11-1-2010

Estimating the Non-Existent Mean and Variance of the F-Distribution by Simulation

Hamid Reza Kamali
Private Scholar, hrkamali@gmail.com

Parisa Shahnazari-Shahrezaei
Firoozkooh Branch Islamic Azad University, Firoozkooh, Iran, parisa_shahnazari@iaufb.ac.ir

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Emerging Scholars
Estimating the Non-Existent Mean and Variance of the F-Distribution by Simulation

Hamid Reza Kamali
Private Scholar

Parisa Shahnazari-Shahrezaei
Firoozkooh Branch Islamic Azad University,
Firoozkooh, Iran

In theory, all moments of some probability distributions do not necessarily exist. In the other words, they may be infinite or undefined. One of these distributions is the F-distribution whose mean and variance have not been defined for the second degree of freedom less than 3 and 5, respectively. In some cases, a large statistical population having an F-distribution may exist and the aim is to obtain its mean and variance which are an estimation of the non-existent mean and variance of F-distribution. This article considers a large sample F-distribution to estimate its non-existent mean and variance using Simul8 simulation software.

Key words: Probability distribution, F distribution, simulation.

Introduction

In practice, it is often necessary to calculate some properties of a statistical population. For this purpose, a random sample is taken from the population, its properties are calculated and these properties are developed to a population. A subset of individuals is selected from within a population to yield some knowledge about the whole population by sampling. The developed properties are approximate and are typically expressed as a confidence interval. Conversely, researchers may seek to calculate some properties of a random sample with the aim of estimating these properties to a population. In this case, the considered properties are calculated for the whole population once and then are used for drawn random samples. The most important statistical properties of a population and random sample are the mean and

Hamid Reza Kamali is a Ph.D. Student in the Department of Industrial Engineering. Email: hrkamali@gmail.com. Parisa Shahnazari-Shahrezaei is a Ph.D. Student in the Department of Industrial Engineering. Email: parisa_shahnazari@iaufb.ac.ir.

variance which are calculated from the probability distribution function of a population. If it is not possible to obtain these properties using a probability distribution function, all individuals of the population should be examined and used to obtain the considered properties. Although this work is easy for small and finite populations, it is not possible for infinite populations and a fairly large random sample is needed for approximate calculations.

F-distribution

Probabilistic behavior of some random variables can be defined as a mathematical function of the value of the considered random variable which is called the probability distribution function. The most important probability distribution functions considered in this article are the Normal, Chi-square, and F distributions.

Equation (1) shows the Normal distribution function (Walpole & Myers, 1993) which includes two parameters $\mu \in R$ and $\sigma \in R^+$. Equations (2) and (3) express its mean and variance, respectively. The values of the Normal random variable, which belong to a real numbers set, have a specific mean and variance.

$$f_N(n) = \frac{1}{\sigma\sqrt{2\pi}} EXP(-\frac{(n-\mu)^2}{2\sigma^2}); \qquad (1)$$

$$-\infty < n < +\infty$$

$$E(N) = \mu \qquad (2)$$

$$Var(N) = \sigma^2 \qquad (3)$$

In the case of $\mu = 0$ and $\sigma = 1$, the Normal distribution is called the Standard Normal distribution and its probability distribution function will be according to Equation (4):

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} EXP(-\frac{z^2}{2}); \qquad (4)$$

$$-\infty < z < +\infty$$

The Chi-square distribution (Walpole & Myers, 1993) has a parameter $v > 0$ which is called degree of freedom. Equations (5) to (7) display its probability distribution function, mean and variance, respectively. The values of a Chi-square random variable which belong to a positive real numbers set also have a specific mean and variance.

$$f_{\chi^2}(x) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}; \qquad (5)$$

$$x > 0$$

$$E(\chi^2) = v \qquad (6)$$

$$Var(\chi^2) = 2v \qquad (7)$$

The F-distribution (Walpole & Myers, 1993) has two parameters $v_1 > 0$ and $v_2 > 0$ which are called the first and second degree of freedom. Its probability distribution function, mean and variance are illustrated by Equations (8) to (10), respectively. The values of an F random variable belong to a positive real numbers set.

$$f_F(w) = \frac{\Gamma[\frac{(v_1+v_2)}{2}](\frac{v_1}{v_2})^{v_1/2} w^{(v_1/2-1)}}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})} (1+\frac{v_1}{v_2})^{-(v_1+v_2)/2};$$

$$w > 0$$

$$(8)$$

$$E(F) = \frac{v_2}{v_2 - 2}; \qquad (9)$$

$$v_2 > 2$$

$$Var(F) = \frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}; \qquad (10)$$

$$v_2 > 4$$

As shown, the F-distribution's mean and variance have not been defined for the second degree of freedom less than 3 and 5, respectively. While confronting these circumstances, the approximate mean and variance of population can be obtained using a fairly large random sample and simulation. The Simul8 software is used in this research to conduct simulations. Because the F-distribution has not been defined in this software, simulation is performed using the relation between F and the Normal distribution. Equation (11) shows that the Chi-square distribution with v degrees of freedom is the distribution of a sum of the squares of v independent Standard Normal random variables (Walpole & Myers, 1993):

$$Z \approx N(0,1);$$

$$X = \sum_{i=1}^{v} Z_i^2 \Rightarrow X \approx \chi^2(v) \qquad (11)$$

Also, the F-distribution with v_1 and v_2 degrees of freedom arises from the ratio of two independent Chi-Square random variables that have been divided on their degrees of freedom according to Equation (12) (Walpole & Myers, 1993):

$$X_1 \approx \chi^2(\nu_1);$$
$$X_2 \approx \chi^2(\nu_2); \qquad (12)$$
$$W = \frac{X_1/\nu_1}{X_2/\nu_2} \Rightarrow W \approx F(\nu_1, \nu_2)$$

Simulation by the Simul8 software

In practice, many problems have probabilistic behavior. One of the most common cases is a queuing problem in which customer arrival rate, service time, and the like are not exact. Queuing theory can be used to solve these types of problems. Sometimes, a complex combination of several queues in exact and/or probabilistic manners with limitations, calculations and different conditions are observed. Examination of such a complex problem by queuing theory is possible theoretically, but nearly impossible practically. Hence, computer and simulation sciences are needed for calculations. Using simulation science, a complex system can be run virtually and its behavior can be forecasted and examined within a reasonable time. The Simul8 software used in this research has a user-friendly graphical aspect and includes the following elements (Simul8 Software, Version 2000):

1) Object. An object is like a customer that moves in a special path to get some services in service centers or work stations and waits in queues until the conditions for service in the next service center are ready. This object may exist in a queue at first or may enter the system through an enter point after the system begins running. The object can have a label. A label is a variable that stores a property of an object.
2) Enter point. New objects come into the system through these points. The object's entry time probability distribution and values of its labels can be nominated.
3) Work station. In this element, the considered object gets service. Service time that can be probabilistic, number of servers, capacity and resources can be assigned. Also, calculations on the model's variables and object's labels can be done in this element.

4) Queue. Objects of the model wait in queues until the conditions to enter the next work station are ready. For example, if the element after the queue is a work station and it needs a resource to work, the object waits in the queue until that resource is ready. The queue capacity can be infinite or bounded.
5) Resource. This element is a variable that obtains the conditions and number of services. The value of each resource can be assigned at first or can be assigned in running time. After the service, the resource value can change to itself or another resource.
6) Exit point. Objects of the model go out of the model from an exit point, for example, the customers that have finished their service.
7) Information store. This element is a variable that stores properties of the model.

Proposed Model

The objective of this article is to obtain the mean of an F-distribution with $1 \leq \nu_1 \leq 30$ and $\nu_2 < 3$ and also the variance of an F-distribution with $1 \leq \nu_1 \leq 30$ and $\nu_2 < 5$. For this purpose, a random sample of size 500 is drawn from a statistical population having an F-distribution. In order to achieve more accurate results, the proposed model is run 100 times and the average of the obtained means and variances are recorded as final results. It should be noted that the proposed model generates Standard Normal random numbers at first and converts them to Chi-square random numbers. Afterwards, Chi-square random numbers are converted to F random numbers for calculations. In the proposed model, queues, work stations, resources and information stores are displayed with Q, W, R and I, respectively. As shown in Figure 1, the proposed model includes three parts. The first and second parts of the model generate the first and second Chi-square random numbers, respectively. Because the performance of both parts is similar, only the first part is described.

In the first part, there is one object in Q1 at first. This object enters W2 while running the model. The service time of W2 is exactly equal to zero. In the other words, W2 is a virtual work

Figure 1: Simulated Model by the Simul8 Software
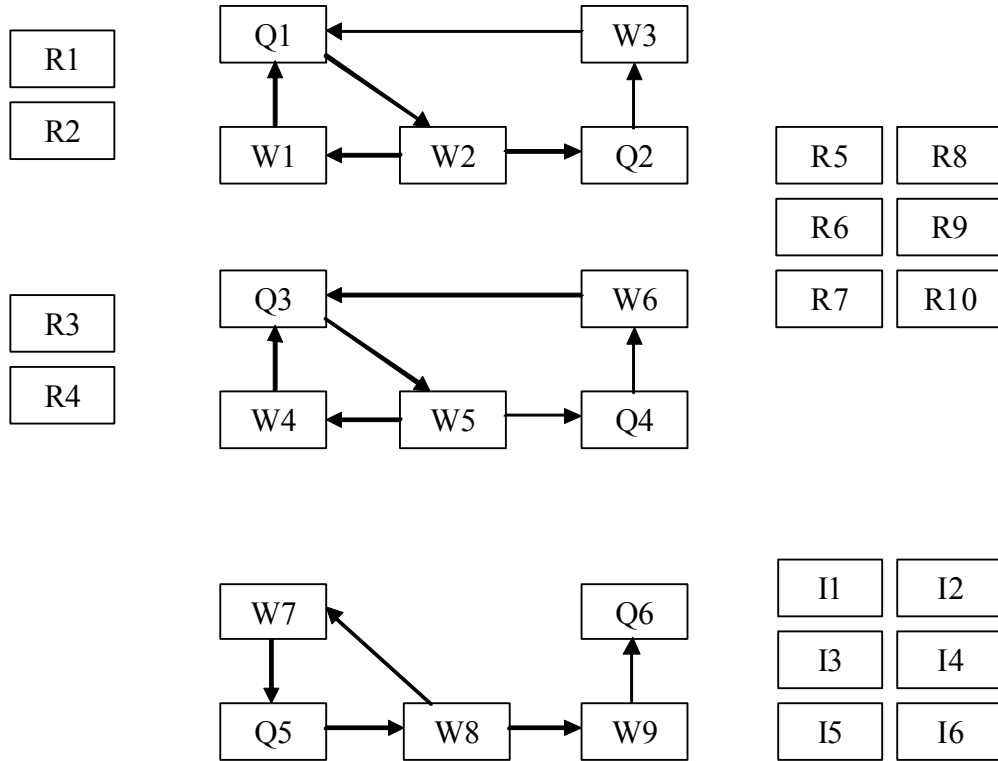
station. In W2, a Standard Normal random number is assigned to the object's label. Afterwards, the object can enter W1 or Q2. In the proposed model, entrance priority belongs to W1, which is also a virtual work station. W1 uses one unit of R1 and then changes it to R2. When R1 finishes, the object cannot go to W1 from W2 and should inevitably enter Q2. In W1, the value of the object's label is squared and added to I1 and it then returns to Q1. The loop Q1-W2-W1 is repeated R1 times and a Chi-square random number with R1 degrees of freedom is generated and stored in I1. The object waits after entering Q2 until the conditions to enter W3 are ready. W3 requires R6 and changes it to R7 after using. The initial value of R6 is zero but after running the model, a value is assigned to it by the third part; R6 is an intermediate between the first and third parts. W3 which is a virtual work station changes the value of R1, that is now equal to zero, to its initial value and also changes the value of R2 and the object's label to zero. In the other words,

the first part of model returns to its initial condition.

In the third part, there is only one object in Q5 at first, the object then enters W8 which is the only real work station of the model and has an exact service time. Because the first and second parts of the model have service times equal to zero they generate Chi-square random numbers I1 with R1 degrees of freedom and I2 with R3 degrees of freedom, respectively, when the object is in W8. Afterwards, the object enters W7 which is a virtual work station. In W7, an F random number is generated by the ratio of I1 and I2 which have been divided on their degrees of freedom (R1 and R3). The generated value and its square are respectively added to I3 and I4 to obtain the sum and sum of squares of values of the F random variable, and then W7 uses one unit of R5 and R8 and changes them to R6 and R9, in that order. Thus, W7 lets the first and second parts of the model generate new Chi-square random numbers. Calculation of sum and sum of squares of values of the F random

variable is repeated R5 times (which is equal to R8 times) in the loop Q5-W8-W7. R5 is considered equal to 500 in the proposed model. After finishing R5 and R8, the object enters W9 from W8. In W9, the values of the sample mean (I5) and sample variance (I6) are calculated using I3 and I4, which are respectively the sum and sum of squares of values of F random variable, according to Equations (13) and (14):

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (13)$$

$$S^2 = \frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)} \qquad (14)$$

By setting resources R1 and R3 equal to the first and second degree of freedom of an F-distribution respectively, mean and variance of a random sample of size 500 (I5 & I6) are calculated. To obtain more accurate results, the proposed model is run 100 times and the average of obtained means and variances are recorded. Table 1 shows the obtained results.

References
Walpole Ronald E. & Raymond H. Myers, Probability and Statistics for Engineers and Scientists, Fifth Edition, Macmillan, 1993.
Manual.pdf Help File, Simul8 Software Version 2000.

Table 1: Calculated Mean and Variance for the F-distribution

F Distribution	Mean		Variance			
V_1	$V_2=1$	$V_2=2$	$V_2=1$	$V_2=2$	$V_2=3$	$V_2=4$
1	11130.21	10.05	8.52E+11	108031.28	3793.61	41.65
2	33169.64	13.75	1.38E+13	541607.53	991.75	1174.13
3	37594.24	14.30	2.38E+13	458742.85	686.37	34.18
4	15012.24	10.74	1.54E+12	112955.47	3717.85	45.92
5	26761.31	12.81	7.69E+12	334859.40	1994.39	74.23
6	23044.24	13.13	4.51E+12	337649.74	5092.49	355.42
7	18210.40	11.22	2.36E+12	165931.92	1215.85	163.34
8	23826.42	13.79	4.96E+12	397856.63	2340.31	155.83
9	22405.18	11.67	3.29E+12	135107.50	3087.51	144.57
10	30296.88	12.90	9.19E+12	220879.02	6534.82	388.49
11	29119.63	11.62	8.05E+12	120507.71	1082.33	251.6
12	28509.47	11.65	6.61E+12	142528.84	2494.88	126.27
13	25922.59	11.60	5.29E+12	130915.62	1814.47	192.06
14	28315.80	12.34	1.07E+13	172780.09	3946.75	93.11
15	21680.74	13.58	3.33E+12	368276.62	1641.08	463.65
16	27846.46	11.80	6.59E+12	146747.04	1627.26	76.08
17	27757.44	13.02	8.41E+12	337102.72	2068.54	248.91
18	25305.75	11.32	4.68E+12	122134.97	634.55	113.93
19	29520.33	13.02	6.81E+12	338522.99	1357.71	171.95
20	24484.70	12.30	5.05E+12	218104.74	1434.49	251.87
21	23815.50	12.92	3.61E+12	245563.95	2027.46	121.14
22	24091.21	12.26	3.81E+12	220336.07	1567	121.76
23	24122.25	13.19	4.21E+12	295272.78	1133.46	150.74
24	24058.93	13.47	3.98E+12	359914.76	1061.62	189.02
25	22481.36	12.01	3.37E+12	191010.65	2112.89	287.59
26	25595.00	12.07	5.59E+12	196676.92	2368.16	157.14
27	27161.81	13.25	5.77E+12	319466.56	2136.64	122.85
28	26124.24	12.42	5.92E+12	230289.00	1710.13	245.72
29	26067.87	12.99	6.83E+12	237958.74	940.05	213.54
30	26485.85	13.58	5.27E+12	370863.43	2165.15	167.76