
Simulation of probability distributions commonly used in hydrological frequency analysis

Ke-Sheng Cheng,* Jie-Lun Chiang and Chieh-Wei Hsu

Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, Taiwan, ROC

Abstract:

Random variable simulation has been applied to many applications in hydrological modelling, flood risk analysis, environmental impact assessment, etc. However, computer codes for simulation of distributions commonly used in hydrological frequency analysis are not available in most software libraries. This paper presents a frequency-factor-based method for random number generation of five distributions (normal, log-normal, extreme-value type I, Pearson type III and log-Pearson type III) commonly used in hydrological frequency analysis. The proposed method is shown to produce random numbers of desired distributions through three means of validation: (1) graphical comparison of cumulative distribution functions (CDFs) and empirical CDFs derived from generated data; (2) properties of estimated parameters; (3) type I error of goodness-of-fit test. An advantage of the method is that it does not require CDF inversion, and frequency factors of the five commonly used distributions involves only the standard normal deviate. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS random number generation; random variable simulation; hydrological frequency analysis; goodness-of-fit test

INTRODUCTION

In various statistical applications, particularly in simulation studies, it is often desired to generate random samples of specified random variables. Recent developments in hydrological modelling, flood risk analysis, environmental impact assessment, etc. have demonstrated the usefulness of random variable simulation (National Research Council, 2000). Although computer codes for simulation of random variables with uniform and normal (or Gaussian) distributions are widely available, simulation of other non-Gaussian continuous random variables frequently used in hydrological study are less common.

Computer simulation of random variables is the task of using computers to generate many random numbers that are independent and identically distributed. It is also known as random number generation (RNG). In fact, these computer-generated random numbers form a deterministic sequence, and the same list of numbers will be cycled repeatedly. This cycle can be made to be so long that the lack of true independence is unimportant (Larget, 2002). Therefore, such computer codes are often termed pseudo-random number generators (PRNGs). There exist mathematical transformation methods to obtain other distributions from uniform variates (Devroye, 1986; Hellekalek, 1997). For this reason, most PRNGs found in software libraries produce uniform random numbers in the unit interval (0, 1). However, transformations from a uniform distribution to the several distributions commonly used in hydrological frequency analysis is not easy, and computer codes for simulation of these distributions are not available in most software libraries. Therefore, the purpose of this paper is to present a method for RNG of five distributions (normal, log-normal, extreme-value type I (EV1), Pearson type III (PT3) and log-Pearson type III (LPT3)) commonly used in hydrological frequency analysis. The method

* Correspondence to: Ke-Sheng Cheng, Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, Taiwan, ROC. E-mail: rslab@ntu.edu.tw

utilizes the frequency factor, which is familiar to hydrologists for transformation from uniform variates to desired distributions.

GENERATING RANDOM NUMBERS: PROBABILITY INTEGRAL TRANSFORMATION AND REJECTION METHODS

Suppose that we are interested in generating n values of a random variable X that has a continuous cumulative distribution function (CDF) $F_X(\cdot)$. A commonly used method for such a purpose is the probability integral transform (PIT) method or the inversion method (Mood *et al.*, 1974).

The PIT method is based on the property that a random variable X with CDF $F_X(\cdot)$ can be transformed into a random variable U with uniform distribution over the interval $(0, 1)$ by defining

$$U = F_X(X) \quad (1)$$

Conversely, if U is uniformly distributed over the interval $(0, 1)$, then $X = F_X^{-1}(U)$ has CDF $F_X(\cdot)$. Thus, to generate a value, say x , of a random variable X having a continuous CDF $F_X(\cdot)$, it suffices to generate a value, say u , of a random variable U that is uniformly distributed over the interval $(0, 1)$. The value x is then obtained by

$$x = F_X^{-1}(u) \quad (2)$$

Another method for generating random deviates is the rejection method (Press *et al.*, 1993), which does not require that the CDF be readily computable, much less the inverse of that function. The rejection method utilizes a comparison function $g(x)$ that lies everywhere above the desired probability distribution $f(x)$. The method first generates a uniform deviate between zero and A , where A is the total area under the comparison function, and uses it to get a corresponding x . Then a second uniform deviate y between zero and $g(x)$ is generated and used to decide whether to accept (if $f(x) \geq y$) or reject (if $f(x) < y$) that x . The non-rejected x values would have the desired distribution $f(x)$.

GENERAL EQUATION FOR HYDROLOGICAL FREQUENCY ANALYSIS

Hydrological frequency analysis is the work of determining the magnitudes of hydrological variables corresponding to a given frequency or recurrence interval. The recurrence interval, also called the return period, is defined as the average interval over a long period of time during which a corresponding magnitude of some hydrological variable is met or exceeded. Rainfall and streamflow are two major types of hydrological variable that are used in frequency analysis. For example, annual maximum flow records are used to estimate the magnitude of flood of 50-year return period Q_{50} , i.e. on average, 1 year of every 50-year sequence is expected to experience a flood of at least Q_{50} .

Two methods of hydrological frequency analysis are commonly applied: the plotting position method and the frequency factor method. The former is a straightforward plotting technique to obtain the CDF by use of certain 'plotting position' formulas (Chow *et al.*, 1988). The frequency factor method is described below.

A random variable X has CDF $F_X(\cdot)$ with mean μ and standard deviation σ . The magnitude of X corresponding to return period T , denoted by x_T , is defined as

$$P(X \geq x_T) = 1/T \quad (3)$$

Chow (1951) proposed the following general equation for hydrological frequency analysis:

$$x_T = \mu + K_T \sigma \quad (4)$$

Table I. Probability density functions and frequency factors of distributions commonly used for hydrological frequency analysis

Distribution, X	Probability density function $f_X(x)$	Frequency factor K_T
Normal	$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$ $-\infty < x < +\infty$	Standard normal deviate z with exceedance probability $1/T$
Log-normal	$f_X(x) = \frac{1}{\sqrt{2\pi}x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right],$ $0 < x < +\infty$ $\mu = e^{\mu_y + \sigma_y^2/2}, \quad \sigma^2 = (e^{\sigma_y^2} - 1)\mu^2$ $\mu_y \text{ and } \sigma_y \text{ are respectively the mean and standard deviation of } Y = \ln X$	$K_T = \frac{\exp\{[\ln(1 + C_V^2)]^{1/2}Z - [\ln(1 + C_V^2)]/2\} - 1}{C_V}$ $C_V = \sigma/\mu, \text{ coefficient of variation of } X$ $z: \text{ standard normal deviate with exceedance probability } 1/T$
EV1	$f_X(x) = \alpha \exp[-\alpha(x - \beta) - e^{-\alpha(x-\beta)}],$ $-\infty < x < +\infty$ $\alpha = \pi/\sqrt{6}\sigma, \quad \beta = \mu - (0.5772/\alpha)$ $\mu \text{ and } \sigma \text{ are respectively the mean and standard deviation of } X$	$K_T = -\frac{\sqrt{6}}{\pi} \left\{ 0.5772 + \ln \left[\ln \left(\frac{T}{T-1} \right) \right] \right\}$
PT3	$f_X(x) = \frac{1}{\alpha\Gamma(\beta)} \left(\frac{x-\varepsilon}{\alpha}\right)^{\beta-1} e^{-[(x-\varepsilon)/\alpha]},$ $\varepsilon \leq x < +\infty$ $\alpha = \sigma/\sqrt{\beta}, \quad \beta = (2/\gamma)^2, \quad \varepsilon = \mu - \sigma\sqrt{\beta}$ $\mu, \sigma \text{ and } \gamma \text{ are respectively the mean, standard deviation and skewness coefficient of } X$	$K_T \approx z + (z^2 - 1)\frac{\gamma}{6} + \frac{1}{3}(z^3 - 6z)\left(\frac{\gamma}{6}\right)^2 - (z^2 - 1)\left(\frac{\gamma}{6}\right)^3 + z\left(\frac{\gamma}{6}\right)^4 - \frac{1}{3}\left(\frac{\gamma}{6}\right)^5$ $z: \text{ standard normal deviate with exceedance probability } 1/T$
LPT3	$f_X(x) = \frac{1}{\alpha x \Gamma(\beta)} \left(\frac{\ln x - \varepsilon}{\alpha}\right)^{\beta-1} e^{-\left(\frac{\ln x - \varepsilon}{\alpha}\right)},$ $\varepsilon \leq \ln x < +\infty$ $\alpha = \sigma_y/\sqrt{\beta}, \quad \beta = (2/\gamma_y)^2, \quad \varepsilon = \mu_y - \sigma_y\sqrt{\beta}$ $\mu_y, \sigma_y \text{ and } \gamma_y \text{ are respectively the mean, standard deviation and skewness coefficient of } Y = \ln X$	Same as K_T of the PT3 distribution (K_T is to be substituted into $y_T = \ln x_T = \mu_y + K_T\sigma_y$)

where K_T , the frequency factor, is a function of T and is distribution specific. Apparently, if X is normally distributed, then the frequency factor K_T corresponds to the standard normal deviate with exceedance probability $1/T$. Frequency factors of commonly used distributions in hydrological frequency analysis have been developed (Kite, 1988). Table I shows the probability density functions and frequency factors of five distributions commonly used for hydrological frequency analysis.

Suppose that a random sample $\{x_1, x_2, \dots, x_n\}$ of a hydrological variable X of known distribution type is available. The magnitude of X corresponding to return period T , x_T , can be estimated by

1. Calculating the sample mean \bar{x} , sample variance s^2 and skewness coefficient $\hat{\gamma}$ from the random sample $\{x_1, x_2, \dots, x_n\}$.
2. Determining the value of frequency factor K_T using the appropriate distribution or equation in Table I. Readers are reminded that frequency factors of random variables of normal, log-normal and PT3 distributions do not explicitly relate to return period T . The relation between K_T and T is embedded in the value of the standard normal deviate z , which satisfies $P(Z \geq z) = 1/T$.
3. x_T is estimated by $\hat{x}_T = \bar{x} + K_T s$.

GENERATING RANDOM NUMBERS USING FREQUENCY FACTORS

As shown in Equation (2), generation of random numbers by the PIT method requires inversion of the CDF $F_X(\cdot)$. It is not always easy to determine the inverse function of $F_X(\cdot)$. As for the rejection method, one needs to choose a comparison function whose indefinite integral is known analytically, and which is analytically invertible. In this section we propose an alternative method that can avoid CDF inversion by using the general equation of frequency analysis.

The CDF of a continuous random variable is a non-decreasing function and $x = F_X^{-1}(u)$ is a one-to-one single-value relation. After a random number, say u , of a uniform distribution over the interval $(0, 1)$ is generated, let us set

$$T = \frac{1}{1-u} \quad (5)$$

This yields

$$P(X \geq x_T) = \frac{1}{T} = 1 - u = P(U \geq u) \quad (6)$$

with T determined by Equation (5). Frequency factor K_T can be calculated using the appropriate distribution or equation in Table I. Finally, the magnitude of x_T is calculated by $x_T = \mu + K_T\sigma$. Similarly, a set of random numbers $\{u_1, u_2, \dots, u_n\}$ of a uniform distribution over interval $(0, 1)$ can be transformed to random numbers $\{x_1, x_2, \dots, x_n\}$ of the desired distribution.

Unlike the PIT method, no CDF inversion is involved in the above calculation, and the proposed method is hereafter referred to as the frequency factor transformation (FQFT) method. Another advantage of the FQFT approach is that, even though there are five types of random variable in Table I, determination of K_T involves only the standard normal deviate z .

TEST AND VALIDATION

In order to demonstrate the applicability of the FQFT approach, random numbers of normal, log-normal, EV1, PT3 and LPT3 distributions are generated and tested. Specific distribution parameters designated for generating random numbers are shown in Table II. For each type of distribution N , random samples (each of size n) were generated and used in subsequent analysis. In this study, the sample size n was set to vary from 50 to 500 in increments of 50 and the number of random samples N was set to 1000 and 10000.

We adopt three means to test the validity of the random numbers generated: (1) graphical comparison of the CDF and the empirical CDF (ECDF) derived from generated data; (2) properties of estimated parameters; (3) type I error of goodness-of-fit (GOF) test.

Graphical comparison of CDF and ECDF

Figure 1 graphically illustrates the closeness of the CDF and ECDF with regard to sample sizes of 50 and 500. Each ECDF in Figure 1 is based on one single random sample of size 50 or 500 and it may change

Table II. Distribution parameters designated for generation of random numbers

Parameter	Distribution				
	Normal	Log-normal	EV1	PT3	LPT3
Mean μ	0	0 ^a	0	0	0 ^a
Standard deviation σ	1	1 ^a	1	1	1 ^a
Coefficient of skewness γ	0	0 ^a	1.1396	1.5	1.5 ^a

^a Parameters are assigned for $Y = \ln X$ when X is a random variable of log-normal or LPT3 distribution.

HYDROLOGICAL FREQUENCY ANALYSIS

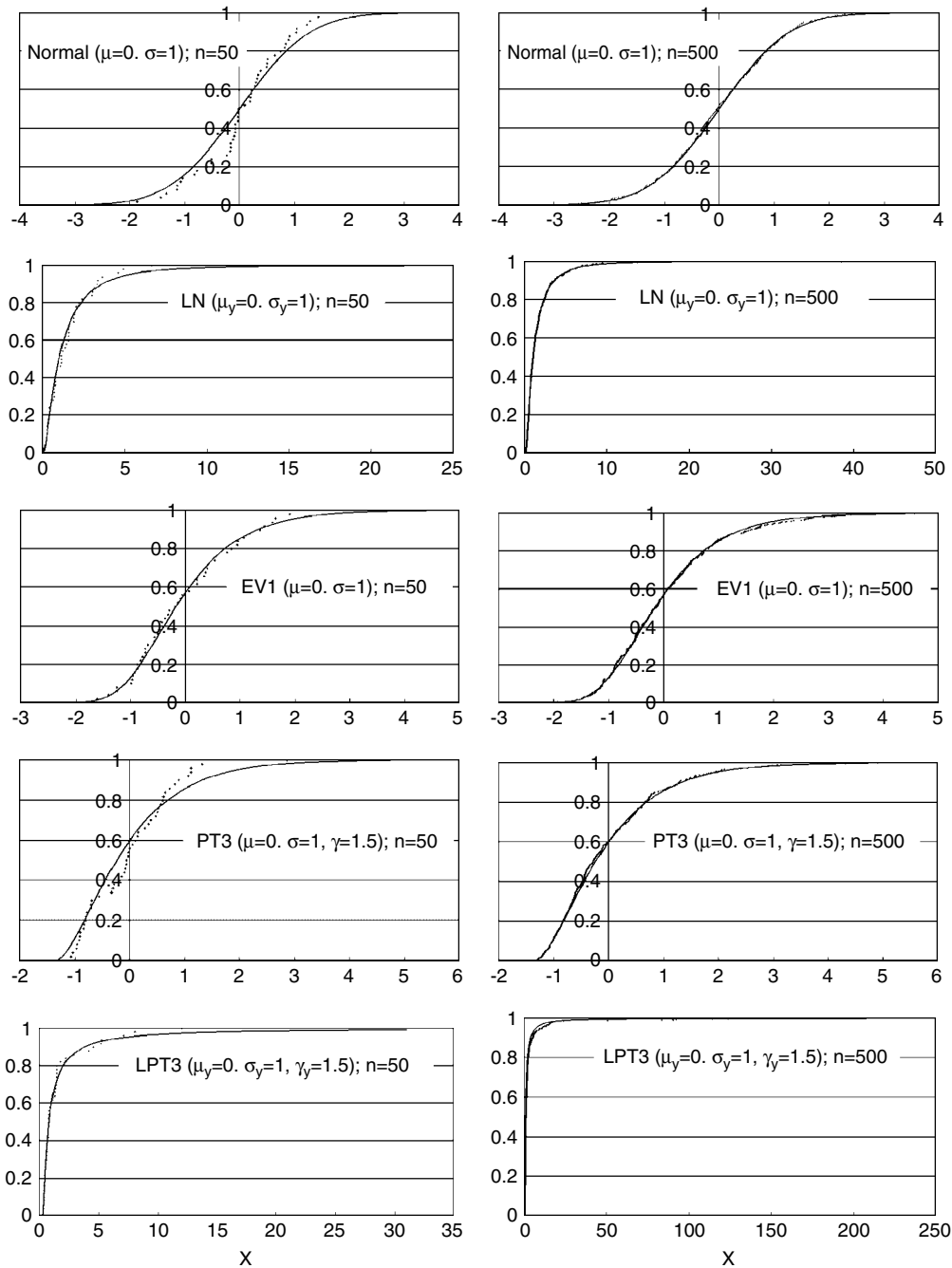


Figure 1. Graphical comparison of ECDF (dots) and CDF (solid curves). (For log-normal and log-Pearson distributions, values of $Y = \ln X$ are used in X -axis)

when another random sample is used. It can be seen that even at sample size of 50 the ECDF is fairly close to the CDF of the designated distribution. At a sample size of 500, all ECDFs become almost indistinguishable from their corresponding CDFs.

Properties of parameter estimators

From each of the N random samples generated, the distribution parameters mean, standard deviation and coefficient of skewness can be estimated by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \tag{7}$$

$$s = \frac{1}{\sqrt{n-1}} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{8}$$

$$\hat{\gamma} = n \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)s^3} \tag{9}$$

The coefficient of skewness is very sensitive to sample size n . Bobee and Robitaille (1977) suggested using the following sample-size-adjusted coefficient of skewness for PT3 and LPT3 distributions

$$\hat{\gamma}' = \hat{\gamma} \frac{[n(n-1)]^{1/2}}{n-2} \left(1 + \frac{8.5}{n} \right) \tag{10}$$

Furthermore, from a total of N random samples, the sample mean and standard deviation of the above estimated parameters were calculated, with respect to sample size n ranging from 50 to 500, and listed in Tables III–VII. Figures 2–4 demonstrate that, with 1000 and 10 000 random samples ($N = 1000$ or 10 000), sample means (the centre line) of the estimated parameters (including mean, standard deviation and coefficient of skewness) are very close to the theoretical values designated for RNG. It is also seen clearly that standard deviations of all parameter estimators decrease with increase of the sample size n , indicating the unbiased nature of the estimator and reduction of uncertainty in parameter estimation. Such characteristics of parameter estimators suggest the random samples generated are indeed from the desired distributions.

Table III. Sample mean and standard deviation of the estimated parameters (mean, standard deviation and coefficient of skewness) with respect to sample size n ranging from 50 to 500 for normal distribution ($\mu = 0, \sigma = 1$)

Estimated parameter	Summary statistic	n (sample size)									
		50	100	150	200	250	300	350	400	450	500
<i>Number of random samples $N = 1000$</i>											
\bar{x}	Mean	0.0025	-0.0053	0.0016	0.0001	-0.0003	0.0011	-0.0008	0.0021	-0.0002	-0.0010
	SD	0.1346	0.1009	0.0802	0.0696	0.0655	0.0603	0.0525	0.0493	0.0479	0.0457
s	Mean	0.9993	0.9955	0.9994	0.9986	1.0010	0.9990	0.9993	0.9989	0.9985	0.9996
	SD	0.1019	0.0683	0.0587	0.0506	0.0449	0.0423	0.0395	0.0354	0.0324	0.0303
$\hat{\gamma}$	Mean	0.0077	-0.0002	0.0066	-0.0008	0.0098	-0.0009	-0.0013	-0.0009	-0.0001	-0.0003
	SD	0.4094	0.2637	0.2094	0.1785	0.1601	0.1495	0.1363	0.1252	0.1187	0.1104
<i>Number of random samples $N = 10\,000$</i>											
\bar{x}	Mean	-0.0012	-0.0003	-0.0010	-0.0006	-0.0004	-0.0005	0.0007	0.0009	-0.0001	0.0004
	SD	0.1422	0.1002	0.0821	0.0714	0.0640	0.0574	0.0536	0.0501	0.0474	0.0443
s	Mean	0.9944	0.9980	0.9987	0.9986	0.9986	0.9982	0.9994	1.0000	0.9989	0.9994
	SD	0.1005	0.0710	0.0577	0.0505	0.0448	0.0406	0.0375	0.0353	0.0331	0.0314
$\hat{\gamma}$	Mean	-0.0057	0.0002	-0.0022	0.0001	-0.0024	-0.0015	0.0004	-0.0025	0.0008	-0.0015
	SD	0.4036	0.2677	0.2133	0.1804	0.1602	0.1453	0.1364	0.1239	0.1178	0.1103

HYDROLOGICAL FREQUENCY ANALYSIS

Table IV. Sample mean and standard deviation of the estimated parameters (mean, standard deviation and coefficient of skewness) with respect to sample size n ranging from 50 to 500 for log-normal distribution ($\mu_y = 0, \sigma_y = 1$)

Estimated parameter	Summary statistic	n (sample size)									
		50	100	150	200	250	300	350	400	450	500
<i>Number of random samples $N = 1000$</i>											
\bar{y}	Mean	-0.0012	-0.0022	0.0028	-0.0047	-0.0011	-0.0016	-0.0014	-0.0017	0.0003	0.0004
	SD	0.1430	0.0999	0.0810	0.0711	0.0607	0.0574	0.0544	0.0484	0.0482	0.0457
s_y	Mean	0.9947	0.9990	0.9976	0.9980	0.9976	0.9990	0.9993	0.9980	0.9979	0.9975
	SD	0.1048	0.0710	0.0587	0.0494	0.0441	0.0419	0.0379	0.0361	0.0337	0.0313
$\hat{\gamma}'_y$	Mean	-0.0019	-0.0100	0.0056	0.0054	-0.0046	0.0015	0.0062	0.0054	0.0037	-0.0042
	SD	0.3987	0.2672	0.2152	0.1849	0.1637	0.1464	0.1322	0.1229	0.1198	0.1102
<i>Number of random samples $N = 10000$</i>											
\bar{y}	Mean	0.0011	0.0000	-0.0004	-0.0010	-0.0002	0.0010	0.0002	-0.0009	-0.0005	0.0000
	SD	0.1430	0.0999	0.0830	0.0719	0.0635	0.0574	0.0532	0.0496	0.0474	0.0448
s_y	Mean	0.9960	0.9959	0.9983	0.9985	0.9986	0.9993	0.9989	0.9994	0.9994	0.9988
	SD	0.1011	0.0718	0.0581	0.0496	0.0447	0.0407	0.0378	0.0356	0.0334	0.0317
$\hat{\gamma}'_y$	Mean	-0.0041	0.0012	0.0006	0.0006	0.0019	0.0005	-0.0011	-0.0003	0.0010	0.0018
	SD	0.4010	0.2664	0.2133	0.1812	0.1599	0.1448	0.1351	0.1247	0.1181	0.1112

\bar{y} , s_y and $\hat{\gamma}'_y$ are respectively the sample estimates of mean, standard deviation and skewness coefficient of $Y = \ln X$ respectively.

Table V. Sample mean and standard deviation of the estimated parameters (mean, standard deviation and coefficient of skewness) with respect to sample size n ranging from 50 to 500 for EV1 distribution ($\mu = 0, \sigma = 1, \gamma = 1.1396$)

Estimated parameter	Summary statistic	n (sample size)									
		50	100	150	200	250	300	350	400	450	500
<i>Number of random samples $N = 1000$</i>											
\bar{x}	Mean	-0.0002	-0.0012	-0.0017	-0.0017	-0.0044	-0.0003	-0.0014	-0.0034	0.0009	0.0028
	SD	0.1417	0.0976	0.0815	0.0713	0.0630	0.0565	0.0515	0.0501	0.0463	0.0461
s	Mean	0.9897	1.0012	0.9956	0.9966	0.9952	0.9985	0.9971	0.9958	0.9994	1.0007
	SD	0.1450	0.1011	0.0863	0.0734	0.0667	0.0595	0.0571	0.0504	0.0487	0.0470
$\hat{\gamma}'$	Mean	1.1481	1.1729	1.1513	1.1394	1.1425	1.1393	1.1270	1.1344	1.1409	1.1402
	SD	0.6397	0.4846	0.4279	0.3654	0.3360	0.3041	0.2584	0.2569	0.2402	0.2364
<i>Number of random samples $N = 10000$</i>											
\bar{x}	Mean	0.0025	-0.0008	-0.0005	-0.0012	-0.0005	0.0005	-0.0006	0.0001	0.0001	0.0001
	SD	0.1419	0.0991	0.0820	0.0704	0.0631	0.0576	0.0532	0.0500	0.0472	0.0445
s	Mean	0.9937	0.9936	0.9976	0.9972	0.9977	0.9984	0.9985	0.9983	0.9993	0.9990
	SD	0.1455	0.1035	0.0851	0.0733	0.0665	0.0604	0.0554	0.0518	0.0490	0.0470
$\hat{\gamma}'$	Mean	1.1766	1.1420	1.1480	1.1391	1.1434	1.1398	1.1345	1.1391	1.1429	1.1397
	SD	0.6342	0.4778	0.4104	0.3479	0.3216	0.2998	0.2687	0.2575	0.2557	0.2353

Type I error of GOF test

Each random sample of size n is generated from a theoretical distribution with designated parameters and the GOF test can be applied to test whether the random sample is drawn from the theoretical distribution. The widely applied chi-square GOF test is adopted in this study.

A random sample x_1, x_2, \dots, x_n consists of n observed values of a hypothesized distribution. These observed values fall into k mutually exclusive categories, and the following statistic

$$T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \tag{11}$$

Table VI. Sample mean and standard deviation of the estimated parameters (mean, standard deviation and coefficient of skewness) with respect to sample size n ranging from 50 to 500 for PT3 distribution ($\mu = 0, \sigma = 1, \gamma = 1.5$)

Estimated parameter	Summary statistic	n (sample size)									
		50	100	150	200	250	300	350	400	450	500
<i>Number of random samples $N = 1000$</i>											
\bar{x}	Mean	-0.0025	-0.0013	0.0022	-0.0023	0.0000	0.0009	0.0021	0.0024	0.0008	-0.0008
	SD	0.1392	0.0976	0.0770	0.0694	0.0621	0.0582	0.0541	0.0498	0.0475	0.0448
s	Mean	0.9830	0.9942	0.9927	0.9945	0.9952	0.9954	0.9989	0.9994	0.9978	1.0002
	SD	0.1666	0.1140	0.0947	0.0825	0.0758	0.0702	0.0637	0.0568	0.0568	0.0529
$\hat{\gamma}$	Mean	1.5675	1.5277	1.5317	1.5526	1.5214	1.5116	1.5284	1.5293	1.5198	1.5390
	SD	0.6938	0.5057	0.4390	0.4125	0.3647	0.3407	0.3283	0.3032	0.2732	0.2738
<i>Number of random samples $N = 10000$</i>											
\bar{x}	Mean	0.0016	-0.0010	0.0004	0.0015	-0.0005	-0.0009	0.0008	0.0010	0.0018	0.0003
	SD	0.1421	0.1007	0.0816	0.0704	0.0626	0.0576	0.0541	0.0497	0.0470	0.0444
s	Mean	0.9876	0.9921	0.9947	0.9977	0.9961	0.9969	0.9971	0.9982	0.9986	0.9976
	SD	0.1620	0.1162	0.0957	0.0831	0.0743	0.0682	0.0636	0.0590	0.0561	0.0526
$\hat{\gamma}$	Mean	1.5690	1.5450	1.5350	1.5364	1.5309	1.5249	1.5265	1.5279	1.5248	1.5250
	SD	0.6561	0.5257	0.4511	0.4050	0.3699	0.3382	0.3221	0.3033	0.2919	0.2745

Table VII. Sample mean and standard deviation of the estimated parameters (mean, standard deviation and coefficient of skewness) with respect to sample size n ranging from 50 to 500 for LPT3 distribution ($\mu_y = 0, \sigma_y = 1, \gamma_y = 1.5$)

Estimated parameter	Summary statistic	n (sample size)									
		50	100	150	200	250	300	350	400	450	500
<i>Number of random samples $N = 1000$</i>											
\bar{y}	Mean	-0.0050	-0.0032	0.0018	0.0001	0.0005	-0.0002	-0.0018	-0.0014	-0.0017	-0.0001
	SD	0.1358	0.1037	0.0801	0.0688	0.0624	0.0572	0.0531	0.0501	0.0461	0.0452
s_y	Mean	0.9821	0.9931	0.9956	0.9929	0.9956	0.9973	0.9935	0.9947	0.9981	0.9993
	SD	0.1608	0.1193	0.0965	0.0844	0.0740	0.0670	0.0632	0.0588	0.0575	0.0522
$\hat{\gamma}'_y$	mean	1.5947	1.5533	1.5283	1.5251	1.5289	1.5222	1.5161	1.5269	1.5279	1.5324
	SD	0.6706	0.4998	0.4629	0.3905	0.3899	0.3265	0.3040	0.2964	0.2941	0.2660
<i>Number of random samples $N = 10000$</i>											
\bar{y}	Mean	0.0018	0.0000	0.0011	0.0017	0.0009	0.0003	0.0010	0.0008	0.0001	0.0007
	SD	0.1400	0.0993	0.0805	0.0710	0.0641	0.0584	0.0534	0.0502	0.0475	0.0450
s_y	Mean	0.9883	0.9928	0.9944	0.9960	0.9968	0.9971	0.9967	0.9979	0.9975	0.9979
	SD	0.1648	0.1168	0.0955	0.0830	0.0751	0.0685	0.0627	0.0593	0.0561	0.0535
$\hat{\gamma}'_y$	Mean	1.5849	1.5419	1.5326	1.5262	1.5283	1.5296	1.5312	1.5285	1.5299	1.5280
	SD	0.6816	0.5160	0.4423	0.3921	0.3644	0.3393	0.3220	0.3046	0.2857	0.2827

\bar{y} , s_y and $\hat{\gamma}'_y$ are respectively the sample estimates of mean, standard deviation and skewness coefficient of $Y = \ln X$ respectively.

has a chi-square distribution, for large n , with $k - 1$ degrees of freedom. In Equation (11), O_i and E_i respectively represent the observed and theoretical expected frequencies falling in the i th category. There are various criteria for determination of sample size n and number of categories k . It is usually felt that n should be large enough that no expected frequency is less than unity and no more than 20% of the expected frequencies are less than five (Milton and Arnold, 2003). Mann and Wald (1942) initiated a study of choice of categories and recommended that the categories be chosen to have equal probabilities under the hypothesized distribution. They found that, for a sample of size n (large) and significance level α , the number

HYDROLOGICAL FREQUENCY ANALYSIS

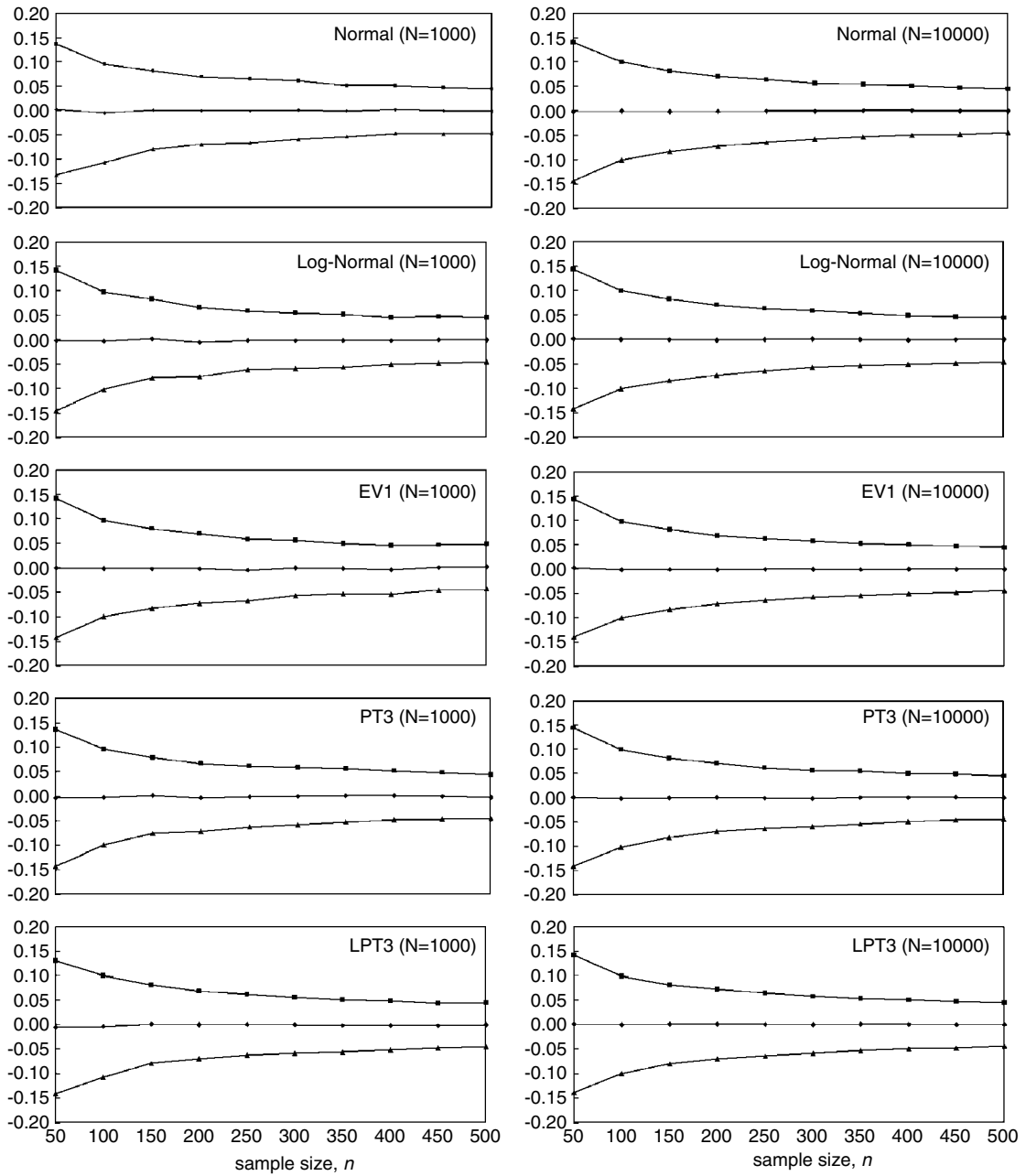


Figure 2. Uncertainty in estimation of mean μ reduces as sample size increases. (The centre line represents mean of \bar{x} and upper and lower lines are one standard deviation away from the centre line)

of equiprobable categories should be approximately

$$k^* = 4 \left[\frac{2n^2}{c(\alpha)^2} \right]^{1/5} \quad (12)$$

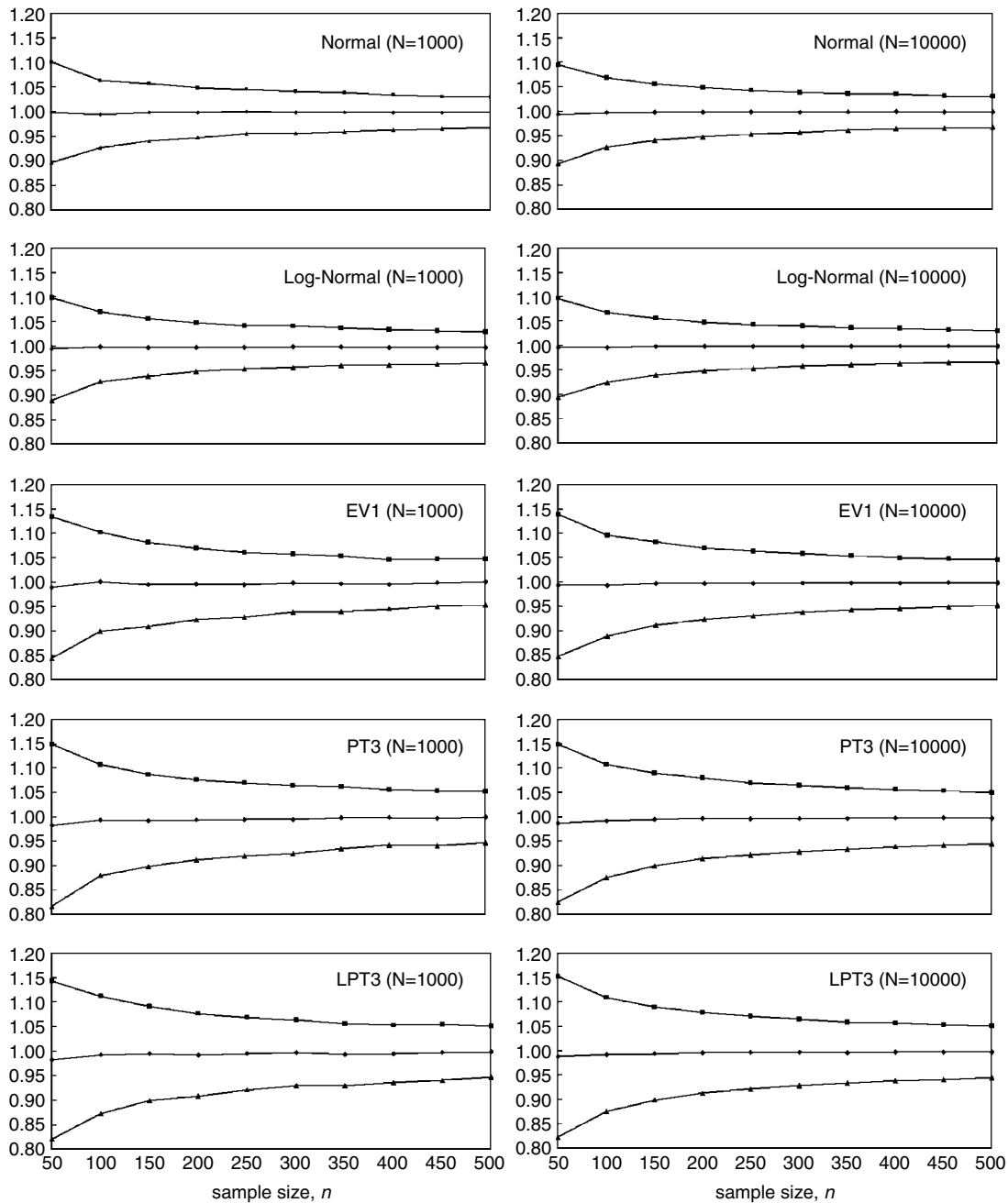


Figure 3. Uncertainty in estimation of standard deviation σ reduces as sample size n increases. (The centre line represents mean of s and upper and lower lines are one standard deviation away from the centre line)

where $c(\alpha)$ is the standard normal deviate with exceedance probability α . D'Agostino and Stephens (1986) further recommended that the number of equiprobable categories should fall between the k^* value determined by Equation (12) for $\alpha = 0.05$ and half that value. Since the value of k^* in Equation (12) increases slowly with α and it overstates the number of categories required, the value of $\alpha = 0.05$ for the $c(\alpha)$ calculation can be

HYDROLOGICAL FREQUENCY ANALYSIS

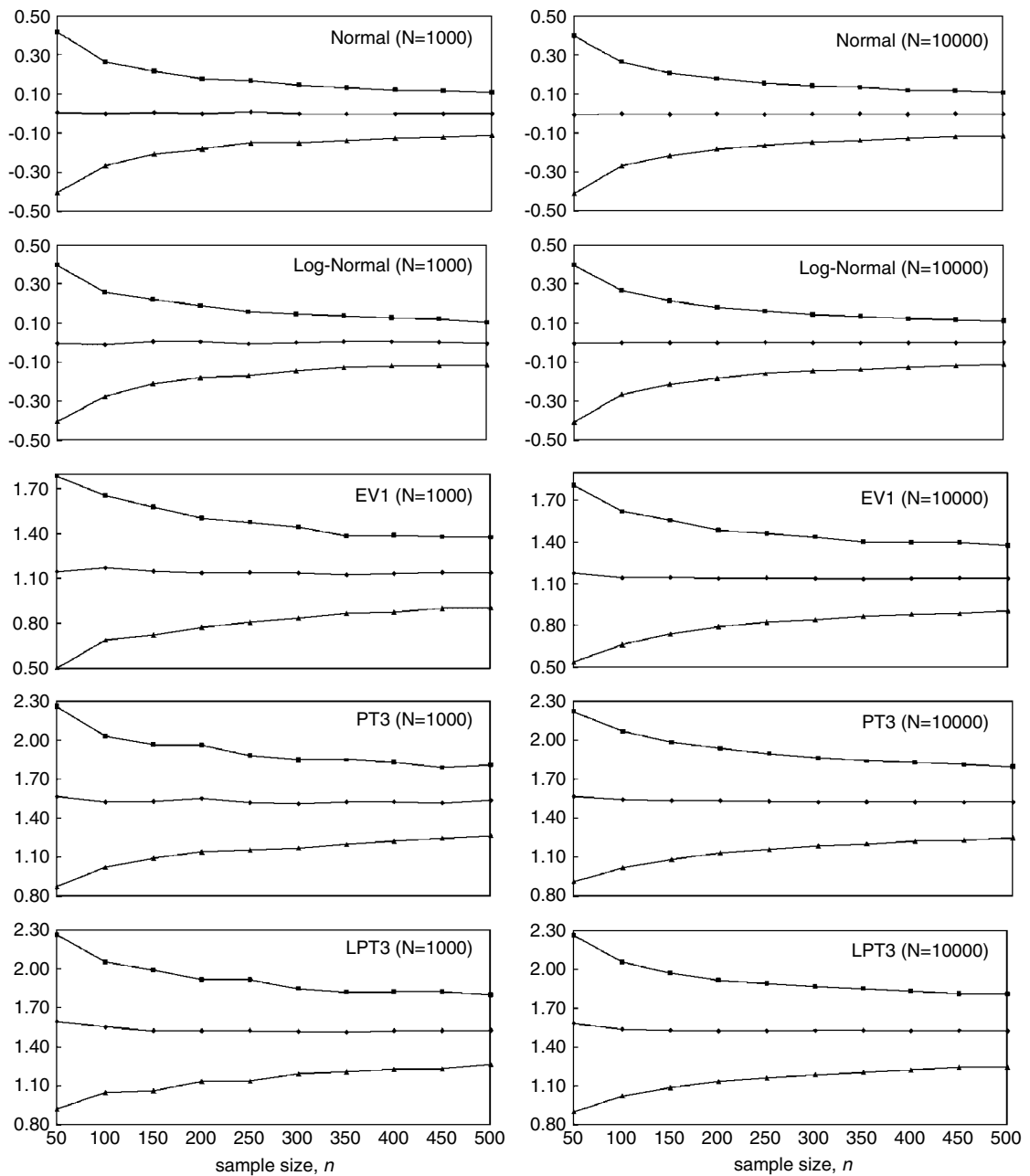


Figure 4. Uncertainty in estimation of skewness coefficient $\hat{\gamma}$ reduces as sample size n increases. (The centre line represents mean of $\hat{\gamma}$ and upper and lower lines are one standard deviation away from the centre line)

used for various levels of significance (D'Agostino and Stephens, 1986). To be more specific, $c(0.05) = 1.645$ and half the value of k^* in Equation (12) is

$$k = 0.5k^* = 2 \left[\frac{2n^2}{c(\alpha)^2} \right]^{1/5} = 1.88n^{2/5} \quad (13)$$

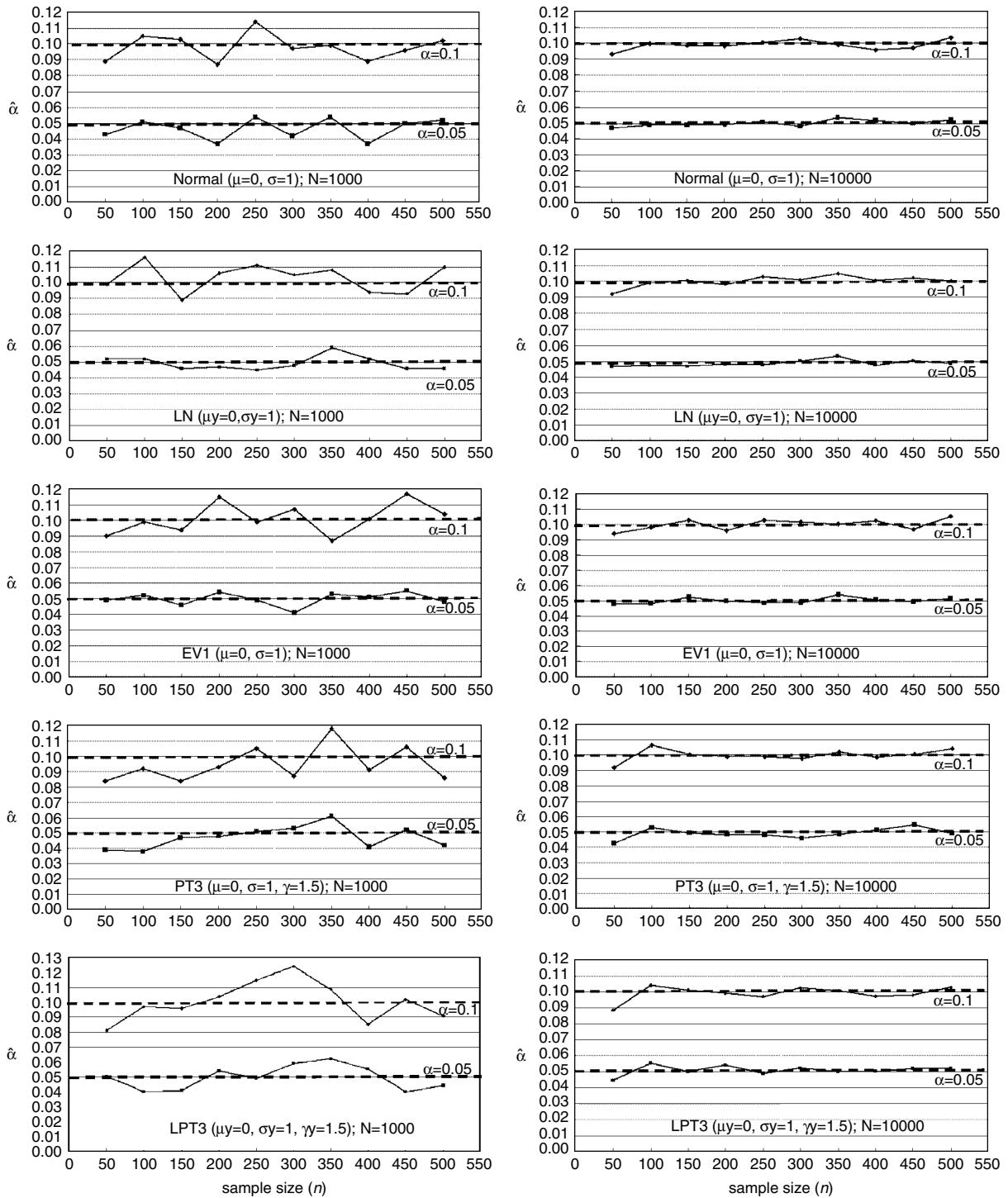


Figure 5. Type I error $\hat{\alpha}$ of chi-square GOF test with respect to sample size n

Therefore, $k \approx 2n^{2/5}$ is a convenient choice for the number of mutually exclusive, equiprobable categories for the chi-square GOF test and is adopted in this study.

The null hypothesis of the chi-square GOF test assumes that the observed sample is drawn from the hypothesized distribution. The null hypothesis is rejected, at level of significance α , if the value of the test statistic T calculated from the random sample x_1, x_2, \dots, x_n exceeds $\chi_{1-\alpha, k-1}^2$, the $(1 - \alpha)$ th quantile of the chi-square distribution with $k - 1$ degrees of freedom. Although all random samples are generated based on a theoretical distribution, there is no guarantee that all random samples will not be rejected by the GOF test, since the level of significance α is imposed. If a random sample is rejected by the GOF test using the theoretical distribution as the hypothesized distribution, then a type I error is conducted. Theoretically, at level of significance α , there will be $100\alpha\%$ of the total number of random samples rejected by the GOF test. In practice, N_r samples out of the totality are rejected and the probability of conducting a type I error is estimated as

$$\hat{\alpha} = N_r/N \quad (14)$$

As N increases, $\hat{\alpha}$ should become increasingly close to the level of significance α . Figure 5 demonstrates type I error $\hat{\alpha}$ of the chi-square GOF test with respect to sample size n . With 1000 random samples, most values of $\hat{\alpha}$ fall between 0.04 and 0.06 for $\alpha = 0.05$, and between 0.09 and 0.11 for $\alpha = 0.10$. As the number of random samples N increases to 10 000, almost all values of $\hat{\alpha}$ are very close to the level of significance α . Such results indicate that random samples generated by the FQFT method do comply with the desired distributions. It can also be seen in Figure 5 that $\hat{\alpha}$ fluctuates steadily with respect to sample size n , indicating that an increase in sample size does not help to stabilize the type I error $\hat{\alpha}$.

CONCLUSIONS

The proposed FQFT method is capable of generating random numbers of five distributions commonly used in hydrological frequency analysis. The ECDFs and the distribution parameters estimated from random samples are very close to, or indistinguishable from, theoretical values. Sample estimates of distribution parameters are unbiased, and estimation uncertainties reduce with increasing sample size. The type I error of the chi-square GOF test on the random samples generated fluctuates slightly around the level of significance. An advantage of the FQFT method is that it does not require CDF inversion and the frequency factor of the five commonly used distributions involves only the standard normal deviate.

ACKNOWLEDGEMENTS

We are grateful to the Council of Agriculture (Taiwan, ROC) for funding a project that led to the initiation of this study.

REFERENCES

- Bobee B, Robitaille R. 1977. The use of the Pearson type III and log Pearson type III distribution revisited. *Water Resources Research* **13**(2): 427–443.
- Chow VT. 1951. A general formula for hydrologic frequency analysis. *Transactions, American Geophysical Union* **32**: 231–237.
- Chow VT, Maidment DR, Mays LW. 1988. *Applied Hydrology*. McGraw-Hill: New York.
- D'Agostino RB, Stephens MA. 1986. *Goodness-of-Fit Techniques*. Marcel Dekker: New York.
- Devroye L. 1986. *Non-Uniform Random Variate Generation*. Springer-Verlag: New York.
- Hellekalek P. 1997. A note on pseudorandom number generators. *EUROSIM—Simulation News Europe* **20**: 6–8.
- Kite GW. 1988. *Frequency and Risk Analysis in Hydrology*. Water Resources Publications.
- Larget B. 2002. *Random number generation*. <http://www.mathcs.duq.edu/larget/math496/random.html> (accessed 31 March 2006).
- Mann HB, Wald A. 1942. On the choice of the number of class intervals in the application of the chi-squared test. *Annals of Mathematical Statistics* **13**: 306–317.

- Milton JS, Arnold JC. 2003. *Introduction to Probability and Statistics*. McGraw-Hill: New York.
- Mood AM, Graybill FA, Boes DC. 1974. *Introduction to the Theory of Statistics*. McGraw-Hill: New York.
- National Research Council. 2000. *Risk Analysis and Uncertainty in Flood Damage Reduction Studies*. National Academy Press.
- Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1993. *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press: 290–296.