CHAPTER 3

AGRICULTURAL METEOROLOGICAL DATA, THEIR PRESENTATION AND STATISTICAL ANALYSIS

This Chapter was written by Henry Hayhoe, Peter Hoefsloot, Lu Houquan, Ray Motha, M. Veerasamy and Jon Wieringa

It was reviewed and edited by Olga Penalba and Matilde Rusticucci with contributions from Zipora Gat

The Chapter was externally co-ordinated by Kees Stigter

1. INTRODUCTION

Agricultural meteorology is an applied science that is the knowledge of weather and climate applied to qualitative and quantitative improvement in agricultural production. It is involved in meteorology, hydrology, agrology, and biology. It requires a diverse multi-disciplined array of data for operational applications and research. Basic agricultural meteorological data are largely the same as those used in general meteorology. These data need to be supplemented with more specific data relating to the biosphere, the environment of all living organisms, and biological data relating to growth and development of these organisms. Agronomic, phenological, and physiological data are necessary for dynamic modeling, operational evaluation, and statistical analyses. Most data need to be processed for generating various products that affect agricultural management decisions, such as cropping, irrigation scheduling etc. The supports from other technologies, e.g. statistics, geographical information and remote sensing are necessary for data processing. Geographical information and remote sensing data such as images of vegetation status and crop damaged by disasters, soil moisture, etc., also should be included in as supplementary data. Derived agrometeorological parameters, such as photosynthetic active radiation and potential evapotranspiration, are often used in agricultural meteorology for both research and operational purposes. On the other hand, many agrometeorological indexes such as drought index, critical point threshold of temperature, and soil water for crop development are also important for agricultural operations. Among the data, weather and climate data play a crucial role in many agricultural decisions.

Agrometeorological information includes not only every stage of growth and development of crops, floriculture, agroforestry, and livestock, but also the technological factors, which impact on agriculture such as irrigation, plan protection, fumigation, and dust spraying. Moreover, agricultural meteorological information play a crucial role in the decision making process for sustainable agriculture and natural disaster reduction with the aim toward preserving natural resources and improving the quality of life.

2. DATA FOR AGRICULTURAL METEOROLOGY

Agrometeorological data are usually provided to Users in a transformed format; for example, rainfall data are presented in pentads or in monthly amounts.

2.1. Nature of the data

Basic agricultural meteorological data may be divided into the following six categories, which include data observed by instruments on the ground and observed by remote sensing.

(a) Data related to the state of the atmospheric environment. These include observations of rainfall, sunshine, solar radiation, air temperature, humidity, and wind speed and direction;

(b) Data related to the state of the soil environment. These include observations of soil moisture, i.e., the soil water reservoir for plant growth and

development. The amount of water available depends on the effectiveness of precipitation or irrigation, and on the soil's physical properties and depth. The rate of water loss from the soil depends on the climate, the soil's physical properties, and the root system of the plant community. Erosion by wind and water depend on weather factors and vegetative cover;

(c) Data related to organism response to varying environments. These involve agricultural crops and livestock, the variety and the state and stages of the growth and development, as well as the pathogenic elements affecting them. Biological data are associated with phenological growth stages and physiological growth functions of living organisms;

(d) Information concerned with the agricultural practices employed. Planning brings together the best available resources and applicable production technologies into an operational farm unit. Each farm is a unique entity with combinations of climate, soils, crops, livestock, and equipment to manage and operate the farming system. The most efficient utilization of weather and climate data for the unique soils on a farm unit will help conserve natural resources, while at the same time, promote economical benefit to the farmer;

(e) Information related to weather disasters and their influence on agriculture; and,

(f) Information related to the distribution of weather and agricultural crops, and geographical information including the digital maps.

(g) Metadata, which describe used observation techniques and procedures.

2.2. Data Collection

The collection of data is very important as it lays the foundation for agricultural weather and climate data systems that are necessary to expedite generation of products, analyses, and forecasts for agricultural cropping decisions, irrigation management, fire weather management, and ecosystem conservation. The impact on crops, livestock, water and soil resources, and forestry must be evaluated from the best available spatial and temporal array of parameters. Agrometeorology is an interdisciplinary branch of science requiring the combination of general meteorological data observations and specific biological parameters. Meteorological data can be considered as typically physical elements that can be measured with relatively high accuracy while other types of observations (i.e., biological or phenological) may be more subjective. In collecting, managing, and analyzing the data for agrometeorological purposes, the source of data and the methods of observation define their character and management criteria. However, a few useful suggestions are listed below:

(a) Original data files, which may be used for reference purposes (the daily register of observations, etc.), should be stored at the observation site; this applies equally to atmospheric, biological, crop, or soil data;

(b) The most frequently used data should be collected at national or regional agrometeorological centers and reside in host servers for network accessibility. However, this may not always be practical since unique agrometeorological data are often collected by stations or laboratories under the control of different authorities (meteorological services, agricultural services, universities, research institutes). Steps should, therefore, be taken to ensure that possible users are aware of the existence of such data, either through some form of data library or computerized documentation, and that appropriate data exchange mechanisms are available to access and share these data;

(c) Data resulting from special studies should be stored at the place where the research work is undertaken, but it would be advantageous to arrange for exchanges of data between centers carrying out similar research work. At the same time, the existence of these data should be publicized at the national level and possibly at the international level, if appropriate, especially in the case of longer series of special observations;

(d) All the usual data-storage media are recommended:

(i) The original data records, or agrometeorological summaries, are often the most convenient format for the observing stations;

(ii) The format of data summaries intended for forwarding to regional or national centers, or for dissemination to the user community, should be designed so that the data may be easily transferred to a variety of media for processing. The format should also facilitate either the manual preparation or automated processing of statistical summaries (computation of means, frequencies, etc.). At the same time, access to and retrieval of data files should be simple, flexible, and reproducible for assessment, modeling, or research purposes;

(iii) Rapid advances in electronic technology facilitate effective exchange of data files, summaries, and charts of recording instruments particularly at the national and international level; and,

(iv) Agrometeorological data should be transferred to electronic media in the same way as conventional climatological data, with an emphasis on automatic processing.

The availability of proper agricultural meteorological data bases is a major prerequisite for studying and managing the processes of agricultural and forest production. The agricultural meteorology community has great interest in incorporating new information technologies into a systematic design for agrometeorological management to ensure timely and reliable data from national reporting networks for the benefit of the local farming community. While much more information has become available to the agricultural user, it is essential that appropriate standards be maintained for basic instrumentation, collection and observations, quality control, and archive and dissemination. After recorded,

collected, and transferred to the data centers, all agricultural meteorological data need to be standardized or treated by some techniques so that it can be used for different purposes. In the data center, the special database is requisite. The database would include meteorological, phenological, edaphic, and agronomic information. The database management and processing, quality controlling, archiving, timely accessing, and dissemination are all important components that will make the information valuable and useful in agricultural research and operational programs.

Having been stored in a data center, the data are disseminated to users. Great strides have been made in the automation age to make more data products available to the user community. The introduction of electronic transfer of data files via Internet using file transfer protocol (FTP) and the World Wide Web (WWW) has advanced this information transfer process to a new level. The WWW allows users to access text, images, and even sound files that can be linked together electronically. The WWW's attributes include the flexibility to handle a wide range of data presentation methods and the capability to reach a large audience. Developing countries have some access to this type of electronic information, but limitations still exist in the development of their own electronically accessible databases. These limitations will diminish as the cost of technology decreases and its availability increases.

2.3. Recording of data

Recording of basic data is the first step for agricultural meteorological data collection. When the environmental factors and other agricultural meteorological elements are measured or observed, they must be recorded in the same media, such as agricultural meteorological registers, diskettes, etc., manually or automatically.

(a) The data, such as the daily register of observations and charts of recording instruments, should be carefully preserved as permanent records. They should be readily identifiable and include the place, date, time of each observation, and the units used.

(b) These basic data should be sent to analysis centers for operational uses, e.g. local agricultural weather forecasts, agricultural meteorological information service, plant-protection treatment, irrigation, etc. The summaries (weekly, 10-day or monthly) of these data should be made regularly from the daily register of observations according to demand of users and then distribute to interested agencies and users.

(c) The data should be recorded by a standard format so that they could be readily transferred to data centers suitable for subsequent automatic processing, so the observers should record all measurements complying with some rules. The data can be transferred to data centers by many ways, such as mail, telephone, telegraph, and fax or Internet, and comsat, in which Internet and comsat are a more efficient approach. After reaching the data centers, data should be identified and processed by special program for facilitating to other users.

2.4. Scrutiny of data and acquisition of metadata

It is very important that all agricultural meteorological data be carefully scrutinized,

both at the observing station and at regional or national centers by subsequent automatic processing of computers. All data should be identified immediately. The code parameters should be specified, such as types, regions, missing values, and possible ranges for different measurements. The quality control should be done according to Wijngaard et al. (2003) and WMO-TD N° 1236 (2004) and the current Climatological Guide. Every code of measurement must be checked to make certain if the measurement is reasonable. If the value is unreasonable, it should be corrected immediately. After being scrutinized, the data could be processed further for different purposes. For ascertaining the quality of observation data and to know about any need to correct or normalize them before analysis, we need metadata. These are details and history of local conditions, instrumentation, operation, data processing and other factors relevant to the observation process. Such metadata should be documented and treated with the same care as the data themselves (see WMO 2003a, 2003b). Unfortunately, observation metadata are often incomplete and poorly organized.

In Chapter 2 of this Guide, essential metadata are specified for individual parameters, and also in section 2.2.5 of Chapter 2 the organization of their acquisition is reviewed. Many metadata can be recorded as simple numbers, e.g. observation heights, but more complex matter like instrument exposure must be recorded as well in some manner which is feasible for the observers and station managers. Acquiring metadata of present observations and inquiring about metadata of past observations is now a major responsibility of data managers. Omission of metadata acquisition implies that their data will have low quality for applications. Optimal setup of a database for metadata is at present still in development, because metadata characteristics are so variable. To be manageable it should not only be efficient for archiving, but also easily accessible for those who are recording the metadata. To allow future improvement and continuing accessibility, good metadata database formats are ASCII, SQL and XML, because these are independent of any presently available computing setup.

2.5. Format of data

The basic data obtained from observing stations, whether specialized or not, are of interest to both scientists and agricultural users. There are a number of established formats and protocols to exchange data. A data format is a documented set of rules for the coding of data in a form for both visual and computer recognition. Its uses can be designed for either or both real-time use and historical or archival data transfer. All the critical elements for identification of data should be covered in the coding, including station identifies, parameter descriptors, time encoding conventions, unit and scale conventions, and common fields.

Large amounts of data are typically required for processing, analysis, and dissemination. It is extremely important that data are in a format being both easily accessible and user friendly. This is particularly true as many data become available

in electronic format. Some software process data in a common form and disseminate to more users, such as NetCDF (network common data form). It is software for array-oriented data access and a library that provides an implementation of the interface (Sivakumar, et al., 2000). The NetCDF software was developed at the Unidata Program Center in Boulder, Colorado, USA. The freely available source can be obtained by anonymous FTP from ftp://ftp.unidata.ucar.edu/pub/netcdf/ or from other mirror sites.

NetCDF package supports the creation, access, and sharing of scientific data. It is particularly useful at sites with a mixture of computers connected by a network. Data stored on one computer may be read directly from another without explicit conversion. NetCDF library generalizes access to scientific data so that the methods for storing and accessing data are independent of the computer architecture and the applications being used. Standardized data access facilitates the sharing of data. Since the NetCDF package is quite general, a wide variety of analysis and display applications can use it. The NetCDF software and documentation may be obtained from the NetCDF website at http://www.unidata.ucar.edu/packages/netcdf/.

2.6. Catalogue of data

Very often, considerable amounts of agrometeorological data are collected by a variety of services. These data sources are not readily publicized or accessible to potential users. So, the users often have great difficulty in discovering whether such data exist. Coordination should therefore be undertaken at the global, regional, and national level to ensure that data catalogues are prepared periodically, giving enough information to users. The data catalogues should include the following information:

(a) The geographical location of each observing site;
(b) The nature of the data obtained;
(c) The location where the data are stored;
(d) The types of file (manuscript, charts of recording instruments, automated weather station, punched cards, magnetic tape, scanned data, computerized digital data); and,
(e) The methods of obtaining the data.

For a more extensive specification of such matters see section 2.2.5 of Chapter 2.

3. DISTRIBUTION OF DATA

3.1. Requirements for research

In order to highlight the salient features of the influence of climatic factors on the growth and development of living things, scientists often have to process a large volume of basic data. These data could be supplied to scientists in the following forms:

(a) Reproductions of original documents (original records, charts of recording instruments) or periodic summaries;
(b) Data sets on server or website ready for processing to different categories, which can be read or viewed on a platform;
(c) Various kinds of satellite digital data and imagery on different regions and different times;

(d) Various basic databases, which can be viewed as reference for research.

3.2. Special requirements of agriculturists

Two aspects of the periodic distribution of agrometeorological data to agricultural users may be considered:

(a) Raw or partially processed operational data supplied after only a short delay (rainfall, potential evapotranspiration, water balance, sums of temperature). These may be distributed:

i. By periodic publications, twice weekly, weekly or at 10-day intervals;
ii. By telephone and note;
iii. By TV special program from regional television station;
iv. By regional radio broadcast; and,
v. By release on agricultural or weather websites.

(b) Agrometeorological or climatic summaries published weekly, 10-day, monthly or annually, containing agrometeorological data (rainfall, temperatures above the ground, soil temperature and moisture content, potential evapotranspiration, sums of rainfall and temperature, abnormal rainfall and temperature, sunshine, global solar radiation, etc.).

3.3. Determining the requirements of users

The agrometeorologist has a major responsibility to ensure that effective use of this information offers an opportunity to enhance agricultural efficiency or to assist agricultural decision making. The information must be accessible, clear, and relevant. However, it is crucial for an agrometeorological service to know who the specific users of information are. The user community ranges from global, national, provincial organizations, and governments to agro-industries, farmers, agricultural consultants, and the agricultural research and technology development communities or private individuals. The variety of agrometeorological information requests emanates from this broad community. So, the agrometeorological service must distribute the appropriate information available at the right time.

Researchers invariably know exactly what agrometeorologica1 data they require for specific statistical analyses, modeling, or other analytical studies. Many agricultural users are often not only unaware of the actual scope of the agrometeorological services available, but have only a vague idea of the data they really need. Frequent contact between agrometeorologists and professional agriculturists, and enquiries through professional associations and among agriculturists themselves or visiting professional websites, can help enormously to improve the awareness of data needs. Sivakumar (1998) presents a broad overview of user requirements for agrometeorological services. Better applications of the type and amount of useful agrometeorological data available and the selection of the type of data to be systematically distributed can be established on that bases. For example, when both the climatic regions and the areas in which different crops are grown are well defined, an agrometeorological analysis can illustrate which crops are most suited to each climate zone. This type of analysis can also show which crops can be adapted to changing climatic and agronomic conditions. These analyses are required

by the agricultural users, they can be distributed by either geographic region, crop region, or climatic region.

3.4. Minimum distribution of agroclimatological documents

Since the large number of potential users of agrometeorological information is so widely dispersed, it is not realistic to recommend a general distribution of data to all users. In fact, the requests for raw agrometeorological data are rare. All of the raw agrometeorological data available is not essential for those directly engaged in agriculture (i.e., farmers, ranchers, foresters). Users generally require data to be processed into an understandable format to facilitate their decision making process. But, the complete data sets should be available and accessible to the technical services, agricultural administrations, and professional organizations. These professionals are responsible for providing practical technical advice concerning the treatment and management of crops, preventive measures, adaptation strategies, etc., based on collected agrometeorological information.

Agrometeorological information should be distributed to all users including:
(a) Agricultural administrations;
(b) Research institutions and laboratories;
(c) Professional organizations;
(d) Private crop and weather services;
(e) Government agencies; and,
(f) Farmers, ranchers, and foresters.

4. DATABASE MANAGEMENT

The management of agroclimatological data in the electronic age has become more efficient. The management to be considered, and already reviewed in this section, is data collection, data processing, quality control, archiving, data analysis and product generation, and product delivery. A wide variety of database choices are available to the agroclimatological user community. Accompanying the agroclimatological databases created, the agrometeorologists and software engineers develop the special software for agroclimatological database management. Thus, a database management system for agricultural applications should be a comprehensive system with the following considerations:

(a) Communication between climatologists, agrometeorologists and agricultural extension personnel must be improved to establish an operational database;

(b) The outputs must be adapted for an operational database in order to support specific agrometeorological applications at a national/regional/global level; and,

(c)Applications must be linked to the Climate Applications Referral System (CARS) project, spatial interpolated databases, and GIS.

Personal computer (PC) is able to produce products formatted for easy reading and presentation generated through simple processors, databases, or spreadsheet applications. However, some careful thought needs to be given to what type of

product is needed, what the product looks like, and what it contains, before the database delivery design is finalized. The greatest difficulty often encountered is how to treat missing data or information (WMO-TD N° 1236, 2004). This process is even more complicated when data from several different data sets such as climatic and agricultural data are combined. Some software for database management, especially the software for climatic database management, provide convenient tools for agrometeorological database management.

4.1. CLICOM Database Management System

CLICOM (CLImate COMputing) refers to the WMO World Climate Data Programme Project with the purpose of coordinating and assisting the implementation, maintenance and upgrading of automated climate data management procedures and systems in WMO Member countries (i.e., National Meteorological and Hydrological Services). The goal of CLICOM is the transfer of three main components of modern technology, *i.e*. desktop computer hardware; database management software, and training in climate data management. CLICOM is a standardized automated database management system (DBMS) software on a personal computer (PC) to introduce a system in developing countries. At the May of 1996, CLICOM version 3.0 was installed in 127 WMO Member countries. Now CLICOM software is available in English, French, Spanish, Czech, and Russian. CLICOM version 3.1 Release 2 was available in January 2000.

CLICOM provides tools to describe and manage the climatological network (i.e, stations, observations, instruments, etc.). It offers procedures to key entry, check and archive climate data, and compute and analyze the data. Typical standard outputs include monthly or 10-day data from daily data; statistics such as means, maximums, minimums, standard deviations; tables; and graphs. Other products, requiring more elaborated data processing, include water balance monitoring, estimation of missing precipitation data, calculation of the return period, and preparation of the CLIMAT message.

The CLICOM is widely used in developing countries. The installation of CLICOM as a data management system in many of these countries has successfully transferred PC's technology, but the resulting climate data management improvements have not yet been fully realized. Station network density as recommended by WMO has not been fully achieved and the collection of data in many countries remains inadequate. However, CLICOM systems are beginning to yield positive results and there is a growing recognition of the operational applications of CLICOM.

There are a number of constraints that have been identified over time and recognized for possible improvement in future versions of the CLICOM system. Among the technical limitations, the list includes (Motha, 2000):

(a)The lack of flexibility to implement specific applications in the agricultural field and/or at a regional/global level;
(b)Lack of functionality in real-time operations;
(c)Few options for file import;
(d)Lack of transparent linkages to other applications;
(e)Risk of many datasets overlapping;

(f)Non-standard geo-referencing system;
(g)Climate data may be stored without the corresponding station information;
(h)The data entry module allows for easy modification, which may destroy existing data.

4.2. Geographic Information System (GIS)

A geographic information system (GIS) is a computer-assisted system for acquisition, storage, analysis, and display of observed data on spatial distribution. GIS technology integrates common database operations such as query and statistical analysis with the unique visualization and geographic analysis benefits offered by mapping overlays. Maps have traditionally been used to explore the earth and its resources. GIS technology takes advantage of computer science technologies, enhancing the efficiency and analytical power of traditional methodologies.

GIS is becoming an essential tool in the effort to understand complex processes at different scales: local, regional, and global. In GIS, the information coming from different disciplines and sources, such as traditional point sources, digital maps, databases, and remote sensing, can be combined in models that simulate the behavior of complex systems.

The presentation of geographic elements is solved in two ways: using x, y coordinates (vectors) or representing the object as variation of values in a geometric array (raster). The possibility to transform the data from one format to the other allows fast interaction between different informative layers. Typical operations include overlaying different thermatic maps, contributing areas and distances, acquiring statistical information about the attributes, changing the legend, scale and projection of maps, and making three-dimensional perspective view plots using elevation data.

The capability to manage this diverse information, analyzing and processing together the informative layers, opens new possibilities for the simulation of complex systems. GIS can be used to produce images, not only maps, but the cartographic products, drawings, animations, or interactive instruments as well. These products allow researchers to analyze their data in new ways, predicting the natural behaviors, explaining events, and planning strategies.

For the agronomic and natural components in agrometeorology, these tools have taken the name Land Information Systems (LIS) (Sivakumar et al., 2000). In both GIS and LIS, the key components are the same: i.e., hardware, software, data, techniques, and technicians. However, LIS requires detailed information on the environmental elements such as meteorological parameters, vegetation, soil, and water. The final product of LIS is often the result of a combination of numerous complex informative layers, whose precision is fundamental for the reliability of the whole system.

4.3. Weather generators (WG)

Weather generators are widely used to generate synthetic weather data, which can be arbitrarily long for input into impact models, such as crop models and hydrological models that are used for assessing agroclimatic long-term risk and

agrometeorological analysis. Weather generators are also the tool to develop future climate scenarios based on GCM simulated or subjectively introduced climate changes for climate change impact models. Weather generators project future changes in means to the observed historical weather series incorporating changes in variability, which is widely used for agricultural impact studies. Daily climate scenarios can be used to study potential changes in agroclimatic resources. Weather generators can calculate agroclimatic indices on the basis of historical climate data and GCM outputs. Various agroclimatic indices can be used to assess crop production potentials and to rate the climatic suitability of land for crops. A methodologically more consistent approach is to use a stochastic weather generator, instead of historical data, in conjunction with a crop simulation model. The stochastic weather generator allows temporal extrapolation of observed weather data for agricultural risk assessment as well as providing an expanded spatial source of weather data by interpolation between the point-based parameters used to define the weather generators. Interpolation procedures can create both spatial input data and spatial output data. The density of meteorological stations is often low, especially in developing countries, and reliable and complete long-term data are scarce. Daily interpolated surfaces of meteorological variables rarely exist. More commonly, weather generators can be used to generate the weather variables in grids that cover large geographic regions and come from interpolated surfaces of weekly or monthly climate variables. From these interpolated surfaces, daily weather data for crop simulation models are then generated using statistical models that attempt to reproduce series of daily data with means and variability similar to what would be observed at a given location.

Weather generators have the capacity to simulate statistical properties of observed weather data for agricultural applications, including a set of agroclimatic indices. They are able to simulate temperature, precipitation, and related statistics. Weather generators typically calculate daily precipitation risk and use this information to guide the generation of other weather variables, such as daily solar radiation, maximum and minimum temperature, and potential evapotranspiration. They also can simulate statistical properties of daily weather series under a changing/changed climate through modifications to the weather generator parameters with optimal use of available information on climate change. For example, weather generators can simulate the frequency distributions of the wet and dry spells fairly well by modifying the four transition probabilities of the second-order Markov chain. Weather generators are generally based on the statistics. For examples, to generate the amount of precipitation on wet days, a two-parameter gamma distribution function is commonly used. The two parameters, a and b, are directly related to the average amount of precipitation per wet day. They can, therefore, be determined with the monthly means for the number of rainy days per month and the amount of precipitation per month, which are obtained either from compilations of climate normal or from interpolated surfaces.

The popular weather generators are WGEN (Richardson, 1984, 1985), SIMMETEO (Geng et al., 1986, 1988), and MARKSIM (Jones and Thornton, 1998; 2000), etc. They are including a first or high order Markov daily generator that requires long-term, at least 5 to 10 years, daily weather data or climate clusters of interpolated surfaces for estimation of theirs parameters. The software allows three

types of input to estimate parameters for the generator:

(1) Latitude and longitude;
(2) Latitude, longitude and elevation;
(3) Latitude, longitude, elevation and long-term monthly climate normals.

5. AGROMETEOROLOGICAL INFORMATION

The impacts of meteorological factors on crop growth and development are consecutive, although sometimes they do not emerge over a short time. The weather and climatological information should vary according to the kind of crop, its sensitivity to the environment factors and water requirements, etc. Certain statistics are important, such as sequences of consecutive days when maximum and minimum temperatures or the amounts of precipitation exceed or are less than certain critical threshold values and the average and extreme dates when these threshold values are reached.

The following are some of the more frequent types of information which can be derived from the basic data:

(a) Air temperature
- (i) Temperature probabilities;
- (ii) Chilling hours;
- (iii) Degree days;
- (iv) Hours or days above or below selected temperatures;
- (v) Interdiurnal variability;
- (vi) Maximum and minimum temperature statistics; and,
- (vii) Growing season statistics. Dates when threshold values of temperature for various kinds of crops growth begin and end.

(b) Precipitation
- (i) Probability of specified amount during a period;
- (ii) Number of days with specified amounts of precipitation;
- (iii) Probabilities of thundershowers;
- (iv) Duration and amount of snow cover;
- (v) Date of beginning and ending of snow cover; and,
- (vi) Probability of extreme precipitation amounts.

(c) Wind
- (i) Wind rose;
- (ii) Maximum wind, average wind speed;
- (iii) Diurnal variation; and,
- (iv) Hours of wind less than selected speed.

(d) Sky cover, sunshine, radiation
- (i) Percent possible sunshine;
- (ii) Number of clear, partly cloudy, cloudy days; and,
- (iii) Amounts of global and net radiation.

(e) Humidity
  (i) Probability of specified relative humidity; and,
  (ii) Duration of specified threshold of humidity with time.

(f) Free water evaporation
  (i) Total amount;
  (ii) Diurnal variation of evaporation;
  (iii) Relative dryness of air; and,
  (iv) Evapotranspiration.

(g) Dew
  (i) Duration and amount of dew;
  (ii) Diurnal variation of dew;
  (iii) Association of dew with vegetative wetting; and,
  (iv) Probability of dew formation with season.

(h) Soil temperature
  (i) Mean and standard deviation at standard depth;
  (ii) Depth of frost penetration;
  (iii) Probability of occurrence of specified temperatures at standard depths; and,
  (iv) Dates when threshold values of temperature (germination, vegetation) are reached.

(i) Weather hazards or extreme events
  (i) Frost;
  (ii) Cold Wave;
  (iii) Hail;
  (iv) Heat Wave;
  (v) Drought;
  (vi) Cyclones;
  (vii) Flood;
  (viii) Rare sunshine; and,
  (ix) Waterlogging.

(j) Agrometeorological observations
  (i) Soil moisture at regular depths;
  (ii) Plant growth observations;
  (iii) Plant population;
  (iv) Phenological events;
  (v) Leaf area index;
  (vi) Above ground biomass;
  (vii) Crop canopy temperature;
  (viii) Leaf temperature; and,
  (ix) Crop root length.

5.1. Forecast information

Operational weather information is defined as real-time data which provide

conditions of past weather (over the previous few days), present weather, as well as predicted weather. It is well known, however, that the forecast product deteriorates with time, so that the longer the forecast period, the less reliable the forecast. Forecasting of agriculturally important elements is discussed in Chapters 4 and 5.

6. STATISTICAL METHODS OF AGROMETEOROLOGICAL DATA ANALYSIS

The remarks set out here are intended to be supplementary to Chapter 5, "The use of statistics in climatology", of the WMO Guide to Climatological Practices and to WMO Technical Note No. 8l, “Some methods of climatological analysis”, which contain advice generally appropriate and applicable to agricultural climatology.

Statistical analyses play an important role in agrometeorology for they provide a means of interrelating series of data from diverse sources, namely biological data, soil and crop data, and atmospheric measurements. Because of the complexity and multiplicity of the effects of environmental factors on the growth and development of living organisms, and consequently on agricultural production, it is sometimes necessary to use rather sophisticated statistical methods to detect the interactions of these factors and their practical consequences.

It must not be forgotten that advice on the long-term agricultural planning, on the selection of the most suitable farming enterprise, on the provision of proper equipment, and on the introduction of protective measures against severe weather conditions all depend to some extent on the quality of the climatological analyses of the agroclimatic and related data, and, hence, on the statistical methods on which these analyses are based. Another point which needs to be stressed is that one is often obliged to compare measurements of the physical environment with biological data, which are often difficult to quantify.

Once the agrometeorological data are stored in electronic form in a file or database, it can be analyzed using a public domain number or commercial statistical software. Some basic statistical analyses can be performed in widely available commercial spreadsheets software. More comprehensive basic and advanced statistical analyses generally require specialized statistical software. Basic statistical analyses include simple descriptive statistics, distribution fitting, correlation analysis, multiple linear regression, nonparametrics, and enhanced graphic capabilities. Advanced software includes linear/non-linear models, time series and forecasting, and multivariate exploratory techniques such as cluster analysis, factor analysis, principal components and classification analysis, classification trees, canonical analysis, and discriminant analysis. Commercial statistical software for PCs would be expected to provide a user-friendly interface with self-prompting analysis selection dialogs. Many software packages include electronic manuals which provide extensive explanations of analysis options with examples and comprehensive statistical advice.

Some commercial packages are rather expensive, but there are some free statistical analysis software which can be downloaded from the web or can be made available upon request. One example of freely available software is INSTAT, which

was developed with applications in agrometeorology in mind. It is a general purpose statistics package for PCs which was developed by the Statistical Service Centre of the University of Reading, England. It uses a simple command language to process and analyze data. The documentation and software can be downloaded from the web. Data for analysis can be entered into a table or copied and pasted from the clipboard. If CLICOM is used as the database management software, then INSTAT, which was designed for use with CLICOM, can readily be used to extract the data and perform statistical analyses. INSTAT can be used to calculate simple descriptive statistics including: minimum and maximum values, range, mean, standard deviation, median, lower quartile, upper quartile, skewness, and kurtosis. It can be used to calculate probabilities and percentiles for standard distributions, normal scores, t-tests and confidence intervals, chi-square tests, and non-parametric statistics. It can be used to plot data, for regression and correlation analysis and analysis of time series. INSTAT is designed to provide a range of climate analyses. It has commands for 10-day, monthly, and yearly statistics. It calculates water balance from rainfall and evaporation, start of rains, degree days, wind direction frequencies, spell lengths, potential ET according to Penman, and crop performance index according to FAO methodology. The usefulness of INSTAT for agroclimatic analysis is illustrated in the publication on the Agroclimatology of West Africa: Niger. The major part of the analysis reported in the bulletin was carried out using INSTAT.

6.1. Series checks

Before selecting a series of values for statistical treatment, the series should be carefully examined for validity. The same checks should be applied to series of agrometeorological data as to conventional climatological data; in particular, the series should be checked for homogeneity and, if necessary, gaps should be filled in. It is assumed that beforehand, the individual values will have been carefully checked (consistency and coherence) in accordance with section 4.3 of the WMO Guide to Climatological Practices.

Availability of good metadata is essential during analysis of the homogeneity of a data series. For example, a large number of temperature and precipitation series were analyzed on homogeneity (Müller-Westermeier 2004). Because in the country of those observations some metadata are archived, the research could show that at least two-thirds of the homogeneity breaks in those series were not due to to climate change, but were caused by instrument relocations, including changes in observation height.

6.2. Climatic scales

In agriculture, perhaps more than in most economic activities, all scales of climate need to be considered (see section 2.1.3):

(a) For the purpose of meeting national and regional requirements, studies on a macroclimatic scale are useful, and may be based mainly on data from synoptic stations. For some atmospheric parameters with little spatial variation--e.g., duration of sunshine over a week or 10-day period--such an analysis is found to be satisfactory;

(b) In order to plan the activities of an agricultural undertaking, or group of undertakings, it is essential, however, to change over to the mesoclimatic or topoclimatic scale, i.e., to take into account local geomorphological features and to use data from an observational network with a finer mesh. These complementary climatological series of data may be for much shorter periods than those used for macroclimatic analyses, provided they can be related to some long reference series;

(c) For bioclimatic research, the physical environment should be studied at the level of the plant or animal or the pathogenic colony itself. Obtaining information about radiation energy, moisture, and chemical exchanges involves handling measurements on the much finer scale of microclimatology.

(d) For research on impact of changing climate, past long-term historical and future climate scenarios should be supposed and extrapolated.

6.2.1. Reference periods

The length of the reference period for which the statistics are defined should be selected according to its suitability for each agricultural activity. Calendar periods of a month or a year are not, in general, suitable. It is often best either to use a reduced timescale or, alternatively, to combine several months in a way that will slow the overall development of an agricultural activity. The following periods are thus suggested for reference purposes:

(a) Ten-day or weekly periods, for operational statistical analyses, e.g., evapotranspiration, water balance, sums of temperature, frequency of occasions when a value exceeds or falls below a critical threshold value, etc. However, data for the weekly period, which has the advantage of being universally adopted for all activities, are difficult to adjust for successive years;

(b) For certain agricultural activities the periods should correspond to phenological stages or to the periods when certain operations are undertaken in crop cultivation. Thus, water balance, sums of temperature, sequences of days with precipitation, or temperature below certain threshold values, etc., could be analyzed for:

(i) The mean growing season;

(ii) Periods corresponding to particularly critical phenological stages;

(iii) Periods during which crop cultivation, plant protection treatment, or preventive measures are found to be necessary.

These suggestions, of course, imply a thorough knowledge of the normal calendar of agricultural activities in an area.

6.2.2. The beginning of reference periods

In agricultural meteorology, it is best to choose starting points corresponding to the biological rhythms, since the arbitrary calendar periods (year, month) do not coincide with these. For example, in temperate zones, the starting point could be autumn (sowing of winter cereals) or spring (resumption of growth). In regions

subject to monsoons or the seasonal movement of the intertropical convergence zone, it could be the onset of the rainy season. It could also be based on the evolution of a significant climatic factor considered to be representative of a biological cycle difficult to assess directly, e.g., summation of temperatures exceeding a threshold temperature necessary for growth.

6.2.3. Analysis of effects of weather

The climatic elements do not act independently on the biological life-cycle of living things: an analytical study of their individual effects is often illusory; handling them all simultaneously, however, requires considerable data and complex statistical treatment. It is often better to try to combine several factors into single agroclimatic indices, considered as complex parameters, which can be compared more easily with biological data.

6.3. Frequency Distributions

When dealing with a large set of measured data, it is usually necessary to arrange it into a certain number of equal groupings, or classes, and to count the number of observations that fall into each class. The number of observations falling into a given class is called the *frequency* for that class. The number of classes chosen depends on the number of observations. As a rough guide, the number of classes should not exceed five times the logarithm (base 10) of the number of observations.

Thus, for 100 observations or more, there should be a maximum of ten classes. It is also important that adjacent groups do not overlap. The result of doing this can be displayed in a grouped frequency table, based on table 1, such as the one depicted in table 2.

Table 1.Climatological series of annual rainfall (mm) for Mbabane (1930-1979)

Year	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
193-	1063	1237	1495	1160	1513	912	1495	1769	1319	2080
194-	1350	1033	1707	1570	1480	1067	1635	1627	1168	1336
195-	1102	1195	1307	1118	1262	1585	1199	1306	1220	1328
196-	1411	1351	1115	1256	1226	1062	1546	1545	1049	1830
197-	1018	1690	1800	1528	1285	1727	1704	1741	1667	1260

Table 2: Frequency Distribution of annual precipitation for Mbabane (1930-1979)

	1 *Group boundaries*	2 *Group limits or class interval*	3 *Mid-mark* x_i	5 *Frequency* f_i	6 *Cummulative frequency* F_i	7 *Relative cumulative frequency(%)*
1	879.5-1029.5	880-1029	954.5	2	2	4
2	1029.5-1179.5	1030-1179	1104.5	8	10	20
3	1179.5-1329.5	1180-1329	1254.5	15	25	50
4	1329.5-1479.5	1330-1479	1404.5	4	29	58
5	1479.5-1629.5	1480-1629	1554.5	10	39	78
6	1629.5-1779.5	1630-1779	1704.5	8	47	94
7	1779.5-1929.5	1780-1929	1854.5	2	49	98
8	1929.5- 2079.5	1930-2079	2004.5	0	49	98
9	2079.5- 2229.5	2080-2229	2154.5	1	50	100
			Total:	50	-	

The table has columns showing *limits* defining classes and another column giving *lower and upper class boundaries* that in turn give rise to *class widths or class intervals*, yet another column gives the *mid-marks* of the classes, and another column gives the totals of the tally known as the *group or class frequencies*.

Another column has entries that are known as the *cumulative frequencies*. They are obtained from the frequency column by entering the number of observations with values less than or equal to the value of the upper class boundary of that group.

The pattern of frequencies obtained by arranging data into classes is called the *frequency distribution* of the sample. The probability of finding an observation in a class can be obtained by dividing the frequency for the class by the total number of observations. A frequency distribution can be represented graphically with a two-dimensional histogram, where the heights of the columns in the graph are proportional to the class frequencies.

Examples using frequency distribution

The probability of an observation falling in class number five is $\frac{10}{50} = 0.2$ or 20%. That is the same thing as saying that the probability of getting between 1480 mm and 1620 mm of rain in Mbabane is 20% or once in five years. The probability of getting less than 1779 mm of rain in Mbabane as in class six is 0.94, found by dividing the cumulative frequency up to this point by 50, the total number of observations or frequencies. This kind of probability is also known as relative cumulative frequency, given as percentage in column seven. From column seven, the probability of getting between 1330 mm and 1929 mm is 98% minus 58%, which is 40%.

Frequency distribution groupings have the disadvantage that when using them, certain information is lost, such as the highest observation in the highest frequency class.

6.3.1 Probability Based On Normal Distributions

A normal distribution is a highly refined frequency distribution having an infinite number of very narrow classes. The histogram from this distribution has smoothed out tops that make a continuous smooth curve, known as a normal or bell curve. A normal curve is symmetric about its centre having a horizontal axis that runs indefinitely both to the left and to the right, with the tails of the curve tapering off towards the axis in both directions. The vertical axis is chosen in such a way that the total area under the curve is exactly 1 (one square unit). The central point on the axis beneath the normal curve is the mean μ and the set of data that produced it has a standard deviation σ. Any set of data that tends to give rise to a normal curve is said to be normally distributed. The normal distribution is completely characterized by its mean and standard deviation. Sample statistics are functions of observed values that are used to infer something about the population from which the values are drawn. The sample mean and sample variance, for instance, can be used as estimates of population mean and population variance, respectively, provided the relationship between these sample statistics and the populations from which the samples are drawn is known. In general, the sampling distribution of means is less spread out than the parent population. This fact is embodied in the ***central limit theorem*** which states that if random samples of size *n* are drawn from a large population(hypothetically infinite), which has mean μ and standard deviation σ, then the theoretical sampling distribution of $\overline{X}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. The theoretical sampling distribution of $\overline{X}$ can be closely approximated by the corresponding Normal curve if *n* is large. Thus, for quite small samples, particularly if we know that the parent population is itself approximately Normal, we can confidently apply the theorem. If we are not sure that the parent population is Normal, we should, as a rule, restrict ourselves to applying the theorem to samples of size ≥ 30.

The standard deviation of a sampling distribution is often called the standard error of the sample statistic concerned. Thus $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ is the standard error of $\overline{X}$.

In order to compare different distributions having different means and different standard deviations with one another they need to be transformed. One way would be to center them about the same mean by subtracting the mean from each observation in each of the populations. This will move each of the distributions along the scale until they are centered about zero, which is the mean of all transformed distributions. However, each distribution will still maintain a different bell-shape.

The Z.score

A further transformation is done by subtracting the mean of the distribution from each observation and dividing by the standard deviation of the distribution, a procedure

known as standardization. The result is a variable *Z*, known as a z -score and having the *standard normal form*:

$$z = \frac{X - \mu}{\sigma}$$

This will give identical bell shaped curves with Normal Distribution around zero mean and standard deviation equal to unit.

The *z-scale* is a horizontal scale set up for any given normal curve with some mean μ and some standard deviation σ. On this scale, the mean is marked 0 and the unit measure is taken to be σ, the particular standard deviation of the normal curve in question. A raw score X can be converted into a *z-score* by the above formula:

For instance, with $\mu = 80$ and $\sigma = 4$, if we want formally to convert the X-score 85 into a *z-score*,

we write

$$z = \frac{X - \mu}{\sigma} = \frac{85 - 80}{4} = \frac{5}{4} = 1.25.$$

The meaning here is that the X-score lies one standard deviation to the right of the mean. If we compute a *z-score* equivalent of X=74, we get

$$z = \frac{X - \mu}{\sigma} = \frac{74 - 80}{4} = \frac{-6}{4} = -1.5.$$

The meaning of this negative z-score is that the original X-score 74 lies one and one-half standard deviations (that is, six units) to the left of the mean. A *z-score* tells how many standard deviations removed from the mean the original *x*-score is, to the right (if *z* is positive) or to the left (if *z* is negative).

There are many different normal curves due to the different means and standard deviations. However, for a fixed mean μ and a fixed standard deviation σ, there is exactly one normal curve having that mean and that standard deviation.

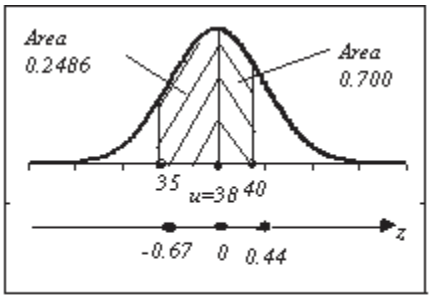
Normal Distributions can be used to calculate probabilities. Since a normal curve is symmetric having a total area of one square unit under it then the area to the right of the mean is half a square unit and similarly to the left. The characteristics of the standard normal distribution are extremely well known, and tables of areas under specified segments of the curve are available in almost all statistical books. The areas are directly expressed as probabilities. The probability of encountering a sample, by random selection from a normal population, whose measurement falls within a specified range, can be found with the use of these tables. The variance of the population must, however, be known. The fundamental idea connected with the area

under a normal curve is that if a measurement X is normally distributed, then the probability that X will lie in some range between a and b on any given occasion is equal to the area under the normal curve between a and b.

To find the area under a normal curve between the mean μ and some x-value, convert the x into a *z-score*. The number indicated is the desired area. If z turns out to be negative, just look it up as if it were positive. If the data are normally distributed, then it is probable that at least 68 percent of data in the series will fall within $\pm 1\sigma$ of the mean, that is $z = \pm 1$. Also, the probability is 95 percent that all data fall within $\pm 2\sigma$ of the mean, or $z = \pm 2$, and 99 percent within $\pm 3\sigma$ of the mean or $z = \pm 3$.

Example using the z-score

a) If the heights of all the rice stalks in a farm are thought to be normally distributed with mean $\mu = 38$ cm and standard deviation $\sigma = 4.5$ cm, find the probability that the height of a stalk taken at random will be between 35 and 40 cm. To solve this problem, we must find the area under a portion of the appropriate normal curve, between x=35 and x=40. (See Figure below). It is necessary to convert these x-values into *z-scores* as follows.



For X=35: $z = \dfrac{X - \mu}{\sigma} = \dfrac{35 - 38}{4.5} = \dfrac{5}{4.5} \cong -0.67.$

For X=40: $z = \dfrac{X - \mu}{\sigma} = \dfrac{40 - 38}{4.5} = \dfrac{2}{4.5} \cong 0.44.$

From the mean out to z=0.44, the area (on the right side of the region in Figure 12-15) from Table 2 is 0.1700.

From the mean out to z=-0.67, the area (on the left side of the region in Figure 12-15) from Table 2 is 0.2486.

It is clear from Figure 12-15 that the shaded area on both sides must be counted, and so we add areas to obtain 0.1700+0.2486=0.4186.

Thus, the probability that a stalk chosen at random will have height X between 35 and 40 cm is 0.4186. In other words, we would expect 41.86% of the paddy field's rice stalks to have heights in that range.

Elements that are not normally distributed may easily be transformed mathematically to the normal distribution, an operation known as normalization. Among the moderate normalizing operators are the square root, the cube root, and the logarithm for positively skewed data such as rainfall. The transformation reduces the higher values by proportionally greater amounts than smaller values.

b) Suppose a population of pumpkins is known to have a Normal Distribution with a mean and standard deviation of its length as 14.2 cm and 4.7 cm respectively. What is the probability of finding, by chance, a specimen shorter than 3 cm? To find the answer, we must convert 3 cm to units of standard deviation and use the Standard Normal Distribution Table.

$$Z = \frac{3.0 - 14.2}{4.7} \approx -2.4 .$$

The probability of finding a variety smaller than -2.4 standard deviations is the cumulative probability to this point; from our table, we can see that it is 0.0082, which is very small indeed. Now, what is the probability of finding one longer than 20 mm? Again, converting to standard normal form:

$$Z = \frac{20.0 - 14.2}{4.7} \approx 1.2$$

Because the total area under the normal distribution curve is 1.00, the probability of obtaining a measurement of 1.2 standard deviations or greater than the mean is the same as 1.0 minus the cumulative probability of obtaining anything smaller. Standard Normal Distribution Table will give us the cumulative probability up to 1.2, which is 0.8849. Therefore, the probability of finding a specimen longer than 20 cm is 1.0000-0.8849=0.1151
or slightly greater than one chance out of ten. Now, compute the probability of finding, at random, a stalk whose length falls in the size range from 15 to 20 cm:

for 15 cm $$Z = \frac{15.0 - 14.2}{4.7} \approx 0.2$$

for 20 cm $$Z = \frac{20.0 - 14.2}{4.7} \approx 1.2$$

6.*3.2. Extreme-value distributions*

Certain crops may be exposed to lethal conditions (frost, excessive heat or cold, drought, high winds, etc.), even in areas where they are commonly grown.

Extreme value analysis typically involves the collection and analysis of annual maxima of parameters that are observed daily such as temperature, precipitation, and wind speed. The process of extreme value analysis involves data gathering, the identification of a suitable probability model, such as the Gumbel distribution or GEV distribution; (Coles, 2001) to represent the distribution of the observed extremes, the estimation of model parameters, and the estimation of the return values for periods of fixed length.

The Gumbel double exponential distribution is the most used for describing extreme values. An event which has occurred *m* times in a long series of *n* independent trials, one per year say, has an estimated probability $p = \frac{m}{n}$; conversely the average interval between recurrences of the event during a long period would be $\frac{n}{m}$; this is defined as the return period *T* where:

$$T = \frac{1}{p}$$

For example, if there is a 5 percent chance that an event will occur in any one year, then its probability of occurrence is 0.05. This can be expressed as an event having a return period of five times in 100 years or once in 20 years. This means that the event is more likely that over a long period of say 200 years, ten events of equal or greater magnitude would have occurred.

For a valid application of extreme value analysis, two conditions must be met:

First, the data must be independent, i.e., the occurrence of one extreme is not linked to the next.

Second, the data series must be trend free and the quantity of data must be big, usually not less than 15 values.

6.3.3. Probability and risk

Frequency distributions, which provide an indication of risk, are of particular interest in agriculture due to the existence of ecological thresholds which, when reached, may result either in a limited yield or in irreversible reactions within the living tissue. Histograms can be fitted to the most appropriate distribution function and used to make statements about probabilities or risk of critical climate conditions such as freezing temperatures or dry spells of more than a specified number of days. Cumulative frequencies are particularly suitable and convenient for operational use in agrometeorology. Cumulative distributions can be used to prepare tables or graphs showing the frequencies of occasions when the values of certain parameters exceed (or fall below) given threshold values during a selected period. If a sufficiently long

series of observations (10 to 20 years) is available, it can be assumed to be representative of the total population, so that mean durations of the periods when the values exceed (or fall below) specified thresholds can be deduced. When calculating these mean frequencies, it is often an advantage to extract information regarding the extreme values observed during the period chosen such as the growing season, growth stage, or period of particular sensitivity. Some examples are:

(a) Threshold values of daily maximum and minimum temperatures, which can be used to estimate the risk of excessive heat or frost and the duration of this risk;
(b) Threshold values of ten-day water deficits, taking into account the reserves in the soil. The quantity of water required for irrigation can then be estimated.
(c) Threshold values of relative humidities from hour or 3-hour observations.

6.3.4. The distribution of sequences of consecutive days

The distribution of sequences of consecutive days in which certain climatic events occur is of special interest to the agriculturist. From such data, one can, for example, deduce the likelihood of being able to undertake cultural operations requiring specific weather conditions and lasting for several days (haymaking, gathering grapes, etc.). The choice of protective measures to be taken against frost or drought may likewise be based on an examination of their occurrence and the distribution of the corresponding sequences. For whatever purpose, the sequences are to be used, it is important to specify clearly the periods to which they refer (also whether or not they are for overlapping periods). Markov chain probability models have frequently been used to estimate the probability of sequences of consecutive days such as wet days or dry days. Under many climate conditions, the probability, for example, of a day being dry is significantly larger if the previous day is known to have been dry. Knowledge of the persistence of weather events such as wet days or dry days can be used to estimate the distribution of consecutive days using a Markov chain. INSTAT includes algorithms to calculate Markov chain models, to simulate spell lengths and estimate probability using climatological data.

6.4. Measuring Central Tendency

One descriptive aspect of statistical analysis is the measurement of what is called *central tendency,* which gives an idea of the average or middle value about which all measurements coming from the process will cluster. To this group belong the mean, the median, and the mode.

Their symbols are as listed below:

$\overline{X}$ – arithmetic mean of a sample

μ – population mean

$\overline{X}w$ – weighted mean

$\overline{X}h$ – harmonic mean

Me – median

Mo – mode

6.4.1. The Mean

While frequency distributions are undoubtedly useful for operational purposes, mean values of the main climatic elements (10-day, monthly, or seasonal) may be used broadly to compare climatic regions. To show how the climatic elements are distributed, these mean values, however, should be supplemented by other descriptive statistics such as the standard deviation, coefficient of variation (variability), quintiles, and extreme values. In agroclimatology, series of observations which have not been made simultaneously may have to be compared. To obtain comparable means in such cases, adjustments are applied to the series so as to fill in any gaps (see WMO Technical Note No. 81). Sivakumar, et al., (1993) illustrate the application of INSTAT in calculating descriptive statistics for climate data and discuss the usefulness of the statistics for assessing agricultural potential. They produce tables for available stations of monthly mean, standard deviation, maximum and minimum for rainfall amounts, and for the number of rainy days. Descriptive statistics are also presented for maximum and minimum air temperatures.

(a) The *arithmetic mean* is the most used measure of central tendency, defined as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} xi \qquad \text{i} = 1,2,...\text{n} \qquad (1)$$

This means adding all data in a series and dividing their sum by the number of data. The mean of the annual precipitation series from table 1 is:

$$\overline{X} = \frac{\sum x}{n} i = 69449/50 = 1388.9 \text{ mm} \qquad (2)$$

The arithmetic mean may be computed using other labour saving methods such as the grouped data technique (WMO-No.100), which estimates the mean from the average of the products of class frequencies and their mid-points.

Another version of the mean is the *weighted mean,* which takes into account the relative importance of each variate by assigning it a weight. An example of the weighted mean is when making areal averages such as yields, population densities or areal rainfall over non-uniform surfaces. The value for each sub-division of the area is multiplied by the sub-division area, and then the sum of the products is divided by the total area. The formula for the weighted mean is expressed as:

$$\overline{X}w = \frac{\sum_{i\equiv 1}^{k} n_i \overline{x_i}}{\sum_{i\equiv 1}^{k} n_i}. \qquad (3)$$

For example, the average yield of maize for the five districts in Ruvuma Region of Tanzania was respectively, 1.5 tons, 2.0 tons, 1.8 tons, 1.3 tons, and 1.9 tons per hectare. The respective areas under maize were, 3000, 7000, 2000, 5000, and 4000. Substituting n1=3000, n2=7000, n3=2000, n4=5000, and n5=4000 into equation (3), we have the over-all mean yield of maize for these 21,000 hectares of land.

$$\overline{X}w = \frac{3000(1.5) + 7000(2.0) + 2000(1.8) + 5000(1.3) + 4000(1.9)}{3000 + 7000 + 2000 + 5000 + 4000} = \frac{33800}{21000} = 1.6$$

In operational agrometeorology, the mean is normally computed for ten-days, known as dekads, as well as for the day, month, year, and longer periods. This is used in agrometeorological bulletins and for describing current weather conditions. At agrometeorological stations where the maximum and the minimum temperatures are read, a useful approximation to the daily mean temperature is given by taking the average of these two temperatures. Such averages should be used with caution when comparing data from different stations as such averages may differ systematically from each other.

Another measure of the mean is the *harmonic mean* defined as *n* divided by the sum of the reciprocals or multiplicative inverses of the numbers

$$\overline{X}h = \frac{n}{\sum_{i=1}^{n} \frac{1}{Xi}}$$

If five sprinklers can individually water a garden in 4 hours, 5 hours, 2 hours, 6 hours, and 3 hours, respectively, the time required for all pipes working together to water the garden is given by

$$\mathrm{t} = \overline{X}h \frac{1}{n} = 46 \text{ minutes and } 45 \text{ seconds.}$$

Means of long-term periods are known as normals. A normal is defined as a period average computed for a uniform and relatively long period comprising of at least three consecutive 10-year periods. A climatological standard normal is the average of

climatological data computed for consecutive periods of 30 years as follows: 1 January 1901 to 31 December 1930, 1 January 1931 to 31 December 1960, etc.

6.4.2. The mode

The mode is the most frequent value in any array. Some series have even more than one modal value. Mean annual rainfall patterns in some sub-equatorial countries have bi-modal distributions, meaning they exhibit two peaks. Unlike the mean, the mode is an actual value in the series. Its use is mainly in describing the average.

6.4.3. The median

The median is obtained by selecting the middle value in an odd-numbered series of variates or taking the average of the two middle-values of an even-numbered series. For large numbers of data it is easiest to obtain a close approximation of their median by graphical or numerical interpolation of their cumulative frequency distribution.

6.5. Fractiles

Fractiles such as quartiles, quintals, and deciles are obtained by first ranking the data in ascending order and then counting an appropriate fraction of the integers in the series (n+1). For quartiles, we divide n+1 by four, for deciles by ten, and for percentiles by a hundred. Thus if $n = 50$, the first decile is the $\frac{1}{10}[n+1]^{\text{th}}$ or the 5.1^{th} observation in the ascending order, the 7^{th} decile is the $\frac{7}{10}[n+1]^{\text{th}}$ in the rank or the 35.7^{th} observation. Interpolation is required between observations. The median is the 50^{th} percentile. It is also the fifth decile and the second quartile. It lies in the third quintile. In agrometeorology, the first decile means that value below which one-tenth of the data falls and above which 9-tenths lie.

6.6. Measuring Dispersion

Other parameters give information about the spread or dispersion of the measurements about the average. These include the range, the variance, and the standard deviation.

6.6.1. The Range

This is the difference between the largest and the smallest values. For instance, the annual range of mean temperature is the difference between the mean daily temperatures of the hottest and coldest months.

6.6.2. The Variance and the Standard Deviation

The variance is the mean of the squares of the deviations from the arithmetic mean. The standard deviation ***S*** is the square root of the variance and is defined as the root-mean-square of the deviations from the arithmetic mean. To obtain the standard deviation of a given sample, the mean $\overline{X}$ is computed first and then the deviations

from the mean $(\overline{X}_i - \overline{X})$:

$$S=\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Alternatively, with only a single computation run summating data values and their squares:

$$S=\sqrt{((\sum(x_i^2) - (\sum(x_i)^2/n)/(n-1))}$$

This standard deviation has the same units as the mean; together they may be used to make precise probability statements about the occurrence of certain values of a climatological series. The influence of the actual magnitude of the mean can be easily eliminated by expressing S as a percentage of the mean to get a dimensionless quantity called the coefficient of variation:

$$Cv = \frac{s}{x} \times 100$$

For comparing values of s between different places, this can be used to provide a measure of relative variability, for such elements as total precipitation.

6.6.3. Measuring Skewness

Yet, others tell us if the population tends to have values straggling out in a tail on one side, a property known as skewness, or asymmetry, i.e., there is a good chance of finding an observation a long way from the middle value on one side but not on the other.

7. DECISION MAKING

7.1. Statistical Inference and Decision Making

Statistical inference is a process of inferring information about a population from the data of samples drawn from it. The purpose of statistical inference is to help a decision-maker to be right more often than not or at least to give some idea of how much danger there is of being wrong when a particular decision is made. It is also meant to ensure that long-term costs through wrong decisions are kept to the minimum.

Two main lines of attacking the problem of statistical inference are available. One is to devise sample statistics which may be regarded as being suitable estimators of

corresponding population parameters. For example, we may use the sample mean $\overline{X}$ as an estimator of the population mean μ, or else we may use the sample median Me. Statistical estimation theory deals with the issue of selecting best estimators.

The steps to be taken to arrive at a decision are as follows:

Step 1. Formulate the null and alternative hypotheses,

Once the null hypothesis has been clearly defined, we may calculate what kind of samples to expect under the supposition that it is true. Then if we draw a random sample, and if it differs markedly in some respect from what we expect, we say that the observed difference is significant; and we are inclined to reject the null hypothesis and accept the alternative hypothesis. If the difference observed is not too large, we might accept the null hypothesis; or we might call for more statistical data before coming to a decision. We can make the decision in a hypothesis test depending upon a random variable known as a *test statistic,* such as z-score used in finding confidence intervals, and we can specify critical values of this, which can be used to indicate not only whether a sample difference is significant but also the strength of the significance.

For instance in a coin experiment to determine if the coin is fair or loaded:

Null Ho: p=0.5 (i.e. the coin is fair)

And alternative H_1: p#0.5 (i.e. the coin is biased)

(Or equivalently H_1: p<0.5 or p>0.5; this is called a two-sided alternative).

Step 2. Choose an appropriate level of significance

We call the probability of wrongly rejecting a null hypothesis the level of significance (α) of the test. We select the value for α first, before carrying out any experiments; the values most commonly used by statisticians are 0.05, 0.01, and 0.001. The level of significance α=0.5 means that our test procedure has only 5 chances in 100 of leading us to decide that the coin is biased if in fact it is not.

Step 3. Choose the sample size n.

It is fairly clear that if bias exists, a large sample will have more chance of demonstrating its existence than a small one. And so, we should make *n* as large as possible, especially if we are concerned with demonstrating a small amount of bias. Cost of experimentation, time involved in sampling, necessity of maintaining statistically constant conditions, amount of inherent random variation, and possible

consequences of making wrong decisions are among the considerations on which the sizes of sample to be drawn depend.

Step 4. Decide upon the test statistic to be used.

We can make the decision in a hypothesis test depending upon a random variable known as a test statistic such as z or t as used in finding confidence intervals. Its sampling distribution, under the assumption that Ho is true, must be known. It can be normal, binomial, or other sampling distributions.

Step 5. Calculate the acceptance and rejection regions

Assuming that the null hypothesis is true, and bearing in mind the chosen values of n and alpha, we now calculate an acceptance region of values for the test statistic. Values outside this region form the rejection region. The acceptance region is so chosen that if a value of the test statistic, obtained from the data of a sample, fails to fall inside it, then the assumption that Ho is true must be strongly doubted. In general, we have a test statistic X, whose sampling distribution, defined by certain parameters such as η and σ, is known. The values of the parameters are specified in the null hypothesis Ho. From integral tables of the sampling distribution we obtain critical values X_1, X_2 such that

$P[X_1<X<X_2] = 1-\alpha$.

These determine an acceptance region, which gives a test for the null hypothesis at the appropriate level of significance (α).

Step 6. Formulate the decision rule.

The general decision rule, or test of hypothesis, may now be stated as follows:

(a) Reject Ho at the α significance if the sample value of X lies in the rejection region (i.e. outside [X_1, X2]). This is equivalent to saying that the observed sample value is significant at the 100α % level.

The alternative hypothesis H_1 is then to be accepted.

(b) *Accept Ho if the sample value of X lies in the acceptance region [X_1, X_2].* (Sometimes, especially if the sample size is small, or if X is close to one of the critical values X_1 and X_2, the decision to accept Ho is deferred until more data is collected.)

Step 7. Carry out the experiment and make the test

The ***n*** trials of the experiment may now be carried out, and from the results, the value

of the chosen test statistic may be calculated. The decision rule described in Step 6 may then be applied. Note: All statistical test procedures should be carefully formulated before experiments are carried out. The test statistic, the level of significance, and whether a one-or two-tailed test is required, must be decided before any sample data is looked at. To switch tests in mid-stream, as it were, leads to invalid probability statements about the decisions made.

7.2. Two-Tailed and One-Tailed Test

If the critical region occupies both extremes of the test distribution, it is called a two-tailed test. If the critical region occurs only at high or low values of the test statistic, such a test is called one-tailed.

This leads to a two-tailed test. The critical region containing 5% of the area of the normal distribution is split into two equal parts, each containing 2.5% of the total area. If the computed value of Z falls into the left-hand region, the sample came from a population having a smaller mean than our known population. Conversely, if it falls into the right-hand region, the mean of the sample's parent population is larger than the mean of the known population. From the standardized normal distribution table (Table.), we find that approximately 2.5% of the area of the curve is to the left of a Z value of -1.9 and 97.5% of the area of the curve is to the left of +1.9.

Once the null hypothesis has been clearly defined, we may calculate what kind of samples to expect under the supposition that it is true. Then, if we draw a random sample, and if it differs markedly in some respect from what we expect, we say that the observed difference is significant; and we are inclined to reject the null hypothesis and accept the alternative hypothesis. If the difference observed is not too large, we might accept the null hypothesis; or we might call for more statistical data before coming to a decision. We can make the decision in a hypothesis test depending upon a random variable known as a test statistic such as z or t as used in finding confidence intervals, and we can specify critical values of this which can be used to indicate not only whether a sample difference is significant but also the strength of the significance.

7.3. Point Estimation

The two population characteristics μ and σ are called parameters of the population, while each of the sample characteristics such as sample mean, $\overline{X}$ and sample standard deviation S is called a sample statistic.

A sample statistic used to provide an estimate of a corresponding population

parameter is called a point estimator. For example, $\overline{X}$ may be used as an estimator of μ, Me may be used as an estimator of μ, S^2 may be used as an estimator of the population variance σ^2.

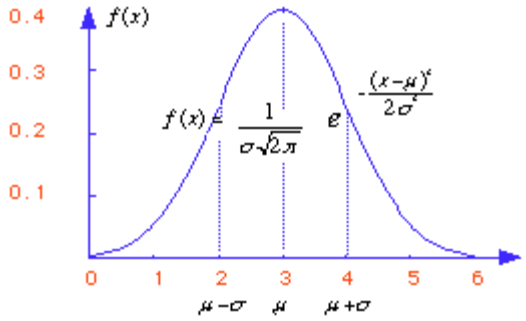
Any one of the statistics mean, median, mode, and mid-interquartile range would seem to be suitable for use as estimators of the population mean μ. In order to pick out the best estimator of a parameter out of a set of estimators, three important desirable properties should be considered. These are unbiasedness, efficiency, and consistency.

7.4. Interval Estimation

Confidence interval estimation is a technique of calculating intervals for population parameters and measures of confidence placed upon them. If we have chosen an unbiased sample statistic *b* as our point estimator of β, the estimator will have a sampling distribution, with mean E(*b*) = β and standard deviation S.D.(*b*) = σ_b. Here the parameter β is the unknown and our purpose is to estimate it. Using the remarkable fact that many sample statistics we use in practice have a Normal or approximately Normal sampling distribution, we can obtain from the tables of the Normal integral, the probability that a particular sample will provide a value of b within a given interval ($\beta - d$) to ($\beta + d$).

This is indicated in the diagram below. Conversely, for a given amount of probability, we can deduce the value *d*. For example, for 0.95 probability, we know from standard Normal tables that $\frac{d}{\sigma_b} = 1.96$. In other words, the probability that a sample will provide a value of *b* in the interval [β -1.96σ_b, β +1.96σ_b] is 0.95. We write this as P[β -1.96σ_b <=b<= β =1.96σ_b]=0.95. After rearranging the inequalities inside the brackets to the equivalent form $[b - 1.96\sigma_b \le b \le \beta + 1.96\sigma_b]$, we get the 95% confidence interval for β, namely the interval [*b*-1.96σ_b, *b*+1.96σ_b]. In general, we express confidence intervals in the form [*b*-z.σ_b, *b*+z.σ_b], where z, the z-score, is the number obtained from tables of the sampling distribution of *b*. This z-score is chosen so that the desired percentage confidence may be assigned to the interval; it is now called the confidence coefficient, or sometimes the critical value. The end points of a

confidence interval are known as the lower and upper confidence limits. The probable error of estimate is half the interval length of the 50% confidence interval, i.e., 0.674 σ.



The most commonly required point and interval estimates are for means, proportions, differences between two means, and standard deviations. The following table gives all the formulae needed for these estimates. The reader should note the standard form of $b \pm z$. σ_b for each of the confidence interval estimators.

For the formulae to be valid, sampling must be random and the samples must be independent. In some cases, σ_b will be known from prior information. Then, the sample estimator will not be used. In each of the confidence interval formulae, the confidence coefficient z may be found from tables of the Normal integral for any desired degree of confidence. This will give exact results if the population from which the sampling is done are Normal; otherwise, the errors introduced will be small if *n* is reasonably large ($n \geq 30$). A brief table of values of z is as follows:

Confidence level	50%	60%	80%	86.8%	90%	92%	93.4%	94.2	95%	95.6%	96%	97.4%	98%
Confidence coefficient z	0.674	0.84	1.28	1.50	1.645	1.75	1.84	1.90	1.96	2.01	2.05	2.23	2.33

What should we do when samples are small? It is clear that the smaller the sample, the smaller amount of confidence we can place on a particular interval estimate. Alternatively, for a given degree of confidence, the interval quoted must be wider than for larger samples. To bring this about, we must have a confidence coefficient which depends upon *n*. We shall use the letter *t* for this coefficient, and give confidence interval formulae for the population mean μ, and for the difference of two population means.

	Confidence interval	Degrees of freedom (υ)

1. Mean μ :	$\bar{x} \pm t.\sigma_{\bar{x}}$	n-1
Difference $\mu_1 - \mu_2$:	$(\bar{x}_1 - \bar{x}_2) \pm t.\sigma(\bar{x}_1 - \bar{x}_2)$	$n_1 + n_2$-2

The reader will note that these are the same as for large samples, except that *t* replaces *z*. When the sample estimators for $\sigma_{\bar{x}}$ and $\sigma_{\bar{x}_1 - \bar{x}_2}$ are used, the correct values for *t* are obtained from what is called the Student *t*-distribution. For convenience, they are related not directly to sample sizes, but to a number known as 'degrees of freedom'; we shall denote this by υ.

An abbreviated table of t-values is given below

υ :	3	4	5	7	9	10	15	20	25	30
90%	2.3	2.13	2.02	1.89	1.83	1.81	1.75	1.72	1.71	1.70
95%	5									
99%	3.18	2.78	2.57	2.36	2.26	2.23	2.13	2.09	2.06	2.04
	5.84	4.60	4.03	3.50	3.25	3.17	2.95	2.85	2.79	2.75

Degrees of freedom υ

7.5. The Z-Test.

The nature of the standard normal distribution allows us to test hypotheses about the origin of certain samples. The test statistic, Z, has a normal frequency distribution which is a standardized normal distribution defined as

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

The observations in the sample were selected randomly from a normal population whose variance is known.

7.6. Tests for Normal Population Means

A random sample size ***n*** is drawn from a Normal population having unknown mean μ and known Standard Deviation σ. The objective is to test the hypothesis Ho: $\mu = \mu'$; i.e., the assumption that the population mean has value μ'.

The variate $Z = \dfrac{\overline{X} - \mu'}{\sigma\sqrt{n}}$ has a standard Normal distribution if Ho is true. We may use

Z (or $\overline{X}$) as the test statistic.

Example 2.

Suppose the shelf life of 1 liter bottles of pasteurized milk is guaranteed to be at least 400 days, with standard deviation 60 days. If a sample of 25 bottles is randomly chosen from a production batch, and after testing the sample mean shelf life of 375 is calculated, should the batch be rejected as not meeting the guarantee?

Solution: Let μ be the batch mean.

Step 1. Null hypothesis *Ho*: $\eta = 400$.

Alternative hypothesis H1: $\eta < 400$ (one-sided: we are only interested in whether or not the mean is up to the guaranteed minimum value).

Steps 2 and 3, *n*=25(given); choose $\alpha = 0.05$.

Step 4. If $\overline{X}$ is the sample mean, the quantity

$$Z = \frac{\overline{X} - 400}{60\sqrt{25}}$$

is a standard normal variate (perhaps approximately) if *Ho* is true. We shall use *Z* as the test statistic.

Step 5. For a one-tailed test, standard normal tables give Z = -1.65 as the lowest value to

be allowed before *Ho* must be rejected, at the 5% significance level. The acceptance region is therefore [-*1.65*, *infinite*].

Step 6. Decision rule:

(a) Reject the production batch if the value of *z* calculated from the sample is less than -1.65.

(b) Accept the batch otherwise.

Step 7. Carry out the test:

From the sample data we find that

$$Z = \frac{375 - 400}{60\sqrt{25}} = -2.083 .$$

Decision: the production batch must be rejected, since -2.083<-1.65. It is highly unlikely that the mean shelf life of milk bottles in the batch will be 400 days or more. The chance that this decision is wrong is smaller than 5%.

Example 2.

A sample of 66 seeds of a certain plant variety were planted on a plot using a randomized block design. Before planting, 30 of the seeds were subjected to a certain heat treatment. The times from planting to germination were observed. The 30 treated seeds took 52 days to germinate, while the 36 untreated seeds took 47 days. If the common standard deviation to germination, for individual seeds, calculated from several thousand seeds, may be taken as 12 days, can it be said that the heat treatment significantly speeds up a seed's germination rate?

From the data given, it is clear that the heat treated seeds had an earlier start in growth. However, we may consider the wider question as to whether heat treated seeds are significantly faster germinating generally than untreated seeds.

The test is as follows:

Step 1. Let μ_A, μ_B be the germination period population means for heat treated and untreated seeds respectively.

Null hypothesis *Ho*: $\mu_A = \mu_B$ (i.e. $\mu_A - \mu_B = 0$)

Alternative hypothesis H_1: $\mu_A > \mu_B$.

We were asked specifically whether the heat treated seeds were faster germinating than the untreated seeds, so we use the one-sided alternative hypothesis.

Step 2 and 3, n_A=30 and n_B=36 (given).

We shall use $\alpha = 0.05$ as the significant level.

Step 4. Test statistic:

We are not given any information other than the two sample means. Even if we were told the individual students' results, we could not use the paired comparison test – there would be no possible reason for linking the results in pairs.

The difference in means ($\bar{x}_A - \bar{x}_B$) is approximately normally distributed, with mean ($\mu_A - \mu_B$) and standard deviation

$$\sigma' = \sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}$$

So we may use as test statistic the standard Normal variate

$$z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sigma'}.$$

And if Ho is true, $\mu_A - \mu_B = 0$; so the test statistic reduces to

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sigma'}.$$

Step 5. Acceptance region:

The critical value of z at 5% level of significance, for a one-tailed test, is 1.65. Therefore, the acceptance region for the null hypothesis is the set of values z less than or equal to 1.65.

Step 6. Decision rule:

(a) If the sample value of z>1.65, conclude that heat treated seeds germinate significantly earlier (at the 5% level) than untreated seeds.
(b) If $Z \leq 1.65$, the germination rates of both heat treated and untreated seeds may well be the same.

Step 7. Carry out the test.

The value of σ is given as 12.

Therefore $\sigma' = \sigma\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}) = 12\sqrt{\frac{1}{30}+\frac{1}{36}} \cong 2.96$

And so the sample value of the test statistic is

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sigma'} = \frac{52-47}{2.96} \cong 1.69.$$

Decision: The heat treated seed is just significantly earlier germinating at the 5% level than the untreated seed.

7.7. The T-Test

The uncertainty introduced into estimates based on samples can be accounted for by using a probability distribution which has a wider spread than the normal distribution. One such distribution is the t-distribution, which is similar to the normal distribution, but dependent on the size of sample taken. When the number of observations in the sample is infinite, the t-distribution and the normal distribution are identical. Tables of the t-distribution and other sample based distributions are used in exactly the same manner as tables of the cumulative standard normal distribution, except that two entries are necessary to find a probability in the table. The two entries are the desired level of significance (α) and the *degrees of freedom* (υ) defined as the number of observations in the sample minus the number of parameters estimated from the sample.

Then for the test statistic we use $t = \frac{\overline{X} - \mu_0}{S\sqrt{n}}$ which has Student-t distribution with *n*-1 degrees of freedom.

Examples using the Z-Test

Example 1

A farmer was found to be selling pumpkins that looked like ordinary pumpkins except that these were very large, the average diameter for ten samples being 30.0 cm. The mean and standard deviation of pumpkins is 14.2 cm and 4.7 cm, respectively. It is intended to test whether the pumpkins that the farmer is selling are ordinary pumpkins.

We hypothesize that the mean of the population from which the farmer's pumpkins was taken is the same as the mean of the ordinary pumpkins by the null hypothesis

$$H_0 : \mu_1 = \mu_0$$

We also must give an alternative hypothesis

$H_0 : \mu_1 \neq \mu_0$ stating that the mean of the population from which the sample was drawn does not equal the specified population mean. If the two parent populations are not the same, we must conclude that the pumpkins that the farmer was selling were not drawn from the ordinary pumpkin population, but from the population of some other genus. We need to specify levels of probability of correctness, or level of significance, denoted by α. Let us take a probability level of 5%; we are willing to risk rejecting the hypothesis when it is correct 5 times out of 100 trials. We must have the variance of the population against which we are checking. We may now set up a formal statistical test in the following manner:

1. The hypothesis and alternative:

$$H_0 : \mu_1 = \mu_0$$

$$H_0 : \mu_1 \neq \mu_0$$

2. The level of significance:

$$\alpha = 0.05$$

3. The test statistic:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

The test statistic, *Z*, has a frequency distribution which is a standardized normal distribution, provided the observations in the sample were selected randomly from a normal population whose variance is known. If we have specified that we are willing to reject the hypothesis of the equality of means when they actually are equal one time out of twenty: that is, we will accept a 5% risk of being wrong. On the standardized normal distribution curve, therefore, we wish to determine the extreme regions which contain 5% of the area of the curve. This part of the probability curve is called the

area of rejection or the critical region. If the computed value of the test statistic falls into this area, we will reject the null hypothesis. The hypothesis will be rejected if the test statistic is either too large or too small. T he critical region, therefore, occupies the extremities of the probability distribution and each sub region contains 2.5% of the total area of the curve.

Working through the grass example, the outline takes the following form:

1. $H_0 : \mu$ of grass = 14.2 mm

   $H_1 : \mu$ of grass $\neq$ 14.2 mm

2. α level = 0.05

3. $Z = \dfrac{30 - 14.2}{4.7/\sqrt{10}} = 10.6$

The computed test value of 10.6 exceeds 1.9, so we conclude that the means of the two populations are not equal, and the grass must represent some genus other than that of ordinary pumpkins.

7.8. Estimators Using Pooled Samples

Let two random samples of sizes n_1, n_2, respectively, be drawn from a large population which has mean μ and variance σ^2. Suppose that the samples yield unbiased estimates, $\overline{X_1}$ and $\overline{X}_2$ of μ and S_1^2, S_2^2 of σ^2. The problem arises of combining these pairs of estimates, to obtain single unbiased estimates of μ and σ^2. The process of combining estimates from two or more samples is known as pooling. The correct ways to pool unbiased estimates of means and variances, to yield single unbiased estimates, are

Means: $\hat{\mu} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

Variances: $\sigma^2 = \dfrac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}$

Example:

A soil scientist made six determinations of the strength of dilute sulfuric acid. His results showed a mean strength of 9.234 with standard deviation 0.12. Using acid from another bottle, he made eleven determinations, which showed mean strength 8.86 with standard deviation 0.21. Obtain 95% confidence limits for the difference in mean strengths of the acids in the two bottles. Could the bottles have been filled from the same source?

Working:

The difference in mean strengths of the acids is estimated by $\bar{x}_1 - \bar{x}_2 =$ 9.234-8.86=0.374.

We first estimate $\sigma(\bar{x}_1 - \bar{x}_2)$, *the standard deviation of the sampling distribution of* $\bar{x}_1 - \bar{x}_2$. *We pool the data from the two samples, thus:*

$$\hat{\sigma}_{\bar{x}1-\bar{x}2} = s.\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = s.\sqrt{\frac{1}{6}+\frac{1}{11}},$$

where $s^2 = \dfrac{(n_1-1)s_1^{\,2} + (n_2-1)s_2^{\,2}}{n_1+n_2-2}$

$$= \frac{5x0.0144+10x0.0441}{6+11-2} = \frac{0.72+0.441}{15} = \frac{1.161}{15} = 0.0774$$

and so $s = 0.2782$

Therefore

$$\hat{\sigma}_{\bar{x}1-\bar{x}2} = 0.2782\sqrt{\frac{17}{66}} = 0.141196$$

With 15 degrees of freedom, the confidence coefficient is t=2.13 for 95% confidence. Therefore the required limits for $\mu_1 - \mu_2$ *are*

$$(\bar{x}_1 - \bar{x}_2) \pm t.\hat{\sigma}_{X1-X2} = 0.374 \pm 2.13x0.141196 = 0.374 \pm 0.300748$$

Thus, the 95% confidence limits for the difference in mean strengths of the acids in the two bottles are 0.0733 and 0.6747. That means that we are 95% confident that the difference in mean strengths of the acids in the two bottles lies between 0.0733 and 0.6747.

7.9. The Paired Comparison Test and The Difference Between Two Means Test

Example: Paired comparison Test.

The yields from two varieties of wheat were compared. The wheat was planted on 25 test plots. Each plot was divided into two equal parts, one part was chosen randomly and planted with the first variety, and the other part was planted with the second variety of wheat. This process was repeated for all the 25 plots. When the crop yields were measured, the difference in yields from each plot was recorded (2^{nd} variety minus first variety). The sample mean plot yield difference was found to be 3.5 ton/ha, and the variance of these differences was calculated to be 16 ton/ha.

(a) Does the 2^{nd} variety produce significantly higher yields than the 1^{st} variety?
(b) Test the hypothesis that the population mean plot yield difference is as high as 5 tons/ha.
(c) Obtain 95% confidence limits for the population mean plot yield difference.

It is clear that there is a good deal of variation in yields from plot to plot. This variation tends to confound the main issue, which is to determine whether yields are increased by using a second variety. This has been eliminated by considering only the change in yields for each plot. If the second variety has no effect, the average change will be zero.

These kind of data, where results are combined in pairs, each pair arising from one experimental unit or having some clear reason for being linked in this way, are analyzed by the paired comparison test. Each pair provides a single comparison as a measure of the effect of the treatment applied (e.g., growing a different variety). Let *D* denote the difference in a given pair of results. *D* will have Normal distribution with mean μ and Standard Deviation σ (both the parameters are unknown in this case).

Example using the t-Test

Step 1.

Null hypothesis Ho: $\mu = 0$ (i.e. the yield of the two wheat varieties are the same).

Alternative hypothesis H_1: $\mu > 0$ i.e. Second variety yields are higher than first variety yields).

H_1 is one sided; we must apply a one-tailed test.

Steps 2 and 3. 25 plots were used, so n=25. We shall use α=0.05

Step 4. The quantity $z = \frac{D-0}{\sigma\sqrt{25}}$ is a standard normal variate, and may be used as the test statistic if σ is known from previous experimentation.

The parameters of a population are rarely known. In our case, σ is not given, so we must estimate it from the sample data.

Step 5. Acceptance region: the critical level of t at the 0.05 level of significance (one-tailed test) is the same as the upper 90% confidence coefficient. As given in Table. With 24 degrees of freedom, this value is 1.71. The acceptance region is, therefore, all values of t from –infinity to 1.171.

Step 6. Decision rule:

(a) If the value of t calculated from the sample is greater than 1.71, we may conclude that the second wheat variety gives higher yields than the first variety.

(b) If the value of t<1.71, we may not reject (at the 5% level) the hypothesis that the observed increases in yield in the second wheat variety were due to chance variation in the experiment.

Step 7. Carry out the test:

From the sample data $t = \frac{D-0}{S\sqrt{n}} = \frac{3.5-0}{6\sqrt{2}} = 2.375$

Decision: since 2.375>1.71, we conclude at the 5% level that the second variety significantly produces higher yields than the 1st variety.

7.10. The Difference Between Two Means

A sampling result, which is frequently used in inference tests, is one concerning the distribution of the difference in means of independent samples drawn from two different populations. Let a random sample of size n_1 be drawn from a population having mean μ_1 and Standard Deviation σ_X; and let an independent sample of size n_2 be drawn from another population having mean μ_Y and Standard Deviation σ_Y.

Consider the random variable $D = \overline{X} - \overline{Y}$; i.e., the difference in means of the two samples. The theorem states that

D has a sampling distribution with mean $\mu_D = \mu_X - \mu_Y$ and variance,

$$\text{Var } (D) = \frac{\sigma_X{}^2}{n_1} + \frac{\sigma_X{}^2}{n_1}$$

7.11. The *F*-Test

It seems reasonable that the sample variances will range more from trial to trial if the number of observations used in their calculation is small. Therefore, the shape of the *F*-distribution would be expected to change with changes in sample size. The degrees of freedom idea comes to mind, except in this situation the *F*-distribution is dependant on two values of γ, one associated with each variance in the ratio. Since the *F*- ratio is the ratio of two positive numbers, the *F*- distribution cannot be negative. If the samples are large, the average of the ratios should be close to 1.0.

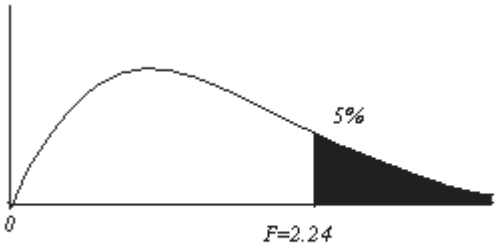
Because the *F*-distribution describes the probabilities of obtaining specified ratios of sample variances drawn from the same population, it can be used to test the equality of variances, which we obtain in statistical sampling.

We may hypothesize that two samples are drawn from populations having equal variances. After computing the *F*- ratio, we then can ascertain the probability of obtaining, by chance, that specific value from two samples from one normal population.

If it is unlikely that such a ratio could be obtained, we regard this as indicating that the samples come from different populations having different variances.

For any pair of variances, two ratios can be computed ($\frac{S_1}{S_2}$ and $\frac{S_2}{S_1}$).

If we arbitrarily decide that the larger variance will always be placed in the numerator, the ratio will always be greater than 1.0 and the statistical tests can be simplified. Only one-tailed tests need be utilized, and the alternative hypothesis actually is a statement that the absolute difference between the two sample variances is greater than expected if the population variances are equal. This is shown in figure XX, a typical *F*-distribution curve in which the critical region or area of rejection has been shaded.



A typical *F*-distribution $\gamma_1 = 10$ and $\gamma_2 = 25$ degrees of freedom, with critical region (shown by shading), which contains 5% of the area under the curve. Critical value of F=2.24.

As an example of an elementary application of the *F*-distribution, consider a comparison between the two sample sets of porosity measurements on soils of two areas of a certain district. We are interested in determining if the variation in porosity is the same in the two areas. For our purposes, we will be content with a level of significance of 5%. That is, we are willing to run the risk of concluding that the porosities are different when actually they are the same one time out of every twenty trials.

The variances of the two samples may be computed by (3.8), when the *F*-ratio between the two may be calculated by

$$F = \frac{S_1^{\,2}}{S_2^{\,2}}$$

where $S_1^{\,2}$ is the larger variance and $S_2^{\,2}$ is the smaller. We now are testing the hypothesis

$$H_0 : \sigma^2{}_1 = \sigma^2{}_2$$

against $H_1 : \sigma^2{}_1 \neq \sigma^2{}_2$

The null hypothesis states that the parent populations of the two samples have equal variances: the alternative hypothesis states that they do not. Degrees of freedom associated with this test are ($n_1 - 1$) for γ_1 and ($n_2 - 1$) for γ_2. The critical value of

F with $\gamma_1 = 9$ and $\gamma_1 = 9$ degree of freedom and a level of significance of 5% (α=0.05) can be found from Table X; that value is 3.18.

The value of F calculated from (3.26) will fall into one of the two areas shown on Fig 3.23. If the calculated value of F exceeds 3.18, the null hypothesis is rejected and we conclude that the variation in porosity is not the same in the two groups. If the calculated value is less than 3.18, we would have no evidence for concluding that the variances are different (determine at α 0.05 if variances are the same).

In most practical situations, we ordinarily have no knowledge of the parameters of the population except for estimates made from samples. In comparing two samples, it is appropriate to first determine if their variances are statistically equivalent. If they appear to be equal and the samples have been selected without bias from a naturally occurring population, you probably are safe in proceeding to additional statistical tests.

The next step in the procedure is to test equality of means. The appropriate test is (3.23)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{Sp\sqrt{(\frac{1}{n_1}) + (\frac{1}{n_2})}}$$

where the quantity Sp, is the pooled estimate of the population standard deviation, based on both samples. The estimate is found from the pooled estimated variance, given by

$$Sp^2 = \frac{(n_1 - 1)S^2{}_1 + (n_2 - 1)S^2{}_2}{n_1 + n_2 - 2}$$

where the subscripts refer, respectively, to the sample from area A and area B of the district.

7.12. RELATIONSHIP BETWEEN VARIABLES

7.12.1. Correlation methods

Correlation methods are used to discover objectively and quantitatively the relationship that may exist between several variables. The correlation coefficient determines the extent to which values of two variables are linearly related; that is, the correlation is high if it can be approximated by a straight line (sloped upwards or downwards). This line is called the regression line. Correlation analysis is especially valuable in agrometeorology, because of the many factors that may be involved,

simultaneously or successively, during the development of a crop and also because for many of them--climatic factors in particular--it is impossible to design accurate experiments, since their occurrence cannot be controlled. There are two sets of circumstances in which, more particularly, the correlation and simple regression method can be used:

(a) In completing climatological series having gaps. Comparisons of data for different atmospheric elements (e.g. precipitation, evapotranspiration, duration of sunshine) allow estimates of the missing data to be made from the other measured elements;

(b) In comparing climatological data and biological or agronomical data, e.g., yields, quality of crops (sugar content, weight of dry matter, etc.).

Care should be exercised in interpreting these correlations. Graphs and scatter plots should be used to give much more information about the nature of the relationship between variables. The discovery of a significant correlation coefficient should encourage the agrometeorologist, in most cases, to seek a physical or biological explanation for the relationship and not just be content with the statistical result.

Having discovered that there is a relationship between variables, one hopes to establish the closeness of this relationship. This closeness of agreement between two or more variables is called correlation. The closeness is expressed by a correlation coefficient whose value lies between +1 (perfect, positive correlation) and -1 (perfect, negative correlation). It is used to measure the linear relationship between two random variables that are represented by pairs of numerical values. The most commonly used formula is:

$$r = \frac{n\sum XiYi - (\sum Xi)(\sum Yi)}{\sqrt{[[n\sum Xi^2 - (\sum Xi)^2][n\sum Yi^2 - (\sum Yi)^2]]}}$$

If the number of pairs is small, the sample correlation coefficient between the two series is subject to large random errors, and in these cases numerically large coefficients may not be significant.

The statistical significance of the correlation may be determined by seeing whether the sample correlation r is significantly different from zero. The test statistic is

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

and t is compared to the tabulated value of Student's t with n-2 degrees of freedom.

7.12.2. Regression

After the strength of the relationship between two or more variables has been quantified, the next logical step is to find out how to predict specific values of one variable in terms of another. This is done by regression models. A single linear regression model is of the form:

$$Y = a + bX$$

where Y is the dependant variable;

X is the independent variable;

a is the intercept on the Y axis;

b is the slope of the regression.

The least squares criterion requires that the line be chosen to fit the data so that the sum of the squares of the vertical deviations separating the points from the line will be a minimum.

The recommended formulae for estimating the two sample coefficients for least squares are:

$$b = \frac{n\sum XiYi - (\sum Xi)(\sum Yi)}{n\sum Xi^2 - (\sum Xi)^2}$$ the slope of the line

$$a = \frac{(\sum Yi)(\sum Xi^2) - (\sum Xi)(\sum XiYi)}{n\sum Xi^2 - (\sum Xi)^2}$$ the y-axis intercept

Example

Compute a and b coefficients of the Angstrom formula

Angstrom's formula:

$$R/RA = a + b\, n/N$$

is used to estimate the global radiation at surface level (R) from the radiation at the upper limit of the atmosphere (RA), the actual hours of bright sunshine (n), and the day length (N). RA and N are taken from appropriate tables or computed; n is an observational value obtained from the Campbell–stokes sunshine recorder.

The data of the example are normals from Lyamungu, Tanzania

(latitude 3˚ 14S, longitude 27˚ 17E, elevation 1250m)

Table II. 2.1

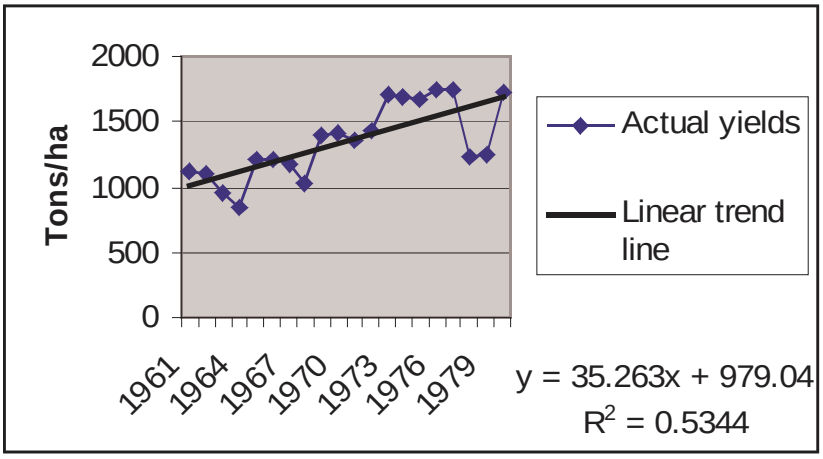
	n/N (X)	R/RA (Y)
J	.660	.620
F	.647	.578
M	.536	.504
A	.366	.395
M	.251	.368
J	.319	.399
J	.310	.395
A	.409	.442
S	.448	.515
O	.542	.537
N	.514	.503
D	. 602	.582

N=12, $\overline{X}$ = •467, σx = •132, $\overline{Y}$ = .487, σy = .081, b = .603, a = .205, r = .973

The regression explains r^2 = 95% of the variance of R/RA, and is significantly below p =0.01.

There are cases where a scatter diagram suggests that the relationship between variables is not linear. This can be turned into a linear regression by taking the logarithms of the relationship if it is exponential or turning it into a reciprocal, if it is square, etc. For example, when the saturation vapour pressure is plotted against temperature, the curve suggests that a function like $y = p.e^{bX}$ could probably be used to describe the function. This is turned into a linear regression *ln* (y)=*ln*(p)+bX, where X is temperature function and y is the saturation vapour pressure. An expression of the form $y = aX^2$ can be turned into a linear form by taking the

$$\text{reciprocal}\frac{1}{y} = \frac{X^{-2}}{a}.$$



7.12.3. Multiple regressions

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent variable. A linear combination of predictor factors is used to predict the outcomes or response factor. For example, multiple regression has been successfully used to estimate crop yield as a function of weather, or to estimate soil temperatures as a function of air temperature, soil characteristic, and soil cover. It has been used to perform a trend analysis of agrometeorological parameter using a polynomial expansion of time. The general linear model is a generalization of the linear regression model, such that effects can be tested for categorical predictor variables, as well as for effects for continuous predictor variables. An objective in performing multiple regression analysis is to specify a parsimonious model whose factors contribute significantly to variation in response. Statistical software such as INSTAT, provide tools to select independent factors for a regression model. They include forward stepwise regression to individually add or delete the independent variables from the model at each step of the regression until the "best" regression model is obtained, or backward stepwise regression to remove the independent variables from the regression equation one at a time until the "best" regression model is obtained. It is generally recommended that one should have at least ten times as many observation or cases as one has variables in a regression model.

Residual analysis is recommended as a tool to assess the multiple regression models and to identify violations of assumptions that threaten the validity of results. Residuals are the deviations of the observed values on the dependent variable from the predicted values. Most statistical software provides extensive residuals analyses, allowing one to use a variety of diagnostic tools in inspecting different residual and predicted values, and thus, to examine the adequacy of the prediction model, the need for transformations of the variables in the model, and the existence of outliers in the data. Outliers (i.e., extreme cases) can seriously bias the results.

7.12.4. Stepwise Regression

This will be explained by using an example for yields. A combination of variables may work together to produce the final yield. These variables could be the annual precipitation, the temperature of a certain month, the precipitation of a certain month, the potential evapotranspiration of a certain month, or the difference between precipitation and potential evapotranspiration for a given month.

In stepwise regression, a simple linear regression for the yield is constructed on each of the variables and their coefficients of determination found. The variable that produces the largest r2 statistic is selected. Additional variables are then brought in one by one and subjected to a multivariate regression with the best variable to see how much that variable would contribute to the model if it were to be included. This is done by calculating the F statistic for each variable. The variable with the largest F statistic, that has a significance probability greater than the specified significance level for entry, is included in the multivariate regression model. Other variables are included in the model one by one. If the partial F statistic of a variable is not significant at a specified level for staying in the regression model, it is left out. Only those variables that have produced significant F statistics are included in the regression. More deeply explanations could be found in Draper and Smith (1981).

7.12.5. Cluster Analysis

Cluster analysis is a technique for grouping individuals or objects into unknown groups. In biology, cluster analysis has been used for taxonomy, to classify living things into arbitrary groups on the basis of their characteristics. In agrometeorology, cluster analysis can be used to analyze historical records of the spatial and temporal variations in pest populations in order to classify regions based on population densities and frequency and persistence of outbreaks. The analysis can be used to improve regional monitoring and control of pest's populations.

Clustering techniques require defining a measure of closeness or similarity of two observations. Clustering algorithms may be hierarchical or nonhierarchical. Hierarchical methods can be either agglomerative or divisive. Agglomerative methods begin by assuming each observation is a cluster and then, through successive steps, combine the closest clusters. Divisive methods begin with one cluster containing all the observations and successively split off cases that are the most dissimilar to the remaining ones. K-mean clustering is a popular nonhierarchical clustering technique. It begins with user-specified clusters and then reassigns data based on the distance from the centroid of each cluster. See von Storch and Zwiers (2001) for more detailed explanations.

7.12.6. Classification trees

The goal of classification trees is to predict or explain responses on a categorical dependent variable. They have much in common with discriminate analysis, cluster analysis, nonparametric statistics, and nonlinear estimation. They are one of the main techniques used in data mining. The ability of classification trees to perform univariate splits, examining the effects of predictors one at a time contributes to their

flexibility. Classification trees can be computed for categorical predictors, continuous predictors, or any mix of the two types of predictors when univariate splits are used. They readily lend themselves to graphical display, which makes them easier to interpret. Classification trees are used in medicine for diagnosis and biology for classification. They have been used to predict levels of winter survival of over wintering crops using weather, and categorical variables related to topography and crop cultivars.

7.13 CLIMATIC PERIODICITIES AND TIME SERIES

Data are commonly collected as time series that is observations made on the same variable at repeated points in time. INSTAT provides facilities for descriptive analysis and display of such data. The goals of time series analysis include identifying the nature of the phenomenon represented by the sequence of observations and predicting future values of the times series. Moving averages are frequently used to 'smooth' a time series so that trends and other patterns are seen more easily. Sivakumar, et al., (1993) present a number of graphs showing the five-year moving averages of monthly and annual rainfall at selected sites in Niger. Most time series can be described in terms of trend and seasonality. When trends, seasonal or other deterministic patterns, have been identified and removed from a series, the interest focuses on the random component. Standard techniques can be used to look at its distribution. The feature of special interest, resulting from the time series nature of the data, is the extent to which consecutive observations are related. A useful summary is provided by the sample autocorrelations at various lags, the autocorrelation at lag m being the correlation between observations m time units apart. In simple applications this is probably most useful for determining whether the assumption of independence of successive observations used in many elementary analyses is valid. The autocorrelations also give an indication of whether more advanced modeling methods are likely to be helpful. The cross correlation function provides a summary of the relationship between two series from which all trend and seasonal patterns have been removed. The lag m cross correlation is defined as the correlation between x and y lagged by m units.

More than any other user of climatic data, the agrometeorologist may be tempted to search for climatic periodicities, which would provide a basis for the management of agricultural production. It should be noted that the *Guide to Climatological Practices* (section 5.3) is more than cautious with regard to such periodicities and that, although they may be of theoretical interest, they have been found to be unreliable, having amplitudes which are too small for any practical conclusions to be drawn.

8. PUBLICATION OF RESULTS

8.1. General methods

For statistical analyses to have practical value they must be distributed to users in a readily understandable format which does not require an advanced knowledge of statistics. Adequate details should be given in each publication to avoid any ambiguity in interpreting the numerical tables or graphs.

8.2 Tables

Numerical tables of frequencies, averages, distribution parameters, return periods of events, etc., should state clearly:

(a) The geographical location (including elevation of the observation site);
(b) The period on which the statistical analysis is based (necessary to estimate how representative the data are);
(c) The number of data (enabling the continuity of the series to be assessed);
(d) The units;
(e) The meaning of any symbols.

For frequency tables, it is better to give relative (percentage) frequencies, to facilitate the comparison of populations consisting of different numbers of observations. In this case, it must be made quite clear whether the percentages refer to the total population or to separate classes.

8.3. Contingency tables

Estimates of the simultaneous occurrence of given values of several elements or events are often needed. The resulting contingency tables should be as simple as possible.

8.4 Graphs

Graphs are used to show, in a concise format, the information contained in numerical tables. They are a useful adjunct to the tables themselves and facilitate the comparison of results. Cumulative frequency curves, histograms, and climograms give a better overall picture than the multiplicity of numerical data obtained by statistical analysis. The scales used on the graph must be specified and their graduations should be shown. Pub1ications intended for wide distribution among agricultural users should not have complicated scales (e.g. logarithmic, Gaussian, etc.) with which the users may be unfamiliar, and which might lead to serious errors in interpreting the data. Furthermore, giving too much information on the same graph and using complicated conventional symbols should be avoided.

8.5 Maps

To present concisely the results of agroclimatological analysis covering an area or region, it is often better to draw isopleths or color classification from the data plotted at specific points. The interpolation between the various locations can be used in a digital map plotted by special plotting tools such as Graph, Grids, Surfer, and GIS. Many climatic parameters useful to agriculture can be shown in this way, for example:

(a) Mean values of climatic elements (temperature, precipitation, evapotranspiration, water balance, radiation balance, etc.);
(b) Frequencies: number of consecutive days without frost, without thawing, without rain, etc.; return periods of atmospheric events;
(c) Dispersion parameters: standard deviations, coefficients of variation;
(d) Agrometeorological indices.

Depending on the scale adopted, this type of supplementary chart can be drawn

more or less taking geomorphological factors into account. However, the users of the charts should be made aware of their generalized nature and, to interpret them usefully, should know that corrections for local conditions must be made. This is particularly important for hilly regions.

8.6. The Agrometeorological Bulletin

Because of the diverse nature of the Users, the content of an Agrometeorological Bulletin (Agmet Bulletin) cannot be standardized. But the basic objectives of all successful Agmet Bulletins are the same: the provision of the right Agmet Information (Agmet Info) to the right Users at the right time. To attain this objective, the following guidelines are suggested. For a complete discussion on the matter, the readers are referred to WMO (2002).

First, it is essential to determine who the Users are. One category of Users may be farmers who need daily information to assist them in day-to-day activities such as sowing, spraying, and irrigating. Another category may be more interested in long-term agricultural decisions such as crop adaptation to weather patterns, or marketing decisions, or modelling.

Second, the Users' requirements must be clearly established, so that the most appropriate information is provided. This is possible only after discussing with them. In most cases, they do not have a clear picture of the type of information which is best suited for their purpose; here, the role of the Agrometeorologist is crucial.

Third, the methods of dissemination of information must be decided upon after consultation with the Users. Some farmers may have full access to Internet, while others have only limited access, and others have no such access to this technology. Obviously, the presentation of data for these categories will not be the same. Furthermore, some information must be provided as quickly as possible, while others may be provided two or three weeks later.

Fourth, it is very important to consider the cost of the Agmet Bulletin that is proposed to the Users, especially in developing countries where the financial burden is getting worst.

8.6.1 Some Examples

Some examples of Agmet Info presentation are given below, to drive home the points made above.

Data in Pentads

Table 1 shows part of an Agmet Bulletin issued by a Government Service in a tropical country where agriculture is an important component of the economy.

AgMet Bulletin in Pentads			
Rainfall Data		**Maximum Temperature**	
Dates	Rainfall amounts in millimetres	Dates	Maximum Temperature (deg. C)

November 2003	Observed values	Normal values	July 2003	Observed values	Normal values
1-5	3.6	4.7	1-5	23.6	23.7
6-10	7.1	4.7	6-10	21.9	23.7
11-15	13.6	4.7	11-15	22.1	23.7
Total RR	24.3	14.2	Mean MxT	22.5	23.7

Table 1: Rainfall amounts (RR) and maximum temperature (MxT) are shown for a given area of a tropical Country. Total rainfall amounts and the mean maximum temperature observed during the three pentads, are compared to their respective normal values.

The above Agmet Bulletin (Table 1) was developed to cater for all crops, ranging from tomatoes to sugarcane. It is issued on a half-monthly basis and is sent to the Users by post and is also available on the website. Bearing in mind the time taken to collect the data, it would not be before, at least, the 20th that the Bulletin would reach the Users. To provide farmers (tomato growers, for example) with data relevant to their day-to-day activities, the Agmet Bulletin is supplemented by daily values of rainfall and maximum and minimum temperatures, which are broadcast on radio and television. Of course, data relevant to different geographical localities can be included.

In AgMet Bulletins, extreme weather events, which are masked by the averaging procedure involved in the calculation of the pentad, must be highlighted, probably in the form of a footnote, to draw the attention of the Users. For example, from Table 1, it can be seen that during the period 6-15 July, the maximum temperature was below the normal by not more than 1.8°C. In fact, during the period 9-12 July, maximum temperature was below the normal by 2.8 to 3.0 degrees Celsius; this can be of importance to both animals and plants.

The presentation of data in this format, together with the broadcast of daily values on the radio and on television, is very effective. It can be used by farmers interested in day-to-day activities and by research workers and model builders. It is suitable for all types of crops, ranging from tomatoes and lettuce to sugar cane and other deep root crops.

Data in 10-day intervals

Based on agrometeorological requirements of a Mediterranean climate with two main seasons and two transitional seasons, the main climatic parameters should be published all the year around. The selection of agromet parameters/indices should be published according to the season and the agricultural situation of the crops, including data representing the various agricultural regions of the country.

The bulletin should include daily data, 10-daily means or totals, and deviation or percent from average. In parameters, such as maximum and minimum temperature and maximum and minimum relative humidity, absolute values of the decade based on a long series of years are also recommended.

The list of the recommended data to be published in the agrometeorological bulletin is as follows:

a) Daily data of maximum and minimum temperature and relative humidity;
b) Temperature near the ground;
c) Soil temperature;
d) Radiation and or Sunshine duration in hours;
e) Class A pan evaporation and/or Penman evapotranspiration;
f) Rainfall amount;
g) Accumulated rainfall from the beginning of the rainy season;
h) Number of rainy days;
i) Accumulated number of dry days since the last rainy day;
j) Number of hours below different temperature thresholds depending on the crop; and,
k) Number of hours below 0°C.

Examples of agrometeorological parameters or indices that should be published are:

a) Accumulated number of dynamic model units since the beginning of the winter is an indication of budbreak in deciduous trees;
b) Accumulated number of units above 13°C since the beginning of spring is an indication of citrus growth; and,
c) Physiological days – the accumulated number of units above 12°C since the beginning of spring is an indication of cotton growth.

Short Range Weather Outlook

With the availability of short range (5 to 10 days) weather forecast on the Internet, provided by World Weather Centres (WWC), many Agrometeorological Services are providing 5 to 10 days weather forecast to farmers. An example is given below.

December 2003	West	North
Friday , 12 th	< 1.0	< 1.0
Saturday, 13th	1.1 - 5.0	< 1.0
Sunday, 14th	5.1 - 25.0	5.1 - 25.0
Monday, 15th	1.1 - 5.0	1.1 - 5.0
Tuesday, 16th	< 1.0	< 1.0

Table 2: Expected rainfall (millimetres) for two rainfed farming areas. The information was released to the Users through e-mail and posted on the website.

This Weather Outlook, based on model output received early on Thursday the 11th from WWC, was released in the afternoon on the 11th December and sent to the Users, through the Farming Centres by e-mail and posted on the website. This Outlook was neither broadcast on the radio nor on the television. The issue of such Weather Outlook is important, but it must be carefully planned, otherwise, it can lead to financial loses, as shown below.

Little rainfall was observed during the first two pentads of December 2003 and

farmers were starting to get worried. The indication that significant rain was expected on Sunday the 14th (Table 2) had given great hope to the farmers and, because it was a weekend, they made plans on Friday to do some field work on Saturday and on Monday. Such plans are costly because it implies the booking of manpower and of transportation, the buying of fertilizers, etc. But model output received on Friday the 12th indicated that the probability of having rain during the following five days was negligible, and in fact, it was not before the 31st of December that significant rainfall was observed.

Here, it is not the validity of the Weather Outlook that is questioned. The point to be noted is that no update of the Outlook could reach the farmers because the Farming Centres were closed for the weekend. If, besides being sent by email and posted on the website, the Outlook was broadcast on radio and television, the update version would have reached the farmers and appropriate measures could have been taken. To avoid similar incidents, it is advisable to decide on the methods of dissemination of information.

Seasonal Forecast

An extract of a seasonal forecast issued, for a country situated in the southern hemisphere, during the first half of October 2003 for summer 2003-2004 (Summer in that country is from November to April) is shown: **"The rainfall season may begin by November. The summer cumulative rainfall amount is expected to reach the long-term mean of 1400 millimeters. Heavy rainfall is expected in January and February 2004."** This seasonal forecast was published in the newspapers and read on the television.

The question is: who are qualified to interpret and use this forecast? Can it be misleading to farmers? To show the problems which such forecast can create, real data for the period October 2003 to January 2004 are presented in Table 3 for an agricultural area.

Rainfall Amounts in millimeters (mm)				
	October 2003	November 2003	December 2003	January 2004
First Half	1.8	4.1	5.2	176.4
Second Half	12.8	35.7	12.8	154.1

Table 3: Rainfall amounts recorded over an agricultural area during the period October 2003 to January 2004. Out of the 35.7 mm of rainfall recorded during the second half of November, 35.0 mm fell during the period 16-25

Given that October and the first half of November 2003 were relatively dry and that a significant amount of rainfall was recorded during the second half of November, and noting that the seasonal forecast opted for normal rainfall during summer and that the rainfall season may start in November, the farmers thought that the rainy season was on. Most of them started planting their crops during the last pentad of November. Unfortunately, the rainfall during the second half of November was a false signal: December was relatively dry. The rainy season started in January 2004.

To avoid seasonal forecasts to fall in the wrong hands, it is not advisable to have them published in the newspapers; these seasonal forecasts must be sent to specialists who are trained to interpret them and should be supplemented by short-range weather forecast.

8.6.2. How Costly Should the AgMet Bulletin Be?

Today, agricultural systems in some Small and Poor Countries (SPC) are in great turmoil. Developed countries have dismantled the safety net, which has been giving some protection to the agricultural products of these SPC. It is in relation to the bleak future of agriculture in these SPC that the question of the cost of the issue of AgMet Bulletin is raised.

Sooner or later, the financial situation in the SPC will not be able to sustain the issue of costly AgMet Bulletin by local personnel. So in these SPC, the agrometeorologists must think carefully about the cost-benefit of the AgMet Bulletin, especially when developed countries are getting ready to propose their services for free (for how long will these be free?).

Already, shipping bulletins, cyclone warnings, and aviation forecasts are being offered for free on a global scale by a few developed countries. But, how long will these services be free? Sooner or later, the SPC will have to pay for these services. It is very important to keep the cost of the AgMet Bulletin to a minimum.

References

A.D. HARTKAMPA, J.W. WHITEA, G. HOOGENBOOMB, 2003. *Comparison of three weather generators for crop modeling: a case study for subtropical environments*, Agricultural Systems, **76**, 539–560.

BESSEMOULIN, G., 1973: *Sur la statistique es valeurs extremes*. Monographie de la Meteorologie Nationale No. 89, Paris, 24 pp.

CARRUTHERS, N. and BROOKS, C.E.P., 1953: *Handbook of statistical methods in meteorology*. Her Majesty's Stationery Office Mo. 538, 412 pp.

COLES, S., 2001: *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics.

DRAPER, N. R & SMITH, H., 1981. *Applied Regression Analysis*. Second Edition, New York: John Wiley & Sons.

GUMBEL, E. J., 1959: *Statistics of extremes*. Columbia University Press, New York, 375 pp.

HUBBARD, K.G. and SIVAKUMAR, M.V.K. (Eds.) 2001. *Automated Weather Stations for Applications in Agriculture and Water Resources Management: Current Use and Future Perspectives*. Proceedings of an International Workshop held in Lincoln, Nebraska, USA, 6-10 March 2000.

MOTHA, R.P. (Ed.), 2000. *Agrometeorological Data Management*. WMO / TD No. 1015, Geneva, Switzerland.

MÜLLER-WESTERMEIER, G., 2004. *Statistical analysis of results of homogeneity testing and homogenisation of long climatological time series in Germany.* Proc. 4th Seminar for Homogenization and Quality control in Climatological databases (Budapest, October 2003), WMO-TD No.1236 (WCDMP-No.56), 25-38.

SIVAKUMAR, M.V.K., MAIDOUKIA, A. and STERN, R.D. 1993. *Agroclimatology of West Africa: Niger.* Information Bulletin no. 5, ICRISAT, India.

SIVAKUMAR, M.V.K., STIGTER, C.J., RIJKS, D. (Eds.) 2000. *Agrometeorology in the 21st Century--Needs and Perspectives*, Agricultural and Forest Meteorology 103, 227 pp.

SIVAKUMAR, M.V.K. U.S. DE, K.C. SIMHARAY, and M. RAJEEVAN (Eds.) 1998. *User Requirements for Agrometeorological Services.* Proceedings of an International Workshop held at Pune, India, 10-14 November 1997.

THOM, H.C.S. 1966. *Some methods of climatological analysis*. WMO Technical Note

No. 81, Geneva, Switzerland.

Von STORCH H and F ZWIERS, 1999. *Statistical Analysis in Climate Research,* Cambridge University Press.

WIERINGA, J., and LOMAS, J. 2001. *Lecture Notes for Training Agricultural Meteorological Personnel.* WMO-No. 551, Geneva, Switzerland.

WIJNGAARD, J.B., KLEIN TANK A.M.G., and KONNEN G.P. 2003. *Homogeneity of 20th century European daily temperature and precipitation series,* International Journal of Climatology, **23**, 679-692.

WMO, 1983. *Guide to Climatological Practices*. WMO-No. 100, Geneva.

WMO, 2002. *Improving Agrometeorological Bulletins*. WMO-TD No. 1108, Geneva.

WMO, 2003a. *Guidelines on climate observation networks and systems* (by N. Plummer, T. Allsopp and J.A. Lopez). WMO-TD No.1185 (WCDMP-No. 52), Geneva.

WMO, 2003 b. *Guidelines on climate metadata and homogenization* (by E. Aguilar, I. Auer, M. Brunet, T.C. Peterson and J. Wieringa). WMO-TD No.1186 (WCDMP-No. 53), Geneva.

WMO, 2004. *Fourth Seminar for homogenization and quality control in climatological databases*. WMO-TD No. 1236, Geneva.