyle → {Red, Thick}];
Show[{gpz, p1},
 PlotLabel → "Annual rain over Seattle in inches. Fitting Normal distribution on data
     using\nmethod of moments to estimate PDF parameters\nEstimated μ=" <>
   ToString[μ] <> "\tEstimated σ=" <> ToString[σ], AxesLabel →
  {"Annual rain in inches", "PDF"}]
```

Out[18]=



Using Random data

Make some random data from Normal and plot its histogram (see appendix for function to make histogram)

In[19]:=
```
sizeOfPopulation = 10 000; nBins = 100;
μ = 1; σ = 2;
population = Table[RandomReal[NormalDistribution[μ, σ]], {i, sizeOfPopulation}];
gz = nmaMakeDensityHistogram[population, nBins];
gpz = GeneralizedBarChart[gz, BarStyle → White, ImageSize → 450];
```

Take a small sample with replacement and obtain the first and second moments from the sample

In[24]:=
```
sampleSize = 30;
sample = RandomSample[population, sampleSize];
m1 = Mean[sample];
m2 = Variance[sample] + Mean[sample]^2;
```

Solve the method of moments equations (this solves equations (2) above)

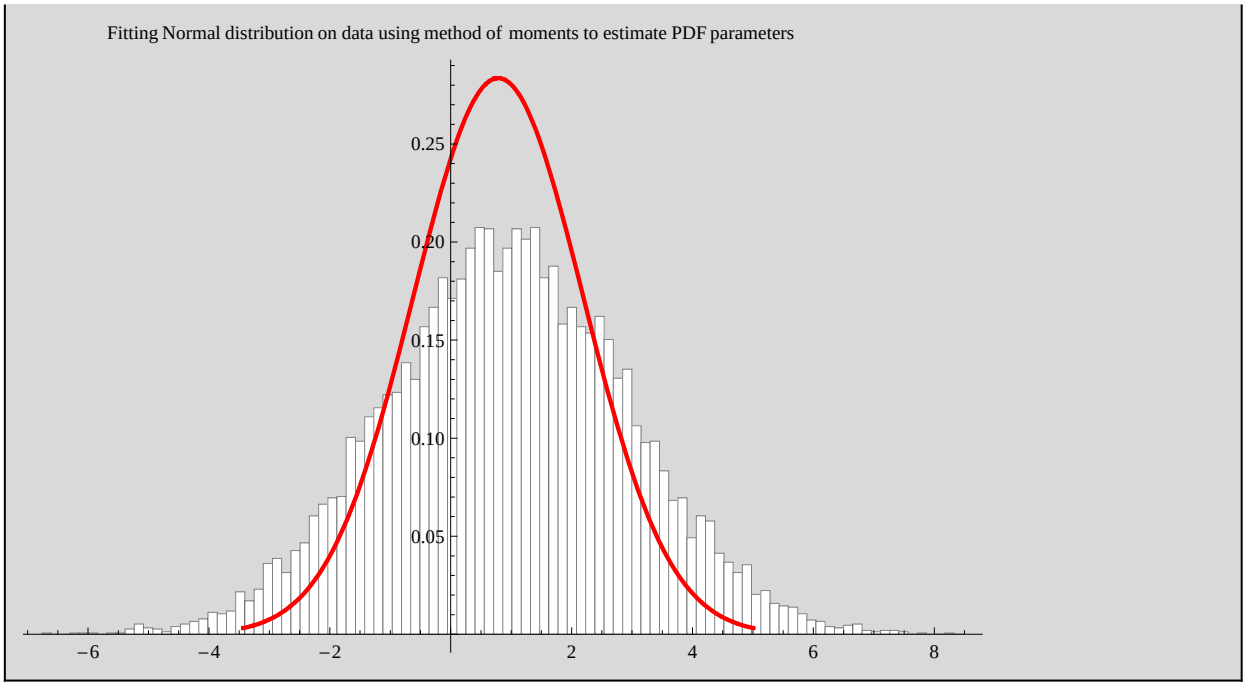
In[28]:=
```
eq1 = μEstimate == m1;
eq2 = varEstimate == m2 - m1^2;
sol = First@Solve[{eq1, eq2}, {μEstimate, varEstimate}]

μ = μEstimate /. sol;
σ = √varEstimate /. sol ;
```

Out[30]=
```
{μEstimate → 0.786502, varEstimate → 1.97899}
```

Plot the fitted PDF using the above estimated parameters

In[35]:=
```
p1 = Plot[PDF[NormalDistribution[μ, σ], x], {x, μ - 3 σ, μ + 3 σ}, PlotStyle → {Red, Thick}];
Show[{gpz, p1},
 AxesLabel → "Fitting Normal distribution on data using method of moments to
    estimate PDF parameters"]
```

Out[36]=



Example 2 fitting to Gamma

Lets try to fit a Gamma on the data to see what we get
Make some random data from Normal and plot its histogram (see appendix for function to make histogram)

In[37]:=
```
sizeOfPopulation = 1000; nBins = 100;
α = 1; λ = 2;
population = Table[RandomReal[GammaDistribution[α, λ]], {i, sizeOfPopulation}];
gz = nmaMakeDensityHistogram[population, nBins];
gpz = GeneralizedBarChart[gz, BarStyle → White, ImageSize → 450];
```

Take a small sample with replacement and obtain the first and second moments from the sample

In[42]:=
```
sampleSize = 100;
sample = RandomSample[population, sampleSize];
m1 = Mean[sample];
m2 = Variance[sample] + Mean[sample]^2;
```

Solve the method of moments equations (this solves equations (5) above)

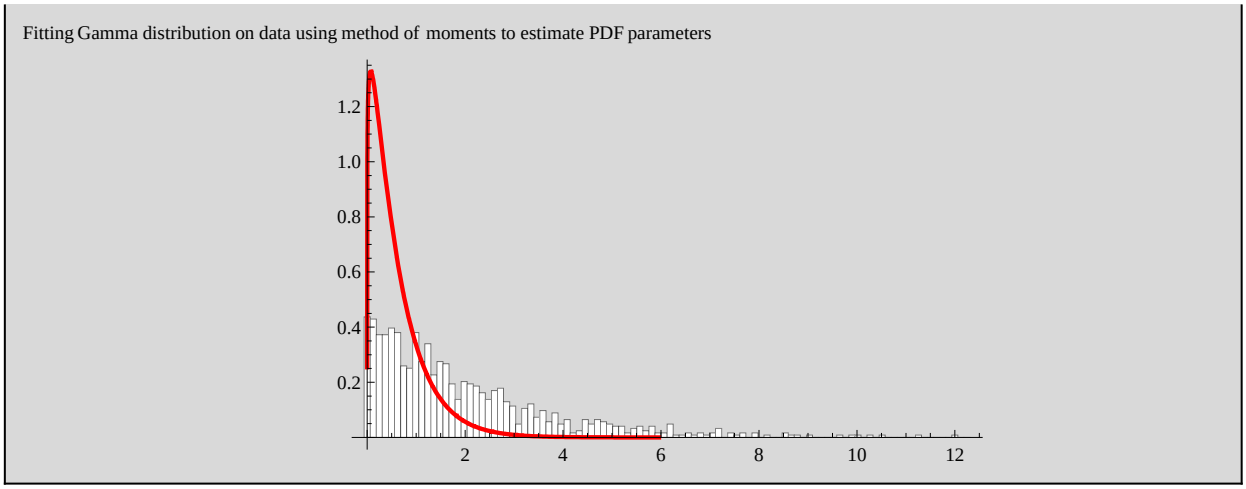
In[46]:=
```
eq1 = alphaEstimate == m1^2/(m2 - m1^2);
eq2 = lambdaEstimate == m1/(m2 - m1^2);
sol = First@Solve[{eq1, eq2}, {alphaEstimate, lambdaEstimate}]
α = alphaEstimate /. sol;
λ = lambdaEstimate /. sol;
std = Sqrt[α/λ^2];
```

Out[48]=
```
{alphaEstimate → 1.13439, lambdaEstimate → 0.536265}
```

Plot the Gamma PDF using the above estimated parameters on top of the data

In[58]:=
```
p1 = Plot[PDF[GammaDistribution[α, λ], x],
   {x, 0 , 3 std }, PlotStyle → {Red, Thick}, PlotRange → All];
Show[{gpz, p1}, AxesLabel → "Fitting Gamma distribution on data
    using method of moments to estimate PDF parameters"]
```

Out[59]=



Appendix

A function to plot histogram

In[1]:=
```
Needs["BarCharts`"]
nmaMakeDensityHistogram[originalData_, nBins_] :=
 Module[{freq, binSize, from, to, scaleFactor, j, a, currentArea},
  to = Max[originalData];
  from = Min[originalData];
  binSize = (to - from) / nBins;
  freq = BinCounts[originalData, binSize];
  currentArea = Sum[binSize * freq[[i]], {i, nBins}];
  freq = freq / currentArea;
  a = from;
  Table[{a + (j - 1) * binSize, freq[[j]], binSize}, {j, 1, nBins}]
 ]
```