HidroEsta, software for hydrological calculations

Máximo Villón-Béjar[1]

1   Telephone: (506) 550-2595. Agricultural Engineer, specialist in Soil and Water Resources Engineering, and Informatics. Professor, School of Agricultural Engineering. Instituto Tecnológico de Costa Rica. Cartago, Costa Rica. E-mail: mvillon@itcr.ac.cr

Keywords

Precipitation; flows; random variables; distributions; statistical parameters; mean; variance; hydrology.

Abstract

This investigation, "HidroEsta, software for Hydrological Calculations" is oriented towards the creation of a computing tool using Visual Basic intended to facilitate and simplify difficult calculations that must be carried out in hydrological studies.

The software allows making calculations of statistical parameters, as well as linear and non-linear, simple and multiple regression calculations, and polynomial regressions. It is also useful for evaluating if a series of data can be fitted to a series of distributions, and calculating design events with a given probability of occurrence, based on a seasonal variation curve or a duration curve; analyzing a storm and calculating maximum intensities based on pluviograms data; calculating flow measurements taken with windmills or current meters; calculating maximum flows with empirical and statistical methods; calculating evapotranspiration and hydric balance.

Several numeric methods were used to solve the equations, selecting the most appropriate one for each case. The results of this work provide civil, agricultural engineers, and agronomists, hydrologists and other specialists who work in this field with a tool that enableZs them to carry out calculations, rapid simulations, and determine flow or precipitation for designs.

I. Introduction

Hydrological studies require analyses of large amounts of hydrometeorological information related to precipitation, flows, temperature, and evaporation, among others.

The data collected represent raw information, but if properly organized and analyzed, they provide hydrologists with a very useful tool which allows them to make decisions in the design of hydraulic structures [5].

To make the calculations, hydrologists have to confront a series of problems:

- Processing the information is highly laborious.
- In most cases, the equations to be solved are very complex and their solution requires using numerical methods.
- Simulations carried out manually are time-consuming, due to the calculations required.

Due to the complexity of information processing and of the calculations, errors can occur, making it useful to have software that provide hydrologists with a tool that enables them to simplify all these processes, and even to simulate their results, allowing them to optimize their design.

Objective

Create software which will facilitate and simplify complex calculations, and the analysis of the abundant information required in hydrological studies.

II. Theoretical basis

Random variables

A random variable [1], is a function *X*, defined on a sample space *S*, which assigns a value to the variable for each point (or each result) of the sample space of an experiment.

Random variables are also known as stochastic variables whose values are real numbers which cannot be predicted with certainty before the phenomenon occurs; i.e., they occur randomly.

Discreet random variable

A random variable *X* is discrete when its values are restricted to a finite or infinite enumerable set.

Example: number of rainy days that occurred during the months of any year.

Continuous random variable

A random variable *X* is continuous when its values are within a continuous range, and may be represented by any whole or decimal number.

Example: the daily flow registered in a measuring station.

Distributions

The behavior of a random variable is described through its probability law, which may in turn be described in several ways. The most common is through the distribution of probabilities of the random variable [4].

Notation:

- *X* $\Rightarrow$ random variable of the function
- *x* $\Rightarrow$ specific value assumed by the random variable
- f(x) $\Rightarrow$ probability density function (PDF) (probability function, probability distribution of *x*)
- F(x) $\Rightarrow$ cumulative function (cumulative distribution function, CDF)

Distribution measurements

Arithmetic mean

Given the composite sample of *n* data x_1, x_2, ..., x_n, the mean is defined as their algebraic sum, divided by the number of data. When the average is calculated for a population, it is denoted by μ, and when defined for a sample, it is denoted by $\overline{x}$.

Mathematically speaking, the average of data that are not grouped is represented by:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where:

m = population mean

$\overline{x}$ = sample mean

x_i = i^{th} value of the mean

n = number of data of the sample or population

The median

The median is a single value of a set of data which measures the central element in the data [5]. This single element of the data is the closest to the middle, or the most central value of a set of numbers. Half of the elements are above this point, and the other half are below it.

If $x_1, x_2, x_3, ..., x_n$ are data organized in increasing or decreasing order, and n is an odd number, the median (Med) is the datum at the center, i.e.:

$$\text{Med} = X_{(n+1)/2}, \quad \text{for an odd number } n$$

if n is an even number, the median is the average of the central numbers, i.e.:

$$Med = \frac{X_{n/2} + X_{n/2+1}}{2}, \quad \text{for an even number } n$$

The mode

Is the value that occurs most frequently in a set of data, denoted by Mo.

For data grouped in class intervals, once the modal class is defined, the mode is calculated using the following equation:

$$Mo = Lm + \frac{d1}{d1 + d2}w$$

where:

Mo = mode

Lm = lower limit of the modal class

d_1 = difference between the frequency of the modal and pre-modal class (the class before the modal class)

d_2 = difference between the frequency of the modal and post-modal class (the class after the modal class)

w = range of the class interval

In general, the modal class is that which has the maximum frequency.

Variance

The variance of the population (σ^2), is defined as the sum of squares of the deviations of the data with respect to the mean, divided by the total number of data, i.e.,:

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

The variance of the sample (S^2) [5], is obtained by dividing the sum of squares of observations of the data with respect to the mean by the total number of observations minus one, i.e.:

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

Standard deviation

The standard deviation is defined as the positive square root of the variance, i.e.:

$$\sigma = \sqrt{\sigma^2} \text{ (for the population)}$$

$$S = \sqrt{S^2} \text{ (for the sample)}$$

Coefficient of variation

This is a relative measurement of dispersion, which is related to the standard deviation and the mean, i.e.:

$$C_V = \frac{S}{\bar{x}}$$

Bias

The bias is the statistics that measures symmetry and asymmetry [2].

The bias (γ) for data from the population is obtained with the following equation:

$$\gamma = \frac{3}{\sigma^3}$$

where:

$$3 = \frac{\sum_{i=1}^{n}(x_i - \ )^3}{n}$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}x_i$$

The bias of data in the sample is obtained with:

$$Cs = \frac{n^2 \, M_3}{(n-1)(n-2) \, S^3}$$

where:

$$M3 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n}$$

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Kurtosis

For data of the population the kurtosis coefficient (k), is defined by the following equation:

$$k = \frac{4}{\sigma^4}$$

where:

$$4 = \frac{\sum_{i=1}^{n} (x_i -)^4}{n}$$

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i$$

The kurtosis coefficient for data in the sample [2], is defined as:

$$C_k = \frac{n^3 \, M_4}{(n-1)(n-2)(n-3) \, S^4}$$

where:

$$M_4 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n}$$

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Estimation of parameters

The parameters of a theoretical distribution are variables that have a defined value for each set of data. Once the parameters are defined, the theoretical distribution is also defined.

In general, a density function or a cumulative distribution function may be expressed as a function of the random variable, and in general as a function of its parameters. For instance, the density function of the normal distribution of random variable *X*, is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x - }{\sigma}\right)^2}$$

where:

μ = location parameter

σ² = scale parameter

To define the function *f(x)*, the parameters μ and σ² must be calculated·

Since all the values of the random variable are usually unknown, the estimation of the parameters is made based on a sample [4].

Methods for the estimation of parameters

To determine the numeric values of the theoretical distribution parameters based on data in the sample, several estimation methods are used [5]; in order of ascending efficiency, they are:

- Graphic
- Method of least squares
- Method of moments (MOM) estimation
- Maximum likelihood

Goodness-of-fit tests

The goodness-of-fit tests consist in verifying, graphically and statistically, if the empirical frequency of the series analyzed fits a previously selected theoretical probability function with parameters estimated based on values of the sample.

The purpose of statistical tests [4], is to measure the certainty obtained when a statistical hypothesis is made for a population, i.e., to evaluate the assumption that a random variable will be distributed according to a certain probability function.

The most used goodness-of-fit tests are:

- Graphical fit
  Chi - square
- Statistical fit
  Smirnov - Kolmogorov

Theoretical distributions

Hydrologists usually have a record of hydro-meteorological data available (precipitation, flows, evapotranspiration, temperatures, etc.). Based on their knowledge of the physical problem, they will choose a probabilistic model to use which will satisfactorily represent the behavior of a variable.

Normal or Gaussian distribution

A random variable X, is said to have a normal distribution [2], if its density function is:

$$f(x) = \frac{1}{\sqrt{2\Pi S}} EXP \left[-\frac{1}{2} \left(\frac{x - \overline{X}}{S} \right)^2 \right]$$

or

$$f(x) = \frac{1}{\sqrt{2\Pi S}} e^{-\frac{1}{2} \left(\frac{x - \overline{X}}{S} \right)^2}$$

   for $-\infty < x < \infty$

where:

   $f(x)$ = normal density function of variable x

   x = independent variable

   $\overline{X}$ = location parameter, equal to the arithmetic mean of x

   S = scale parameter, equal to the standard deviation of x

   EXP = exponential function with base e (napierian logarithms).

When the random variable X, is normally distributed with mean $\sigma = \overline{X}$ and variance ($\sigma^2 = S^2$), it is denoted as:

   $X \sim N(\overline{X}, S^2)$

Parameter log-normal distribution

The random variable X, is positive and the lower limit x_O does not appear.

The random variable: $Y = lnX$, is normally distributed with mean γy and variance $\sigma^2 y$.

These parameters are used to specify that the distribution is logarithmic, given that the mean and variance of X may also be used.

It is said that a random variable X, has a 2-parameter log-normal distribution [5], if its density function is:

$$f(x) = \frac{1}{x\sqrt{2\Pi}\sigma_y} e^{-\frac{1}{2}\left(\frac{\ln x - \mu_y}{\sigma_y}\right)^2}$$

for $0 < x < \infty$

$$x \sim \log N\left(\mu_y, \sigma_y^2\right)$$

Where μ_y, σ_y, are the media and the standard deviation of the natural logarithms of x, i.e., of lnx, and represent respectively the scale and the shape parameters of the distribution.

Parameter gamma distribution

It is said that a random variable *X*, has a 2-parameter gamma distribution [5], if its probability density function is:

$$f(x) = \frac{x^{\gamma-1} e^{-\frac{x}{\beta}}}{\beta^\gamma \Gamma(\gamma)}$$

for:

$0 \leq x < \infty$

$0 < \gamma < \infty$

$0 < \beta < \infty$

Where:

γ = shape parameter (+)

β = scale parameter (+)

$\Gamma(\gamma)$ = complete gamma function, defined as:

$\Gamma(\gamma) = \int_0^\infty x^{\gamma-1} e^{-x} dx$ which converges if $\gamma > 0$

Log-Pearson type III distribution

It is said that a random variable *X*, has a log-Pearson type II [5] distribution if its probability density function is:

$$f(x) = \frac{(\ln x - x_0)^{\gamma-1} e^{-\frac{\ln x - x_0}{\beta}}}{x\beta^\gamma \Gamma(\gamma)}$$

for:

$x_0 \leq x < \infty$

$-\infty < x_0 < \infty$

$0 < \beta < \infty$

$0 < \gamma < \infty$

where:

x_0 = position parameter

β = scale parameter

γ = shape parameter

Gumbel distribution

The Gumbel distribution [5], is one of the extreme value distributions, also known as the Extreme Value Type I, Type I Fisher-Tippett or double exponential distribution.

The cumulative distribution function of the Gumbel distribution, has the form:

$F(x) = EXP(-EXP(-(x - \mu)/\alpha))$

or

$$F(x) = e^{-e^{-\frac{(x-\mu)}{\alpha}}}$$

for: $-\infty < x < \infty$

where:

$0 < \alpha < \infty$, is the scale parameter

$-\infty < \mu < \infty$, is the position parameter, also called the central value or mode

Deriving the cumulative distribution function with respect to x, the density function of probability is obtained, i.e.:

$$f(x) = \frac{dF(x)}{dx}$$

$$f(x) = \frac{1}{\alpha} EXP\left[-\frac{(x-\mu)}{\alpha} - EXP\left[-\frac{(x-\mu)}{\alpha}\right]\right]$$

or

$$f(x) = \frac{1}{\alpha} e^{-\frac{x-\mu}{\alpha} - e^{-\frac{x-\mu}{\alpha}}}$$

for:

$-\infty < x < \infty$

Log-Gumbel distribution

The cumulative distribution function of the Gumbel distribution [5], has the form:

$$F(x) = e^{-e^{-\frac{(x-\mu)}{\alpha}}}$$

for: $-\infty < x < \infty$

where:

$0 < \alpha < \infty$, is the scale parameter

$-\infty < \mu < \infty$, is the position parameter, also called the central value or mode

If the variable x is replaced by lnx in the equation the cumulative function of the log-Gumbel distribution, or Fréchet distribution, is obtained.

Hydrometry is the branch of Hydrology that studies runoff measurement. Another term commonly used for the same purpose is *flow measurement*. To estimate a flow is to determine by measurement the flow that passes by a given section at a given moment [6].

Maximum flows

To design [6]:

- The dimensions of a channel
- Drainage system
  - agricultural
  - airport
  - city
  - road
- Channeling walls to protect cities and plantations
- Sewers
- Emergency spillways
- Bridge span

The design flow must be calculated or estimated, which in these cases are the maximum flows.

Some methods to calculate the maximum flow are:

- Direct method
- Empirical methods
- Curve number method
- Statistical methods
- Hydrological methods

Evapotranspiration

These are the total losses: evaporation from the evaporating surface (soil and water) + plant transpiration.

The term *potential evapotranspiration* [6] was introduced by Thornthwaite, and is defined as the total loss of water that would occur if there was never a lack of water in the soil to be used by vegetation.

Consumptive use is defined as the sum of evapotranspiration and water directly used to build plant tissue.

Since the amount of water used to build tissues is insignificant compared to that consumed in evapotranspiration, it may be said that:

Consumptive use » evapotranspiration

In irrigation projects, it is important to make initial calculations of the water required by crops. These water needs, which will be met through irrigation, are the evapotranspiration or consumptive use.

Methods based on meteorological data are used to calculate these amounts of water, of which the most well known are the Thornthwaite and Blaney –Clidde methods.

Materials and methods

Visual Basic, version 6.0 was used to develop the software, which permits the creation of 32-bit applications.

Numerical methods are be used to solve the equations, such as:

- Newton Raphson algorithm
- Secant method
- Graphical integration
- Lagrange interpolation
- Romberg's algorithm

To calculate theoretical cumulative distributions such as the Normal, Log-Normal, and Gamma, and as to estimate their inverse values, series development is used.

Many solutions of the equations are found using nomograms and tables. In the process of computation, they should be transformed into equations; the following will be used to determine the most appropriate equations:

- Simple linear correlations
- Simple non-linear correlations
- Multiple linear correlations
- Multiple non-linear correlations

Results

The product of this work is HidroEsta. The following figure shows one of its screens, with the calculation of the normal distribution:
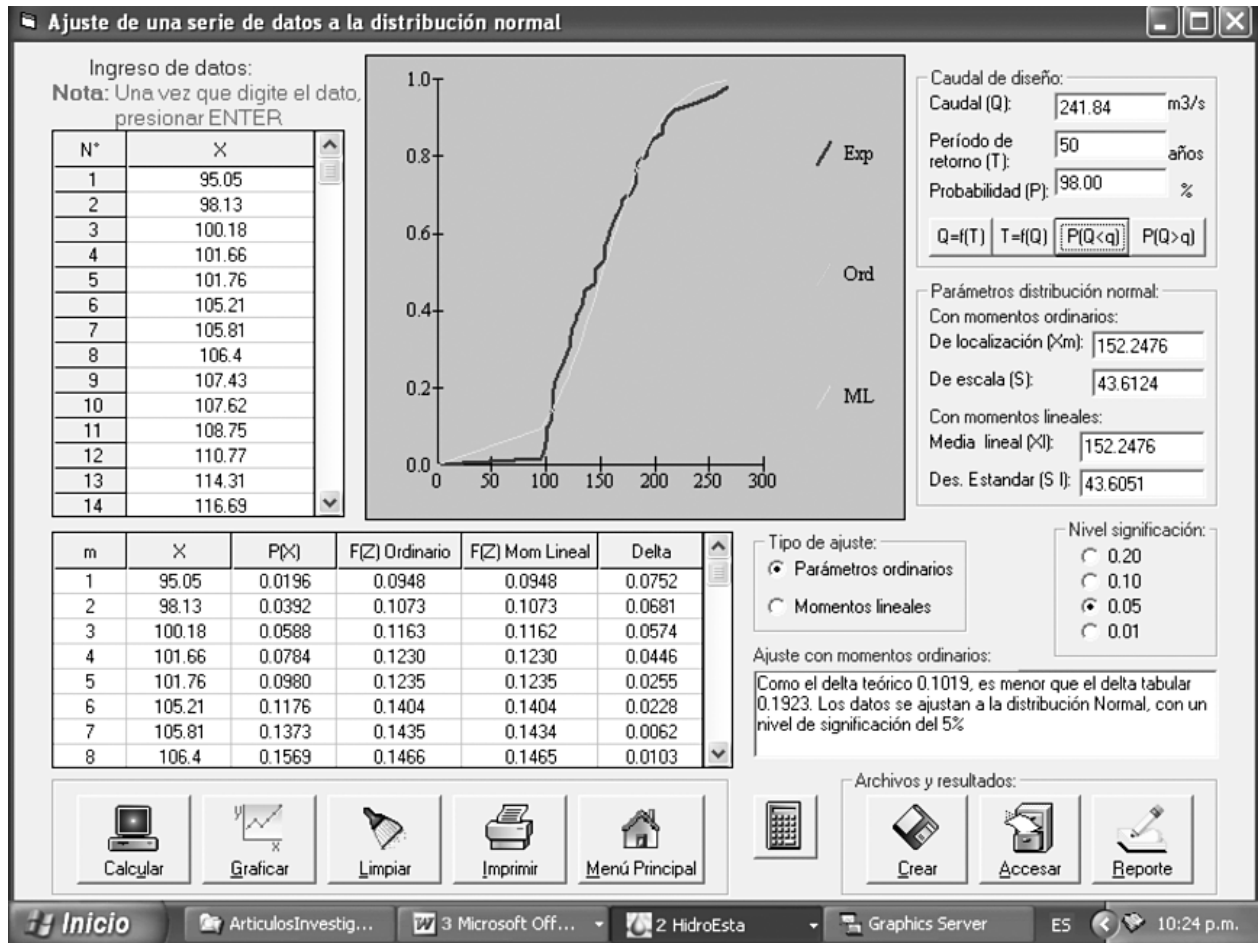
Discussion

The results calculated using the application are in all cases more approximate than those obtained with nomograms.

HidroEsta represents a contribution to the simplification of hydrological studies. It is important because:

- It provides an innovative and easy-to use tool for civil engineers, agricultural engineers, agronomical engineers, and other specialists who work in fields related to hydrological studies.

- It simplifies processing of abundant information and complicated calculations.

- Based on the information obtained, it allows simulation of design parameters for structures to be constructed.

- It greatly reduces calculation time.

- It permits obtaining optimal and economical designs.

Acknowledgement

Bibliographic references

[1]   Haan, Ch. *Statistical methods in hydrology*. Iowa, State University Press, U.S.A., 1977.

[2]   Kite, G.W. *Frequency and risk analyses in hydrology.* Water Resources Publications, Michigan, U.S.A. 1977.

[3]   Maidment, David. *Handbook of hydrology*. McGraw-Hill. New York – USA. 1993.

[4]   Villón, Máximo. *Diseño de capacidad de embalses por el método experimental – teoría de range*. Tesis para optar por el grado de Magíster Scientiae en Ingeniería de Recursos de Agua y Tierra. Universidad Nacional Agraria "La Molina". Lima – Perú, 1983.

[5]   Villón, Máximo. *Hidrología estadística*. Taller de Publicaciones. Instituto Tecnológico de Costa Rica. Cartago – Costa Rica. 2001.

[6]   Villón, Máximo. *Hidrología*. Taller de Publicaciones. Instituto Tecnológico de Costa Rica. Cartago – Costa Rica. 2002.