A New Statistical Test for Analyzing Skew Normal Data

Hassan Elsalloukh, Ph.D.
Associate Professor of Statistics
Department of Mathematics and Statistics
University of Arkansas at Little Rock

COMPSTAT2010
August 24, 2010
Paris, France

Overview

- Motivation
- Azzalini's Class of Skew Distributions
- A New Density Function within Azzalini's Class of Skew Distributions
- A Score Test for Detecting Non-Normality within the New Density Function
- Applications
  - Volcanoes Height Example
  - Rainfall Example
- Summary

Motivation

- The celebrated Gaussian distribution has been known since at least a century before Gauss (1809) popularized it.
- It is the most well-known and widely used probability density function and has the form:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\theta)^2}{2\sigma^2} \right) \quad , -\infty < x < \infty$$

- It became more important because of the central limit effect discovered by De Moivre (1733).

- This distribution appears in probability process, and in the theories and methods of univariate and multivariate, parametric and non-parametric , frequentist and Bayesian statistics.
- Yet there have always been doubts and reservations and criticisms about the unqualified use of Normality
- This reflected in the quote by Geary (1947) “Normality is a myth; there never was, and never will be, a normal distribution”.

- The normal distribution is symmetric and not practical for modeling skewed data.
- During the last decade, there has been a growing interest in the construction of flexible parametric classes of distributions that are asymmetric.
- Various practical applications require models for data exhibiting a unimodal but skew distributions
- The skewed and kurtotic distributions are useful for data modeling,

- Such distributions are useful for data modeling including environmental and financial data that often do not follow the normal law
- One can introduce skewness into a symmetric distribution in many ways
- One generalization of the normal distribution was proposed by O'Hagan and Leonard (1976).
- This generalization was used for Bayesian analysis of normal means

- It was also investigated in detail by Azzalini (1985, 1986), who defined a skew-normal distribution as

$$f(x) = 2\phi(x)\Phi(\lambda x)$$

- Runnenburg (1978) devised a different way of introducing skewness into a symmetric distribution.

- By splicing two half-normal distributions with different scale parameters

- Mudholkar and Hutson (2000) found that this idea could be re-expressed in terms of an explicit skewness parameter ε.

- Mudholkar and Hutson (2000) called their probability density function the Epsilon-Skew-Normal family (ESN) :

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\begin{cases}\exp\left(\dfrac{-(x-\theta)^2}{2(1-\varepsilon)^2\sigma^2}\right), & for\ x\geq\theta\\ \exp\left(\dfrac{-(x-\theta)^2}{2(1+\varepsilon)^2\sigma^2}\right), & for\ x<\theta.\end{cases}$$

- Where the parameters are $-\infty<\theta<\infty$, $\sigma>0$, and $-1<\varepsilon<1$

- This density resembles the normal family members in many ways and includes the normal family when $\varepsilon = 0$.
- Note that the limiting cases of this density as epsilon goes to + or – 1 are the well-known half normal distributions
- This family is convenient for Bayesian analysis of normal means

- In this research, Azzalini's new skew normal distribution is modified leading to a new class of asymmetric distributions.
- A new score test is derived for detecting non-normality within the new class of asymmetric distributions.
- Then, the new score test is applied on an example of a real data set within the new class of asymmetric distributions to detect non-normality
- Maximum likelihood estimators are used to fit the data with a skew distribution and compared to studies in which researchers used the normal distribution.

Azzalini's Class of Skew Distributions

- Azzalini introduced the skew-normal class of distributions, as a class or family able to reflect varying degrees of skewness

- One such class of distributions was defined by Azzalini as a skew-normal random variable Z with a skewness parameter λ; with a density function

$$\phi(Z;\lambda) = 2\phi(z)\Phi(\lambda z) \qquad (-\infty < z < \infty),$$

- that is, Z is SN(λ) with $-\infty < Z < \infty$, where φ and Φ are the standard normal density and distribution functions, respectively

- One limitation of SN(λ) family is that the parameter λ can produce only tails thinner than the normal distribution.

- However, we are often interested in analyzing data from heavy-tailed distributions.

- Azzalini suggested a class of densities, which includes the normal family and allows thick tails, that is,

$$g(y;\omega) = C_{\omega} \exp\left\{-\frac{|y|^{\omega}}{\omega}\right\} \qquad (-\infty < y < \infty),$$

- where ω is a positive tail weight parameter and

$$C_\omega = \left\{ 2\omega^{\frac{1}{\omega-1}}\Gamma\left(1/\omega\right) \right\}^{-1}$$

- The density g(y,2) is the N(0,1) and g(y,1) is the Laplace density. As ω→∞, g(y, ω) converges to the uniform density on (-1,1)

- Azzalini introduces skewness in g(y, ω) in the form of

$$2G(\lambda y)g(y;\omega)$$

Where $\psi = \frac{\omega}{2}$

- The choice of G is the distribution function of

$$\operatorname{sgn}(U)\left|\sqrt{\psi}U\right|^{\frac{1}{\psi}}$$

where $U \sim N(0,1)$.

- Therefore, the density that was considered is

$$h(y) = 2C_{2\psi} \exp\left\{-\frac{|y|^{2\psi}}{2\psi}\right\} \Phi\left\{\operatorname{sgn}(\lambda y)\frac{|\lambda y|^{\psi}}{\sqrt{\psi}}\right\}$$

- Many choices of *G* and g(y, ω) are possible. The choices that are considered in this paper are modified to produce a new density function of the form

$$h(y_i) = 2G(\lambda u_i)g(u_i;\alpha),$$

where

$$u_i = \frac{y_i - \mu}{\sigma}, \quad g(u_i \mid \alpha) = w(\alpha)\sigma^{-1}\exp\left\{-c(\alpha)\left|u_i\right|^{\frac{2}{1+\alpha}}\right\},$$

$$c(\alpha) = \Gamma\left[\frac{3(1+\alpha)}{2}\right]^{\frac{1}{1+\alpha}}\Gamma\left[\frac{(1+\alpha)}{2}\right]^{-\frac{1}{1+\alpha}}, \quad w(\alpha) = (1+\alpha)^{-1}\left\{\Gamma\left[\frac{3(1+\alpha)}{2}\right]\right\}^{\frac{1}{2}}\left\{\Gamma\left[\frac{(1+\alpha)}{2}\right]\right\}^{\frac{-3}{2}},$$

and $G(\lambda u_i \mid \alpha) = \Phi\left[\operatorname{sgn}(\lambda u_i)\sqrt{1+\alpha}\left|\lambda u_i\right|^{\frac{1}{1+\alpha}}\right]$, for $\lambda \geq 0$, Note that when

$\lambda=0$ and $\alpha=0$, h(y) reduces to a standard normal.

A Score Test for Detecting Non-Normality within the New Density Function

- The problem of testing hypotheses of univariate normality of a set of observations has been of interest to experimenters for many years
- As a result, many test statistics have been suggested as possible solutions to the testing-normality problem.
- One such is the score test or Lagrange multiplier test
- A score test of normality within the family of new skew distributions are developed now.

- Since the score test testing procedure requires estimation only under the null hypothesis, an asymptotically unbiased test of the normality assumption

$$H_0: \lambda=0 \text{ and } \alpha=0 \text{ vs. } H_A: \lambda\neq 0 \text{ and } \alpha\neq 0$$

can be easily constructed.

- Let $y_1, \ldots, y_n$ be random variables from a new skew distribution then the test statistic is

$$\Lambda = \left[\frac{\partial L(\hat{\varphi})}{\partial \alpha} \quad \frac{\partial L(\hat{\varphi})}{\partial \lambda}\right] \begin{pmatrix} E\left[\frac{\partial L(\hat{\varphi})}{\partial \alpha}\right]^2 & E\left[\frac{\partial L(\hat{\varphi})}{\partial \alpha}\frac{\partial L(\hat{\varphi})}{\partial \lambda}\right] \\ E\left[\frac{\partial L(\hat{\varphi})}{\partial \alpha}\frac{\partial L(\hat{\varphi})}{\partial \lambda}\right] & E\left[\frac{\partial L(\hat{\varphi})}{\partial \lambda}\right]^2 \end{pmatrix}^{-1} \begin{bmatrix} \frac{\partial L(\hat{\varphi})}{\partial \alpha} \\ \frac{\partial L(\hat{\varphi})}{\partial \lambda} \end{bmatrix}$$

$$= \frac{\left[\frac{n}{2} - .8648186\sum_{i=1}^{n}\hat{u}_i^2 + \sum_{i=1}^{n}\hat{u}_i^2 \ln|\hat{u}_i|\right]^2}{.2011014n} + \frac{\left[\sum_{i=1}^{n}\hat{u}_i\right]^2}{n} = \hat{\xi}_1 + \hat{\xi}_2,$$

where $u_i \sim N(\mu, \sigma)$. Note that as n→∞, the asymptotic distribution of Λ is chi-square with two degrees of freedom,

- Thus, the null hypothesis is rejected if

$$\Lambda < \chi^2_{(2,1-\alpha/2)} \quad \text{or} \quad \Lambda > \chi^2_{(2,\alpha/2)}$$

- The first part of the test statistic, ξ_1, measures kurtosis and the second part, ξ_2, measures the skewness of the distribution of interest.

- We now present two examples

Application

Example 1

- The score test computations are used on the heights of 219 of the world's volcanoes (Source: National Geographic Society and the World Almanac 1966, pp. 282-283)

- Figure 1 shows an exploratory data analysis in the form of a stem-and-leaf plot.

- The basic descriptive statistics for the volcano heights Y are: the sample mean $\bar{Y}$ =70.246, the standard deviation S=43.018, the median=65.000, and the coefficient of skewness b_1 =0.840. This coefficient indicates that Y is asymmetric.

Figure 1. Heights of 219 of the world's volcanoes

Stem	Leaf	#
19	03379	5
18	5	1
17	29	2
16	25	2
15	667	3
14	00	2
13	03478	5
12	11244456	8
11	0112334669	10
10	0112233445689	13
9	000123344556779	15
8	122223335679	12
7	00001112334555678889	20
6	0011445566677889	18
5	00112223445566666677799	23
4	01112333334446788999999	22
3	0112244555566678899	18
2	00112224445566677889999	22
1	0001366799	10
0	25666799	8

----+----+----+----+---

Multiply Stem.Leaf by 10**+1

- Therefore, the score test Λ was calculated for the volcano heights using SAS IML, Λ=.0635.

- Since Λ falls in the rejection region, at the 5% significant level, we conclude that the data does not come from a symmetric normal distribution; indeed it can be modeled using the asymmetric distribution

- The MLEs are:

$$\hat{\mu} = 41.134 \qquad \hat{\sigma} = 40.350 \qquad \hat{\lambda} = 0.7$$

- These estimators were used to provide a better fit for the data as shown in Figure 2.

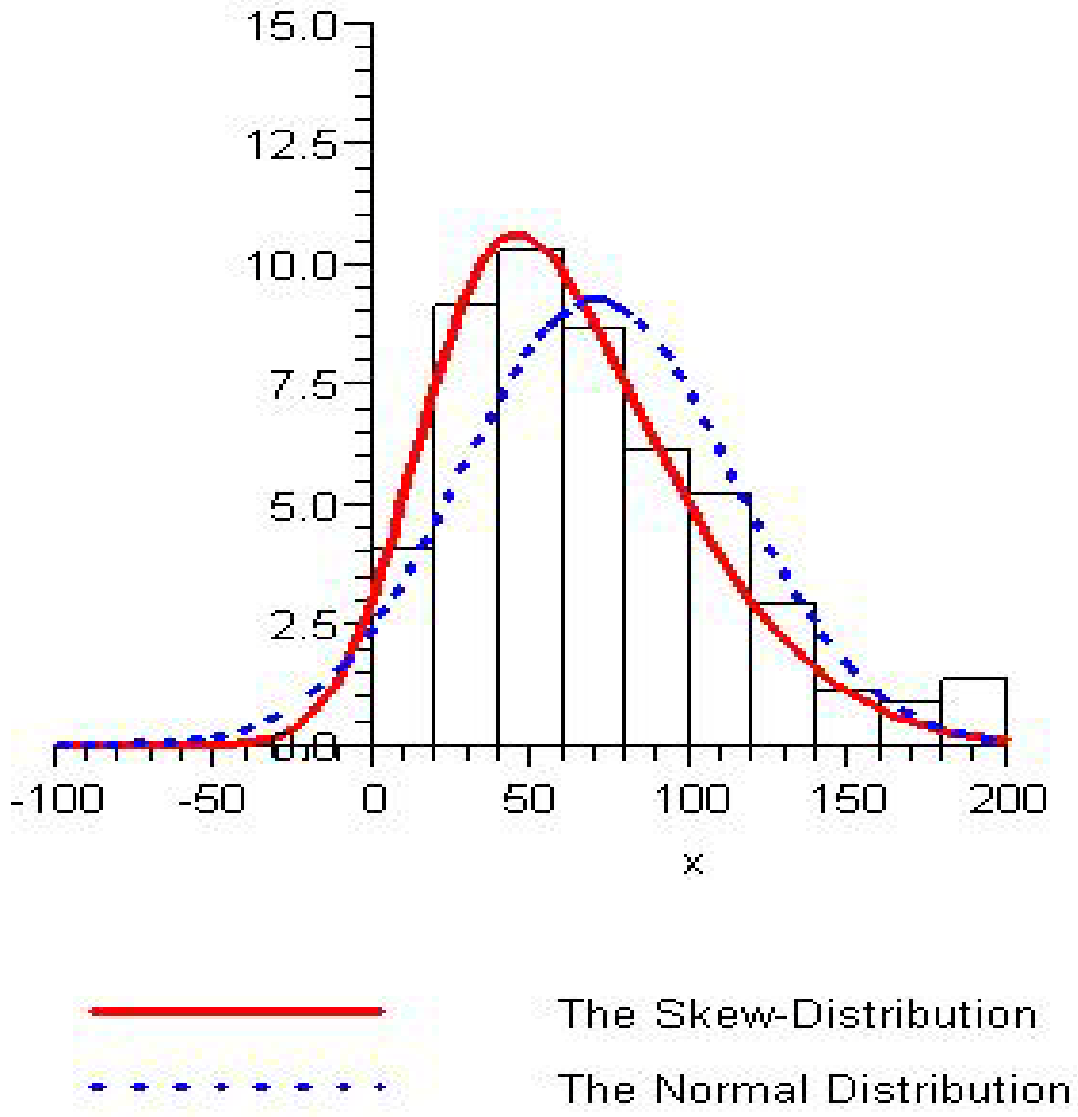

Figure 2. The normal and skew-normal for heights the world's volcanoes

Example 2

- Now the score test computations are used Daily rainfall in millimeters over a 47 year period in Turramurra, Sydney, Australia

- Figure 3 shows an exploratory data analysis in the form of a stem-and-leaf plot.

- The basic descriptive statistics for the volcano heights Y are: the sample mean $\bar{Y}$ =1369.106, the standard deviation S=693.670, the median=1331, and the coefficient of skewness b_1 =1.295. This coefficient indicates that Y is asymmetric.

Figure 3. Daily Rainfall in Millimeters

Stem	Leaf	#
38	3	1
36		
34		
32		
30		
28		
26	582	3
24	4	1
22		
20	0	1
18	0567	4
16	2248	4
14	07466	5
12	333679	6
10	149	3
8	4558116689	10
6	8025	4
4	58688	5

----+----+----+----+

Multiply Stem.Leaf by 10**+2

- Therefore, the score test Λ was calculated for rainfall data using SAS IML, Λ=.0365.

- Since Λ falls in the rejection region, at the 5% significant level, we conclude that the data does not come from a symmetric normal distribution; indeed it can be modeled using the asymmetric distribution

- The MLEs are:

$$\hat{\mu} = 1100.356 \qquad \hat{\sigma} = 580.230 \qquad \hat{\lambda} = 0.8$$

- These estimators were used to provide a better fit for the data as shown in Figure 4.

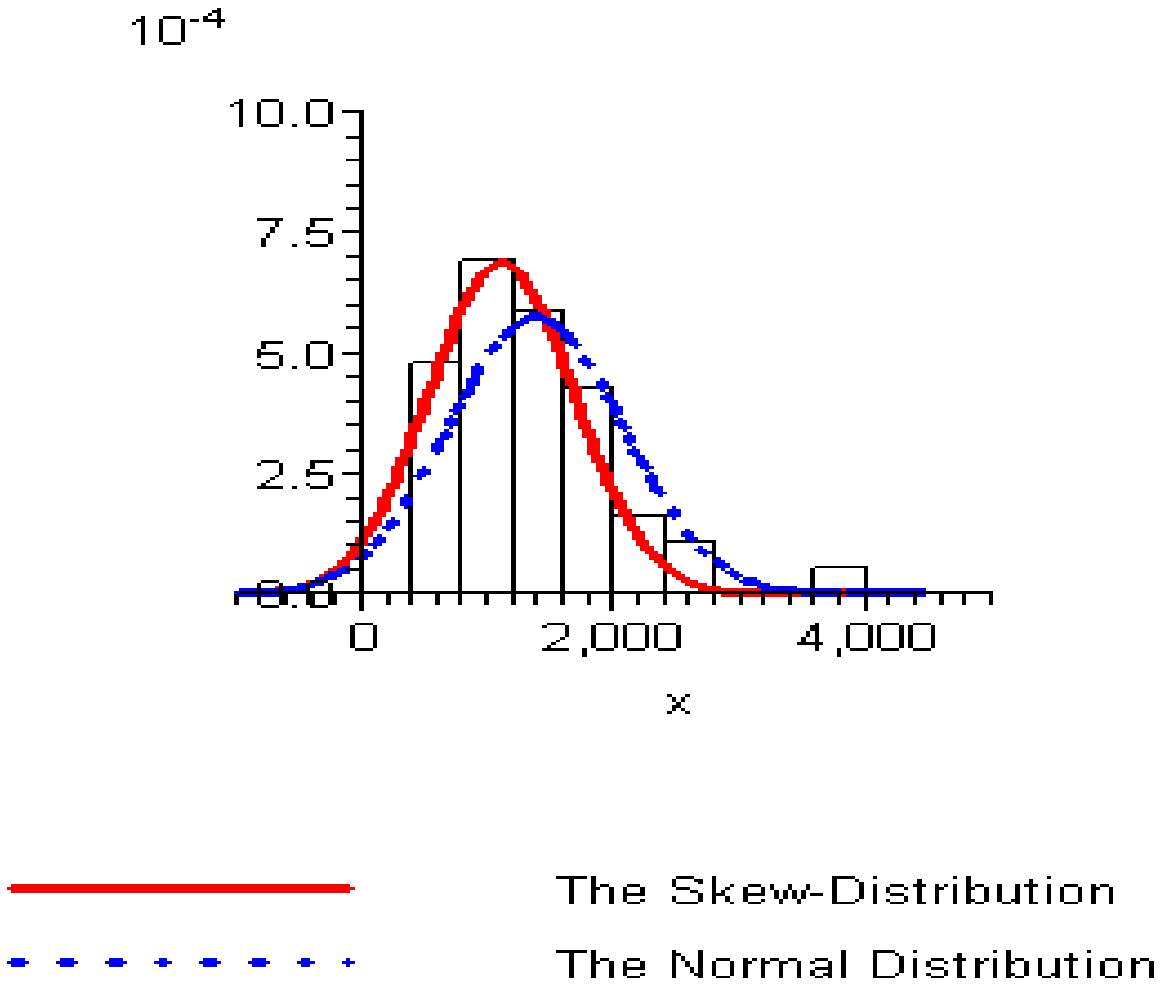

Figure 4. The normal and skew-normal for rainfall data

Summary

- Azzalini's new skew normal distribution is modified leading to a new class of asymmetric distributions Azzalini's Class of Skew Distributions
- A new score test is derived for detecting non-normality within the new class of asymmetric distributions
- The score test was applied on Volcanoes Height Rainfall Examples
- The test score provided more accurate and better fits in both examples
- This research can also be generalized to derive a score test for testing near-Laplace data using the new density function

Questions?