TEMPORAL AND SPATIALLY HETEROGENEOUS FINITE LENGTH RUNS ANALYSIS

Zekai Şen

Istanbul Technical University, Civil Engineering Faculty, Hydraulics Division, Maslak 34469, İstanbul, Turkey. zsen@itu.edu.tr

Abstract – Various problems in physics and engineering lead to the problem of run-length probability distribution function (pdf) in a finite time series. In this study to find the pdf of run-lengths in such a series, first infinite sequence properties are reviewed and then finite series run length pdf is derived on the basis of simple set theory. This paper presents the derivation of exact run length pdf in finite length dependent series. In the derivation, two different definitions of runs are considered as the integration method for infinite series and combinatorial analysis for finite time series. The analytical derivations are solved numerically and the results are presented in forms of cumulative pdf, expectation, variance and higher order moment changes with the sample lengths. On the other hand, homogenous run properties based on Bernoulli trials are used in many physical and engineering applications for many decades. Heterogeneous regional Bernoulli trial probability distribution model is not available so far in applications and numerical calculations. Herein, a plausible, rational and logical mathematical derivation of the heterogeneous case is derived, which reduces to the classical homogeneous Bernoulli trial case. This paper provides regional probabilistic success and failure period areal coverage modeling, which is useful for temporal and spatial pattern recognition of spatial risk predictions and parameter assessments. The basis of the methodology is mutually exclusive and independent sub-areal (site) success and failure occurrences' heterogeneous probabilities.

Key Words: Bernoulli trials, combinatorial, finite sample, heterogeneity, random field, runs.

1. INTRODUCTION

From the statistical point of view, a run is defined by Mood (1940) as a succession of similar kind of elements proceeded and succeeded by different kinds. The number of elements in such a run is referred to as its length. The theory of runs has been successfully applied to recurrent events by Feller (1957).

Supposing that x_1, x_2, x_3,......, x_n in a sequence of independent random variables and x_0 is a preselected value. It is then possible to truncate the given sequence at x_0 level. Such a truncation procedure leads to a new sequence of two basic elements, namely, a success when $x_i - x_0 > 0$ and a failure if $x_1 - x_0 < 0$. Moreover, these two basic elements are, in fact, the Bernoulli trials. In the classical statistics literature, a success run of lengths s is defined as an uninterrupted sequence of either s or at least s successes. Therefore, a run length has an integer value. Various statistical properties of runs or their functions can be employed in many engineering problems. For instance, in water resources engineering x_0 might represent the demand necessary for water supply

to an urban area or agricultural need for water. In addition, the length of a success run corresponds to wet spells whereas a failure run-length is representative of dry spells. During the subsequent sections in this paper Mood's basic definition of runs will be adopted.

The main purpose of this paper is to present a rigorous analytical methodology to model theoretically the spatial or temporal heterogeneous probabilities in a given region or during certain time duration. The formulation in this study reduces exactly to the well-known Binomial pdf, and it empowers the meteorologists and hydrologists alike to model areal or temporal drought, flood, and wet or dry spell occurrences even in the most complicated set of heterogeneous probability of occurrences.

2. RUNS OF INFINITE SEQUENCES

So far in the application of runs theory two different approaches have been followed in studies of run lengths, namely, the integration and the combinatorial approaches.

It appears that the integration approach refers to runs of infinite population as employed by [2]. If the sequence of random variables is stationary and ergodic then any run length is synonymous with the first run. The relation between the probability mass, $P\{s = j\}$ of a given sequence run-length, $\{s = j\}$, and the probability distribution function $P\{s > j\}$ is given as [2]

$$P\{s = j\} = P\{s \geq j\} - P\{s \geq j+1\} \tag{1}$$

However, the following relationship for the r-th order moment about the origin has been presented.

$$E\left(s^r\right) = \sum_{j=1}^{\infty} j^r P\{s = j\} = \sum_{j=1}^{\infty}\left[j^r - (j-1)^r\right]P\{s \geq j\} \tag{2}$$

In general, for stationary and ergodic stochastic processes it is possible to write

$$P\{s \geq j\} = P\{j\} + \sum_{k=1}^{\infty} P\{k, j\} \tag{3}$$

where $P\{j\}$ is the probability of j successive successes and, $P\{k,j\}$ is the joint probability of the simultaneous occurrence of j successes to be followed by k successive failures. In general, the computation of $P\{k,j\}$ can be achieved through the multiple integration of the joint probability distribution function (PDF), $f(x_1, x_2,......., x_{k+j})$, of random variables $x_1, x_2, x_3,...., x_{k+j}$ which can be written as

$$\mathbf{P\{k, j\}} = \int_{-\infty}^{x_0} .... \int_{-\infty}^{x_0} \int_{x_0}^{+\infty} .... \int_{x_0}^{+\infty} \mathbf{f}\left(\mathbf{x_1, x_2,....x_{k+j}}\right)\mathbf{dx_1 dx_2......dx_{k+j}} \tag{4}$$

In the case of dependent random variables, equation (4) has been evaluated in terms of tetrachoric series expansion [3]. However, for independent and identically distributed random variables, equation (4) simplifies down to k+j factors which render the multiple integration into its simplest form as

$$P\{k,j\} = \prod_{i=1}^{k} \int_{-\infty}^{x_0} f(x_i)dx_i \prod_{i=k+1}^{k+j} \int_{x_0}^{+\infty} f(x_i)dx_i \qquad (5)$$

or in terms of success and failure probabilities which are expressed as

$$p = \int_{x_0}^{+\infty} f(x_i)dx_i$$

and

$$q = 1-p = \int_{-\infty}^{x_0} f(x_i)dx_i$$

respectively. With these two last expressions, equation (5) takes the form as

$$P\{k,j\} = q^k p^j \qquad (6)$$

On the other hand, for $k = 0$

$$P\{j\} = p^j \qquad (7)$$

Furthermore, the substitutions of equations (6) and (7) into equation (3) leads to

$$P\{s \geq j\} = p^{j-1} \qquad (8)$$

or from equation (1)

$$P\{s = j\} = qp^{j-1} \qquad (9)$$

By making use of this probability mass, the expected and then the variance values of success runs in an infinite sequence can be obtained as [2]

$$E(s) = \frac{1}{q} \qquad (10)$$

and

$$V(s) = \frac{p}{q^2} \qquad (11)$$

respectively. It is also possible to extent the above mentioned procedure to cover the dependent random variables. Various properties of run-lengths for the Markov process have been already derived [4].

3. RUNS OF FINITE SEQUENCES

In the case of finite sequences the runs have been examined by means of the combinatorial analysis [5]. It is possible through this analysis to find the total number of either success or failure runs regardless of their lengths and the longest run-length of any given type.

However, the combinatoric analysis fails to yield any statistical moments such as the expected value of say, success run-length in a finite sequence. The following methodology which is based on the integration approach has been developed in order to evaluate the probability mass of a run-length in finite sequences.

Let $S_i = \{s = i\}$ be an event denoting the occurrence of s successive runs of length i in an infinite sequence of Bernoulli trials. It is then possible to verify from equation (9) that the union of all these events is

$$\bigcup_{i=1}^{\infty} E_i = S \tag{12}$$

where S is a universal set with probability equal to one. Furthermore, all these events are mutually exclusive , that is

$$E_i \bigcap_{i \neq j}^{\infty} E_j = \Phi \ (i, j = 1, 2, \dots, \infty) \tag{13}$$

in which Φ denotes the empty set with zero probability. However, for a finite sequence of length n, no run-length greater than n can occur. Hence, the sample space is restricted by n. Therefore,

$$\bigcup_{i=1}^{n} E_i = \Omega \tag{14}$$

where Ω represents a subsample space contained in the universal set S. Since the elementary events E_i are mutually exclusive, the following statement can be written for the probability of the subsample, $P\{\Omega\}$ as,

$$P\{\Omega\} = \sum_{i=1}^{n} P\{E_i\} \leq 1 \tag{15}$$

In order to render the subsample space to a proper one over which the mass of probabilities of elementary events is equal to one, it is necessary to reappraise the probabilities of N_i's as $P\{E|\Omega\}$ which transforms equation (15) into an equality form by defining a new set denoted as Ω' :

$$P(\Omega') = \sum_{i=1}^{n} P\{E_i / \Omega\} = 1 \tag{16}$$

where

$$P\{E_i / \Omega\} = \frac{P\{E_i \bigcap \Omega\}}{P\{\Omega\}} \tag{17}$$

Since, E_i is an element of the set, Ω, it is, therefore, known from set theory that $P\{E \cap \Omega\} = P\{E_i\}$ the substitution of which into equation (17) gives.

$$P\{E_i / \Omega\} = \frac{P\{E_i\}}{P\{\Omega\}} \tag{18}$$

which is fundamental in evaluation probability mass of run-length in finite sequences. In the case of independently and identically distributed random variables equation (15) can be evaluated by considering equation (9) which yields

$$P\{\Omega\} = 1 - q^n \tag{19}$$

Furthermore, equation (18) can be written more explicitly as

$$P\{s = i / n\} = \frac{P\{s = i\}}{1 - q^n}$$

or

$$P\{s = i / n\} = \frac{qp^{n-1}}{1 - q^n} \tag{20}$$

where n has been employed instead of Ω, since it is the only parameter that characterizes Ω.

It is convenient to interpret equation (20) as the truncated form of the probability mass function given by equation (9) at a truncation level equal to n. On the other hand, the expected value of success run-length in a finite sequence with n elements can be found by subtracting equation (20) from equation (2) with r = 1 which after some algebra yields

$$E(s / n) = \frac{1}{q} - \frac{np^n}{1 - q^n} \tag{21}$$

where the second term on the right hand side becomes zero when n goes to infinity, hence equation (21) reduces to equation (10). In a similar manner, the variance of s can be evaluated by finding the second order original moment which becomes

$$E\left(s / n^2\right) = \frac{1 - p^n - 2nqp^n - n^2 qp^n}{q} + 2p\frac{1 - p^n - nqp^n}{q^3} \tag{22}$$

or the variance can be estimated as

$$V(n / s) = \frac{1 - p^n - 2nqp^n - n^2 qp^n}{q} + 2p\frac{1 - p^n - nqp^n}{q^3} - \left(\frac{1}{q} - \frac{np^n}{1 - q^n}\right)^2 \tag{23}$$

which for $n \rightarrow \infty$ reduces to equation (11)

The calculation of higher order moments than two is possible but requires very tedious algebraic work. Perhaps, it is better to find these methods numerically on a digital computer. For instance, the m-th order moment can be evaluated by

$$E\left(s^m / n\right) = \sum_{i=1}^{n} i^m P\{s = i / n\}$$

All the above referred formula is applicable for failure runs provided that p and q in the formulations are interchanged.

4. RUNS OF DEPENDENT VARIABLES IN A FINITE SEQUENCE

So far in the runs theory applications two different approaches are followed for run-lengths. These are the integration and the combinatorial approaches. This approach refers to runs of infinite population. If the sequence of random variables is stationary and ergodic then any run-length is synonymous with the first run. In general, the relation between the probability mass, P{s = j} of a given sequence run-length, {s = j}, and the probability distribution function (pdf), P{s > j}, is given as [2],

$$P\{s=j\}=P\{s\geq j\}-P\{s\geq j+1\} \tag{24}$$

This gives rise to the following relationship for the r-th order moment about the origin,

$$E\left(s^r\right)=\sum_{j=1}^{\infty}j^r P\{s=j\}=\sum_{j=1}^{\infty}\left[j^r-(j-1)^r\right]P\{s\geq j\} \tag{25}$$

In general, for stationary and ergodic stochastic processes one can write,

$$P\{s\geq j\}=P\{j\}+\sum_{k=1}^{\infty}P\{k,j\} \tag{26}$$

where $P\{j\}$ is the probability of j successive successes and, $P\{k,j\}$ is the joint probability of the simultaneous occurrence of j successes followed by k successive failures. The computation of $P\{k,j\}$ can be achieved through the multiple integration of the joint pdf, $f(x_1, x_2,......., x_{k+j})$, of random variables $x_1, x_2, x_3,...., x_{k+j}$, which can be written as,

$$P\{k,j\}=\int_{-\infty}^{x_0}(k)\int_{-\infty}^{x_0}\int_{x_0}^{+\infty}(j)\int_{x_0}^{+\infty}f\left(x_1,x_2,.\ .\ .\ ,x_{k+j}\right)dx_1 dx_2 \ldots dx_{k+j} \tag{27}$$

In the case of dependent random variables, Eq. (27) has been evaluated in terms of tetrachoric series expansion [3]. Factorization for dependent and identically distributed random variables simplifies Eq. (27) down to $(k + j)$ factors as follows [4],

$$P\{k,j\}=P\{x_1<x_0\}\prod_{i=2}^{k-1}P\{x_i<x_0\mid x_{i-1}<x_0\}P\{x_{k+1}>x_o\mid x_k>x_0\}\prod_{i=k+2}^{k+j}P\{x_i>x_0\mid x_{i-1}>x_0\} \tag{28}$$

Furthermore, each conditional probability in this expression is simplified in terms of success and failure probabilities as,

$$P\{x_i>x_0\mid x_{i-1}>x_0\}=r$$

$$P\{x_i>x_0\mid x_{i-1}\leq x_0\}=\frac{p}{q}(1-r)$$

$$P\{x_i\leq x_0\mid x_{i-1}\leq x_0\}=1-\frac{p}{q}(1-r) \tag{29}$$

$$P\{x_i\leq x_0\mid x_{i-1}>x_0\}=(1-r)$$

where p and q are the success and failure probabilities which are expressed as,

$$p=\int_{x_0}^{+\infty}f(x_i)dx_i$$

and

$$q=1-p=\int_{-\infty}^{x_0}f(x_i)dx_i$$

respectively. With these two last expressions, Eq. (28) takes the form as,

$$P\{k, j\} = p(1-r)\left[1 - \frac{p}{q}(1-r)\right]^{k-1} r^{j-1} \tag{30}$$

On the other hand, for k = 0

$$P\{j\} = pr^{j-1} \tag{31}$$

Furthermore, the substitutions of Eqs. (30) and (31) into Eq. (26) lead to,

$$P\{s \geq j\} = r^{j-1} \tag{32}$$

from Eq. (1)

$$P\{s = j\} = (1-r)r^{j-1} \tag{33}$$

By making use of this probability mass, the expected and then the variance values of success runs in an infinite sequence can be obtained as [2],

$$E(s) = \frac{1}{1-r} \tag{34}$$

and

$$V(s) = \frac{r}{(1-r)^2} \tag{35}$$

respectively. It is also possible to reduce the above mentioned procedure to cover the independent random variables, when substitution of p instead of r.

5. COMBINATORIAL METHOD AND FINITE SEQUENCE

In finite sequence cases, the runs are examined by the combinatorial analysis [5]. It is possible through this analysis to find the total number of either success or failure runs regardless of their lengths and the longest run-length of any given type [6]. However, the combinatorial analysis fails to yield any statistical moments such as the expected value of say, success run-length in a finite sequence. The following methodology which is based on the integration approach is developed in order to evaluate the probability mass of a run-length in finite sequences, which helps to calculate the statistical moments.

Let $S_i = \{s = i\}$ be an event denoting the occurrence of s successive runs of length i in an infinite sequence of Bernoulli trials. It is then possible to verify from Eq. (33) that the union of all these events is

$$\bigcup_{i=1}^{\infty} E_i = S \tag{36}$$

where S is a universal set with probability equal to one. Furthermore, all these events are mutually exclusive, that is

$$E_i \bigcap_{i \neq j}^{\infty} E_j = \Phi \ (i, j = 1, 2, \ldots, \infty) \tag{37}$$

in which Φ denotes the empty set with zero probability. However, for a finite sequence of length n, run-lengths greater than n cannot occur, because the sample space is restricted by n, therefore,

$$\bigcup_{i=1}^{n} E_i = \Omega \tag{38}$$

where Ω represents a sub-sample space in the universal set S. Since the elementary events E_i are mutually exclusive, the following statement can be written for the probability of the sub-sample, $P\{\Omega\}$ as,

$$P\{\Omega\} = \sum_{i=1}^{n} P\{E_i\} \le 1 \tag{39}$$

In order to render the sub-sample space to a proper one over which the mass of elementary events' probabilities is equal to one, it is necessary to reappraise the probabilities of N_i's as $P\{E|\Omega\}$, which transforms Eq. (39) into an equality form by defining a new set as Ω'.

$$P(\Omega') = \sum_{i=1}^{n} P\{E_i | \Omega\} = 1 \tag{40}$$

where

$$P\{E_i | \Omega\} = \frac{P\{E_i \bigcap \Omega\}}{P\{\Omega\}} \tag{41}$$

Since, E_i is an element in the set, Ω, it is, therefore, known from set theory that $P\{E \cap \Omega\} = P\{E_i\}$, the substitution of which into Eq. (41) gives.

$$P\{E_i | \Omega\} = \frac{P\{E_i\}}{P\{\Omega\}} \tag{42}$$

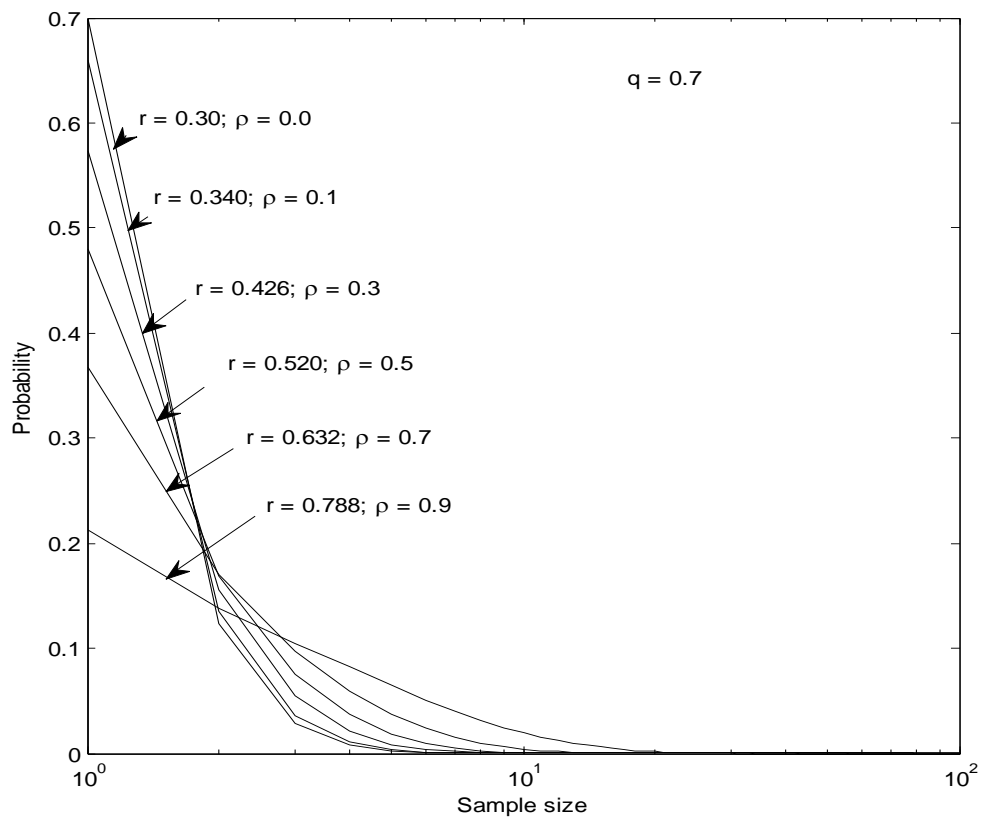this is fundamental in evaluation probability mass of run-length in finite sequences.
In the case of independently and identically distributed random variables Eq. (39) can be evaluated by considering Eq. (33) which yields,
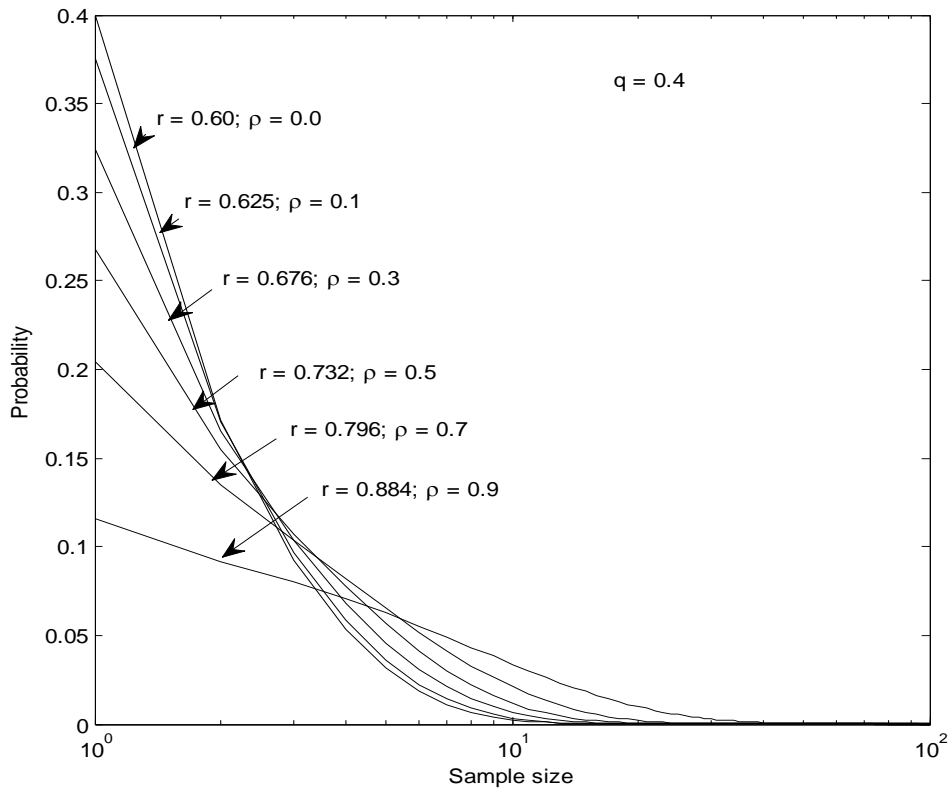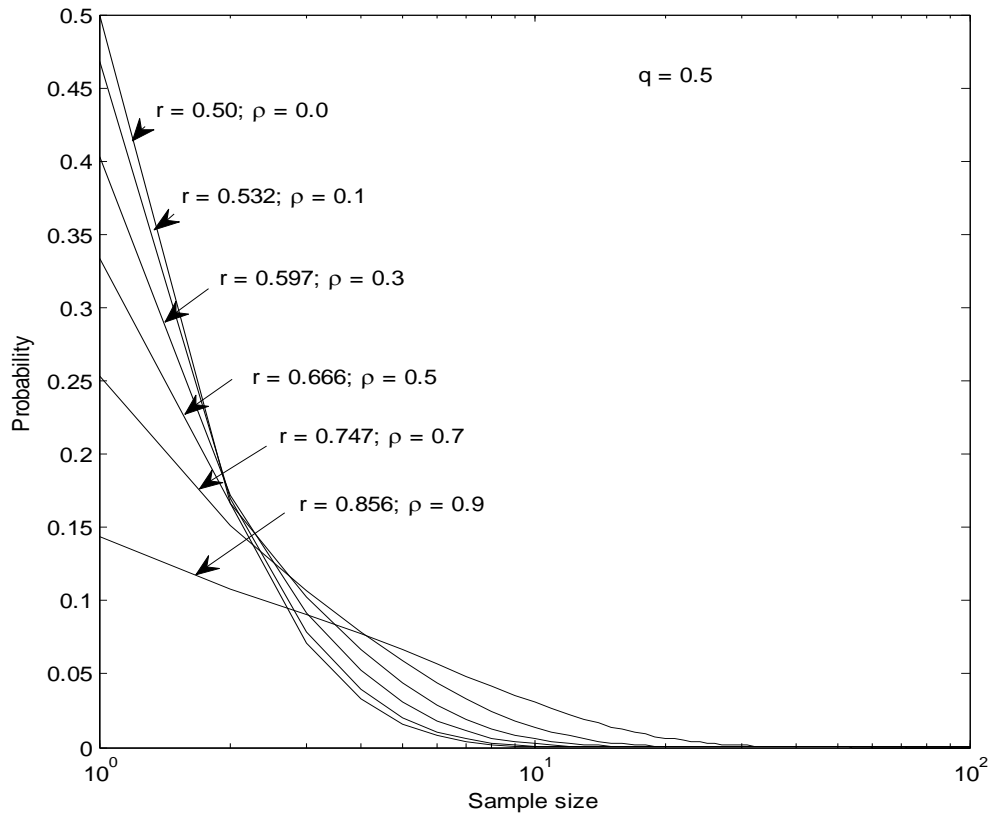
$$P\{\Omega\} = 1 - (1-r)^n \tag{43}$$

Furthermore, Eq. (42) can be written more explicitly as,

$$P\{s = i | n\} = \frac{(1-r)r^{i-1}}{1-(1-r)^n} \tag{44}$$

where n has been employed instead of Ω, since it is the only parameter that characterizes Ω. It is convenient to interpret Eq. (44) as the truncated form of the probability mass function given by Eq. (33) in a finite sample size n. The graphical representation of Eq. (44) for different sample sizes and r values [4] are given in Figure 1.

q = 0.5

r = 0.50; ρ = 0.0
r = 0.532; ρ = 0.1
r = 0.597; ρ = 0.3
r = 0.666; ρ = 0.5
r = 0.747; ρ = 0.7
r = 0.856; ρ = 0.9

Probability

Sample size

q = 0.4

r = 0.60; ρ = 0.0
r = 0.625; ρ = 0.1
r = 0.676; ρ = 0.3
r = 0.732; ρ = 0.5
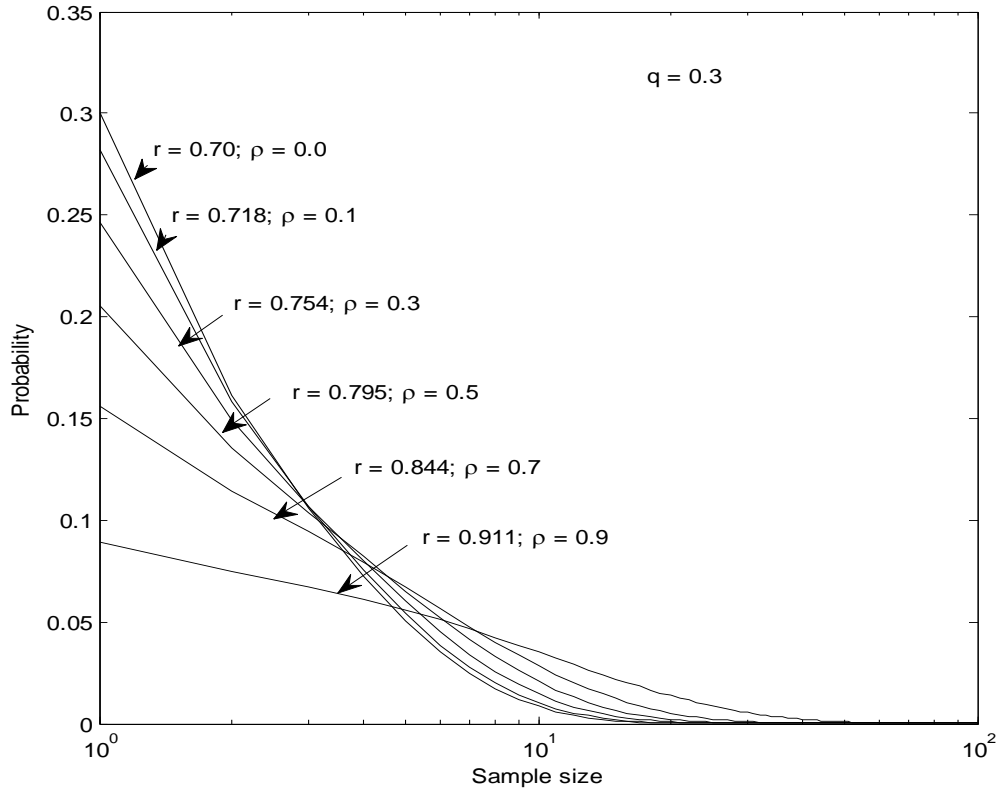r = 0.796; ρ = 0.7
r = 0.884; ρ = 0.9

Probability
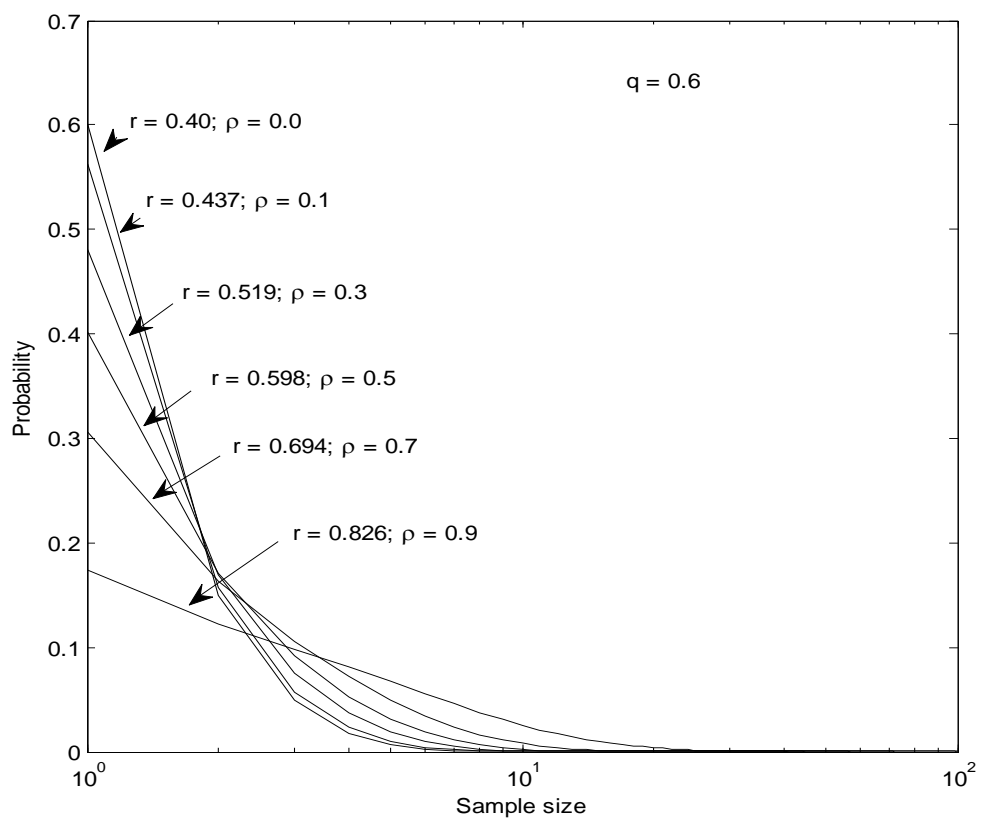
Sample size

Figure 1. Run-length probability variations by sample size

6. FINITE LENGTH RUN-LENGTH STATISTICAL

On the other hand, the expected value of success run-length in a finite sequence with n elements can be found by substitution of Eq. (33) into Eq. (25) with r = 1, which after some algebra yields

$$E(s \mid n) = \frac{nr^{n+1} - n^r - r^n + 1}{(1-r)\left[1 - (1-r)^n\right]} \tag{45}$$

where the second term on the right hand side becomes zero when n goes to infinity, hence Eq. (45) reduces to Eq. (34). The expected value variation with the sample size is presented in Figure 2.
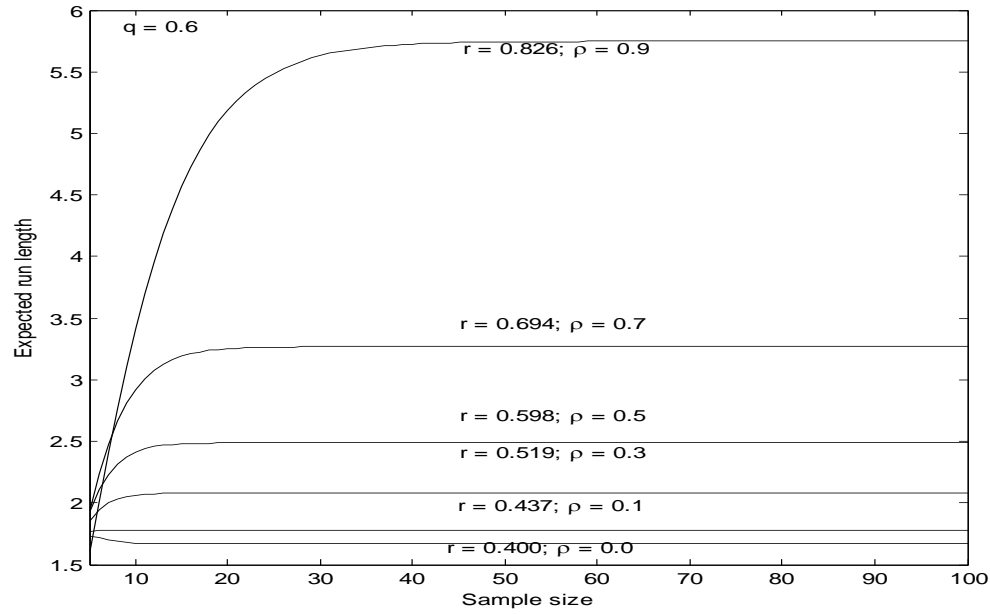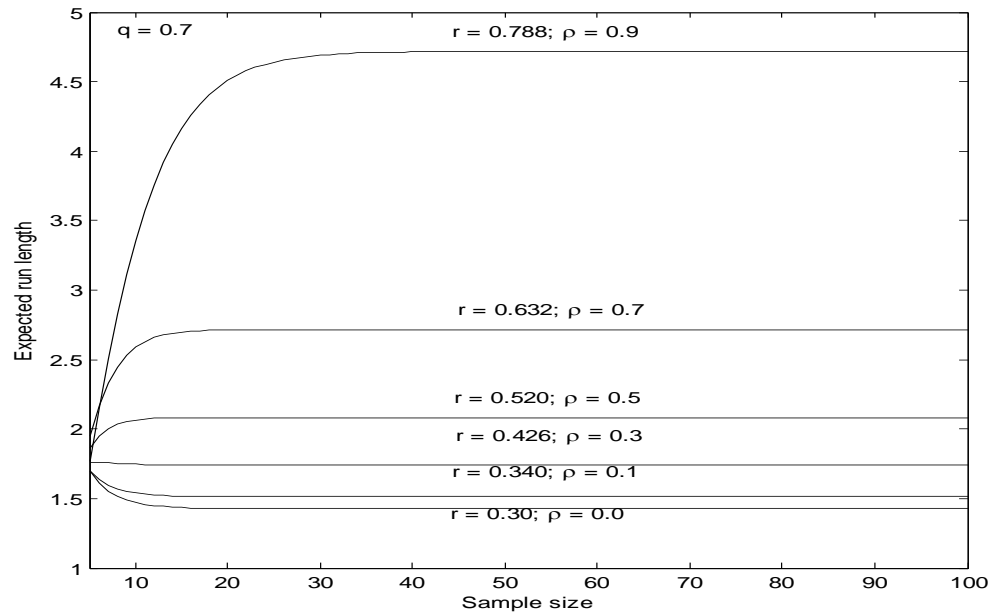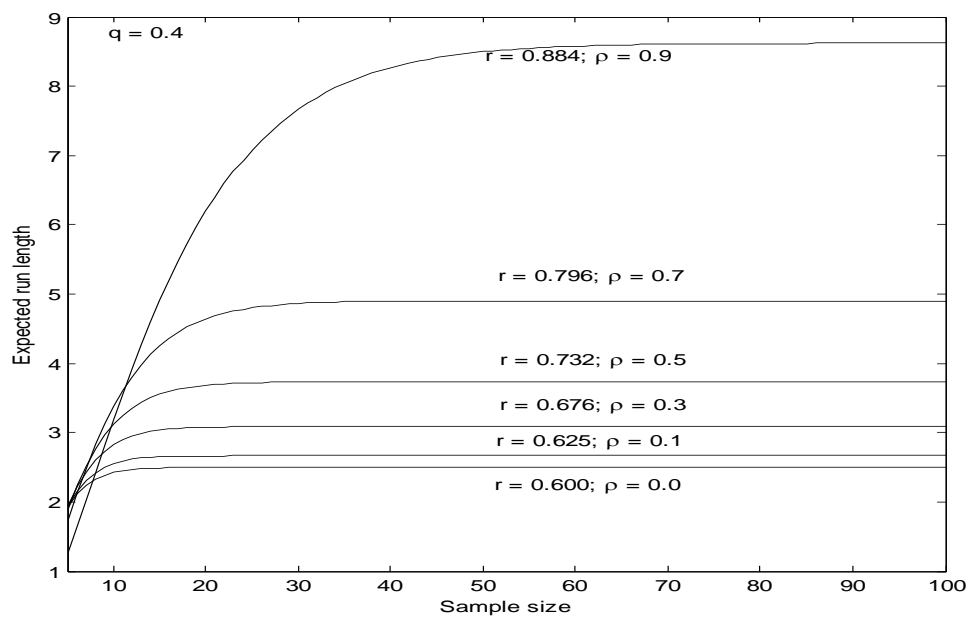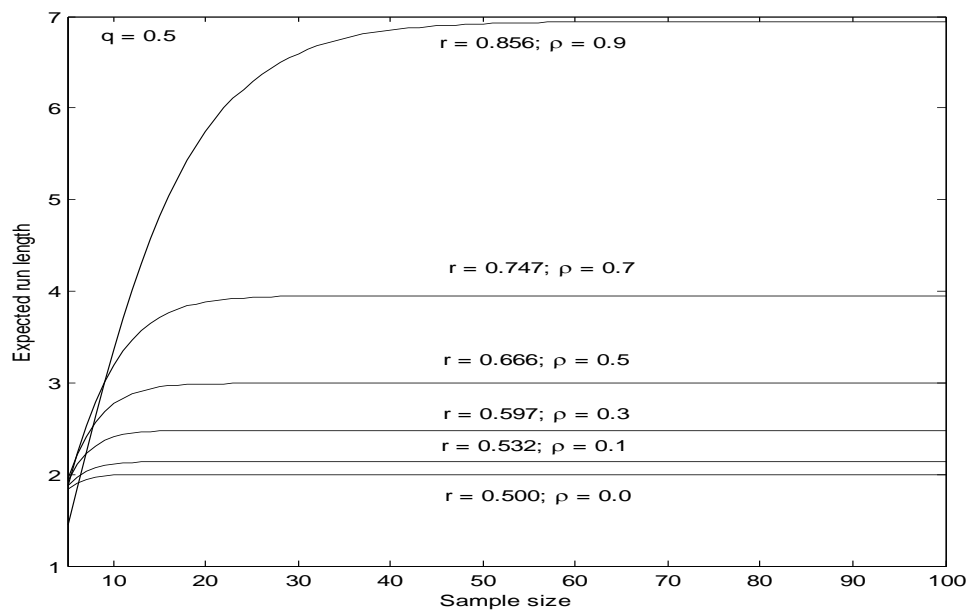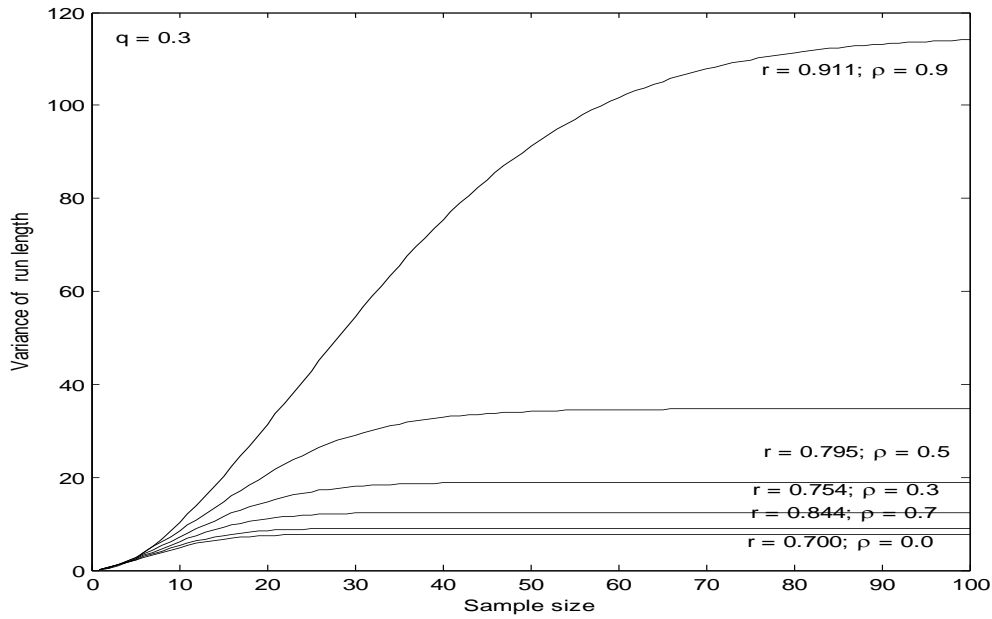
Figure 2. Run-length expectation variations by sample size

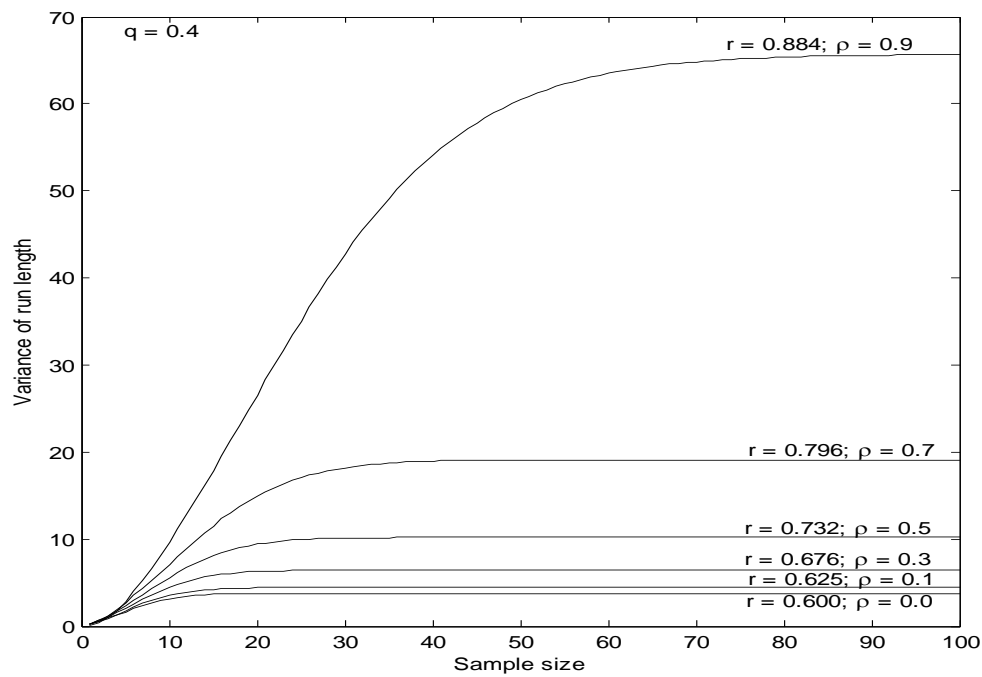In a similar manner, the variance of s can be evaluated by finding the second order original moment which becomes,

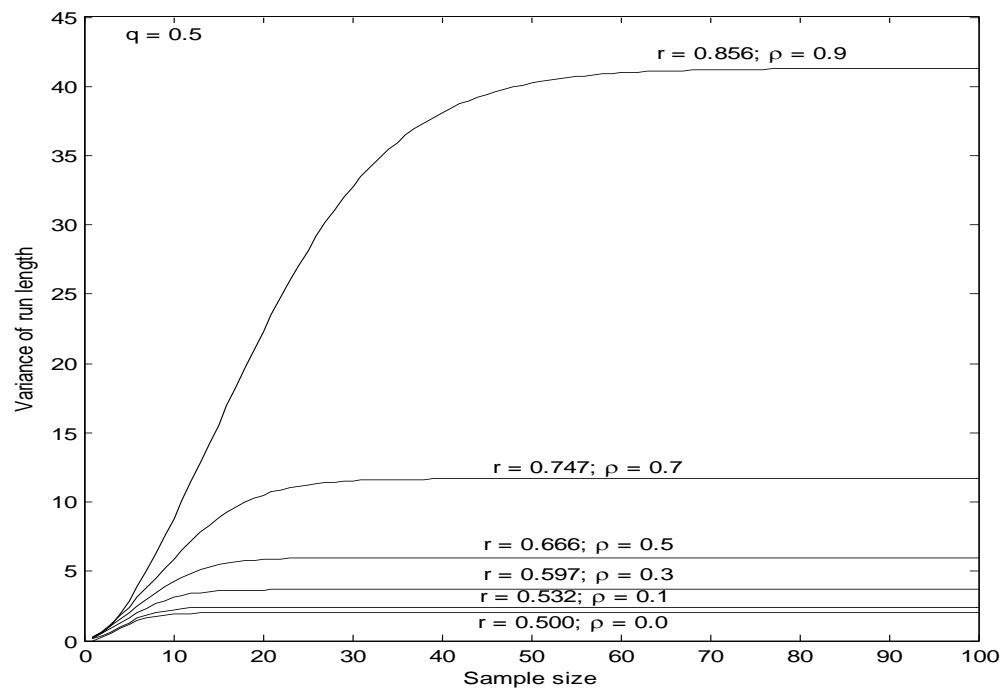$$E\left(s|n^2\right) = \frac{-2n^2 r^{n+1} + n^2 r^{n+2} + n^2 r^n - 2nr^{n+1} + r^{n+1} + 2nr^n + r^n - r - 1}{(r-1)^3} \quad (46)$$

On the other hand, the variance can be estimated as,

$$V(n\,|\,s) = \frac{+2n^2 r^{n+1} - n^2 r^{n+2} - n^2 r^n + 2nr^{n+1} - r^{n+1} - 2nr^n - r^n + r + 1}{(1-r)^2\left[1-(1-r)^n\right]} - E^2(n|s) \quad (47)$$

which for $n \to \infty$ reduces to Eq. (36). Variation change with the sample size is presented in Figure 3.
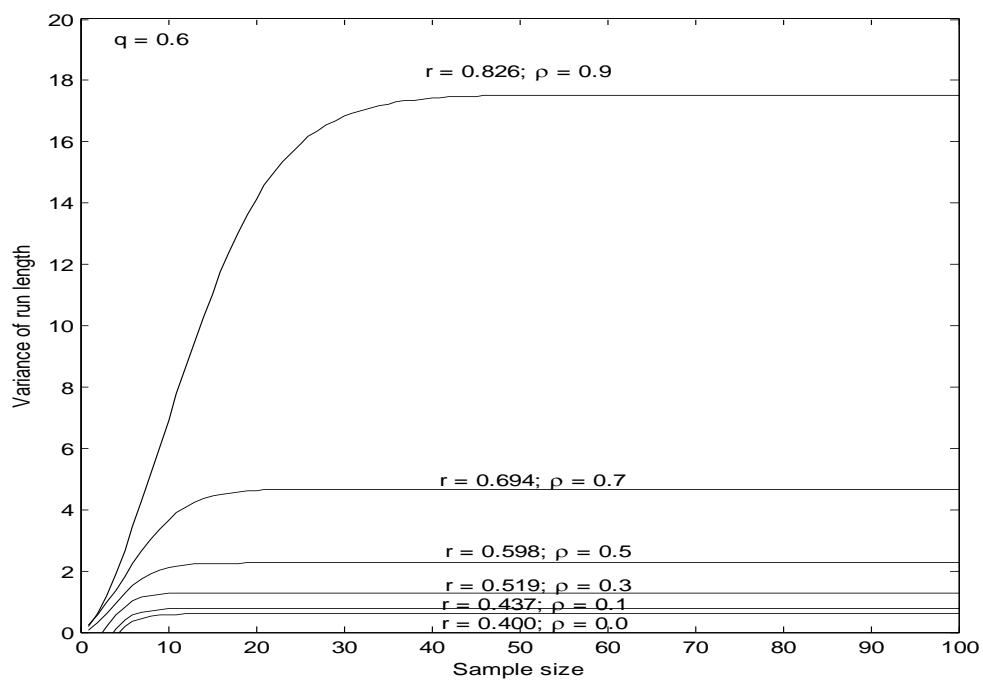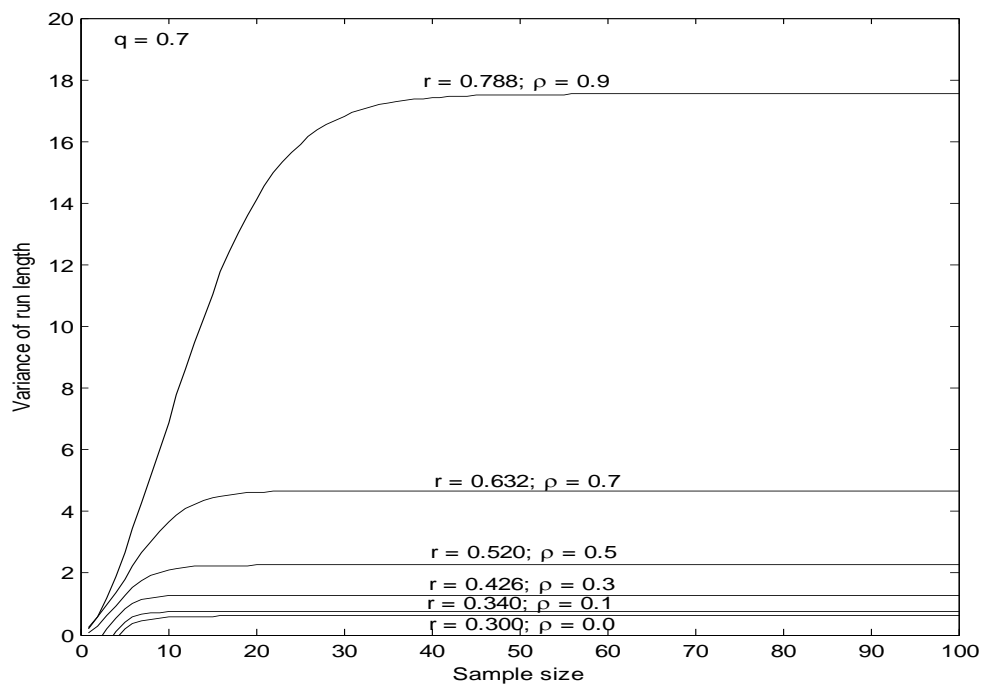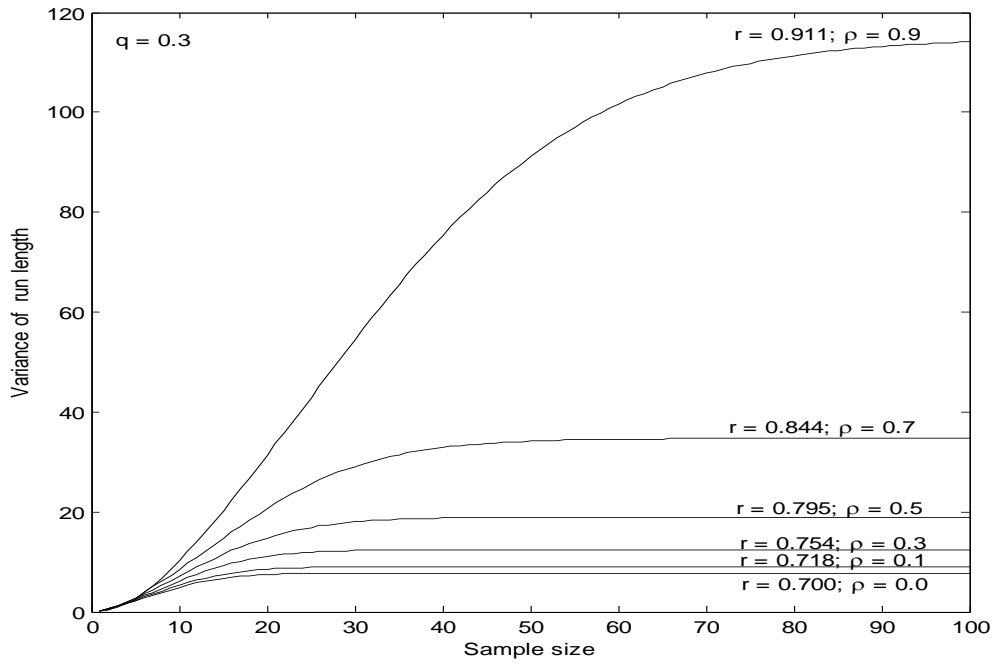
Figure 3. Run-length variation changes by sample size

The calculation of higher order moments than two is possible but requires very tedious algebraic work. Perhaps, it is better to find these methods numerically on a digital computer. For instance, the m-th order moment can be evaluated by

$$E\left(s^m \mid n\right) = \sum_{i=1}^{n} i^m P\{s = i \mid n\} \tag{48}$$

The third and fourth order expansion of this series with the substitution of Eq. (44) in this last expression is given in the Appendix. All the above referred formula is applicable for failure runs provided that p and q in the formulations are interchanged.

7. SPATIALLY HETEROGENEOUS RUNS

Natural phnomenon evolves temporally as well as spatially and they are recorded through the measuring instruments at a set of pointstemporally. Such a space-time distribution is referred to as a field quantity. Since natural phenomenon cannot be predicted with certainty they are assumed to be random and hence, it is necessary to study a new class of fields, i.e., random fields. Quantitatively, a random field is characterized as $\xi(s, t)$ and called a random function at a point, s, and time, t. It is obvious that random field concept is generalization of a stochastic process for which time is the scale variable. The expression "the random function of the coordinate (s, t)" must be understood in the sense that at each point (s, t) of four-dimensional space-time, the value $\xi(s, t)$ is a random variable, and consequently, cannot be predicted exactly. In addition, the values $\xi(s, t)$ are subject to a certain law of probability. A complete description of a random field can be

achieved by constructing all the finite dimensional pdf of the field at different points in space. A random field ξ (s) at a fixed time instant with the given finite dimensional pdf is called homogeneous, if this function is invariant with respect to a shift of the system of points. The same random field is called statistically homogeneous and isotropic (if it is homogeneous in the sense indicated above), while the pdf's are invariant with respect to an arbitrary rotation of the system of points such as of a solid body and to a mirror reflection of this system with respect to the arbitrary plane passing through the origin of the coordinate system. In other words, statistical moments depend upon the configuration of the system of points for which they are formed, but not upon the position of the system in space [7].

In this section precipitation phenomenon is modeled as a random field where time-wise probabilities at a fixed site are referred to as the PoP and spatial probabilities for a fixed time instant are coverage probabilities (CP). The areal probability is the fraction of the area hit by rainfall. It does not provide a means for depicting which of the subareas is going to be hit with precipitation event. However, it simply represents the estimate of what the fractional coverage area will be.   By definition the PoP at any desired threshold value, x_0, such as standard 0.01 inches is equivalent to the exceedence probability of this value. If the pdf of precipitation at site i is denoted by $f_i(X)$ then the PoP, p_i, at this site is

$$p_i = \int_{x_0}^{\infty} f_i(X)dX \tag{49}$$

Even in the heterogeneous point probability case one can simple calculate the average CP, $\bar{p}$, for a given number, n, of sites as,

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{50}$$

In the special case where the point probability is homogeneous over the forecast area, $\bar{p}$ is equal to the common value of the point probability. However, in heterogeneous case it does not mean any physical correspondence.

7.1   Theoretical treatment

In the most general case, none of the sites have equal PoP's which implies that the random field is heterogeneous. It is quite likely that probabilities might vary from place to place even within a single area. In practice, in addition to the spatial correlation variations, the following three situations give rise to heterogeneous and anisotropic random fields. These are;

(i) Identical pdf's of precipitation at different sites but non-uniform threshold levels,

(ii) Non-identical pdf of precipitation at sites but uniform threshold level,

(iii) Non-identical pdf's at sites and non-uniform threshold levels.

Let the PoP and its complementary probability at j-th site within a region of n sites be given by p_j and q_j, (j=1, 2,...., n), respectively. They are point-wise mutually exclusive and complementary, i.e., $p_j + q_j = 1$. The CP, $P(C=i|n)$ including i sites can be evaluated through enumeration technique after the application of summation and multiplication rules of probability theory.   For the derivation of this expression the conceptual model can be

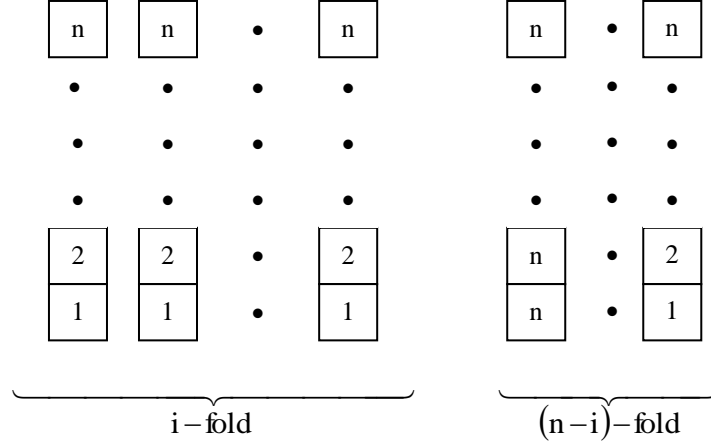visualized as in Figure 4, where there are n mutually exclusive sites represented in square boxes.

Figure 4. Probability derivation conceptual model

Each box has two mutually independent and exhaustive events as p_j and q_j for site j. If out of n sites, i sites are required to have precipitation and the remaining (n-i) sites without, then the whole possible cases (alternative sequences) can be categorized into two mutually exclusive groups. The first group includes i cases with n possibility in each site and the second (n-i) sites again each with n possibilities. Since the PoP on each site is independent from others then the probability of precipitation occurrence collectively can be found from the multiplication rule of independent events as $p_1.p_2.p_3 \ldots p_i$ for one pattern. In case of i sites with precipitation will have $i(i-1)(n-2) \ldots 1 = i!$, mutually exclusive alternatives each with probability of precipitation. Mutually exclusive events imply summation in the probability theory, and therefore, the successive summation terms express all the possible, collective and exhaustive joint probabilities of i sites. The second part on the right hand side in Figure 1 corresponds to the joint probability of the remaining (n − i) sites each belonging to respective pattern of joint PoP occurrence at i precedent pattern. Again the independence principle of the probability theory provides the multiplication of the remaining non-precipitation occurrence sites.

The explanation in the last paragraph can be translated into a mathematical formulation as follows:

$$P(C=i \mid n)=\frac{1}{i!}\sum_{\substack{j_1=1}}^{n}\sum_{\substack{j_2=1 \\ j_2 \neq j_1}}^{n}\ldots\ldots\sum_{\substack{j_i=1 \\ j_i \neq j_k \\ k=1,2,\ldots i-1}}^{n} p_{j_1}p_{j_2}\ldots\ldots p_{j_i}\prod_{\substack{m=1 \\ m \neq j_l \\ l=1,2,\ldots i}}^{n} q_m \qquad (51)$$

$$\underbrace{\hspace{5cm}}_{i-\text{fold}}\qquad\underbrace{\hspace{3cm}}_{(n-i)-\text{fold}}$$

where the i summation terms in the first horizontal big bracket includes all the possible combinations of i precipitation occurrences at n sites, whereas the second horizontal big bracket implies the multiplication term corresponding to possible non-precipitation combinations.

For heterogeneous PoP's the term in the first horizontal big bracket simplifies to $n(n-1) \ldots (n-i+1)p^i$ and the in the next horizontal big bracket multiplication yields to q^{n-i}, hence Eq.(51) reduces to,

$$P(C=i) = \binom{n}{i} p^i q^{n-i} \tag{52}$$

This is the well-known Binomial distribution with two-stage Bernoulli trials [2].

7.2 . Numerical applications

The first numerical application will be the confirmation of the basic expression (Eq. 51) with the well-known Binomial pdf. For this purpose, a Matlab program is developed for heterogeneous point probabilities of different sizes. In Figure 5, Eqs. (51) and (52) results are plotted versus each other at all probability levels. The scatter of points for different sample sizes fall exactly on the 45° line confirming the validity of heterogeneous probability formulation, as in Eq. (51).
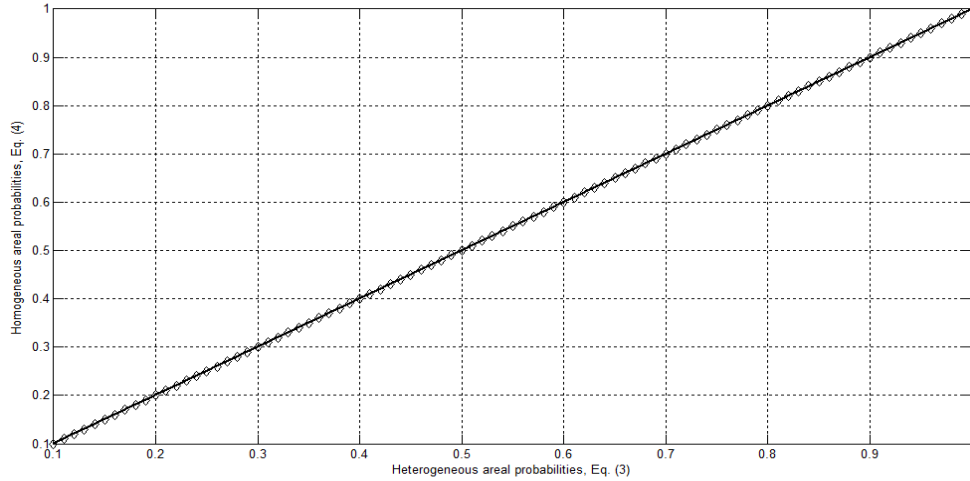

Figure 5. Comparisons of Eqs. (3) and (4)

After the validation of Eq. (51), it is now possible to use this formulation for any set of heterogeneous probability sequences in time or groups within an area. Prior to real data application, it is helpful to look at various coverage probability variations with number of sites and the number of wet spells among such sites. For instance, Figure 6 shows the areal or temporal heterogeneous probabilities sets {0.1, 0.2, 0.3, 0,4, 0.5, 0.6, 0.7, 0.8, 0.9}. Although the probability values are in a systematic sequence in this set, they may occupy any location site among the available number of locations. In this figure n = 12 sites are considered and the variation of coverage probabilities are calculated according to Eq. (51)

if 1, 2, . . , 8 sites are subject to wet spells. The coverage probability indicates the probability of any 1, 2, . . . , 8 sites to be under the influence of wetness. These wet sites are randomly distributed in a given area of 12, say, meteorology stations, or any 1, 2, . . . , 8 wet sites among the available 12 time periods.
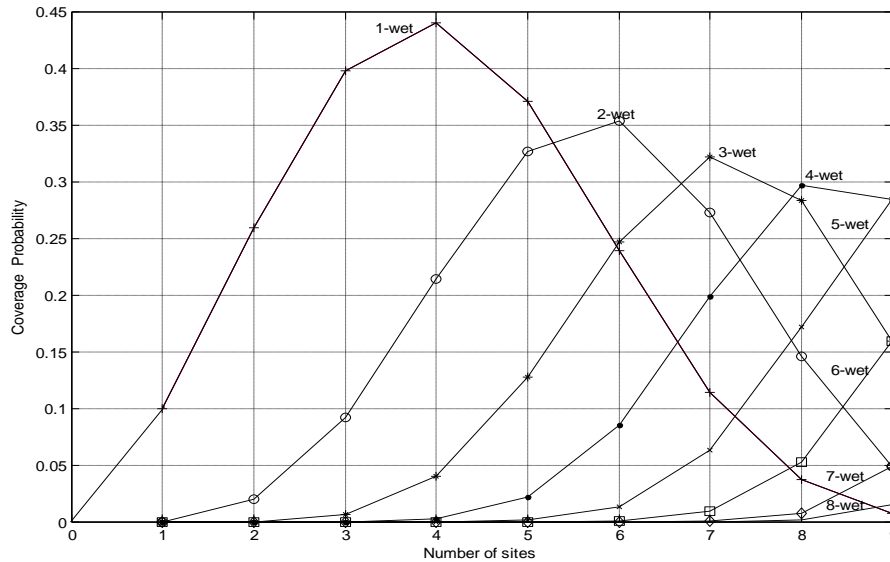
Figure 6. Some heterogeneous ACP variation

Another example is given in Figure 7 for n = 9 sites and with the same probability of set as in the previous example. The comparison of these two figures indicates that as the number of sites increases the CP also increases.
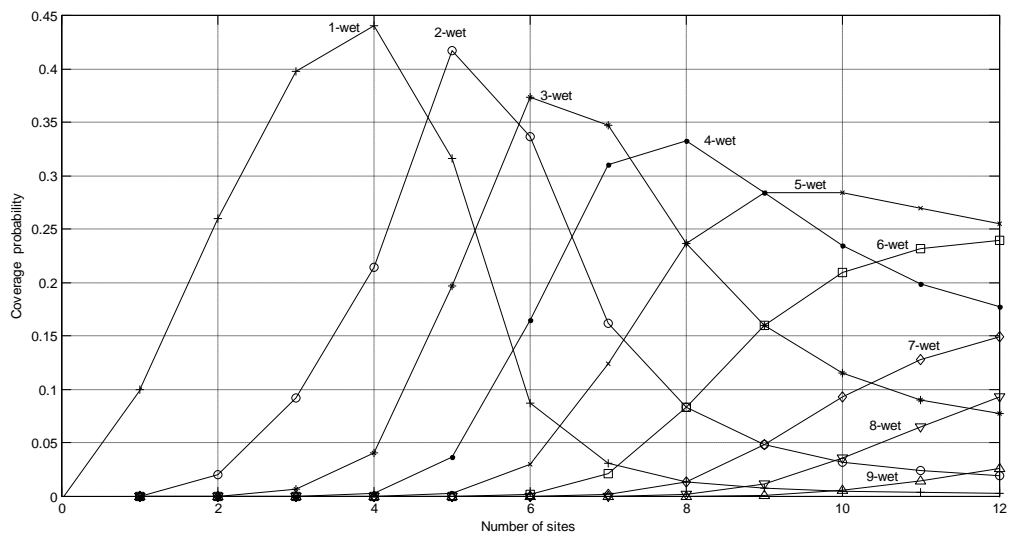
Figure 7. Some heterogeneous ACP variation

8. CONCLUSIONS

Run-length probability distribution function (pdf) is derived for stationary stochastic dependent processes in finite lengths. The necessary analytical derivations are achieved through the combinatorial methodology.

Analytical derivations of various point precipitation including probability of precipitation (PoP) concept and the areal or temporal coverage probabilities conditioned on a certain number of wet sites (precipitation occurrence more than a given threshold level) have been presented for heterogeneous and homogeneous precipitation evaluations over a region. In this paper, the general formulation of heterogeneous probability areal or temporal coverage probability calculation is developed based on the random field concept, where the temporal and areal precipitation occurrences are assumed independent from site to another. This general formula reduces exactly to homogeneous case, where the well-known Binomial probability distribution is valid. The application of the suggested formulation is shown first for its confirmation with the Binomial pdf for homogeneous probabilities values and it is seen that the general formula reduces to the classical Binomial pdf exactly. Theoretically its application is presented a set of probabilities with 9 and 12 sites. The factual data application is shown for 6 meteorology stations in and around Istanbul City, Turkey. It is hoped that in the future theoretical methodology developed herein can be extended to areally dependent precipitation events.

9. REFERENCES

1. A. M. Mood, The distribution theory of runs, *Annals of Mathematical Statistics* **11**, 427-432, 1940.

2. W. Feller, *An Introduction to Probability Theory and Its Application* **1**, John Wiley and Sons, Inc., New York, 1957.

3. N. G. Kendall, Relations connected with the tetrachoric series and its generalization, *Biometrika* **32**, 196, 1941.

4. Z. Şen, Wet and dry periods of annual flow series. ASCE, *Journal of Hydraulics Divisio* **102**, HY10, 1503-1514, 1976.

5. J. Wishart and H.O. Hirschfeld, A theory concerning the distribution of joints between segments, *Journal of the London Mathematical Society* **11**, 227, 1936.

6. Z. Şen, Probabilistic modeling of crossing in small samples and application of runs to hydrology, *Journal of Hydrology* **124**, 345-362,1991.

7. Z. Şen, Regional drought and flood frequency analysis: Theoretical considerations. *Journal of Hydrology* **46**, 265-279, 1980.

APPENDIX

$$\sum_{i=1}^{n} i^3 P\{s=i\mid n\} = \left\{ \frac{1}{(1-r)^3 \left[1-(1-r)^n\right]} \right\} (3n^3 r^{n+1} + 6n^2 r^{n+1} - 4r^{n+1} - 3n^3 r^{n+2} - 3n^2 r^{n+2} + 3nr^{n+2} -$$

$$r^{n+2} + n^3 r^{n+3} - n^3 r^n - 3n^2 r^n - 3nr^n - r^n + r^2 + 4r + 1)$$

$$\sum_{i=1}^{n} i^4 P\{s=i\mid n\} = \left\{\frac{1}{(1-r)^4\left[1-(1-r)^n\right]}\right\}(-4n^4r^{n+1}-12n^3r^{n+1}-6n^2r^{n+1}+12nr^{n+1}+11r^{n+1}+6n^4r^{n+2}+12n^3r^{n+2}-$$

$$6n^2r^{n+2}-12nr^{n+2}+11r^{n+2}-4n^4r^{n+3}-4n^3r^{n+3}+6n^2r^{n+3}-4nr^{n+3}+$$

$$r^{n+3}+n^4r^{n+4}+n^4r^n+4n^3r^n+6n^2r^n+4nr^n+r^n-r^3-11r^2-11r-1)$$