# On estimation of the exponential distribution parameter under conditions of small statistics and observation interval.

Zlokazov V.B.
FLNP, JINR, RUSSIA.
E-mail: zlokazov@nf.jinr.ru

Some typical problems:

- Is a newly synthesized chemical element really existing, and, if so what halflive it has and how long should be the observation interval?.
- There is a complicated chain of "mother - daughter" decays. Are the times of the obsevation of each enough to determine their halflives?
- Extreme case. An object is supposed to be unstable but none event of its decay was registered. What can we say here about this?

Let events $\xi$ be given, subject to

$$P(\xi < t) = 1 - exp(-t/T), \;\; t \in [0, \infty), \;\; (1)$$

and let the interval of observation be a time interval $[0, B]$, and $Q = (t_1, t_2, ..., t_m)$ - a set of observed $\xi$ in the interval $[0, B]$;

The common goal of the analysis of the data $Q$ is the evaluation of $T$. If $B = \infty$, then the usual estimatators (maximum likelihood, moment, etc.) give the sample mean as the best estimate of $T$, which is efficient, consistent, unbiased, sufficient, etc. due to the excelent asymptotic properties of the sample means. However, a specific feature of innovative physical experiments is the fact that often the quantity $B$ is much smaller than the quantity $T$. And the size of the sample $m$ is often very small and doesn't increase. This diminishes the information volume of the data $Q$.

First, let us find the expectation and the variance of $\xi$ if the observation interval is only $[0, B]$. For the probability density we have

$$p(t) = \begin{cases} e_t/(T \cdot (1 - e_B)) & \text{if } t \in [0, B]; \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

If $B << T$ $p(t) = e_t/B \cdot \chi_{[0,B]}(t)$,
where $e_t = exp(-t/T)$.

Omitting the further details, we have

$$\hat{E}\xi = T - B \cdot e_B/(1 - e_B) \quad (3)$$

$$\hat{V}\xi = T^2 - B^2 \cdot e_B/(1-e_B) + B^2(e_B/(1-e_B))^2 \quad (4)$$

In all the cases the likelihood function is

$$L(T) = \prod_{j=1}^{m} p(t_j), \quad (5)$$

and we can try to get the maximum likelihood estimates of $T$ from maximum of (5) with respect to $T$. It is obvious, that if $B << T$ the maximum is at $T = \infty$, if, at least, one $t_j$ gets in the interval $[0, B]$.

Let us assume that at $t = 0$ there were $N_0$ possible events, and let us introduce an additional quantity:
$n$ - the number of events, registered in $[0, t]$;
We got an additionald parameter to be estimated $= N_0$.
$n$ is described by the multibinomial distribution $F(n = k, t)$, However, under very common conditions (transitivity and homogenuity of the distribution function, and rareness of the events) $F$ will asymptotically tend to the Poisson distribution

$$F(n = k, t) = (at)^k/k!\, exp(-at), \;\; k = 0, 1, ..., \infty, \tag{6}$$

the parameter of which also can be determined on the basis of our initial exponential distribution

$$at = N_0 \cdot (1 - e_t), \tag{7}$$

The quantity $at$ is the mean (expectation) and the variance of $n$ at the same time. For $[0, B]$ $\hat{E}n = \hat{V}n = N_0 \cdot (1 - e_B)$.

The Poisson distribution is more suitable for analytical operations.
Still, under poor statistics and/or small observation interval $[0, B]$, $B << T$) the chances of a successful solution of this problem are very small.
The poor statistics means, in particular, the smallness of $aB$ and its equavalence to the mean quadratic deviation of $n$, which is $\sqrt{aB}$;
The distribution function of $n$ on condition that the interval of the observation is $[0, B]$, does not depend explicitely on $B$:

$$P(n = k) = (at)^k/k!\ exp(-at) \quad \text{if } t \in [0, B]$$

For $m$ events in $[0, B]$ the equation of the maximum likelihood is $m/a - B = 0$;
from this we get: $\hat{a} = m/B$.

One can try to use the moment estimator for the evaluation of $T$ from (3) and (4). The mean of the sample $Q$ will belong to $[0, B]$ and will be strongly biased with respect to the true value $T$. The equation for the moment estimates is:

$$\sum_{j=1}^{m} t_j/m = T - B \cdot e_t/(1 - e_t); \qquad (8)$$

Already at $T > 4B$ this function is practically constant, and there are no chances again to find the root of (8) with the acceptable accuracy.

The moment estimate of $N_0$ is obtained from

$$m = N_0(1 - e_B)$$

If $B << T$, $\hat{N}_0 = mT/B$, i.e., depends on unreliable estimate of $T$.

We can proceed as follows. Let us denote the length of the observation interval as $2B$, and introduce two random quantities: $n_1$ and $n_2$ - sums of registered decays in the intervals $[0, B]$ and $[B, 2B]$, respectively. It is obvious that

$$\hat{E}n_1 = N/(1 - e_B),\ \hat{E}n_2 = N/(e_B - e_{2B})$$

Let $r = \frac{n_1}{n_2}$. We have

$$\hat{E}r = exp(B/T). \tag{9}$$

On the basis of (9) one can build an estimator of $T$:

$$\hat{T} = B/ln(r). \tag{10}$$

The practical use of (10) is not successful in all the cases: the probability that $n_1 = n_2$ is not equal to zero, and it means that the expectation of (10) is not bounded.

It is obvious that for the analysis only those samples are admissible which more or less look as exponential curves. For instance, we can make use of some criterion for testing the statistical significance of an inequality $n_2 < n_1$, e.g. this one:

$$n_1 > n_2 + k \cdot \sigma(n_2). \qquad (11)$$

where $k$ is any number, and $\sigma$ - deviation function.

For the Poisson distributed $n_2$ we have $\sigma(n_2) = \sqrt{(}N_0(e_B - e_{2B}))$, and using this from (11) we can derive the formula

$$N \geq k^2 \cdot e_B/(1 - e_B)^3, \qquad (12)$$

and from it the restrictions on

- the level of the statistics $N_0$ at $B/T$ given;
- the length of the observation interval $B$ at $N_0$ given,

which provide for the success of the analysis of such data.

In order to determine $B$ it is necessary to solve the cubic equation $N_0(1-z)^3 = k^2 z$, i.e. find its positive root $z_0$, and then determine $B$ from the condition $exp(B/T) = z_0$: i.e. $B = T \cdot ln(z_0)$.

Strictly speaking, the condition like (12) is necessary also for the above - considered estimation of $T$ at the known $N$, since with a non zero-valued probability the sample can fail to contain any registered decays. But it is a theme of a special consideration [2].

If $T >> B$ the formula 12) looks simpler: $N_0 \geq k^2((T/B)^3 - (T/B)^2)$.

Thus, we see that for a successful estimation of the parameter $T$ requirements on $N_0$ and the ratio $T/B$ are very severe. In the innovative experiments the investigator has often no possibility to control both the factors. A question arises: what can be done in this case?

Here an idea of an estimate of a lower parameter bound instead of parameter itself is very fruitful. The lower parameter bound is a quantity, which with a certain (calculable) probability is less than $T$, but greater than the length of the observation interval $B$.
In our case such an estimator can be obtained, e.g., from such a consideration: the estimator $\hat{T} = B/ln(\frac{n_1}{n_2})$ is used only if the condition $n_1 > n_2 + k \cdot \sigma(n_2)$ with a given $k$ holds.
Such estimates of $T$ are normally lower than the true parameter value since they are based only on data with a sufficiently stiff slope and will range between some minimal $T_l$ and the true value of $T$. Just this $T_l$ can be taken as the lower bound of the parameter $T$.

Dividing of the event set into the 2 parts is only a particular case of dividing it in several parts, and summing events in this parts with a goal to check whether the obtained histogram statistically can be described by the distribution function

$$P(\xi < t) = 1 - exp(-t/T),$$

and if so then the analysis of the data has chances to be succesful. Certainly, it also gives a lowered estimate of $T$ parameter.
2 parts are preferable because all this is appropriate for the case when the data statistics is small.

We can carry out several tests to show how the formulae

$$\hat{T} = B/ln(\frac{n_1}{n_2})$$

on condition $n_1 > n_2 + k \cdot \sigma(n_2)$ work. A random number generator subject to an exponential distribution $P(t, T_0)$ produced series of decays within an interval $[0, B_0]$ $B_0 << T_0$. Then for a given series and $n_1, n_2$ the above condition was checked and, if satisfied, the above estimate of $T_0$ was built.
$B_0 = 20$, $T_0 = 200$. The number $N_0$ at these $B_0, T_0$ should be greater than 1050.

The following table contains the results obtained. Average value of $T_{est} = 157$, $\sigma = 163$.

| N1 | N2 | $n_1/n_2$ | $T_{est}$ |
|---|---|---|---|
| 40 | 33 | 1.22 | 103.91 |
| 198 | 161 | 2.92 | 96.64 |
| 327 | 263 | 3.95 | 91.78 |
| 336 | 312 | 1.36 | 269.51 |
| 389 | 336 | 2.89 | 136.46 |
| 468 | 403 | 3.24 | 133.66 |
| 484 | 442 | 2.00 | 220.08 |
| 644 | 558 | 3.64 | 139.43 |
| 641 | 521 | 5.26 | 96.44 |
| 664 | 619 | 1.81 | 284.59 |
| 731 | 635 | 3.81 | 141.96 |
| 777 | 693 | 3.19 | 174.66 |

The tests showed that if $B << T$, and the statistics is small, the practical chances to get a good estimate of the parameter $T$ are very small. An alternative to this method is the regression analysis: building a distribution $s(t)$ from the events $t_j,\ j = 1, m$ and fitting it by the curve $Nexp(-B/T)$, where $N, T$ are parameters of interest. If the above conditions hold, this method gives very lowered estimates of the parameter $T$ too. Besides, this method needs large statistics of the events $t_j$.

Summarizing one can say that the situation $B << T$ is a bigger evil than the small statistics of data: even if the statistics is large, still the quality of the estimates of the parameter $T$ will be very poor. In geology and other sciences often the estimation of half-lifes of the processes which last millions of years must use the apriori knowledge of the parameter $N_0$ - some value of $N$ at the time $t = 0$ and then if the observation time was $B$, which equals only to years one can get rather reliable estimate of $T$ from

$$n = N_0 exp(-B/T)$$

where $n$ is the number of the events $T_j$ in the interval $[0, B]$. Otherwise, the estimates of $T$ will be completely unreliable.