PROBABILITIES OF RUNS OF CONSECUTIVE DRY DAYS IN WEATHER PHENOMENA

by

JAI PRAKASH SINGHAL

B.Sc., Meerut College, Meerut, India, 1956
M.Sc., Meerut College, Meerut, India, 1958

A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
MANHATTAN, KANSAS

1961

TABLE OF CONTENTS

INTRODUCTION

In this thesis, the probability that a given day of the climatological year will be a dry day is defined as the ratio of the number of years in which the given day is dry to the total number of years considered in the sequence under study. A dry day is a 24 hour period (midnight to midnight) with precipitation <.20 inch.

One of the main aims of research in Meteorology is to determine the nature of the relation between the microclimate and crop growth. Like rainfall, sunshine and dry weather play an important role in affecting the fluctuations in the microclimate of a crop. Bright sunshine is of immense value during harvest season. Continuous rainfall or cloudiness in the month of crop ripening is conducive to the spread of some diseases in standing crops. This may also result in a loss of yield for certain crops. In the process of drying hay, it is important that there should be enough dry weather to get the hay dried. Hence, if the chances of runs of varying lengths of dry days are known, the hay could be cut at a more favourable time thus reducing crop damage. Farmers look for dry periods for sowing seeds. A heavy rain after seeds are sown may result in a loss of seed from washing. If favourable weather could be predicted for certain activities, losses could be reduced. In this way the importance of dry days can be seen to be of immense value. This is why, the problem of estimation of probabilities of runs of consecutive dry days has been studied.

This study is confined to three Kansas stations; namely Garden City, Manhattan and Columbus. Of these, Garden City and Columbus

are respectively the driest and wettest stations.

This thesis is written with two main purposes. The first is to establish a suitable mathematical model which expresses the probability of a dry day as a function of time t. The second purpose is to indicate how one might compute the probabilities of runs of consecutive dry days beginning with any day of the year. Each of the three stations has been dealt with separately.

To obtain the probability of runs of consecutive dry days, the fundamental " Multiplication theorem of probabilities " has been utilized. If there exists day to day independence in weather phenomena, the problem of determining probabilities of runs of consecutive dry days will be much simpler than if such independence does not exist. Therefore, it was necesary to investigate the problem of independence of weather phenomena from day to day. Data for this study is based on records for the 58 year period from 1900 to 1957. These data have been put on IBM punch cards and they form a portion of the weather library maintained at Kansas State University.

To obtain the number of consecutive dry days in a sequence the cards for days with precipitation $\geq .20$ inch were selected by use of an IBM sorter and the resulting data was listed in order of months and days using an IBM tabulating machine. The number of years in which a sequence of given length had one or more wet days were counted. This number was subtracted from 58 to get the number of sequences of consecutive dry days. In this way the number of years in which there was one dry day, two consecutive dry days, three consecutive dry days in a given sequence was counted. The probability

of a wet day or days is equally important and might also have been studied. Because of the time factor, this study of sequences of consecutive wet days could not be made. Such a project is anticipated for future study.

DEFINITIONS, RELATIONS, AND MODELS

Definitions

In the following definitions, relations, and models t will range over the values 1,2,3, ... , 365.

Definition 1. $N(t)$ = Number of years in which the t^{th} day was a dry day.

Definition 2. $N(t,t-1)$ = Number of years in which both the t^{th} and $(t-1)^{st}$ days were dry days.

Definition 3. $N(t,t-1,t-2)$ = Number of years in which the t^{th}, $(t-1)^{st}$ and $(t-2)^{nd}$ days were dry days.

Definition 4. $N(t-1)$ = Number of years in which the $(t-1)^{st}$ day was a dry day.

Definition 5. $N(t-1,t-2)$ = Number of years in which the $(t-1)^{st}$ and $(t-2)^{nd}$ days were dry days.

Definition 6. $P_E(t)$ = Empirical probability that the t^{th} day will be a dry day.

Definition 7. $p(t)$ = Least square estimate of the probability that the t^{th} day will be a dry day.

Definition 8. $P_E(t,t-1)$ = Empirical probability (joint) that the t^{th} and $(t-1)^{st}$ days will be dry days.

Definition 9. $P_E(t/t-1)$ = Empirical conditional probability that the t^{th} day will be a dry day given that the $(t-1)^{st}$ day was dry.

Definition 10. $p(t/t-1)$ = Least squares estimate of the conditional probability that the t^{th} day will be a dry day given that the $(t-1)^{st}$ day was dry.

Definition 11. $P_E(t/t-1,t-2)$ = Empirical conditional probability that the t^{th} day will be a dry day given that $(t-1)^{st}$ and $(t-2)^{nd}$ days were dry.

Definition 12. $P_E(t,t-1,t-2)$ = Empirical joint probability that the t^{th}, $(t-1)^{st}$ and $(t-2)^{nd}$ days will be dry days.

Definition 13. $p(t/t-1,t-2)$ = Least squares estimate of the conditional probability that the t^{th} day will be a dry day given that the $(t-1)^{st}$ and $(t-2)^{nd}$ days were dry.

Definition 14. $P_E(t,t-1,\ldots\ldots,t-k)$ = Empirical joint probability that the t^{th}, $(t-1)^{st}$, ..., $(t-k)^{th}$ days will be dry days, where $k=1,2,\ldots,n-1$.

Definition 15. $p_n(t)$ = Estimated joint probability that the t^{th}, $(t+1)^{st}$,.....,$(t+n-1)^{th}$ days will be dry days.

Definition 16. $\bar{p}(t)$ = True probability that the t^{th} day will be a dry day, and definitions analogous to definitions 6,8,9,11,12,14 may be defined in a similar way.

Definition 17. $P_n(t)$ = True joint probability that n consecutive days will be dry, counting forward and including the t^{th} day of the year.

Relations

From probability theory it is known, that

$$P(t,t-1) = P(t-1).P(t/t-1) \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$P(t,t-1,t-2) = P(t-2)P(t-1/t-2)P(t/t-1,t-2) \quad \ldots\ldots\ldots(2)$$

In general,

$$P(t,t-1,\ldots\ldots,t-k) = P(t-k)P(t-k+1/t-k)P(t-k+2/t-k+1,t-k) \ldots\ldots\ldots\ldots P(t/t-1,\ldots\ldots,t-k) \quad \ldots\ldots(3)$$

If the probability that the t^{th} day is dry is independent of the weather on the $(t-1)^{st}$, $(t-2)^{nd}$, ..., $(t-k)^{th}$ days then eqn. (3) becomes,

$$P(t,t-1,\ldots\ldots,t-k) = P(t-k)P(t-k+1)\ldots\ldots\ldots P(t-1)P(t) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

If the probability that the t^{th} day is dry is dependent on the weather on the $(t-1)^{st}$ day but independent of the weather on $(t-2)^{nd}$, $(t-3)^{rd}$, ..., $(t-k)^{th}$ days, then eqn. (3) becomes,

$$P(t,t-1,\ldots\ldots,t-k) = P(t-k)P(t-k+1/t-k)P(t-k+2/t-k+1)\ldots \ldots\ldots\ldots P(t-1/t-2)P(t/t-1) \quad \ldots\ldots\ldots(5)$$

If the probability that the t^{th} day is dry is dependent on the weather on the $(t-1)^{st}$ and $(t-2)^{nd}$ days but independent of the weather on the $(t-3)^{rd}$, $(t-4)^{th}$,....., $(t-k)^{th}$ days then eqn. (3) becomes,

$$P(t,t-1,\ldots\ldots\ldots,t-k) = P(t-k)P(t-k+1/t-k)P(t-k+2/t-k+1,t-k) \ldots\ldots\ldots P(t/t-1,t-2) \quad \ldots\ldots\ldots\ldots(6)$$

If the probability that the t^{th} day is dry is dependent on the weather on the $(t-2)^{nd}$ day but independent of the weather on $(t-1)^{st}$, $(t-3)^{rd}$, $(t-4)^{th}$,.....,$(t-k)^{th}$ days then eqn. (3) becomes,

$$P(t,t-1,.....,t-k) = P(t-k)P(t-k+1)P(t-k+2/t-k).....$$
$$....P(t-1/t-3)P(t/t-2).....(7)$$

In general, relation (3) can be expressed in many other forms depending upon the type of dependence. Also, from equations (1) and (2) it can be shown that

$$P(t/t-1) = \frac{P(t,t-1)}{P(t-1)}$$

$$P(t/t-1,t-2) = \frac{P(t,t-1,t-2)}{P(t-1,t-2)}$$

Models

One of the reasons for selecting polynomial models to estimate $p(t)$, $p(t/t-1)$ and $p(t/t-1,t-2)$ was ease of computation. Additionally initial plots of $P_E(t)$ and $P_E(t/t-1)$, for the three stations, suggested fitting a 4^{th} degree polynomial in t. Choice of a 4^{th} degree polynomial allowed for the existence of two maximum points and one minimum point. Finally, polynomial models were used because maximum, minimum and inflexion points could be easily determined by the methods of elementary calculus. The following mathematical models were used for the estimated probabilities $p(t)$, $p(t/t-1)$ and $p(t/t-1,t-2)$.

Model 1.

$$p(t) = a_0 + a_1t + a_2t^2 + a_3t^3 + a_4t^4$$
$$p(t/t-1) = b_0 + b_1t + b_2t^2 + b_3t^3 + b_4t^4$$
$$p(t/t-1,t-2) = c_0 + c_1t + c_2t^2 + c_3t^3 + c_4t^4$$

Model 2.

$$p(t) = a_o + a_1 t + a_3 t^3 + a_4 t^4$$
$$p(t/t-1) = b_o + b_1 t + b_3 t^3 + b_4 t^4$$
$$p(t/t-1,t-2) = c_o + c_1 t + c_3 t^3 + c_4 t^4$$

Model 3.

$$p(t) = a_o + a_1 t + a_2 t^2 + a_3 t^3$$
$$p(t/t-1) = b_o + b_1 t + b_2 t^2 + b_3 t^3$$
$$p(t/t-1,t-2) = c_o + c_1 t + c_2 t^2 + c_3 t^3$$

Model 4.

$$p(t) = a_o + a_1 t + a_2 t^2 + a_4 t^4$$
$$p(t/t-1) = b_o + b_1 t + b_2 t^2 + b_4 t^4$$
$$p(t/t-1,t-2) = c_o + c_1 t + c_2 t^2 + c_4 t^4$$

In all these mathematical models t, t^2, t^3, t^4, were the independent variables and $p(t)$, $p(t/t-1)$ and $p(t/t-1,t-2)$ the dependent variables. The constants $a_o, a_1, \ldots\ldots, a_4$; $b_o, \ldots\ldots, b_4$; $c_o, \ldots\ldots, c_4$ were determined by the method of least squares.

Model (1) is different from models (2), (3) and (4) in that the removal of the t^2, t^4, t^3 terms respectively from model (1) produced models (2), (3), (4). The removal of the t^2 and t^3 terms changes the location of maximum, minimum and inflexion points. The removal of the t^4 term in the model does not simply change the location of tese points but also reduces the number of maximum or minimum points by 1 but also reduces the number of inflexion points by 1. The selection of a model to estimate $P(t)$, $P(t/t-1)$ and $P(t/t-1,t-2)$ for each station was decided on the basis of statistical tests of significance and the magnitude of the standard errors of the coefficients. The model finally chosen was that which

had the smallest standard errors of the coefficients.

The following models were used to estimate $p_n(t)$, given $p(t)$, $p(t/t-1)$, and $p(t/t-1,t-2)$.

Model 5.

$$p_n(t) = p(t).p(t+1).p(t+2).....p(t+n-1)$$

Model 6.

$$p_n(t) = p(t).p(t+1/t).p(t+2/t+1).....$$
$$.....p(t+n-1/t+n-2)$$

Model 7.

$$p_n(t) = p(t).p(t+1/t).p(t+2/t+1,t).....$$
$$.....p(t+n-1/t+n-2,t+n-3)$$

Models (5), (6), (7) are essentially the same as relations (4), (5) and (6) except that in models (5), (6), (7) the n consecutive dry days are counted forward and include the t^{th} day while in relations (4), (5), (6) the n consecutive dry days are counted backward and included the t^{th} day. In model (5) one assumes that P(t) does not depend upon the weather on the $(t-1)^{st}$, $(t-2)^{nd}$ etc. days. In model (6) P(t) is assumed to depend upon the weather on the $(t-1)^{st}$ day but independent of the weather on the $(t-2)^{nd}$, $(t-3)^{rd}$ etc. days. In model (7) P(t) is assumed to depend upon the weather on the $(t-1)^{st}$ and $(t-2)^{nd}$ days but is independent of the weather on the $(t-3)^{rd}$, $(t-4)^{th}$ etc. days. The selection of a model to estimate $P_n(t)$ was decided on the basis of statistical tests of significance.

STATISTICAL METHODS

If one assumes that the probability that the t^{th} day will be dry will not depend upon the weather, dry or wet, on the $(t-3)^{rd}$, $(t-4)^{th}$ etc. days, one of the models (5), (6), (7) should be used to compute $p_n(t)$. The selection of the model depends upon whether the probability $p(t)$ is dependent of the weather, dry or wet, on the $(t-1)^{st}$ and $(t-2)^{nd}$ days or independent of the weather on these days. The empirical probabilities $P_E(t)$, $P_E(t,t-1)$ and $P_E(t,t-1,t-2)$ were determined by counts for each of the three stations. The empirical conditional probabilities $P_E(t/t-1)$, and $P_E(t/t-1,t-2)$ were determined by using the relations (1) and (2) respectively. The coefficients of correlation (r's) were computed for the following pairs of factors.

$P_E(t)$ and $P_E(t/t-1)$

$P_E(t)$ and $P_E(t/t-1,t-2)$

$P_E(t/t-1)$ and $P_E(t/t-1,t-2)$

These correlation coefficients are given in the table 1. The test of the hypothesis, that these several r's estimate $\rho = 1$, was made. The test of the hypothesis, that these several r's estimate a common ρ was also made. This method of testing the hypothesis about the several r's estimating a common ρ was devised by R.A. Fisher (1921). According to this method, the several r's were transformed to a quantity, z, distributed almost normally with variance,

$$\sigma_z^2 = \frac{1}{n-3} ,$$

" practically independent of the value of the correlation in the population from which the sample was drawn ". The relation of z to r is given by

$$z = \frac{1}{2}\left[\log_e(1+r) - \log_e(1-r)\right]$$

The computations for each of the three stations are given in tables 2, 3, 4, and 5.

If, the several correlation coefficients (r's) estimate $\rho = 1$, model (5) would be used to compute $p_n(t)$. In case the several correlation coefficients estimate different ρ's, the model (7) should be used to compute $p_n(t)$. In using model (5) only p(t) is needed, while in using model (7), p(t), p(t/t-1), and p(t/t-1,t-2) are needed. Models (1) and (2) were fitted to data obtained for $P_E(t)$, $P_E(t/t-1)$, and $P_E(t/t-1,t-2)$ for data from stations at Garden City and Manhattan. Models (1), (3), and (4) were fitted to data obtained for $P_E(t)$, $P_E(t/t-1)$, and $P_E(t/t-1,t-2)$ for the columbus station.

The constants in the above mentioned models were determined by the method of least squares. The model, which had the minimum standard errors for it's coefficients, was selected for each of the three stations. The different models, which were fitted to the data for the three stations, are given in the tables 6, 7, and 8 together with the standard errors of the corresponding coefficients. The dates where maxima, minima, and inflexion points occur were also estimated from the models using well known calculus methods. The estimated dates where maxima, minima, and inflexion points occur are found in the table 9.

RESULTS AND DISCUSSION

Various correlation coefficients (r's) are given in table 1. The 99 % confidence intervals for the correlation between $P_E(t)$ and $P_E(t/t-1)$ does not include the value $\rho=1$. This implies that the probability that the t^{th} day will be a dry day depends upon the weather on the $(t-1)^{st}$ day. The 99 % confidence interval for the correlation between $P_E(t/t-1)$ and $P_E(t/t-1,t-2)$ does not include the value $\rho=1$, indicating that the probability that the t^{th} day will be a dry day depends upon the weather on the $(t-1)^{st}$ and $(t-2)^{nd}$ days. These results, therefore, suggest the use of model (7) to compute $p_n(t)$. The computations for the confidence intervals for the value of ρ from the sample values r's are given in the table 2.

Further, it was tested whether the different correlation coefficients (r's) estimated the same ρ. The computations are given in tables 3, 4, and 5. It was observed that the different correlation coefficients did not estimate the same ρ. The correlations between $P_E(t)$ and $P_E(t/t-1)$, $P_E(t/t-1)$ and $P_E(t/t-1,t-2)$ are greater than the correlation between $P_E(t)$ and $P_E(t/t-1,t-2)$. The correlation between $P_E(t/t-1)$ and $P_E(t/t-1,t-2)$ is greater than the correlation between $P_E(t)$ and $P_E(t/t-1)$. These facts suggest the possibility that the correlation between $P_E(t)$ and the probabilities $P_E(t/t-1,t-2,t-3)$, $P_E(t/t-1,t-2,t-3,t-4)$, etc. will go on decreasing. These facts also suggest the possibility of increasing correlations between $P_E(t/t-1,t-2)$ and $P_E(t/t-1,t-2,t-3)$, $P_E(t/t-1,t-2,t-3)$ and $P_E(t/t-1,t-2,t-3,t-4)$, etc.. Because

of lack of time, these correlation coefficients could not be determined. It is hoped that these correlation coefficients can be determined in future study.

In fitting the mathematical models (1) and (2) to the data for Garden City and Manhattan, it was observed that the addition, in sequence, of t, t^2, t^3, t^4 in model (1) and t, t^3, t^4 in model (2), after the preceding one had been added, was significant. Within a station, both the models reduced the total sum of squares by almost the same percentage. Model (2) was selected for both stations on the basis of the size of the standard errors of the coefficients.

For the Columbus station, the addition of the further term t^4 after t, t^2, t^3 were added, was found nonsignificant at the 5% level. However, the three models (1), (3) and (4) reduced the total sum of squares by almost the same percentages. Model (4) was selected because it yielded the smallest standard errors for it's coefficients.

The estimated dates where maxima, minima and inflexion points occured, were almost in agreement. A change in the model to estimate P(t), P(t/t-1) and P(t/t-1,t-2) tends to change the dates where maxima, minima and inflexion points occur.

Weather phenomena under study in this investigation showed cyclic behaviour. This was apparent from the observed data and the fitted functions. Ideally, the values for p(1), p(2), p(3),.... ..... should be respectively equal to p(366), p(367), p(368) ..... which was not the case with these models. This suggests that in future studies use might be made of Fourier's series to estimate P(t), P(t/t-1), and P(t/t-1,t-2). Fourier's

series provide periodic functions and the period in this case will be a year. In this way, the cyclic behaviour in weather phenomena could be incorporated into the models, although there may be some sacrifice in goodness of fit.

SUMMARY

In this thesis, the probability that a given day of the climatological year will be a dry day is defined as the ratio of the number of years in which the given day is dry to the total number of years considered in the sequence under study. A dry day is a 24 hour period (midnight to midnight) with precipitation $< .20$ inch. The importance of dry weather can be seen to be of immense value in certain activities. The problem of estimation of probabilities of runs of consecutive dry days has been studied. This study is confined to three Kansas stations; namely Garden City, Manhattan and Columbus. In the thesis, suitable mathematical models have been established which expresses the probability of a dry day as a function of time t. Additionally, probability models are given for computing the probabilities of runs of consecutive dry days beginning with any day of the year. Data for this study is based on records for the 58 year period from 1900 to 1957.

It was assumed that the probability that the t^{th} day will be dry does not depend upon the weather, dry or wet, on the $(t-3)^{rd}$, $(t-4)^{th}$ etc. days. The empirical probabilities $P_E(t)$ the probability that the t^{th} day will be a dry day, $P_E(t/t-1)$ the probability that the t^{th} day will be a dry day given that the $(t-1)^{st}$ day is dry,

$P_E(t/t-1,t-2)$ the probability that the t^{th} day will ba dry given that the $(t-1)^{st}$ and $(t-2)^{nd}$ days are dry, were determined by counts. The correlation coefficients (r's) for the following pairs of factors

$$P_E(t) \text{ and } P_E(t/t-1)$$
$$P_E(t) \text{ and } P_E(t/t-1,t-2)$$
$$P_E(t/t-1) \text{ and } P_E(t/t-1,t-2)$$

were respectively determined to be (.947, .921, .973), (.952, .901, .953), (.910, .862, .952) for Garden City, Manhattan and Columbus respectively. In the tests of hypothesis by the method due to R.A.Fisher, it was observed that the above correlation coefficients did not estimate $\rho = 1$. It implies that the probability that the t^{th} day will be dry depends upon the weather on $(t-1)^{st}$ and $(t-2)^{nd}$ days. It was also observed that the different correlation coefficients did not estimate the same ρ. The following model was used to estimate $p_n(t)$ the probability of n consecutive dry days counting forward and including the t^{th} day;

$$p_n(t) = p(t).p(t+1/t).p(t+2/t,t+1).....p(t+n-1/t+n-3,t+n-2)$$

The differences in the correlation coefficients (r's) suggested the possibility that the correlation between $P_E(t)$ and any of the probabilities $P_E(t/t-1,t-2,t-3)$, $P_E(t/t-1,t-2,t-3,t-4)$ etc. will go on decreasing. It also suggested the possibility of increasing correlations between $P_E(t/t-1,t-2)$ and $P_E(t/t-1,t-2,t-3)$, $P_E(t/t-1,t-2,t-3)$ and $P_E(t/t-1,t-2,t-3,t-4)$, etc.

The probabilities p(t), p(t/t-1) and p(t/t-1,t-2) were estimated by the following models;

Garden City:

$$p(t) = .99920 - .1239\times10^{-2}t + .28796\times10^{-7}t^3 + .55003\times10^{-10}t^4$$

$$p(t/t-1) = 1.00115 - .11562\times10^{-2}t + .27327\times10^{-7}t^3 - .52599\times10^{-10}t^4$$

$$p(t/t-1,t-2) = .99914 - .11411\times10^{-2}t + .27810\times10^{-7}t^3 - .54288\times10^{-10}t^4$$

Manhattan:

$$p(t) = .98684 - .16022\times10^{-2}t + .34651\times10^{-7}t^3 - .63941\times10^{-10}t^4$$

$$p(t/t-1) = .98732 - .14776\times10^{-2}t + .33080\times10^{-7}t^3 - .62046\times10^{-10}t^4$$

$$p(t/t-1,t-2) = .98570 - .14279\times10^{-2}t + .31603\times10^{-7}t^3 - .58796\times10^{-10}t^4$$

Columbus:

$$p(t) = .94324 - .18931\times10^{-2}t + .76734\times10^{-5}t^2 - .20750\times10^{-10}t^4$$

$$p(t/t-1) = .95319 - .18796\times10^{-2}t + .79161\times10^{-5}t^2 - .23213\times10^{-10}t^4$$

$$p(t/t-1,t-2) = .95664 - .19671\times10^{-2}t + .83538\times10^{-5}t^2 - .25257\times10^{-10}t^4$$

The coefficients in the above models were determined by the method of least squares. The estimated dates where maxima, minima, and inflexion points occured, were determined to be;

Garden City: (minimum: June 16, Maximum: Dec. 25, Inflexion: Jan. 15, Oct. 3)

Manhattan: (Minimum: June 22, Maximum: Jan.6, Inflexion: Jan. 15, Oct. 12)

Columbus: (Minimum: May 31, Maximum: Dec. 25, Inflexion: Sept, 19)

Ideally, the values for p(1), p(2), p(3),..... should be respectively equal to p(366), p(367), p(368), ..., which was not the case with these models. This suggests that in future studies use might be made of Fourier's series to estimate p(t), p(t/t-1), and p(t/t-1,t-2). Fourier's series provide periodic functions and the period in this case will be a year. In this way,

the cyclic behaviour in weather phenomena could be incorporated into the models, although there may be some sacrifice in goodness of fit.

Table 1. Correlation coefficients for the three stations.

Station	Factor	$P_E(t)$	$P_E(t/t-1)$	$P_E(t/t-1,t-2)$
Garden City				
	$P_E(t)$	1.000	.947	.921
	$P_E(t/t-1)$	.947	1.000	.973
Manhattan				
	$P_E(t)$	1.000	.952	.901
	$P_E(t/t-1)$	.952	1.000	.953
Columbus				
	$P_E(t)$	1.000	.910	.862
	$P_E(t/t-1)$	.910	1.000	.952

Table 2. The confidence limits for the value of ρ in the population from the sample value r for the three stations.

Station	r	z	$\sigma_z^2 = 1/n-3$	$(t_{.01})\sigma_z$	Confidence interval z	Confidence interval r
Garden City						
	.947	1.797	.0525	.1354	1.662-1.932	.935-.959
	.921	1.593	.0525	.1354	1.458-1.728	.902-.939
	.973	2.149	.0525	.1354	2.014-2.284	.965-.979
Manhattan						
	.952	1.851	.0525	.1354	1.716-1.986	.937-.963
	.901	1.483	.0525	.1354	1.348-1.618	.874-.924
	.953	1.858	.0525	.1354	1.723-1.993	.938-.964
Columbus						
	.910	1.533	.0525	.1354	1.398-1.668	.886-.931
	.862	1.300	.0525	.1354	1.165-1.435	.823-.893
	.952	1.850	.0525	.1354	1.715-1.985	.937-.963

Since the confidence intervals for each r does not include $\rho = 1$, The hypothesis $\rho = 1$ is rejected.

Table 3. Test of hypothesis of common ρ for Garden City.

Factors	n	$1/\sigma_z^2$	r	z	(n-3)z	$(n-3)z^2$
$P_E(t)$ and $P_E(t/t-1)$	365	362	.947	1.797	650.514	1168.973
$P_E(t)$ and $P_E(t/t-1,t-2)$	365	362	.921	1.593	576.666	918.629
$P_E(t/t-1)$ and $P_E(t/t-1,t-2)$	365	362	.973	2.149	777.938	1671.789
			2.841	5.539	2005.118	3759.391

Average r = .947

Average z = 1.846

$= 3759.391 - 2005.118 \times 1.846$

$= 57.282$

$Pr(\chi^2 = 57.282) < .005$

Table 4. Test of hypothesis of common ρ for Manhattan.

Factors	n	$1/\sigma_z^2$	r	z	(n-3)z	$(n-3)z^2$
$P_E(t)$ and						
$P_E(t/t-1)$	365	362	.952	1.851	670.062	1240.2847
$P_E(t)$ and						
$P_E(t/t-1,t-2)$	365	362	.901	1.483	536.846	796.1426
$P_E(t/t-1)$ and						
$P_E(t/t-1,t-2)$	365	362	.953	1.858	672.596	1249.6833
			2.806	5.192	1879.504	3286.1106

Average r = .935

Average z = 1.731

$= 3286.1106 - 1879.504 \times 1.731$

$= 33.3282$

$Pr(\chi^2 = 33.3282) < .005$

Table 5. Test of hypothesis of common ρ for Columbus.

Factors	n	$1/\sigma_E^2$	r	z	(n-3)z	$(n-3)z^2$
$P_E(t)$ and $P_E(t/t-1)$	365	362	.910	1.533	554.946	850.732
$P_E(t)$ and $P_E(t/t-1,t-2)$	365	362	.862	1.300	470.600	611.780
$P_E(t/t-1)$ and $P_E(t/t-1,t-2)$	365	362	.952	1.850	669.700	1238.945
			2.724	4.683	1695.246	2701.457

Average r = .908

Average z = 1.561

$= 2701.457 - 1695.246 \times 1.561$

$= 55.178$

$Pr(\chi^2 = 55.178) < .005$

Table 6. Least squares estimates of the coefficients for Garden City

Model	Factor	R^2	a_0	a_1	a_2	a_3	a_4
1	p(t)	.50822	.99498	$-.10332\times10^{-2}$	$-.23554\times10^{-5}$	$.38325\times10^{-7}$	$-.67558\times10^{-10}$
	s_{a_k}			$.34420\times10^{-3}$	$.3816\times10^{-5}$	$.15660\times10^{-7}$	$.21220\times10^{-10}$
	p(t/t-1)	.46654	.99116	$-.66952\times10^{-3}$	$-.55706\times10^{-5}$	$.49863\times10^{-7}$	$-.82292\times10^{-10}$
	s_{a_k}			$.34910\times10^{-3}$	$.38710\times10^{-5}$	$.15880\times10^{-7}$	$.21530\times10^{-10}$
	p(t/t-1,t-2)	.45458	.99305	$-.84476\times10^{-3}$	$-.33926\times10^{-5}$	$.41534\times10^{-7}$	$-.72372\times10^{-10}$
	s_{a_k}			$.34970\times10^{-3}$	$.38780\times10^{-5}$	$.15910\times10^{-7}$	$.21570\times10^{-10}$
2	p(t)	.50770	.99920	$-.12390\times10^{-2}$		$.28796\times10^{-7}$	$-.55003\times10^{-10}$
	s_{a_k}			$.85478\times10^{-4}$		$.26002\times10^{-8}$	$.60465\times10^{-11}$
	p(t/t-1)	.46347	1.00115	$-.11562\times10^{-2}$		$.27327\times10^{-7}$	$-.52599\times10^{-10}$
	s_{a_k}			$.86899\times10^{-4}$		$.26434\times10^{-8}$	$.61470\times10^{-11}$
	p(t/t-1,t-2)	.45342	.99914	$-.11411\times10^{-2}$		$.27810\times10^{-7}$	$-.54288\times10^{-10}$
	s_{a_k}			$.86908\times10^{-4}$		$.26437\times10^{-8}$	$.61476\times10^{-11}$

Table 7. Least squares estimates of the coefficients for Manhattan.

Model	Factor	R^2	a_o	a_1	a_2	a_3	a_4
1	p(t)	.55538	.97733	$-.11389\times10^{-2}$	$-.53031\times10^{-5}$	$.56104\times10^{-7}$	$-.92208\times10^{-10}$
	s_{a_k}			$.41195\times10^{-3}$	$.45677\times10^{-5}$	$.18739\times10^{-7}$	$.25402\times10^{-10}$
	p(t/t-1)	.52195	.97607	$-.09300\times10^{-2}$	$-.62689\times10^{-5}$	$.58441\times10^{-7}$	$-.95461\times10^{-10}$
	s_{a_k}			$.40433\times10^{-3}$	$.44831\times10^{-5}$	$.19392\times10^{-7}$	$.24932\times10^{-10}$
	p(t/t-1,t-2)	.48459	.97782	$-.10441\times10^{-2}$	$-.43939\times10^{-5}$	$.49378\times10^{-7}$	$-.82216\times10^{-10}$
	s_{a_k}			$.42604\times10^{-3}$	$.47239\times10^{-5}$	$.19380\times10^{-7}$	$.26270\times10^{-10}$
2	p(t)	.55371	.98684	$-.16022\times10^{-2}$		$.34651\times10^{-7}$	$-.63941\times10^{-10}$
	s_{a_k}			$.10244\times10^{-3}$		$.31162\times10^{-8}$	$.72464\times10^{-11}$
	p(t/t-1)	.51935	.98732	$-.14776\times10^{-2}$		$.33080\times10^{-7}$	$-.62046\times10^{-10}$
	s_{a_k}			$.10063\times10^{-3}$		$.30611\times10^{-8}$	$.71183\times10^{-11}$
	p(t/t-1,t-2)	.48335	.98570	$-.14279\times10^{-2}$		$.31603\times10^{-7}$	$-.58796\times10^{-10}$
	s_{a_k}			$.10587\times10^{-3}$		$.32206\times10^{-8}$	$.74892\times10^{-11}$

Table 8. Least squares estimates of the coefficients for Columbus.

Model	Factor	R^2	a_0	a_1	a_2	a_3	a_4
1	$\bar{p}(t)$	.37327	.95063	$-.22766\times10^{-2}$	1.22451×10^{-5}	$-.19020\times10^{-7}$	$.48310\times10^{-11}$
	s_{a_k}			$.44508\times10^{-3}$	$.49350\times10^{-5}$	$.20246\times10^{-7}$	$.27444\times10^{-10}$
	p(t/t-1)	.33342	.96181	$-.23268\times10^{-2}$	1.32471×10^{-5}	$-.22179\times10^{-7}$	$.66172\times10^{-11}$
	s_{a_k}			$.46211\times10^{-3}$	$.51239\times10^{-5}$	$.21021\times10^{-7}$	$.28494\times10^{-10}$
	p(t/t-1,t-2)	.31060	.97057	$-.26896\times10^{-2}$	1.69664×10^{-5}	$-.35832\times10^{-7}$	$.22935\times10^{-10}$
	s_{a_k}			$.49221\times10^{-3}$	$.54575\times10^{-5}$	$.22390\times10^{-7}$	$.30350\times10^{-10}$
3	p(t)	.37321	.94938	$-.22087\times10^{-2}$	1.14125×10^{-5}	$-.15484\times10^{-7}$	
	s_{a_k}			$.22150\times10^{-3}$	$.14053\times10^{-5}$	$.02524\times10^{-7}$	
	p(t/t-1)	.33332	.96009	$-.22338\times10^{-2}$	1.20660×10^{-5}	$-.17336\times10^{-7}$	
	s_{a_k}			$.22998\times10^{-3}$	$.14591\times10^{-5}$	$.26207\times10^{-8}$	
	p(t/t-1,t-2)	.30950	.96460	$-.23672\times10^{-2}$	1.30135×10^{-5}	$-.19045\times10^{-7}$	
	s_{a_k}			$.24513\times10^{-3}$	$.15552\times10^{-5}$	$.27934\times10^{-8}$	
4	p(t)	.37173	.94324	$-.18931\times10^{-2}$	$.76734\times10^{-5}$		$-.20750\times10^{-10}$
	s_{a_k}			$.17724\times10^{-3}$	$.82010\times10^{-6}$		$.34255\times10^{-11}$
	p(t/t-1)	.33136	.95319	$-.18796\times10^{-2}$	$.79161\times10^{-5}$		$-.23213\times10^{-10}$
	s_{a_k}			$.18408\times10^{-3}$	$.85180\times10^{-6}$		$.35577\times10^{-11}$
	p(t/t-1,t-2)	.30569	.95664	$-.19671\times10^{-2}$	$.83538\times10^{-5}$		$-.25257\times10^{-10}$
	s_{a_k}			$.19646\times10^{-3}$	$.90909\times10^{-6}$		$.37970\times10^{-11}$

Table 9. Estimated dates where maxima, minima and inflexion points occur for the three stations.

	Minimum		Maximum		Inflexion	
Station	t	Day	t	Day	t	Day
Garden City	153	June 16	345	Dec. 25	0	Jan. 15
					262	Oct. 3
Manhattan	159	June 22	357	Jan. 6	0	Jan. 15
					271	Oct. 12
Columbus	137	May 31	345	Dec. 25	248	Sept. 19

ACNOWLEDGMENTS

The author wishes to take this opportunity to express his sincere thanks to his major professor, Dr. Arlin M. Feyerherm, for his interest and skillful guidance not only throughout the writing of this thesis but also throughout this portion of the writer's graduate sudy. Appreciation must also be shown to Dr. L. D. Bark, of the Physics department, for providing the data which made this thesis possible. The author is also indebted to Dr. S.T.Parker, of the Mathematics department, for his helpful suggestions in the use of IBM 650 and its complimentary equipment.

REFERENCES

Anderson, R.L., Bancroft, T.A.
Statistical theory in research. New York: McGraw Hill Book Company, Inc., 1952.

Bala Subramaniyan, C., Jayaraman, M.V.
A note on duration of sunshine at Coimbatore. Indian Jour. Met. Geophysics. 4:107-110. 1953.

Byerly, W.E.
Elements of Differential Calculus. Boston: GINN, HEATH & Co. 1884.

Deming, W. Edwards.
Statistical adjustment of data. New York: John Wiley & Sons, Inc., 1943.

Ezekiel, M.
Methods of correlation analysis. New York: John Wiley & Sons, Inc., 1941.

Fisher, Ronald, A.
Statistical methods for Research Workers. 12th ed. Edinburgh: Oliver and Boyd. 1954

Gumbel, E.J.
Statistics of Extremes. New York: Columbia University Press. 1958.

Lawrence, E.N.
Application of Mathematical Series to the frequency of weather spells. London: Meteorological Magazine. 83:195-200. 1954.

Panofsky, Hans. A., Brier, Glenn. W.
Some applications of Statistics to Meteorology. Pennsylvania: The Pennsylvania State University. 1958.

Penquite, Robert.
Thesis submitted for the degree Doctor of Philosophy, Iowa State College. 1936.

Rambhadran, V.K.
Statistical study of the persistency of rain days during Monsoon season at Poona. Indian Jour. Met. Geophysics.5:48-55. 1954.

Sasuly, Max.
Trend analysis of Statistics. Washington: The Brookings Institution. 1934.

Shea, John D., Birge, Raymond. T.
A rapid method of calculating the least squares solution of a polynomial of any degree. University of California Publications in mathematics. 2(5). 1927.

Snedecor, George. W.
Statistical methods applied to experiments in agriculture and biology. 5th ed. Ames, Iowa: Iowa State College Press. 1956.

Srinivassan, T.R.
Sequence of months with rain above and below average in Rayalseema considered in relation to the theory of probability. Delhi: Indian Jour. Met. Geophysics. 5:230-238. 1954.

Williamson, Benjamin.
An elementary treatise on the differential calculus. 8th ed. New York and London: Longmans, Green and Co., 1895.

PROBABILITIES OF RUNS OF CONSECUTIVE DRY DAYS IN WEATHER PHENOMENA

by

JAI PRAKASH SINGHAL

B.Sc., Meerut College, Meerut, India, 1956
M.Sc., Meerut College, Meerut, India, 1958

AN ABSTRACT OF A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
MANHATTAN, KANSAS

1961

In this thesis, the probability that a given day of the climatological year will be a dry day is defined as the ratio of the number of years in which the given day is dry to the total number of years considered in the sequence under study. A dry day is a 24 hour period (midnight to midnight) with precipitation $\leq$ .20 inch. The problem of estimation of probabilities of runs of consecutive dry days has been studied. This study is confined to three Kansas stations; namely Garden City, Manhattan and Columbus. The purpose of this thesis is to establish a suitable mathematical model which expresses the probability of a dry day as a function of time t. The second purpose is to indicate how one might compute the probabilities of runs of consecutive dry days beginning with any day of the year. Data for this study is based on records for 58 year period from 1900 to 1957.

The probabilities, $P_E(t)$ the empirical probability that the t^{th} day will be dry, $P_E(t/t-1)$ the empirical probability that the t^{th} day will be dry given that $(t-1)^{st}$ day is dry, $P_E(t/t-1,t-2)$ the empirical probability that the t^{th} day will be dry given that the $(t-1)^{st}$ and $(t-2)^{nd}$ days are dry, were determined by counts. The correlation coefficients (r's) between the factors $P_E(t)$ and $P_E(t/t-1)$, $P_E(t)$ and $P_E(t/t-1,t-2)$, $P_E(t/t-1)$ and $P_E(t/t-1,t-2)$ for Garden City, Manhattan and Columbus were found to be (.947, .921, .973), (.952, .901, .953) and (.914, .862, .952) respectively. The 99 % confidence intervals for these correlations did not include $\rho = 1$. This suggested the use of the model

$$p_n(t) = p(t).p(t+1/t).p(t+2/t,t+1).....p(t+n-1/t+n-3,t+n-2)$$

to compute $p_n(t)$ the probability of n consecutive dry days. It was also observed that the several correlation coefficients did not

estimate the same ρ. In the above models p(t), p(t/t-1) and p(t/t-1,t-2) were estimated by the following;

Garden City:

$$p(t) = .99920 - .1239\times10^{-2}t + .28796\times10^{-7}t^3 - .55003\times10^{-10}t^4$$

$$p(t/t-1) = 1.00115 - .11562\times10^{-2}t + .27327\times10^{-7}t^3 - .52599\times10^{-10}t^4$$

$$p(t/t-1,t-2) = .99914 - .11411\times10^{-2}t + .27810\times10^{-7}t^3 - .54288\times10^{-10}t^4$$

Manhattan:

$$p(t) = .98684 - .16022\times10^{-2}t + .34651\times10^{-7}t^3 - .63941\times10^{-10}t^4$$

$$p(t/t-1) = .98732 - .14776\times10^{-2}t + .33080\times10^{-7}t^3 - .62046\times10^{-10}t^4$$

$$p(t/t-1,t-2) = .98570 - .14279\times10^{-2}t + .31603\times10^{-7}t^3 - .58796\times10^{-10}t^4$$

Columbus:

$$p(t) = .94324 - .18931\times10^{-2}t + .76734\times10^{-5}t^2 - .20750\times10^{-10}t^4$$

$$p(t/t-1) = .95319 - .18796\times10^{-2}t + .79161\times10^{-5}t^2 - .23213\times10^{-10}t^4$$

$$p(t/t-1,t-2) = .95664 - .19671\times10^{-2}t + .83538\times10^{-5}t^2 - .25257\times10^{-10}t^4$$

The coefficients in the above models were determined by the method of least squares. The estimated dates of maxima, minima and inflexion points were found to be:

Garden City:(Minimum: June 16, Maximum: Dec. 25, Inflexion: Jan. 15, Oct. 3).

Manhattan:(Minimum: June 22, Maximum: Jan. 6, Inflexion: Jan. 15, Oct. 12).

Columbus:(Minimum: May 31, Maximum: Dec. 25, Inflexion: Sept.19)

using well known calculus methods.