Seminars in Applied Statistics for Radiation Cytogenetics and Biodosimetry

Volodymyr Vinnikov
S.P. Grigoriev Institute
for Medical Radiology and Oncology
of the National Academy
of Medical Science of Ukraine

Seminar II. Description of datasets

DISCLAIMER

This is how I work with data. Why I do it in this way? – Because I learned it from either textbooks or my teachers. For detailed explanations or alternative methods, please, refer to textbooks on statistics or to professional mathematicians.

Sometimes I’ll talk about very simple or apparent things, essentially based on common sense and common knowledge.

Contents. Descriptive statistics for the presentation of research data in Radiation Cytogenetics and Biodosimetry. Types of data. Variation statistics: distributions, mean, mode, median, quartiles, dispersion, standard deviation, standard error, confidence intervals. ~~“Normalization” of data.~~ Qualitative data: frequencies. Cytogenetic damage yield: individual data *versus* group data.

Virtual dataset: results of the cytogenetic investigation in a group of occupationally exposed persons

Types of data

28 individuals. 9♀ & 19 ♂. Age ranged 22 – 75 years. Physical radiation dose records are available for 21 individuals. Cytogenetic analysis performed with numbers of metaphases scored per individual ranging from 50 to 1000.

Quantitative data

- Discrete values = count data (0, 1, 2, 3, ….): cells scored, aberrations found.

- Continuous values (results of measurements): physical dose in mGy, concentrations in mg/ml, etc. Yields and frequencies.

Qualitative data

- Categories (A, B, C): professional groups (admins, guards, workers); mitoses (M_1, M_2, M_3, M_{4+})

- Alternatives (“yes” or “no”): ♀ or ♂; monitored (Dosimetry +) or non-monitored (Dosimetry -) workers; aberrant cells or non-aberrant cells.

Assembling a group dataset

Step 1. Estimate individual aberration yields:
N – cells scored;
X – aberrations found;
X / N = Y (Yield per cell)

Step 2. Build variation series for aberration yield (Y); sort from low to high

Nr	Individial	Cells scored	Aberrations found	Yield
1	AAA	1000	0	0.0000
2	AAB	750	0	0.0000
3	AAC	500	0	0.0000
4	AAD	500	1	0.0020
5	AAE	800	2	0.0025
6	AAF	100	0	0.0000
7	GAA	100	1	0.0100
8	GAB	100	2	0.0200
9	GAC	100	0	0.0000
10	GAD	150	1	0.0067
11	GAE	200	1	0.0050
12	GAF	50	1	0.0200
13	WAA	1000	0	0.0000
14	WAB	1000	1	0.0010
15	WAC	1000	2	0.0020
16	FPA	200	57	0.2850
17	WAD	1200	3	0.0025
18	WAE	800	4	0.0050
19	RAA	500	4	0.0080
20	RAB	500	3	0.0060
21	RAC	500	2	0.0040
22	FPB	200	69	0.3450
23	RAD	500	1	0.0020
24	RAE	200	0	0.0000
25	RAF	1000	7	0.0070
26	RAG	236	5	0.0212
27	RTA	117	3	0.0256
28	RTB	67	1	0.0149

Nr	Individial	Cells scored	Aberrations found	Yield
1	AAA	1000	0	0.0000
13	WAA	1000	0	0.0000
2	AAB	750	0	0.0000
3	AAC	500	0	0.0000
24	RAE	200	0	0.0000
6	AAF	100	0	0.0000
9	GAC	100	0	0.0000
14	WAB	1000	1	0.0010
15	WAC	1000	2	0.0020
4	AAD	500	1	0.0020
23	RAD	500	1	0.0020
17	WAD	1200	3	0.0025
5	AAE	800	2	0.0025
21	RAC	500	2	0.0040
18	WAE	800	4	0.0050
11	GAE	200	1	0.0050
20	RAB	500	3	0.0060
10	GAD	150	1	0.0067
25	RAF	1000	7	0.0070
19	RAA	500	4	0.0080
7	GAA	100	1	0.0100
28	RTB	67	1	0.0149
8	GAB	100	2	0.0200
12	GAF	50	1	0.0200
26	RAG	236	5	0.0212
27	RTA	117	3	0.0256
16	FPA	200	57	0.2850
22	FPB	200	69	0.3450

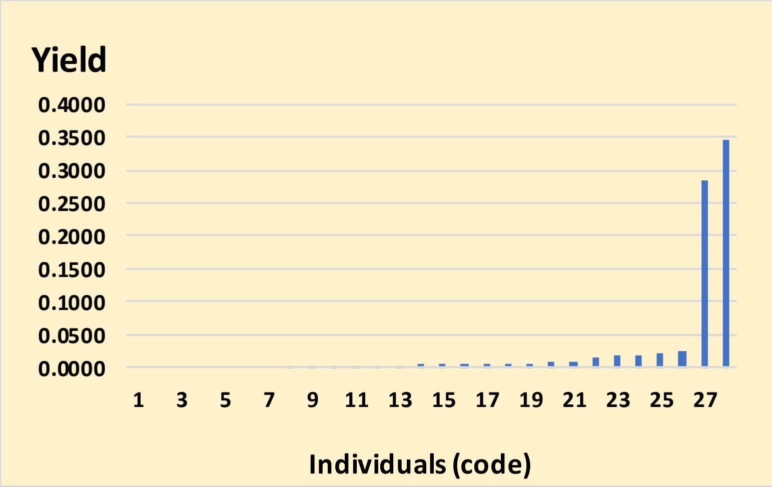
Is the group homogeneous in regard of individual aberration yields?

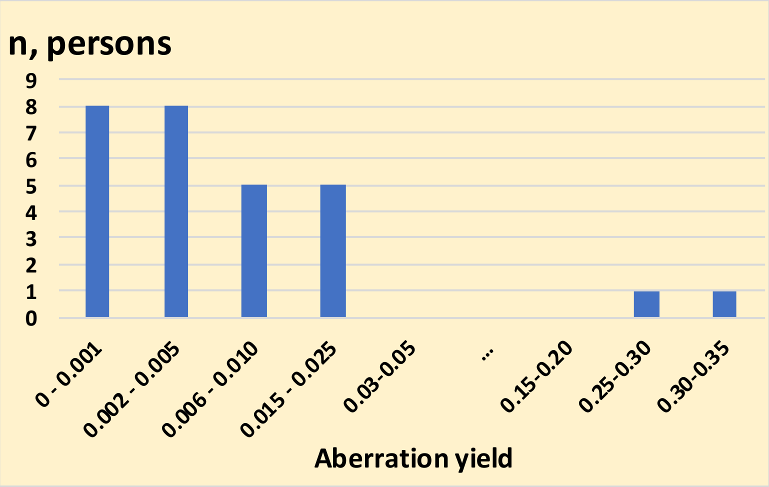
Testing a group for homogeneity

Step 1. Build a distribution of individual aberration yields and check it for the presence of “abnormal” values.

Step 2. Calculate a <u>weighted</u> mean of Y for the group (Y_{wm}).

Step 3. Perform a rough χ^2 test, using weighted group mean as an expected value.





Individual statistical weight is the number of cells scored for person (N_i), divided by the mean number of cells scored per person in the group. However, a final formula is simple:

$Y_{wm} = (Y_1 x N_1 + Y_2 x N_2 + \ldots + Y_i \times N_i) / (N_1 + N_2 + \ldots N_i) = X_{total} / N_{total}$.

In our group $X_{total} = 171$, $N_{total} = 13370$; $Y_{wm} = 0.0128$.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f - f')^2}{f'} = \sum_{i=1}^{k} \left(\frac{d^2}{f'} \right)$$

f – observed number of aberrations;

f’ – expected number of aberrations;

$f' = Y_{wm} \times N_i$

d – difference between observed and expected numbers;

Σ is arithmetic sum.

χ^2 test

Nr	Individial	Cells scored (N_i)	Aberrations observed	Yield	Expected aberrations Y_{wm} x N_i	Difference (d) between observed and expected	d^2	d^2 / Expected
1	AAA	1000	0	0.0000	12.8	-12.8	163.84	12.8
13	WAA	1000	0	0.0000	12.8	-12.8	163.84	12.8
2	AAB	750	0	0.0000	9.6	-9.6	92.16	9.6
3	AAC	500	0	0.0000	6.4	-6.4	40.96	6.4
24	RAE	200	0	0.0000	2.56	-2.56	6.5536	2.56
6	AAF	100	0	0.0000	1.28	-1.28	1.6384	1.28
9	GAC	100	0	0.0000	1.28	-1.28	1.6384	1.28
14	WAB	1000	1	0.0010	12.8	-11.8	139.24	10.87813
15	WAC	1000	2	0.0020	12.8	-10.8	116.64	9.1125
4	AAD	500	1	0.0020	6.4	-5.4	29.16	4.55625
23	RAD	500	1	0.0020	6.4	-5.4	29.16	4.55625
17	WAD	1200	3	0.0025	15.36	-12.36	152.7696	9.945938
5	AAE	800	2	0.0025	10.24	-8.24	67.8976	6.630625
21	RAC	500	2	0.0040	6.4	-4.4	19.36	3.025
18	WAE	800	4	0.0050	10.24	-6.24	38.9376	3.8025
11	GAE	200	1	0.0050	2.56	-1.56	2.4336	0.950625
20	RAB	500	3	0.0060	6.4	-3.4	11.56	1.80625
10	GAD	150	1	0.0067	1.92	-0.92	0.8464	0.440833
25	RAF	1000	7	0.0070	12.8	-5.8	33.64	2.628125
19	RAA	500	4	0.0080	6.4	-2.4	5.76	0.9
7	GAA	100	1	0.0100	1.28	-0.28	0.0784	0.06125
28	RTB	67	1	0.0149	0.8576	0.1424	0.020278	0.023645
8	GAB	100	2	0.0200	1.28	0.72	0.5184	0.405
12	GAF	50	1	0.0200	0.64	0.36	0.1296	0.2025
26	RAG	236	5	0.0212	3.0208	1.9792	3.917233	1.296753
27	RTA	117	3	0.0256	1.4976	1.5024	2.257206	1.507215
16	FPA	200	57	0.2850	2.56	54.44	2963.714	1157.701
22	FPB	200	69	0.3450	2.56	66.44	4414.274	1724.326
	Total	13370	171	Y_{wm} =0.0128				Σ=2991.476
						Degrees of freedom		
						k = n-1		27
						Chi-squared critical		40,11
							Exact p	0

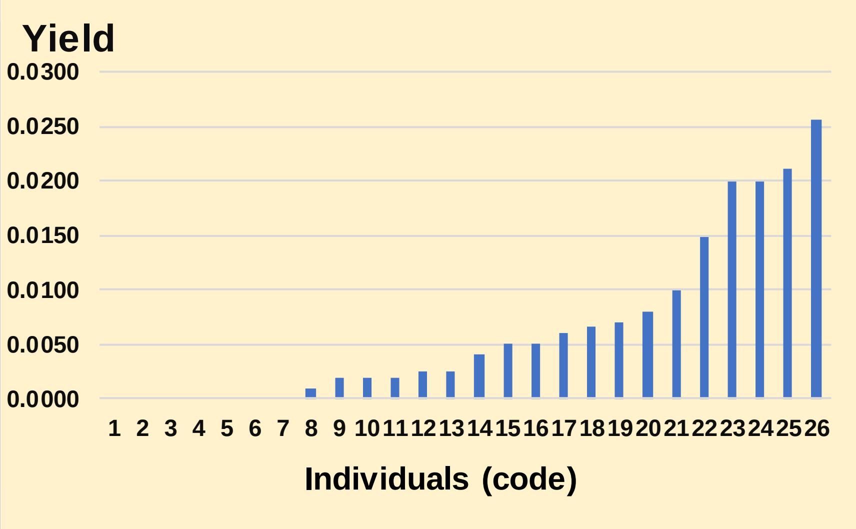
The resultant $\chi^2 > \chi^2_{stat}$, thus our group is not homogenous.
Probably, because of two individual cases with very high aberration yields (FPA-16 and FPA-22).
Let's remove them.

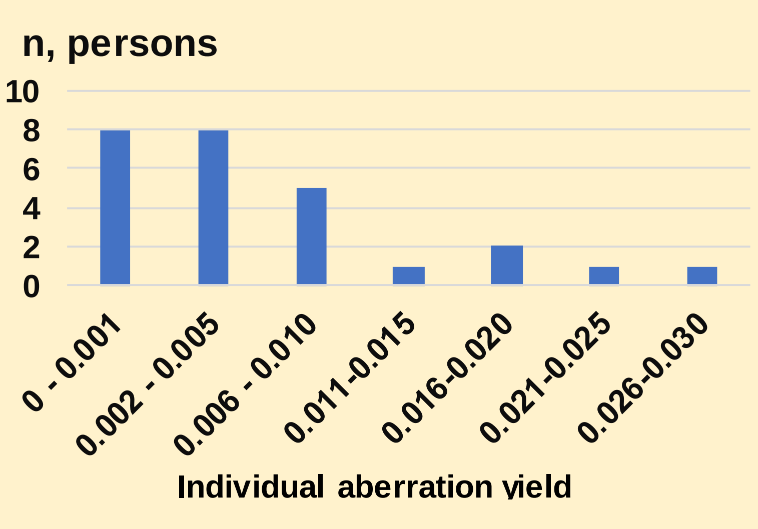
Repeat testing in the corrected group

Step 1. Build a distribution of individual aberration yields.

Step 2. Calculate <u>new</u> weighted mean of Y for the group (Y_{wm}).

Step 3. Perform a rough χ^2 test again, using new weighted group mean as an expected value.





Now the distribution looks smoother. The range of values became much smaller. No gaps between classes.

After removing two cases (FPA-16 and FPA-22) we have new values in our group:

$n = 26$; $X_{total} = 45$; $N_{total} = 12970$; $Y_{wm} = X_{total} / N_{total} = 0.0035$

In future data analysis two cases with very high aberration yields (FPA-16 and FPA-22) will be presented separately, on individual basis.

χ^2 test repeated

Nr	Individial	Cells scored (N_i)	Aberrations observed	Yield	Expected aberrations Y_{wm} x N_i	Difference (d) between observed and expected	d^2	d^2 / Expected
1	AAA	1000	0	0.0000	3.5	-3.5	12.25	3.5
13	WAA	1000	0	0.0000	3.5	-3.5	12.25	3.5
2	AAB	750	0	0.0000	2.625	-2.625	6.890625	2.625
3	AAC	500	0	0.0000	1.75	-1.75	3.0625	1.75
24	RAE	200	0	0.0000	0.7	-0.7	0.49	0.7
6	AAF	100	0	0.0000	0.35	-0.35	0.1225	0.35
9	GAC	100	0	0.0000	0.35	-0.35	0.1225	0.35
14	WAB	1000	1	0.0010	3.5	-2.5	6.25	1.785714
15	WAC	1000	2	0.0020	3.5	-1.5	2.25	0.642857
4	AAD	500	1	0.0020	1.75	-0.75	0.5625	0.321429
23	RAD	500	1	0.0020	1.75	-0.75	0.5625	0.321429
17	WAD	1200	3	0.0025	4.2	-1.2	1.44	0.342857
5	AAE	800	2	0.0025	2.8	-0.8	0.64	0.228571
21	RAC	500	2	0.0040	1.75	0.25	0.0625	0.035714
18	WAE	800	4	0.0050	2.8	1.2	1.44	0.514286
11	GAE	200	1	0.0050	0.7	0.3	0.09	0.128571
20	RAB	500	3	0.0060	1.75	1.25	1.5625	0.892857
10	GAD	150	1	0.0067	0.525	0.475	0.225625	0.429762
25	RAF	1000	7	0.0070	3.5	3.5	12.25	3.5
19	RAA	500	4	0.0080	1.75	2.25	5.0625	2.892857
7	GAA	100	1	0.0100	0.35	0.65	0.4225	1.207143
28	RTB	67	1	0.0149	0.2345	0.7655	0.5859903	2.498892
8	GAB	100	2	0.0200	0.35	1.65	2.7225	7.778571
12	GAF	50	1	0.0200	0.175	0.825	0.680625	3.889286
26	RAG	236	5	0.0212	0.826	4.174	17.422276	21.09234
27	RTA	117	3	0.0256	0.4095	2.5905	6.7106903	16.38752
	Total	12970	45	Y_{wm} =0.0035				Σ=77.66566
						Degrees of freedom		
						k = n-1		25
						Chi-squared critical		37,65
						Exact p		2.64E-07

The resultant χ^2 still exceeds χ^2_{stat}, thus our group is still not fully homogenous. However, there are no apparently “abnormal” cases. Let’s take a risk to characterize it as an entity.

Variation statistics – trying to quantify the parameters of the distribution.

Mean is the most important parameter.

Unweighted mean yield of aberrations:

$$Y_{uwm} = (Y_1 + Y_2 + \ldots + Y_i\ ) / n =$$

$$= 0.1654 / 26 = 0.0064$$

Individual statistical weight is the number of cells scored for person (N_i), divided by the mean number of cells scored per person in the group (N_{mean} =12970 / 26 = 498.85)

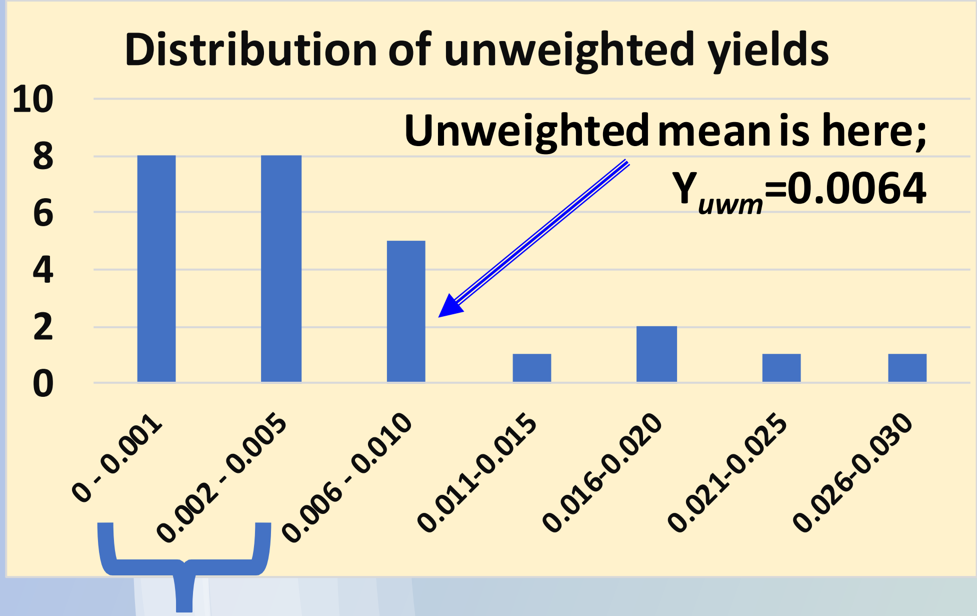
Weighted mean yield of aberrations:

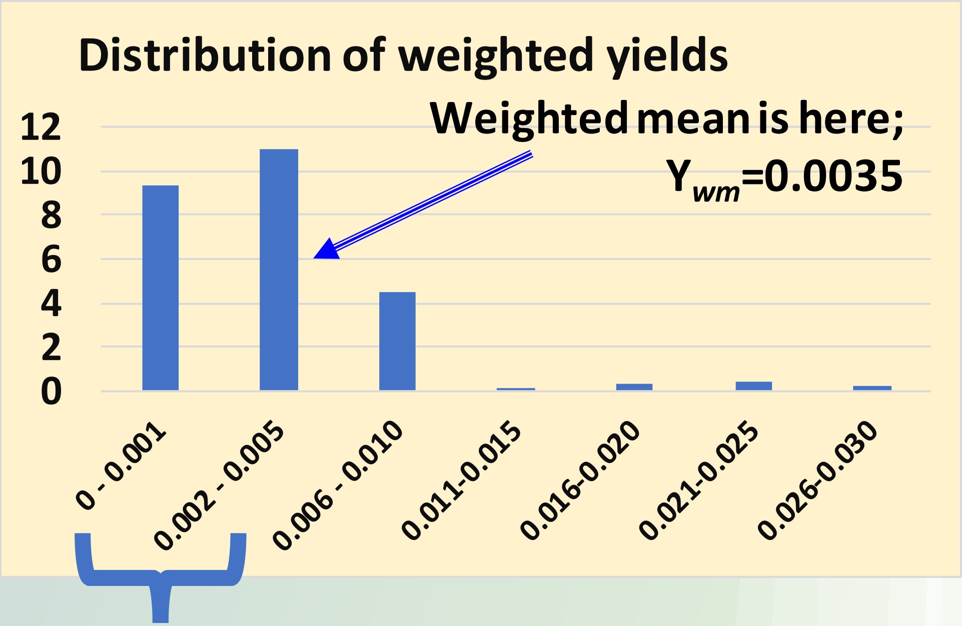
$$Y_{wm} = (Y_1 \times N_1/N_{mean} + Y_2 \times N_2/N_{mean} + \ldots + Y_i \times N_i/N_{mean}\ ) / n =$$

$$= 0.0902 / 26 = 0.0035$$

Code	Individial	Cells scored	Aberrations found	Yield	Weight, f_i	Yield * Weight
1	AAA	1000	0	0.0000	2.0046	0.0000
13	WAA	1000	0	0.0000	2.0046	0.0000
2	AAB	750	0	0.0000	1.5035	0.0000
3	AAC	500	0	0.0000	1.0023	0.0000
24	RAE	200	0	0.0000	0.4009	0.0000
6	AAF	100	0	0.0000	0.2005	0.0000
9	GAC	100	0	0.0000	0.2005	0.0000
14	WAB	1000	1	0.0010	2.0046	0.0020
15	WAC	1000	2	0.0020	2.0046	0.0040
4	AAD	500	1	0.0020	1.0023	0.0020
23	RAD	500	1	0.0020	1.0023	0.0020
17	WAD	1200	3	0.0025	2.4056	0.0060
5	AAE	800	2	0.0025	1.6037	0.0040
21	RAC	500	2	0.0040	1.0023	0.0040
18	WAE	800	4	0.0050	1.6037	0.0080
11	GAE	200	1	0.0050	0.4009	0.0020
20	RAB	500	3	0.0060	1.0023	0.0060
10	GAD	150	1	0.0067	0.3007	0.0020
25	RAF	1000	7	0.0070	2.0046	0.0140
19	RAA	500	4	0.0080	1.0023	0.0080
7	GAA	100	1	0.0100	0.2005	0.0020
28	RTB	67	1	0.0149	0.1343	0.0020
8	GAB	100	2	0.0200	0.2005	0.0040
12	GAF	50	1	0.0200	0.1002	0.0020
26	RAG	236	5	0.0212	0.4731	0.0100
27	RTA	117	3	0.0256	0.2345	0.0060
	Total	12970	45	0.1654	26	0.0902

Analysis of the distribution of individual yields





Modal classes = most frequent cases

Modal classes = most frequent cases

Distribution can be expressed as proportions of cases in each class.

Proportion or Frequency (*f*) = Number of cases in the certain class (*m*) / Total number of people in the group (*n*)

Comparison of the modal and middle classes show the asymmetry of our distribution.

Such an asymmetry is all'right. Modal classes are in the range of low values. Mean is shifted to the right from the modal values.

Poisson distribution of aberration yields in the studied group

If an equal number of cells would have been scored for each person, then the observed frequency distribution can be easily compared to a distribution, expected from a certain statistical model, e.g. Poisson distribution, using χ^2 test.
If there would be no statistical difference between our observed distribution of individual yields and the Poisson model, then we may apply parameters of the Poisson statistics in our future calculations.

Poisson frequency:
Probability of observing X aberrations, if the mean yield is $\bar{Y}$.

$$f = \frac{\bar{Y}^X}{X!} e^{-\bar{Y}}$$

Mean yield must be expressed as mean number of aberrations found per person.

$$\hat{Y} = \frac{\text{Total number of aberrations } (X_{total})}{\text{Total number of persons } (n)} = \frac{45}{26} = 1.7308$$

Number of cases in the certain X class (*m*) =
= Frequency (*f*) x Total number of cases (*n*)

Also, values of the class (X) must be transformed into the normalized weighted yield per mean number of cells scored per person: $\hat{Y}_{nw} = X_i / (N_{total} / n)$

Mean = Ŷ = 1.7308; n = 26						
X	X!	$\hat{Y}^X$	$e^{-\hat{Y}}$	Poisson frequency, *f*	Cases, *f* x n	Normalized yield class
0	1	1.0000	0.1771	0.1771	4.6059	0.000
1	1	1.7308	0.1771	0.3066	7.9717	0.002
2	2	2.9956	0.1771	0.2653	6.8986	0.004
3	6	5.1846	0.1771	0.1531	3.9799	0.006
4	24	8.9734	0.1771	0.0662	1.7221	0.008
5	120	15.5309	0.1771	0.0229	0.5961	0.010
6	720	26.8804	0.1771	0.0066	0.1720	0.012
7	5040	46.5237	0.1771	0.0016	0.0425	0.014
8	40320	80.5218	0.1771	0.0004	0.0092	0.016
9	362880	139.3646	0.1771	0.0001	0.0018	0.018

Normalization of the aberration yield in the studied group

Nr	Individial	Cells scored	Aberrations found	Yield	Weight, f	$\hat{Y}_{nw}$ = Yield x Weight
1	AAA	1000	0	0.0000	2.00462606	0
13	WAA	1000	0	0.0000	2.00462606	0
2	AAB	750	0	0.0000	1.503469545	0
3	AAC	500	0	0.0000	1.00231303	0
24	RAE	200	0	0.0000	0.400925212	0
6	AAF	100	0	0.0000	0.200462606	0
9	GAC	100	0	0.0000	0.200462606	0
14	WAB	1000	1	0.0010	2.00462606	0.0020046
15	WAC	1000	2	0.0020	2.00462606	0.0040093
4	AAD	500	1	0.0020	1.00231303	0.0020046
23	RAD	500	1	0.0020	1.00231303	0.0020046
17	WAD	1200	3	0.0025	2.405551272	0.0060139
5	AAE	800	2	0.0025	1.603700848	0.0040093
21	RAC	500	2	0.0040	1.00231303	0.0040093
18	WAE	800	4	0.0050	1.603700848	0.0080185
11	GAE	200	1	0.0050	0.400925212	0.0020046
20	RAB	500	3	0.0060	1.00231303	0.0060139
10	GAD	150	1	0.0067	0.300693909	0.0020046
25	RAF	1000	7	0.0070	2.00462606	0.0140324
19	RAA	500	4	0.0080	1.00231303	0.0080185
7	GAA	100	1	0.0100	0.200462606	0.0020046
28	RTB	67	1	0.0149	0.134309946	0.0020046
8	GAB	100	2	0.0200	0.200462606	0.0040093
12	GAF	50	1	0.0200	0.100231303	0.0020046
26	RAG	236	5	0.0212	0.47309175	0.0100231
27	RTA	117	3	0.0256	0.234541249	0.0060139
	Total	12970	45	0.0035	26	0.0902082
						Mean=0.0035

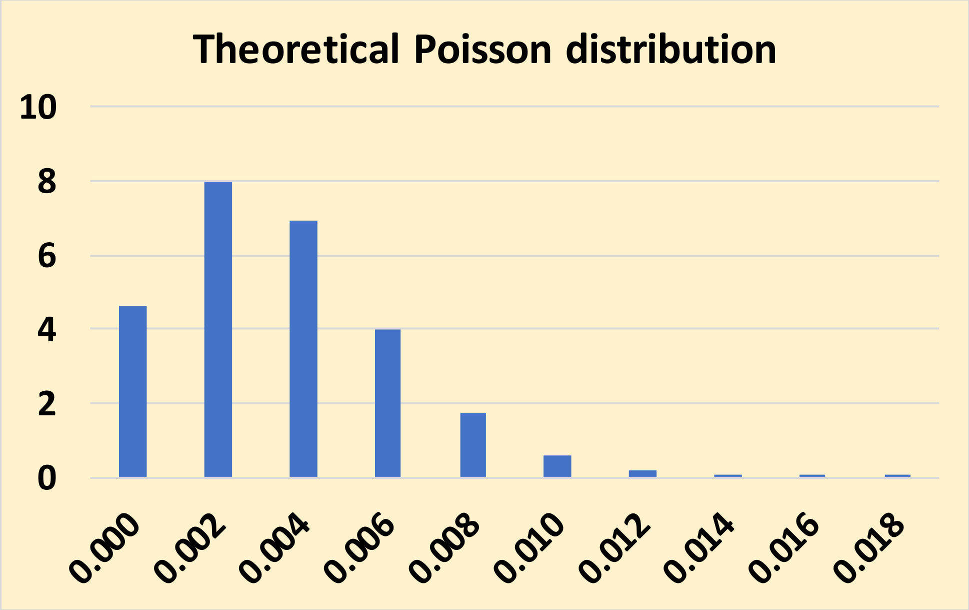
Sort in ascending order

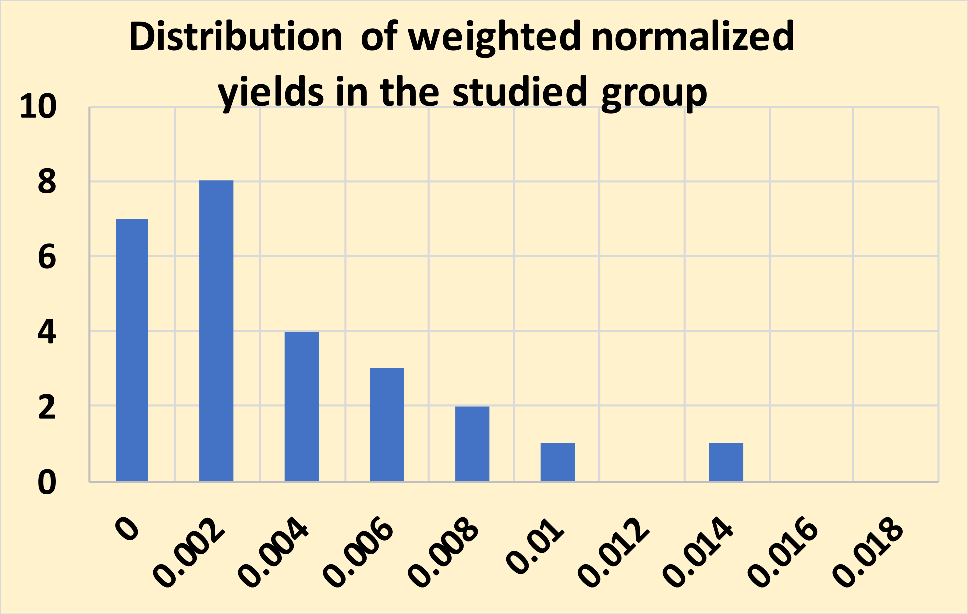
$\hat{Y}_{nw}$ Sorted	X normalized, weighted
0.0000	0.0000
0.0000	0.0000
0.0000	0.0000
0.0000	0.0000
0.0000	0.0000
0.0000	0.0000
0.0000	0.0000
0.0020	1.0000
0.0020	1.0000
0.0020	1.0000
0.0020	1.0000
0.0020	1.0000
0.0020	1.0000
0.0020	1.0000
0.0020	1.0000
0.0040	2.0000
0.0040	2.0000
0.0040	2.0000
0.0040	2.0000
0.0060	3.0000
0.0060	3.0000
0.0060	3.0000
0.0080	4.0000
0.0080	4.0000
0.0100	5.0000
0.0140	7.0000
Σ=0.0902	Σ=45.0000
	Mean = 1.7308

Normalized weighted number of aberrations found = Normalized weighted yield $\hat{Y}_{nw}$ x Mean number of cells scored per person

Normalised yield class	Cases
0.000	7
0.002	8
0.004	4
0.006	3
0.008	2
0.010	1
0.012	0
0.014	1
0.016	0
0.018	0

Comparison of the weighted normalized distribution of aberration yield in the studied group and the Poisson model





The distribution of normalized weighted individual yields in our group is similar by its shape to Poisson distribution.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f - f')^2}{f'} = \sum_{i=1}^{k} \left(\frac{d^2}{f'} \right)$$

f – observed number of cases;
f' – expected Poisson number of cases;
d – difference between observed and expected numbers.

Yield class	Observed, f	Poisson, f'	d = f – f'	d²	d² / f'
0.000	7	4.6059	2.3941	5.7317	1.2444
0.002	8	7.9717	0.0283	0.0008	0.0001
0.004	4	6.8986	-2.8986	8.4019	1.2179
0.006	3	3.9799	-0.9799	0.9602	0.2413
≥0.008	4	2.5419	1.4581	2.1261	0.8364
Total	26	25.998	0.0020	17.2207	3.5401
Degrees of freedom k = Nr of classes – 2 = 5 – 2 = 3					
		Chi-squared critical for k=3 χ^2_{st} =7.82			
		Exact p = 0.3156 >0.05			

Characteristics of the aberration yield in the studied group

No much use of the median beyond symmetric bell-shaped, normal distribution. It can be applied for large groups in population studies.

Quartiles divide the total **range** of values in 4 equal parts. The 1st Quartile cuts off 25 % of values from the left (from zero towards median).

The 2nd Quartile cuts off 50 %, i.e. it is a Median *per se*.

The 3rd Quartile cuts off 25 % of values from the right (from maximum towards median).

Thus, the 1st and the 3rd quartiles are the middles of the ranges of values, which a lower or higher than the median. 50 % of the total **range** of values are in between the 1st and the 3rd quartiles.

$$Q_i = x_н + \lambda \left(\frac{K - \Sigma f_i}{f_Q} \right)$$

x_H – the lower boarder of the class, containing the quartile;
K = n*i, *where* i=0.25, 0.50 or 0.75 for Q_1, Q_2 and Q_3, respectively;
Σf_i – the sum of accumulated frequencies before this class;
λ – the width of the class; f_Q – frequency in the class, which contains the quartile.

Yied Class	Frequency, f_i (weighted)	Accumulated Σf_i
0	7.3169	7.3169
0.0001-0.001	2.0046	9.3215
0.0011-0.002	4.0093	13.3308
0.0021-0.003	4.0093	17.3401
0.0031-0.004	1.0023	18.3424
0.0041-0.005	2.0046	20.347
0.0051-0.006	1.0023	21.3493
0.0061-0.007	2.3053	23.6546
0.0071-0.008	1.0023	24.6569
0.0081-0.009	0	24.6569
0.0091-0.010	0.2005	24.8574
0.0011-0.011	0	24.8574
0.0111-0.012	0	24.8574
0.0121-0.013	0	24.8574
0.0131-0.014	0	24.8574
0.0141-0.015	0.1343	24.9917
0.0151-0.016	0	24.9917
0.0161-0.017	0	24.9917
0.0171-0.018	0	24.9917
0.0181-0.019	0	24.9917
0.0191-0.020	0.3007	25.2924
0.0201-0.021	0.4731	25.7655
0.0211-0.022	0	25.7655
0.0221-0.023	0	25.7655
0.0231-0.024	0	25.7655
0.0241-0.025	0	25.7655
0.0251-0.026	0.2345	26

Q_1
K = n * 0.25 = 6.50
Yield Class 0.000
x_H = 0.000; λ =0.001
Σf_i =0; f_Q = 7.3169
Q_1 = 0.0009

Q_2 = Median
K = n*0.5 = 13
Yield Class 0.002
x_H = 0.0011; λ =0.001
Σf_i =9.3215; f_Q = 4.0093
Q_2 = Me = 0.0020

Q_3
K= n * 0.75 =19.50
Yield Class 0.005
x_H = 0.0041; λ =0.001
Σf_i =18.3424; f_Q = 2.0046
Q_3 = 0.0047

Age characteristic of the studied group

Nr	Individial	Age, y
9	GAC	22
4	AAD	24
7	GAA	25
25	RAF	26
20	RAB	27
24	RAE	27
21	RAC	28
22	FPB	28
23	RAD	30
5	AAE	33
6	AAF	34
2	AAB	36
13	WAA	38
11	GAE	40
26	RAG	41
27	RTA	42
12	GAF	42
3	AAC	43
19	RAA	44
28	RTB	48
1	AAA	50
15	WAC	51
14	WAB	54
10	GAD	55
18	WAE	57
8	GAB	59
17	WAD	62
16	FPA	75

Mean age 40.8 years (for n=28!!!)

Interval 5 years	Persons, f_i	Accumulated Σf_i
22-26	4	4
27-31	5	9
32-36	3	12
37-41	3	15
42-46	4	19
47-51	3	22
52-56	2	24
57-61	2	26
62-66	1	27
67-71	0	0
72-76	1	28

$$Q_i = x_н + \lambda \left(\frac{K - \Sigma f_i}{f_Q} \right)$$

x_H – the lower boarder of the class, containing the quartile;
K = n**i*, *where i*=0.25, 0.50 or 0.75 for Q_1, Q_2 and Q_3, respectively;
Σf_i – the sum of accumulated frequencies before this class;
λ – the width of the class;
f_Q – frequency in the class, which contains the quartile.

Q_1
K = n * 0.25 = 7.0
Yield Class 27-31
$x_H = 27$; $\lambda = 5$
$\Sigma f_i = 4$; $f_Q = 5$
Q_1 = 30.0 years

Q_2 = Median
K = n*0.5 = 14.0
Yield Class 37-41
$x_H = 37$; $\lambda = 5$
$\Sigma f_i = 12$; $f_Q = 3$
Q_2 = Me = 40.3

Q_3
K= n * 0.75 =21.0
Yield Class 47-51
$x_H = 47$; $\lambda = 5$
$\Sigma f_i = 19$; $f_Q = 3$
Q_3 = 50.3

Characteristics of the aberration yield in the studied group

Dispersion around the mean $\bar{x}$

$$s_x^2 = \frac{\sum_i^k f_i (x_i - \bar{x})^2}{n-1}$$

Standard deviation

$$s_x = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{n-1}}$$

Standard error of the mean

$$SE = \sqrt{\frac{s_x^2}{n-1}}$$

Confidence intervals of the mean $\bar{x}$

$$\bar{x} - \frac{t s_x}{\sqrt{n}} \leqslant \mu \leqslant \bar{x} + \frac{t s_x}{\sqrt{n}}$$

For the mean group yield 0.0035 and n=26

P=0.95 t = 1,96 CL_{lw}=0.00167 CL_{up}=0.00533

P=0.99 t=2,58 CL_{lw}=0.00109 CL_{up}=0.00591

P=0.999 t=3.29 CL_{lw}=0.00043 CL_{up}=0.00657

Patient's code	Individial	Cells scored	Aberrations found	Yield	Difference with mean	Square difference	Weight, f_i	Square difference x Weight
1	AAA	1000	0	0.0000	-0.0035	0.00001225	2.00462606	2.45567E-05
13	WAA	1000	0	0.0000	-0.0035	0.00001225	2.00462606	2.45567E-05
2	AAB	750	0	0.0000	-0.0035	0.00001225	1.50346955	1.84175E-05
3	AAC	500	0	0.0000	-0.0035	0.00001225	1.00231303	1.22783E-05
24	RAE	200	0	0.0000	-0.0035	0.00001225	0.40092521	4.91133E-06
6	AAF	100	0	0.0000	-0.0035	0.00001225	0.20046261	2.45567E-06
9	GAC	100	0	0.0000	-0.0035	0.00001225	0.20046261	2.45567E-06
14	WAB	1000	1	0.0010	-0.0025	0.00000625	2.00462606	1.25289E-05
15	WAC	1000	2	0.0020	-0.0015	0.00000225	2.00462606	4.51041E-06
4	AAD	500	1	0.0020	-0.0015	0.00000225	1.00231303	2.2552E-06
23	RAD	500	1	0.0020	-0.0015	0.00000225	1.00231303	2.2552E-06
17	WAD	1200	3	0.0025	-0.0010	0.000001	2.40555127	2.40555E-06
5	AAE	800	2	0.0025	-0.0010	0.000001	1.60370085	1.6037E-06
21	RAC	500	2	0.0040	0.0005	0.00000025	1.00231303	2.50578E-07
18	WAE	800	4	0.0050	0.0015	0.00000225	1.60370085	3.60833E-06
11	GAE	200	1	0.0050	0.0015	0.00000225	0.40092521	9.02082E-07
20	RAB	500	3	0.0060	0.0025	0.00000625	1.00231303	6.26446E-06
10	GAD	150	1	0.0067	0.0032	1.0028E-05	0.30069391	3.01529E-06
25	RAF	1000	7	0.0070	0.0035	0.00001225	2.00462606	2.45567E-05
19	RAA	500	4	0.0080	0.0045	0.00002025	1.00231303	2.02968E-05
7	GAA	100	1	0.0100	0.0065	0.00004225	0.20046261	8.46955E-06
28	RTB	67	1	0.0149	0.0114	0.00013054	0.13430995	1.75327E-05
8	GAB	100	2	0.0200	0.0165	0.00027225	0.20046261	5.45759E-05
12	GAF	50	1	0.0200	0.0165	0.00027225	0.1002313	2.7288E-05
26	RAG	236	5	0.0212	0.0177	0.00031281	0.47309175	0.000147988
27	RTA	117	3	0.0256	0.0221	0.00049023	0.23454125	0.000114978
	Total	12970	45	0.0035		0.0016746	26	0.000544917
						6.4408E-05	Dispersion	2.17967E-05
						0.00802545	St Deviation	0.004668692
						0.00160509	St Error	0.000933738

Unweighted = incorrect

Characteristics of the qualitative indices in the studied group

Proportion = Frequency (*f*) =
Number of cases in the certain class (*m*) /
Total number of people in the group (*n*)

Standard Error for the Frequency

$$SE = \sqrt{\frac{f(1-f)}{n}}$$

Proportion of people with Y > Control level (Control = 0.0010 per cell)
***f* = m / n = 18 / 26 = 0.6923**
SE = √ (0.6923 x (1 - 0.6923) / 26) = 0.0905

Proportion of people with Y > [2 x Control level]
***f* = m / n = 15 / 26 = 0.5769**
SE = √ (0.5769 x (1 - 0.5769) / 26) = 0.0969

Proportions and SE can be x100 to obtain the estimates in %.

Characteristics of the aberration yield in the studied group

Mean aberration yield and its standard error are 0.0035±0.0009 per cell.

95% Confidence Limits for the mean are 0.0017 – 0.0053 per cell.

Dispersion of individual yields inside the group is smaller than Mean.

Modal classes in the distribution of individual yields

are 0 – 0.001 and 0.002 – 0.005 per cell.

Median value is 0.0020 per cell;

quartiles Q1 = 0.0009 and Q3 = 0.0047 per cell.

Among the studied persons the elevated aberration yield was observed in 69 ± 9 % cases, and the yield, exceeding the control level more than twice, was observed in 58 ± 10 % cases.

Individual cases

Patients	Cells scored	Aberrant cells	Aberration per cell distribution						ΣAbs
			0	1	2	3	4	5	
16FPA	200	39	161	25	11	2	1	0	57
22FBP	200	60	140	52	7	1	0	0	69

Dispersion

$$s_x^2 = \frac{\sum_i^k f_i (x_i - \bar{x})^2}{n-1}$$

Standard Error

$$SE = \sqrt{\frac{s_x^2}{n-1}}$$

In Poisson distribution Dispersion ≈ Mean !

16FPA

Weight, f_i is the number of cells in this Ab/Cell category

Abs/Cell	Cells	Aberrations in these cells	Difference with mean (d)	d^2	$d^2 \times f_i$
0	161	0	-0.2850	0.081225	13.077225
1	25	25	0.7150	0.511225	12.780625
2	11	22	1.7150	2.941225	32.353475
3	2	6	2.7150	7.371225	14.74245
4	1	4	3.7150	13.801225	13.801225
5	0	0	4.7150	22.231225	0
6	0	0	5.7150	32.661225	0
7	0	0	6.7150	45.091225	0
8	0	0	7.7150	59.521225	0
9	0	0	8.7150	75.951225	0
10	0	0	9.7150	94.381225	0
Total	200	57			86.7550
	Yield	0.2850		Dispersion	0.4360
	Y=X / N			St Error	0.0468
				σ^2 / Y	1.5297

22FBP

Weight, f_i is the number of cells in this Ab/Cell category

Abs/Cell	Cells	Aberrations in these cells	Difference with mean (d)	d^2	$d^2 \times f_i$
0	140	0	-0.3450	0.119025	16.6635
1	52	52	0.6550	0.429025	22.3093
2	7	14	1.6550	2.739025	19.173175
3	1	3	2.6550	7.049025	7.049025
4	0	0	3.6550	13.35903	0
5	0	0	4.6550	21.66903	0
6	0	0	5.6550	31.97903	0
7	0	0	6.6550	44.28903	0
8	0	0	7.6550	58.59903	0
9	0	0	8.6550	74.90903	0
10	0	0	9.6550	93.21903	0
Total	200	69			65.1950
	Yield	0.3450		Dispersion	0.3276
	Y=X / N			St Error	0.0406
				σ^2 / Y	0.9496

Individual cases

Aberrant cells

16FPA Y = X / N = 39 / 200 = 0.1950

22FBP Y = X = N = 60 / 200 = 0.3000

Standard Error for the Frequency

$$SE = \sqrt{\frac{f(1-f)}{n}}$$

Y ± SE = 0.1950 ± 0.0281

Y ± SE = 0.3000 ± 0.0325

Papworth's u-test

$$u = (\sigma^2/y - 1)\sqrt{\frac{N-1}{2(1-1/X)}}$$

16FPA σ^2=0.4360; Y=0.2850; X=57; N=200 → u = 5.3304

22FBP σ^2=0.3276; Y=0.3450; X=69; N=200 → u= -0.5064

In Poisson distribution Dispersion ≈ Mean !

Papworth's u-test > |1.96| indicates the non-random aberration-per-cell distribution.
u > 1.96 points at the overdispersion, which happens due to non-random occurrence of DNA breaks.

Patients	Cells scored	Aberrant cells	Aberration per cell distribution						ΣAbs	Yield ± SE	σ^2 / Y	u
			0	1	2	3	4	5				
16FPA	200	39	161	25	11	2	1	0	57	**0.2850 ± 0.0468**	**1.530**	**5.330**
22FBP	200	60	140	52	7	1	0	0	69	**0.3450 ± 0.0406**	**0.950**	**-0.506**

Individual cases

95% Confidence Limits for Poisson Statistics: X – number of aberrations found.
(Taken from IAEA, 2011)

TABLE 8. THE POISSON UPPER AND LOWER 95% CONFIDENCE LIMITS ON OBSERVED NUMBERS (X) OF DICENTRICS
(adapted from [133])

X	Lower	Upper	X	Lower	Upper	X	Lower	Upper	X	Lower	Upper
0	0	3.285	26	16.77	37.67	52	38.165	66.76	78	61.9	96.06
1	0.051	5.323	27	17.63	38.165	53	39.76	68.1	79	62.81	97.545
2	0.355	6.686	28	19.05	39.76	54	4.094	69.62	80	62.81	99.17
3	0.818	8.102	29	19.05	40.94	55	40.94	71.09	81	63.49	99.17
4	1.366	9.598	30	20.335	41.75	56	41.75	71.28	82	64.95	100.32
5	1.97	11.177	31	21.36	43.45	57	43.45	72.66	83	66.76	101.71
6	2.613	12.817	32	21.36	44.26	58	44.26	74.22	84	66.76	103.315
7	3.285	13.765	33	22.945	45.28	59	44.26	75.49	85	66.76	104.4
8	3.285	14.921	34	23.76	47.025	60	45.28	75.785	86	68.1	104.58
9	4.46	16.768	35	23.76	47.69	61	47.025	77.16	87	69.62	105.905
10	5.323	17.633	36	25.4	48.74	62	47.69	78.73	88	71.02	107.32
11	5.323	19.05	37	26.31	50.42	63	47.69	79.98	89	71.09	109.11
12	6.686	20.335	38	26.31	51.29	64	48.74	80.25	90	71.28	109.61
13	6.686	21.364	39	27.735	52.15	65	50.42	81.61	91	72.66	110.11
14	8.102	22.945	40	28.97	53.72	66	51.29	83.14	92	74.22	111.44
15	8.102	23.762	41	28.97	54.99	67	51.29	84.57	93	75.49	112.87
16	9.598	25.4	42	30.02	55.51	68	52.15	84.67	94	75.49	114.84
17	9.598	26.306	43	31.675	56.99	69	53.72	86.01	95	75.785	114.84
18	11.177	27.735	44	31.675	58.72	70	54.99	87.48	96	77.16	115.605
19	11.177	28.966	45	32.28	58.84	71	54.99	89.23	97	78.73	116.93
20	12.817	30.017	46	34.05	60.24	72	55.51	89.23	98	79.98	118.35
21	12.817	31.675	47	34.665	61.9	73	56.99	90.37	99	79.98	120.36
22	13.765	32.277	48	34.665	62.81	74	58.72	91.78	100	80.25	120.36
23	14.921	34.048	49	36.03	63.49	75	58.72	93.48	101	81.61	121.06
24	14.921	34.665	50	37.67	64.95	76	58.84	94.23	102	83.14	122.57
25	16.768	36.03	51	37.67	66.76	77	60.24	94.705	103	84.57	123.77

The CL for X should be divided by the number of cells scored

For the studied group
X = 45 Y=0.0035
CL_{lw} = 32.28/12970 = 0.0025
CL_{up} = 58.84/12970 = 0.0045

For case 22FBP
X=69 Y=0.0345
CL_{lw} = 53.72/200 = 0.2686
CL_{up} = 86.01/200 = 0.4301

For case 16FPA
The overdispersion!!!
Poisson CL are non-applicable.

Relative accuracy of the measurement of aberration yield

Accuracy = SE / Y
Yield Y = X / N
SE = $\sqrt{(\sigma^2 / N)}$
$\sigma^2 \approx$ Y = X / N
Accuracy = ($\sqrt{X}$) / X

For a single measurement
X=1 Y←1.0 ± 1.0 (100%)
X=2 Y←2.0 ± 1.4 (70%)
X=3 Y←3.0 ± 1.7 (58%)
…
X=10 Y←10.0 ± 3.2 (32%)
…
X=20 Y←20.0 ± 4.5 (22%)
…
X=50 Y←50.0 ± 7.1 (14%)
…
X=100 Y←100.0 ± 10.0 (10%)

Rule Nr 1: Don’t consider aberration yield as something fixed. It is induced with some probability and detected with some probability. Therefore, it must be presented and treated in a manner that takes into account its probabilistic (stochastic) nature.

How we should treat cases, in which 0 (zero) aberrations were found during the analysis?
Is zero a “true” zero? The question can be addressed via probability analysis. The answer is as follows.
The most probable value is zero. However, mean value in the probability distribution is ≈ 0.7.
Upper 95%-confidence interval is ≈3.

Why 95% confidence interval for observed zero aberrations doesn’t allow excluding as many as 3?
If you are interested in the answer, please, refer to the seminar, which will include the Bayesian approach.

Thank you for your attention!