Correlation and regression analysis of the mean and standard deviation of samples of two from a gamma population

Şerafettin Çabukoglu *

School of Business
North Carolina Central University
1801 Fayetteville Street
Durham, North Carolina 27707
U.S.A.

Abstract

It is well-known that a necessary and sufficient condition that a probability distribution be normal is that its mean and standard deviation be independent. Beyond this, little seems to be known about the relationship between the mean $\bar{x}$ and standard deviations $s$ of samples drawn from non normal populations. The exponential distribution and more generally the gamma distribution adequately fit many distributions arising in technological application of statistics, particularly in the fields of reliability studies, life testing, queuing theory, wheatear analysis, and medicine. Inference for parameters of more than two gamma distributions is quite rare in the literature. Tripathi et al (1993) proposed a test for parameters of $m \geq 2$ gamma distributions based on a generalized minimum Chi-square procedure. For gamma populations with integer shape parameters $m \geq 2$, this paper derives, among other things, the regression function and the conditional distribution function of $s$ given $X_n F(s | \bar{x})$. The latter is particularly important in determining whether the correct gamma distribution (as identified by its shape parameter) is used in connection with its (frequent) application to real world problems in diverse fields. For example, the gamma density function is used both as a p.d.f. and as a Bayesian prior density function in reliability analysis (Mann et al (1974), p. 127, p. 379). Mooley (1973) discusses the gamma model for summer monsoon rainfall in millimeters (See Bowman and Shenton (1988), p. 90.). Bordi et al (2001) discussed using gamma distribution for drought monitoring in the Mediterranean area. Masyma and Kuroiwa (1951) give data on the sedimentation rate during the period of normal pregnancy, and fit the gamma distribution to the data. (See Bowman and Shenton (1988), p. 90.) Amorosa (1925) used the gamma distribution in analyzing the distribution of income. And of course there are many other applications. But whatever the area or nature of application, the proper use of the gamma distribution depends heavily upon using the correct value of the shape parameter, which can be simply and easily tested by using the conditional distribution

*E-mail: scabukoglu@wpo.nccu.edu

Journal of Statistics & Management Systems
Vol. 9 (2006), No. 2, pp. 473–484
© Taru Publications
function of $s$ given $\bar{x}$ to compare the value of $s$ relative to that of $\bar{x}$. Furthermore, when using analytical models involving the mean and standard deviation of gamma distribution. It is necessary to know whether the correlation between them is sufficiently high that it cannot be required. Other interesting results obtained in this paper from the derived distribution $g(\bar{x}, s)$, are:

1. the regression function is linear.
2. the scedastic function is quadratic.
3. for any integer value of the shape parameter $m$, the coefficient of variation for density function of $\bar{x}$ is $1/(2m)^{1/2}$.
4. the correlation ratios for the shape parameters $m = 2, 3$ are, respectively, 0.70076 and 0.48564.

The author conjectures that values of the correlation ratios decrease with increasing values of $m$, but this has not been proven.

**Keywords**: Joint distribution, conditional distribution, scedastic function, regression function, correlation ratio.

1. **Introduction**

It is well-known that a necessary and sufficient condition for a distribution to be normal is that its mean and standard deviation be independent (Geary (1936)). Beyond this, little seems to be known about their relationship. For example, for what statistical distribution is the relationship linear, and what is the magnitude of the correlation?

The simplest gamma distribution — but nevertheless an important one — is the one for which the shape parameter has the value one. It is usually referred to as the exponential distribution, and is commonly used in life testing and in the reliability analysis of electronic components and systems. It has been discussed in some detail by Çabukoğlu (2004). Johnson and Kotz (1970) provide a detailed review of the gamma distribution. Since extension of the analysis to cover the more general complex cases with shape parameter values greater than one unduly lengthens the paper. It was felt advisable to defer their treatment to a later paper. In the literature, inference for parameters of more than two distributions is rare. Bhattacharya (2002) tests general linear combinations for gamma distributions with parameters $m \geq 3$ against inequality restrictions.
This paper analyzes in considerable detail the relationship between the sample mean $\bar{x}$ and standard deviation $s$ for samples of two items drawn at random from a gamma distribution, by deriving and studying their joint distribution $g(\bar{x}, s)$. Knowledge of the joint distribution enables one to derive various relative functions and statistics, including:

(a) the marginal distributions of $\bar{x}$ and $s$ and their moments.
(b) the coefficient of variation (c.v.) $\sigma_{\bar{x}}/E[\bar{x}]$.
(c) the conditional density and conditional distribution functions: $f(s | \bar{x})$ and $F(s | \bar{x})$.
(d) the regression function $E(s | \bar{x})$ of $s$ on $\bar{x}$ and the associated scedastic function $V(s | \bar{x})$.
(e) the expected value $E[V(s | \bar{x})]$ of the scedastic function $V(s | \bar{x})$.
(f) the correlation ratio for the regression of $s$ on $\bar{x}$:
$$r(s, \bar{x}) = \left[1 - \{E[V(s | \bar{x})]/\sigma_s^2\}\right]^{1/2}.$$ 

The derivation of the joint distribution and an explanation of how these relative functions and statistics stem therefrom will now be explained.

2. Derivation of the joint distribution $g(s | \bar{x})$

The function
$$f(y) = \frac{1}{\Gamma(a)b^a} y^{a-1}e^{-y/b}, \quad a, b > 0, \quad 0 \leq y < \infty$$
is a two-parameter gamma probability density function (p.d.f) with shape parameter $a$ and scale parameter $b$. Since changing $b$ merely changes the scale on the two axes, one can, without loss of generality, limit the analysis to the standardized form.
$$f(x) = \frac{1}{\Gamma(a)} x^{a-1}e^{-x}, \quad a \geq 1, \quad 0 \leq x < \infty.$$ 
That is, the shape parameter completely determines the behavior of the gamma p.d.f. In this analysis, the shape parameter $a$ has integer values, which is often the case in practical applications.

Consider, the sample of two independent items $x_i, i = 1, 2$ drawn at random from the gamma population,
$$f(x) = \frac{1}{\Gamma(a)} x^{a-1}e^{-x}, \quad a > 1, \quad 0 \leq x < \infty.$$ (1)
The sample mean is
\[ \bar{x} = \frac{x_1 + x_2}{2}. \]  
(2)

If \( x_1 \geq x_2 \), the sample standard deviation is
\[ s = \frac{x_1 - x_2}{2}, \]  
whereas if \( x_1 < x_2 \), the sample standard deviation is
\[ s = \frac{x_2 - x_1}{2}. \]  
(4)

Addition and subtraction of (2) and (3) lead to the transformation
\[ x_1 = \bar{x} + s, \quad x_2 = \bar{x} - s, \quad x_1 \geq x_2, \]  
(5)

while addition and subtraction of (2) and (4) yield the transformation
\[ x_1 = \bar{x} - s, \quad x_2 = \bar{x} + s, \quad x_1 < x_2. \]  
(6)

The Jacobian of the transformations (5) and (6) are, respectively,
\[ J_1 = \begin{vmatrix} \frac{\partial x_1}{\partial \bar{x}} & \frac{\partial x_1}{\partial s} \\ \frac{\partial x_2}{\partial \bar{x}} & \frac{\partial x_2}{\partial s} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2 \]
and
\[ J_2 = \begin{vmatrix} \frac{\partial x_1}{\partial \bar{x}} & \frac{\partial x_1}{\partial s} \\ \frac{\partial x_2}{\partial \bar{x}} & \frac{\partial x_2}{\partial s} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2. \]

Application of the transformations (5) and (6) to the joint density function
\[ f(x_1, x_2) \] leads to the joint density function \( g(\bar{x}, s) \). Specifically,
\[ f(x_1)f(x_2)dx_1dx_2 = f(\bar{x} + s)f(\bar{x} - s) | J_1 | d\bar{x}ds \
+ f(\bar{x} - s)f(\bar{x} + s) | J_2 | d\bar{x}ds \]
\[ = \frac{1}{\Gamma(a)} \left[ (\bar{x} + s)^{a-1}(\bar{x} - s)^{a-1}e^{-(\bar{x}+s)-(\bar{x}-s)}2d\bar{x}ds \right] \]
\[ + (\bar{x} - s)^{a+1}(\bar{x} + s)^{a-1}e^{-(\bar{x}-s)-(\bar{x}+s)}(2)d\bar{x}ds \].

That is,
\[ g(\bar{x}, s) = \frac{1}{\Gamma(a)} \left[ 4(\bar{x}^2 - s^2)^{a-1}e^{-2\bar{x}} \right], \quad 0 \leq s \leq \bar{x} < \infty. \]  
(7)
In practice, when \( a \) is a positive integer \( m \geq 2 \),

\[
g(x, s) = \frac{4e^{(-2\bar{x})}}{(\Gamma(m))^2} \sum_{j=1}^{m-1} (-1)^j \binom{m-1}{j} (\bar{x}^2)^{m-1-j} (s^2)^j,
\]

\[
0 \leq s \leq \bar{x} < \infty.
\]  

(8)

(The letter \( m \) is here used instead of \( n \), since \( n \) is usually used to denote sample size. In the paper, \( x \) and \( s \) are restricted to samples of size two.)

3. Analysis of the regression system of \( s \) on \( \bar{x} \)

In this section, the regression system of \( s \) on \( \bar{x} \) is analyzed. Specifically, the marginal density functions, scedastic function, conditional density function, and their means and variances are determined, the regression function being equivalent to the mean of the conditional distribution of \( s \) on \( \bar{x} \). The correlation ratio can then be evaluated.

3.1 The marginal density functions

The marginal density function \( f_1(\bar{x}) \) is by definition

\[
f_1(\bar{x}) = \int_0^{\bar{x}} g(\bar{x}, s) ds,
\]

where \( g(\bar{x}, s) \) is given by (8). Clearly, the resultant integral is of the form

\[
f_1(\bar{x}) = c_m \bar{x}^{2m-1} e^{-2\bar{x}}, \quad c_m = a \text{ constant}
\]  

(9a)

and since \( f_1(\bar{x}) \) is a p.d.f, \( c_m = 2^{2m}/(2m-1)! \), so that (10) becomes

\[
f_1(\bar{x}) = 2^{2m}/(2m-1)! \bar{x}^{2m-1} e^{-2\bar{x}}, \quad m=2,3,\ldots, \quad 0 \leq \bar{x} < \infty.
\]

(10)

The \( K \)th moment of \( f_1(\bar{x}) \) about the origin is

\[
E[\bar{x}^K] = \frac{2^{2m}}{(2m-1)!} \int_0^{\infty} \bar{x}^{2m+K-1} \exp(-2\bar{x}) d\bar{x}
\]

\[
= (2m + K - 1)!/( (2m - 1)! 2^K), \quad K = 0,1,2,\ldots.
\]

In particular,

\[
E[\bar{x}] = m, \quad m = 2,3,4,\ldots,
\]

and

\[
\sigma_{\bar{x}}^2 = m/2, \quad m = 2,3,4,\ldots.
\]
From the above results, it also follows that $m \geq 2$ and coefficient of variation $\sigma_s/E[\bar{X}] = 1/(2m)^{1/2}$. These results are not surprising, since (10) is equivalent to a gamma distribution.

Similarly the marginal density function $f_2(s)$ is a p.d.f. and consists of the integral

$$f_1(s) = \int_s^\infty g(x,s)dx,$$

which from (8) becomes

$$f_2(s) = \frac{4}{((m-1)!)^2} \int_s^\infty \sum_{j=0}^{m-1} (-1)^j \binom{m-1}{j} (s^2)^j (\bar{x}^2)^{m-1-j} \times \exp(-2\bar{x})d\bar{x}, \quad m = 2, 3, \ldots \quad (12)$$

Specifically,

$$f_2(s) = \frac{4}{((m-1)!)^2} \sum_{j=0}^{m-1} (-1)^j \binom{m-1}{j} (s^2)^j \times \frac{(2m-2j-2)!}{2^{m-2j-1}} \exp(-2s). \quad (13)$$

The evaluation of (12) and its moments utilizes integrals of the form

$$\int_s^\infty s^r \exp(-2s)ds, \quad r = 0, 1, 2, \ldots \quad (14)$$

where evaluation for integer values $r > 0$ is considerably simplified by application of the recursive formula

$$\int_s^\infty x^r \exp(2x)dx = \frac{s^r}{2} + r/2 \int_s^\infty x^{r/2} \exp(-2x)d\bar{x}, \quad r = 0, 1, 2, \ldots \quad (15)$$

(See, e.g., CRC Basic Statistical Tables (1971), p. 287.) In this paper, only the mean $E[s]$ and variance $\sigma_s^2 = E[s^2] - (E[s])^2$ are involved.

### 3.2 The regression function of $s$ on $\bar{x}$

As is well known (Hoel (1962), p. 192) the regression function of $s$ on $\bar{x}$ is based on the conditional density function $s$ on $\bar{x}$, denoted by $f(s | \bar{x})$, and defined (Hoel (1962), p. 193) as the quotient density function

$$f(s/\bar{x}) = g(\bar{x},s)/(f_1(x) \quad (16)$$
where $g(\bar{x}, s)$ is the joint density function (8) and $f_1(\bar{x})$ is the marginal density function (10).

Specifically,

$$f(s | \bar{x}) = \frac{(2m - 1)!}{2^{2m-2}((m - 1)!)^2} \sum_{j=0}^{m-1} (-1)^j \left( \frac{m - 1}{j} \right) (s^2)^j / \bar{x}^{2j+1},$$

$$m = 2, 3, 4, \ldots , (17)$$

which is, of course a p.d.f. over the range $0 \leq s \leq \bar{x}$. The regression function of $s$ on $\bar{x}$ is by definition,

$$E[s | \bar{x}] = \int_0^{\bar{x}} sf(s | \bar{x}) ds$$

$$= \frac{(2m - 1)!}{2^{2m-2}((m - 1)!)^2} \left[ \sum_{y=0}^{m-1} (-1)^j \left( \frac{m - 1}{j} \right) / (2j + 2) \right] \bar{x},$$

$$m = 2, 3, 4, \ldots . (18)$$

From (18), it is clear that the regression function of $s$ on $\bar{x}$ is linear.

More generally, the $k$th moment of conditional density function $f(s | \bar{x})$ about the origin is

$$E[s^K | \bar{x}] = \int_0^{\bar{x}} s^K f(s | \bar{x}) ds$$

$$= \frac{(2m - 1)!}{2^{2m-2}((m - 1)!)^2} \left[ \sum_{y=0}^{m-1} (-1)^j \left( \frac{m - 1}{j} \right) / (2j + K + 1) \right] \bar{x}^K,$$

$$m = 2, 3, 4, \ldots . (19)$$

4. Correlation analysis

The correlation analysis for the mean and standard deviation of samples of size two drawn from the gamma population (1) involves:

1. the scedastic function which characterizes the variance of the conditional density function (17) over the range $0 \leq s \leq \bar{x}$.

2. the expected value of the scedastic function in (a) over the entire range $0 \leq \bar{x} < \infty$.

3. the variance $\sigma_s^2 = E[s^2] - (E[s])^2$ of the density function $f_2(s)$ as given by (13).

Once these are obtained, the correlation ratio can be evaluated, as will now be shown.
4.1 The scedastic function and its expected value

The scedastic function is defined as

\[ V(s | \bar{x}) = \int_{0}^{\bar{x}} (s - E[s | \bar{x}])^2 f(s | \bar{x}) ds, \]  

and represents the variance of the \( s \) values for a specific value of \( \bar{x} \). See Kendall, Stuart and Ord (1991, p. 996; 1986, p. 524, footnotes). Note that the value of \( s \) for any sample cannot exceed that of the sample mean \( \bar{x} \). Specifically since \( f(s | \bar{x}) \) and \( E(s | \bar{x}) \) are given, respectively by (17) and (18), (20) is expressible sum-wise as

\[ V(s | \bar{x}) = \int_{0}^{\bar{x}} \left[ s - \frac{(2m - 1)!}{2^{2m-2}((m-1)!)^2} \left( \sum_{j=0}^{m-1} (-1)^j \binom{m-1}{j} \right) \frac{1}{s^{2j+2}} \right] ds \]  

\[ m = 2, 3, 4, \ldots \]  

(21)

The variance as given by (21) is of the form \( V(s | \bar{x}) = c_m \bar{x}^2 \), where \( c_m \) is a constant whose value depends upon the value of \( m \). But more importantly, the value of the variance \( V(s | \bar{x}) \) depends also upon the value of \( \bar{x} \). Since \( V(s | \bar{x}) \) does not have the same value over all values \( 0 \leq \bar{x} < \infty \), the regression system of \( s \) on \( \bar{x} \) is said to be heteroscedastic, rather than homoscedastic.

The remaining statistic needed to evaluate the correlation ratio is the expected value of the scedastic function, which by definition is

\[ E[V(s | \bar{x})] = \int_{0}^{\infty} V(s | \bar{x}) f_1(\bar{x}) d\bar{x}, \]  

(22)

where \( f_1(\bar{x}) \) and \( V(s | \bar{x}) \) are given, respectively, by (9) and (21), and the required integration is carried out in a straightforward manner.

4.2 The correlation ratio

In this paper, the correlation ratio (Cramer (1946), pp. 280–281) is given by the formula

\[ r(s, \bar{x}) = [1 - E[V(s | \bar{x})/\sigma_s^2]]^{1/2}, \]  

(23)

where \( V(s | \bar{x}) \) and \( \sigma_s^2 \) are, respectively, the expected value of the variance of the scedastic function and the value of the variance of the marginal
density function of $s$. The functions and statistics necessary for evaluating (23) were given earlier in this paper, and will not be repeated here.

The correlation ratio (23) measures the extent of the dependence existing between the mean and standard deviation of sample drawn at random from a gamma population with shape parameter $m$. When using analytical models involving the mean and standard deviation of gamma distributions, it is necessary to know whether the correlation between them is sufficiently high that it cannot be ignored.

5. The conditional distribution function

One important application of the material presented in this paper is the use of the conditional distribution function $F(s \mid \bar{x})$. Clearly, this is readily obtained from the conditional probability density function (17) previously derived. Specifically,

$$F(s, \bar{x}) = \Pr[s < S \mid \bar{x}] = \int_0^S f(s \mid \bar{x}) \, ds, \quad 0 \leq S \leq \bar{x}, \quad (24)$$

which enables one to test the hypothesis that a specific sample was drawn from a gamma population with a specific shape parameter, as is shown subsequently in Section 7.

6. Numerical structure of the regression and correlation model

This section presents a numerical description of the structural elements of the regression and correlation model for gamma distributions with shape parameters $m = 2, 3$.

**Gamma distribution with shape parameter $m = 2$**

1. $g(\bar{x}, s) = 4(\bar{x}^2 - s^2) \exp(-2\bar{x})$
2. $f_1(\bar{x}) = (8/3)\bar{x}^3 \exp(-2\bar{x})$
3. $E[\bar{x}] = 2$
4. $\sigma^2_{\bar{x}} = 1$
5. $E[\bar{x}] / \sigma^2_{\bar{x}} = 2$
6. c.o.v. $= \sigma_{\bar{x}} / E[\bar{x}] = 1/2$
7. $f_2(s) = (2s + 1) \exp(-2s)$
8. $E[s] = 3/4$
\( \sigma_s^2 = 7/12 \)  
\( f(s | \bar{x}) = (3/2)(1/\bar{x} - s^2/\bar{x}^3) \)  
\( E[s | \bar{x}] = (3/2)\bar{x}^3 \)  
\( V(s | \bar{x}) = (19/320)x^2 \)  
\( E[V(s | \bar{x})] = 19/64 \)  
\( r(s, \bar{x}) = (55/112)^{1/2} \)

**Gamma distribution with shape parameter** \( m = 3 \)

\( g(\bar{x}, s) = (x^4 - 2x^2s^2 + s^4) \exp(-2\bar{x}) \)
\( f_1(\bar{x}) = (8/15)x^5 \exp(-2\bar{x}) \)
\( E[\bar{x}] = 3 \)
\( \sigma_s^2 = 3/2 \)

\( E[\bar{x}] / \sigma_s^2 = 2 \)

\( c.v. = \sigma_s / E[\bar{x}] = \sqrt{(1/2)}/2 \)
\( f_2(s) = (s^2 + (3/2)s + 3/4) \exp(-2s) \)
\( E[s] = 15/16 \)

\( \sigma_s^2 = 159/256 = 0.62109375 \)
\( f(s | \bar{x}) = (15/8)(1/\bar{x} - 2(s^2/\bar{x}^3) + s^4/\bar{x}^5 \)
\( E[s | \bar{x}] = (5/16)\bar{x} \)
\( V(s | \bar{x}) = (81/1792)x^2 = 0.0452009x^2 \)
\( E[V(s | \bar{x})] = 0.474609358 \)
\( r(s, \bar{x}) = 0.485642959 \)

7. **Testing hypotheses via the conditional distribution function**

As has already been seen, the conditional density function \( f(s | \bar{x}) \) is an important element in the development of the regression function and correlation ratio. Surprisingly, the conditional distribution function \( F(s | \bar{x}) \) provides an amazingly simple means of testing the hypothesis that a given sample was drawn from a gamma population characterized by a shape parameter with a specific value \( m \geq 2 \). The implementation of the test is greatly simplified by the constraint \( s \leq \bar{x} \).
Quite naturally, the conditional distribution function $F(s \mid \bar{x})$ is defined as

$$F(S < K\bar{x} \mid \bar{x}) = \Pr[s \leq K\bar{x} \mid \bar{x}] = \int_0^{K\bar{x}} f(s \mid \bar{x}) \, dx, \quad K \leq 1$$

That is, the conditional distribution function is

$$F(S < K\bar{x} \mid \bar{x}) = \frac{15}{8}[K - (2/3)K^3 + K^5/5], \quad K \leq 1. \quad (25)$$

Hence, when testing the above hypothesis at a specific level of significance $0 < \alpha < 1$, based on an observed sample in which $s = K_0\bar{x}$, $K_0 < 1$, one simply substitutes the value $K = K_0$ in (25) and rejects the hypothesis when the resulting value is less than $\alpha/2$ or greater than $1 - \alpha/2$.

As an example, suppose one wishes to test, at the $\alpha = 0.05$ level of significance the null hypothesis that a random sample of two items with $\bar{x}_0 = 1.2$ and $s = 1.02 = 0.85\bar{x}_0$ was randomly drawn from gamma population with shape parameter $m = 3$. To carry out this test, one would substitute $K = 0.85$ in (25). The resultant value is $F(s \leq 0.85\bar{x}_0 \mid \bar{x}_0) = 0.9924$. Since 0.9924 > 0.975, one rejects the null hypothesis $F(s \mid \bar{x})$.

Acknowledgements. The author is grateful to Professor Norman L. Johnson of University North Carolina-Chapel Hill for his comments and criticisms.

References


Received March, 2005