

## NETA 1.0 HELP

### What is NETA

NETA is a freeware PC-based computer program that will run in the Windows environment. It was developed as a tool to calculate the 95% confidence limits for events that are distributed according to the Neyman type A law (Neyman 1939).

### How to run NETA

NETA is easy to use: it has only one window where the user enters the observed distribution.

Following entry of the distribution the user presses the COUNT button. NETA verifies if the distribution is Poissonian by the u-test as described by Edwards et al. (1979). The result of this test is given in the report that is accessible via the Report/view report menu. The value of the index of dispersion (var/mean) is displayed in the window Index of dispersion. In the case of a Poissonian distribution the 95% confidence limits (LCL - lower confidence limit and UCL - upper confidence limit) are calculated for a Poisson distribution as described by Deperas et al. (2007). When the distribution is not Poissonian NETA verifies if it is a Neyman type A distribution by performing a chi-square goodness-of-fit test. The result of this test is shown in the window chi2 G-o-F to Neyman. The confidence limits are calculated when the distribution is a Neyman type A. If the distribution is neither Poisson nor Neyman, an error message is displayed. In addition to the confidence limits NETA also gives some information about the statistics of the entered distribution. The entered data can be saved as a .dat or .txt file via the menu File/Save data as... The saved files can be loaded via the menu File/select file. The results can be printed or saved via the menu Report/view report.

The CLEAR DATA button is used to delete the data: NETA is now ready to perform a new calculation.

### The NETA algorithm

#### The sampling distribution of the mean

The Neyman type A distribution is discrete. Neyman himself defined it as a generalization of the Poisson probability distribution (Neyman 1939). The Neyman type A distribution is an example of compound distribution, e.g. a distribution  $P(X)$  of a random variable  $X$  which is a sum of  $Y$  variables  $Z_i$  ( $i = 1, 2, 3, \dots, Y$ ) following the same distribution  $P(Z)$  and the number  $Y$  is also a random variable with distribution  $P(Y)$ . If the distributions  $P(Z)$  and  $P(Y)$  are both Poissonian, the distribution  $P(X)$  is called "compound Poisson-Poisson" or Neyman type A distribution.

Let  $M$  be the random variable corresponding to the number of chromosomal aberrations per cell. Let  $E(M)$  and  $V(M)$  be the mean and the variance of a population, respectively. Let also  $m$  be the estimated mean of  $M$  and  $s^2$  be the estimated variance of  $M$ , both normalized for one hundred cells.

Then, if  $M$  follows the Neyman type A law, the probability to obtain  $k$  chromosomal aberrations in the cell is:

$$p(M = k) = \sum_{n=0}^{n=\infty} \frac{e^{-\lambda} \lambda^n}{n!} \cdot \frac{e^{-n\mu} (n\mu)^k}{k!} = P(\lambda, \mu, k) \quad (1)$$

where  $\mu = \frac{V(M)}{E(M)} - 1$  and  $\lambda = \frac{E(M)}{\mu}$ . The term  $\frac{V(M)}{E(M)}$  is called the index of dispersion (ID).

The characteristic function  $\phi(t)$  of the Neyman type A random variable  $M$  is given by

$$\varphi(t) = \exp[-\lambda + \lambda \exp(\mu(e^{it} - 1))] \quad (2)$$

Hence the characteristic function  $\Phi(t)$  of a random variable  $S_r$  being a sum of  $r$  Neyman type A variables is

$$\Phi(t) = \varphi^r(t) = \exp[-r\lambda + r\lambda \exp(\mu(e^{it} - 1))]. \quad (3)$$

Thus the distribution of the variable  $S_r$  is simply the same one, as for the random variable  $M$  (Eq. 1), but with the modified parameter  $\lambda$ :

$$p(S_r = d) = P(r\lambda, \mu, d), \quad (4)$$

where  $d = 0, 1, 2, \dots$ .

Since the mean  $m = \frac{S_r}{r}$  of a random sample of size  $r$  is an estimator of  $E(M)$ , the sampling distribution of the mean for a fixed  $V(M)$  is given by

$$p(m = \frac{d}{r}) = P(r\lambda, \mu, d). \quad (5)$$

where  $\mu = \frac{s^2}{m} - 1$  and  $\lambda = \frac{m}{\mu}$ .

Hence

$$p(m = \frac{k}{r}) = \sum_{n=0}^{\infty} \frac{e^{-r\lambda} (r\lambda)^n}{n!} \cdot \frac{e^{-n\mu} (n\mu)^k}{k!}. \quad (6)$$

The determination of the confidence interval for  $E(M)$  is based on the idea introduced by Jerzy Neyman. The first step of the method consists in searching for the integers  $N_1$  and  $N_2$ , such as

$$\sum_{k=N_1}^{k=N_2} p(m = \frac{k}{r}) = 1 - \alpha, \quad (7)$$

where  $1 - \alpha$  is the confidence level (the confidence level can also be written as  $100(1 - \alpha)\%$  confidence level). More precisely, the integers  $N_1$  and  $N_2$  must be such as

$\sum_{k=0}^{N_1} p(m = \frac{k}{r}) = \frac{\alpha}{2}$  and  $\sum_{k=0}^{N_2} p(m = \frac{k}{r}) = 1 - \frac{\alpha}{2}$ . Then the lower and upper confidence limits for the mean for the fixed variance  $V(M)$  or, in other words, for the fixed index of dispersion are given by  $m_{low} = \frac{N_1}{r}$  and  $m_{up} = \frac{N_2}{r}$ , respectively. It would be

enough to determine the confidence interval, provided that the exact value of the variance is known. Otherwise, the influence of the uncertainty of that parameter has to be taken into account. This results in broadening of the confidence interval.

The calculation algorithm

A significant part of the algorithm is the computing of the Neyman type A distribution. Here, the main problem is the infinite sum in the formula. Thus, in order to perform a computation, an approximation must be made, even for the reason of the limited precision of a computer representation of numbers. The infinite sum is replaced by a finite sum, from  $n=0$  to  $n=C$  and

$$C = \max \left\{ n : \left( \frac{e^{-r\lambda} (r\lambda)^n}{n!} \cdot \frac{e^{-n\mu} (n\mu)^k}{k!} \right) < \delta \right\}, \quad (8)$$

where  $\delta$  is the minimal nonzero value in a computer representation of numbers or an arbitrary precision limitation and the expression in parentheses is simply a term of the infinite series in 6.

Hence, the approximation of sample distribution 6 is

$$p\left(m = \frac{k}{r}\right) = \sum_{n=0}^{n=C} \frac{e^{-r\lambda} (r\lambda)^n}{n!} \cdot \frac{e^{-n\mu} (n\mu)^k}{k!}. \quad (9)$$

The Neyman type A distribution depends on two parameters,  $\lambda$  and  $\mu$ . These parameters are functions of the mean  $E(M)$  and the variance  $V(M)$  of the random variable  $M$ . The estimation  $m$  of the mean is given by  $m = \sum_{i=0}^{i_{\max}} p_i x_i$  and the estimation

$s^2$  of the variance is calculated with the König theorem <sup>(15)</sup>:  $s^2 = \sum_{i=0}^{i_{\max}} p_i x_i^2 - m^2$ , where

$x_i$  is a value of the random variable  $M$ ,  $p_i$  is the frequency of  $x_i$  (i.e. the value for  $x_i$  of an empirical distribution).

Then, the Neyman type A distribution as the sampling distribution of the mean for the fixed  $s^2$  can be approximated with

$$p\left(m = \frac{k}{r}\right) = \sum_{n=0}^{n=C} \frac{e^{-r\lambda_s} (r\lambda_s)^n}{n!} \cdot \frac{e^{-n\mu_s} (n\mu_s)^k}{k!}, \quad (10)$$

where  $\mu_s = \frac{s^2}{m} - 1$  and  $\lambda_s = \frac{m}{\mu_s}$  are obtained from a sample.

A simple computation algorithm using formula 10 for determination of lower and upper confidence limits for the mean with the fixed variance (namely the sample variance) may be written as follows:

F = 0

N = 0

while (F <  $\frac{\alpha}{2}$ )

F = F +  $p\left(m = \frac{N}{r}\right)$

N = N + 1

endwhile

$m_{\text{low}} = \frac{N-1}{r}$

while (F ≤ (1 -  $\frac{\alpha}{2}$ ))

F = F +  $p\left(m = \frac{N}{r}\right)$

N = N + 1

endwhile

$m_{\text{up}} = \frac{N-1}{r}$

Numerical limitation of the algorithm

A problem occurs for large values of the confidence limits. By construction the algorithm is convergent and provides values for confidence limits on the condition that the values of the numerical function stay in the range of the machine on which it is running. We discovered that the algorithm enters an infinite loop (leading to a hang-up) when the number of samples is greater than 127. For all tested

distributions, the problem occurred for number of samples bigger than 127. An analysis of the algorithm variables showed that the values taken by the p function become decrease continuously until the null value is reached. Once this occurs it is impossible for the F value to increase to  $1-\alpha/2$ , and the process stays locked in the infinite loop mentioned above. This computer artefact is linked to the limited precision of number representation used for numerical calculations. It was mentioned above when introducing the cut-off value C in the formula 10.

The problem can be solved by reducing the precision of the computer calculations by assuming that "zero is greater than zero" or simply by applying the central limit theorem according to which, when the number of observations is high enough, the distribution of the mean of observed events tends to follow the Gaussian law {Réfrégier, 2002 2108 /id}. In such cases it is possible to estimate the confidence intervals for the mean  $m$ , when the variance is unknown, with

$$[LCL, UCL] = \left[ \bar{x} - 1.96 \frac{\sqrt{s^2}}{\sqrt{N}}, \bar{x} + 1.96 \frac{\sqrt{s^2}}{\sqrt{N}} \right],$$

where  $N$  is the number of cells in the sample,  $s^2$  is the estimated variance and 1.96 is the t value read in the Student table for  $N > 120$  and the confidence level equal to 0.95.

#### References

Deperas J., Szłuińska M., Deperas-Kaminska M., Edwards, A., Lloyd D., Lindholm C., Romm H., Roy L., Moss R., Morand J., and Wojcik A. (2007) CABAS - a freely available PC program for fitting calibration curves in chromosome aberration dosimetry. Radiation Protection Dosimetry, in press.

Edwards A. A., Lloyd D. C., and Purrot R. J. (1979). Radiation induced chromosome aberrations and the Poisson distribution. Radiation and Environmental Biophysics 16: 89-100.

Neyman J. (1939). On a new class of "contagious" distribution, applicable in entomology and bacteriology. Am. Math. Stat. 10: 35-55.

Réfrégier P. (2002) Théorie du bruit et applications en physique, Hermès Sciences Publications, Paris.