

## On cumulative frequency/probability distributions and confidence intervals.

R.J. Oosterbaan

A summary of [www.waterlog.info/pdf/binomial.pdf](http://www.waterlog.info/pdf/binomial.pdf)

Frequency analysis is the analysis of how often, or how frequent, the value of an observed phenomenon occurs in a certain range with an upper and lower limit. It applies to a record of length  $N$  of observed data  $X_1, X_2, X_3, \dots, X_N$  on a variable phenomenon  $X$ .

The record may be time-dependent (e.g. rainfall measured in one spot) or space-dependent (e.g. crop yields in an area) or otherwise.

The cumulative frequency  $M_{X_r}$  of a reference value  $X_r$  is the number of observed values  $X$  in the range less than or equal to  $X_r$ .

The relative cumulative frequency  $F_c$  can be calculated from:  $F_c = M_{X_r} / N$ , where  $N$  is the number of data. Briefly this expression can be noted as:  $F_c = M / N$

A cumulative probability distribution (CPD) is a distribution of the probability  $P_c$  that an event ( $X$ ) is smaller than (or at most equal to) a reference value  $X_r$ . This probability is called cumulative probability.

The CPD can be estimated from the observed cumulative frequency distribution (CFD), which is then considered as a sample from the CPD. The cumulative probability is then defined as  $P_c = F_c$ .

When  $M = N$ , i.e.  $X = X_{\max}$  (the maximum value in the range), one finds  $F_c = 1$ . To create the possibility that there may be  $X$  values larger than  $X_{\max}$ , the value of  $F_c$  can be redefined as  $F_c = M / (N + 1)$ .

When the observed data of  $X$  are arranged in ascending order, the minimum first and the maximum last, and  $R_i$  is the rank number of the observation  $X_i$ , where the affix  $i$  indicates the serial number in the range of ascending data, then the cumulative frequency is  $F_c = R_i / (N + 1)$  and the estimated cumulative probability will be  $P_c = R_i / (N + 1)$ .

In literature, there are mathematical expressions of many different probability distributions and CPD's. One can try to fit a CPD to an observed CFD, optimizing the parameters of the CPD so that the sum of the squares of the differences  $P_c$  and  $F_c$  is minimum. The fitted CPD is then an estimate of the true CPD from which the CFD is considered a sample.

In theory, a sample is subject to a sampling error. When taking different samples, one may find different fitted CFD's (figure 1, next page). Therefore the estimated CPD may deviate from the true CPD from which the CFD is a sample. To estimate the possible error, a confidence analysis can be made.

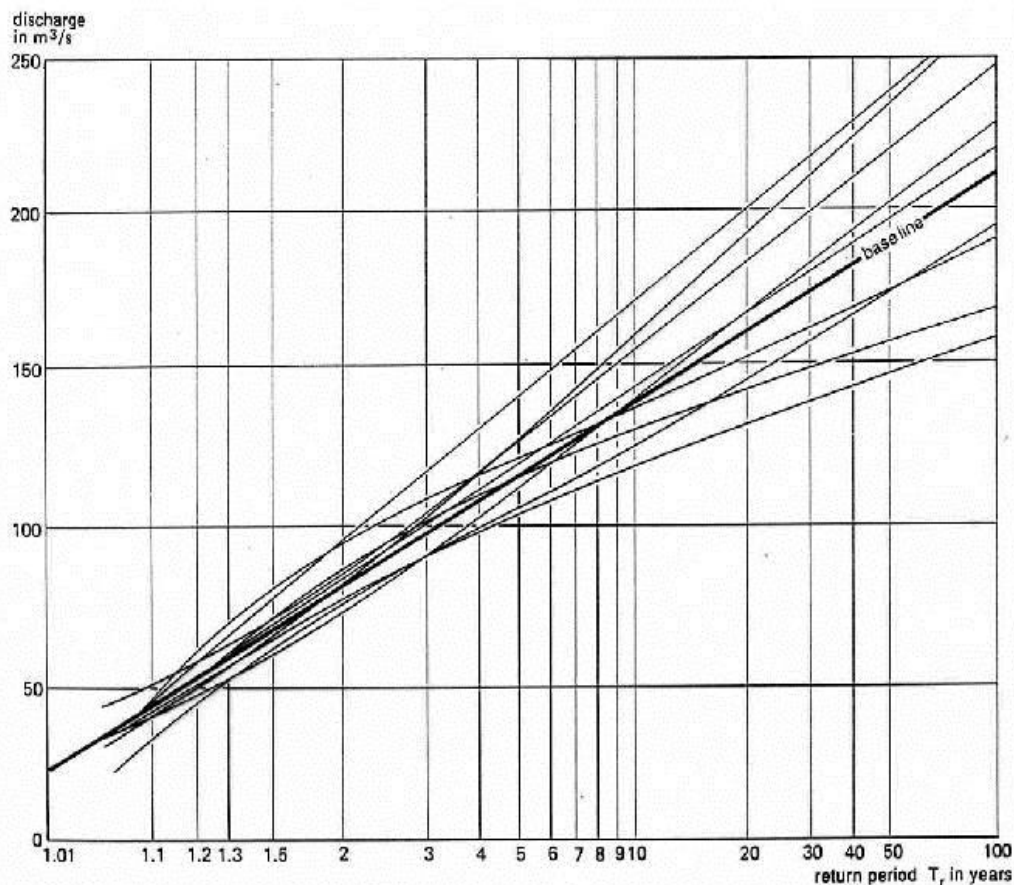


Figure 6.2 Frequency curves for different 50-year sample periods derived from the same base distribution (after Benson 1960)

Figure 1. Copy of figure 6.2 in <http://www.waterlog.info/pdf/freqtxt.pdf> illustrating different outcomes of samples from the same probability distribution

An observed value  $X$  can either be greater than a reference value  $X_r$  ( $X > X_r$ ) not greater than  $X_r$  ( $X \leq X_r$ ). The well known binomial probability distribution (BPD) deals with phenomena that either occur or do not occur. It gives the probability that an event actually occurs  $M$  times in a sample of size  $N$  when the probability of occurrence is  $P_0$ . Therefore, BPD can be used to estimate confidence intervals for the CPD estimated from a CFD.

An example: assuming the value  $X_r$  has a true cumulative probability  $P_{ct}$ , then one can estimate with the BPD what the probability is that, in a sample of size  $N$ , the number of occurrences is  $M+1, M+2 \dots N$  and  $M, M-1, M-2, \dots 0$ . Also, one can find the probability of the number of occurrences greater than a relatively large number  $M_g (>M)$  and less than a relatively small number  $M_s (\leq M)$ . Calling these probabilities  $P_{mg}$  and  $P_{ms}$  respectively, one can obtain the confidence interval  $P_{mg} - P_{ms}$  for  $M_g$  and  $M_s$ .

For example, assuming  $P_{mg} = 95\%$  and  $P_{ms} = 5\%$ , while the corresponding  $M$  values are  $M_{gc}$  and  $M_{sc}$ , one finds a 90% confidence interval as  $M_{gc} - M_{sc}$ . Converting these value into relative frequencies  $F_{gc} = M_{gc}/(N+1)$  and  $F_{sc} = M_{sc}/(N+1)$ , one obtains the confidence

interval in terms of cumulative of frequencies as  $F_{gc} - F_{sc}$ , which can serve as an estimate for the confidence interval in terms of  $P_c$  as  $P_{cg} - P_{cs}$ , where  $P_{cg}$  is the upper and  $P_{cs}$  the lower confidence limit.

The disadvantage of the BPD is its discontinuity (figure 2, next page), and the values of  $M_g$  and  $M_s$  are whole numbers (integers), so that the value of  $P_{mg}$  and  $P_{ms}$  for real values with a decimal part of must be estimated by interpolation (figure 3). This problem may be solved as follows.

The standard deviation of the binomial distribution as a fraction of  $N$  (the relative standard deviation) equals

$$SDb = \sqrt{P_c(1-P_c)/N} \quad (\text{eqn. 1})$$

When  $P_c = 0.5$ , the BPD is symmetrical and coincides with the corresponding normal distribution with mean  $N.P_c$  and a standard deviation equal to  $SDb$  (figure 4). Hence, here the confidence limits can be approximated by the same method used for the normal distribution as

$$P_{cg} = P_c + t.SDb \quad (\text{eqn. 2})$$

and

$$P_{cs} = P_c - t.SDb \quad (\text{eqn. 3})$$

where  $t$  is the value of the variable in Student's distribution corresponding to the level of certainty one wishes to achieve. For example, when  $N > 10$ , the  $t$ -value for 90% certainty is close to 1.7

A calculator for the  $t$ -distribution can be downloaded from <http://www.waterlog.info/t-tester.htm> and a calculator for the normal distribution from <http://www.waterlog.info/normdis.htm> With these one can verify that for larger values of  $N$  ( $>10$ ) the  $t$ -values and the  $x$ -values of the normal distribution at the same certainty level do not differ much relatively.

At  $P_c$  values other than 0.5, the binomial distribution becomes skew (figure 2). A simple method to take the skewness into account for the determination of the confidence limits is by making use of the  $P_c$  values themselves as a weight factor as follows:

$$\text{Upper confidence limit: } P_{cg} = P_c + 2(1-P_c).t.SDb = P_c + 3.4(1-P_c)SDb \quad (\text{eqn. 4})$$

$$\text{Lower confidence limit: } P_{cs} = P_c - 2P_c.t.SDb = P_c - 3.4 P_c.SDb \quad (\text{eqn. 5})$$

For higher values of  $P_c$  (e.g.  $P_c = 0.9$ ) the addition to  $P_c$  to obtain the  $P_{cg}$  becomes small while the subtraction from  $P_c$  to obtain the  $P_{cs}$  is getting larger. For lower values of  $P_c$ , the reverse is true. For  $P_c = 0.5$ , both equations give the same result and they equal the

expressions (eqn. 2 and eqn. 3) given before. Since the maximum value of  $P_c$  is unity, the value of  $P_{cg}$  can never become greater than unity, as required, and since the minimum value of  $P_c$  equals zero the value of  $P_{cs}$  can never be less than zero, as required.

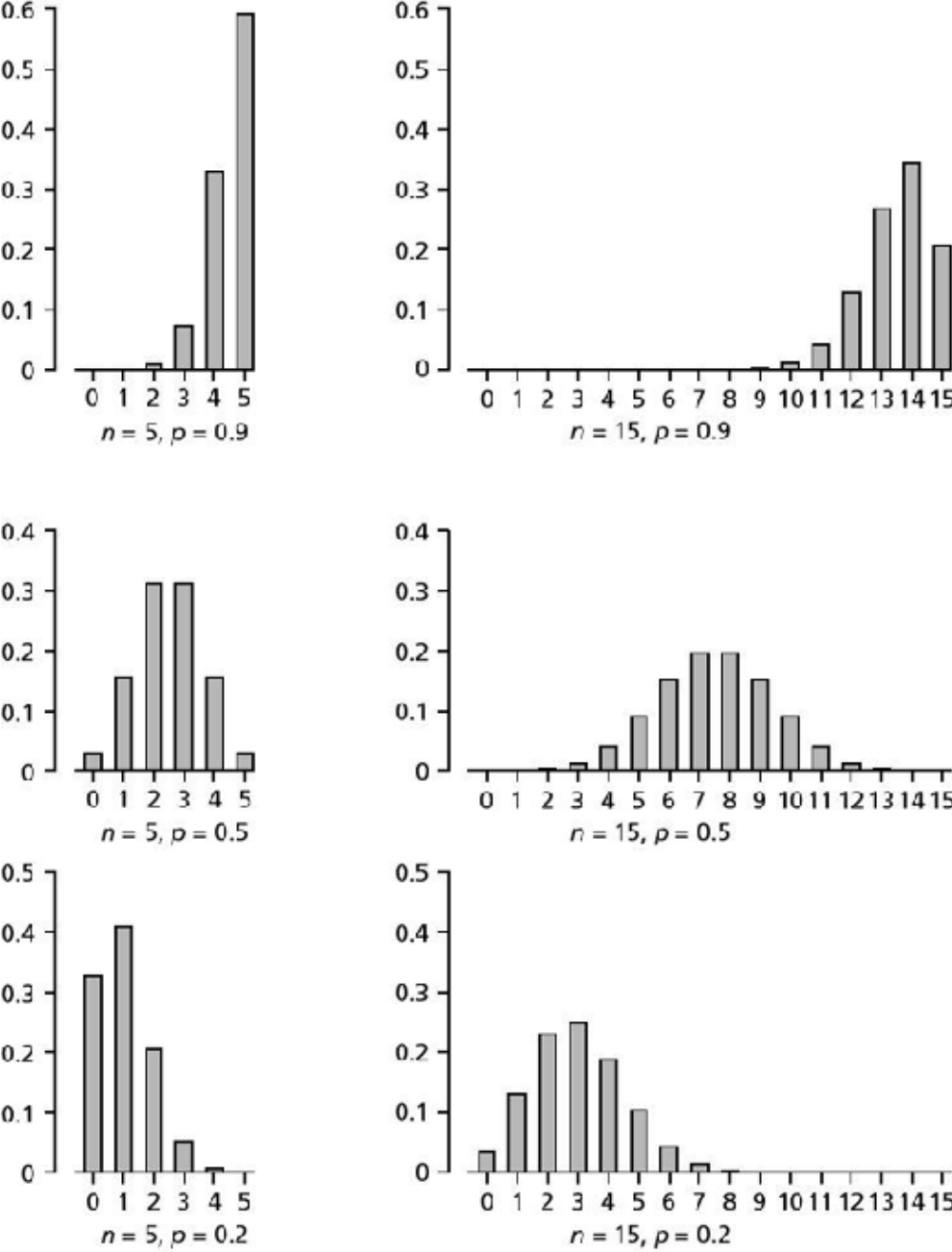


Figure 2. The binomial distribution is symmetrical when  $P_c$  ( $p$  in the figure) is 0.5 as can be seen in the middle two graphs. It is skew to the left at higher  $p$  values (top two graphs) and skew to the right at lower  $p$  values (bottom two figures)

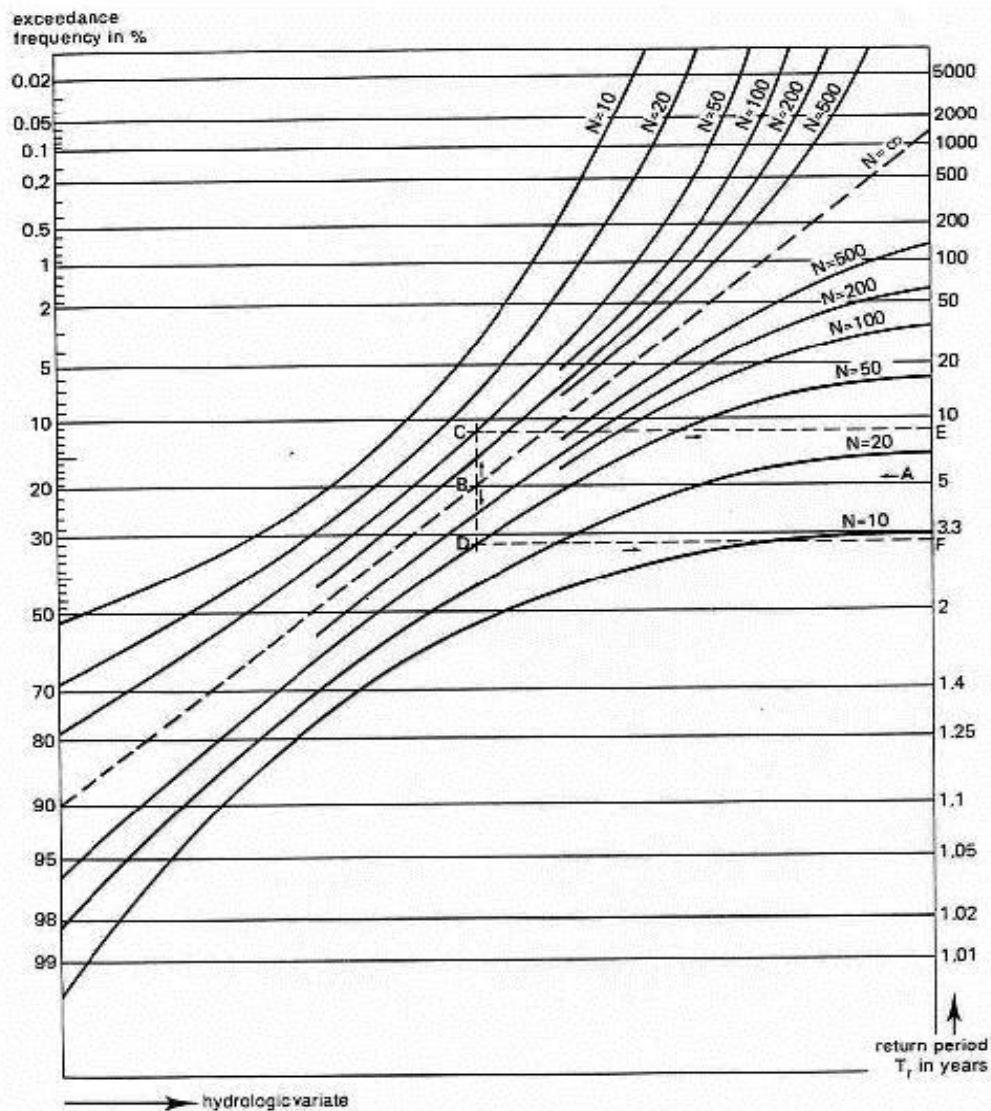


Figure 6.3 90% Confidence belts of frequencies for different values of sample size N

Figure 3. Copy of Figure 6.3 in <http://www.waterlog.info/pdf/freqtxt.pdf> illustrating a graphical procedure to arrive at 90% binomial confidence intervals.

Example of the use of figure 3. It is based on return periods ( $T_r$ ), but it can also be used for exceedance frequency ( $F_e = 1 - \text{cumulative frequency}$ ).  $T_r = 1/F_e$ .

- Enter the graph on the vertical axis with a return period of  $T_r = 5$ , (point A), and move horizontally to intersect the baseline, with  $N = \infty$ , at point B;
- Move vertically from the intersection point (B) and intersect the curves for  $N = 50$  to obtain points C and D;
- Move back horizontally from points C and D to the axis with the return periods and read points E and F;
- The interval from E to F is the 90% confidence interval of A, hence it can be predicted with 90% confidence that  $T_r$  is between 3.2 and 9 years. Nomographs for confidence intervals other than 90% can be found in literature (e.g. in Oosterbaan 1988).

Oosterbaan, R.J. 1988. Frequency predictions and their binomial confidence limits. In: Economic Aspects of flood control and non-structural measures, Proceedings of the Special Technical Session of the International Commission on Irrigation and Drainage (ICID), Dubrovnik, pp. 149-160.

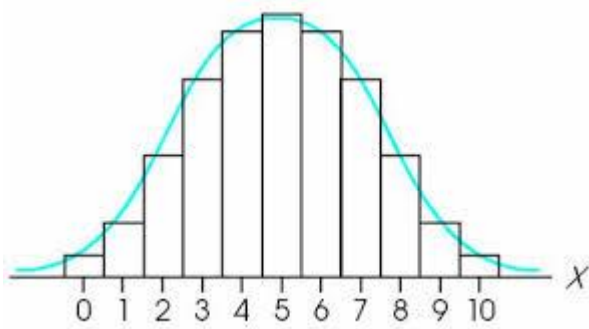


Figure 4. The symmetrical but discontinuous (stepwise) binomial distribution and the corresponding continuous (smooth) normal distribution with the same mean and standard deviation (blue curve).

### Confidence in interval of the return period.

The return period  $T$  is defined as

$$Tr = 1/(1-Pc) \quad (\text{eqn. 6})$$

Mathematically the upper ( $Ut$ ) and lower ( $Lt$ ) confidence limits of the return period  $T$  could be expressed as

$$Ut = 1 / (1 - Pcg) \quad (\text{eqn. 7})$$

$$Lt = 1 / (1 - Pcs) \quad (\text{eqn. 8})$$

where  $Pcg$  and  $Pcs$  are as defined in eqn. 4 and 5.

However, this leads to  $Ut$  values closer to the calculated/expected  $Tr$  values than  $Lt$ . One would expect a much higher  $Ut$  value, because the probability distribution of the return period should be skewed to the right. Therefore, arbitrarily, the following adjustment can be made:

$$D = Ut - Lt \text{ (giving the entire confidence range)} \quad (\text{eqn. 9})$$

$$Ut = Tr + 0.67 D \quad (\text{eqn. 10})$$

$$Lt = Tr - 0.33D \quad (\text{eqn. 11})$$

### Confidence interval of the $X$ value at a certain cumulative probability.

The cumulative probability ( $Pc$ ) can be written as a function  $f$  of  $X$ :

$$Pc = f(X) \quad (\text{eqn. 12})$$

Hence, X can be written as the inverse function  $f'$  of  $P_c$ :

$$X = f'(P_c) \quad (\text{eqn. 13})$$

With this inverse function, the upper confidence limit  $U_x$  and the lower limit  $L_x$  of X at  $P_c$  can be calculated as:

$$U_x = f'(P_{cg}) \quad (\text{eqn. 14})$$

$$L_x = f'(P_{cs}) \quad (\text{eqn. 15})$$

where  $P_{cg}$  and  $P_{cs}$  are as defined in eqn. 4 and 5.