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 IV. Assessment of the effect of acting factor(s)

Contents:

ANOVA & Kruskal–Wallis *H* test.

Estimation of the induced effect in radiation cytogenetics: Detection versus measurement of the background-doubling impact

Correlation analysis: Pearson's linear correlation coefficient & Spearman's rank correlation coefficient. Other correlation methods.

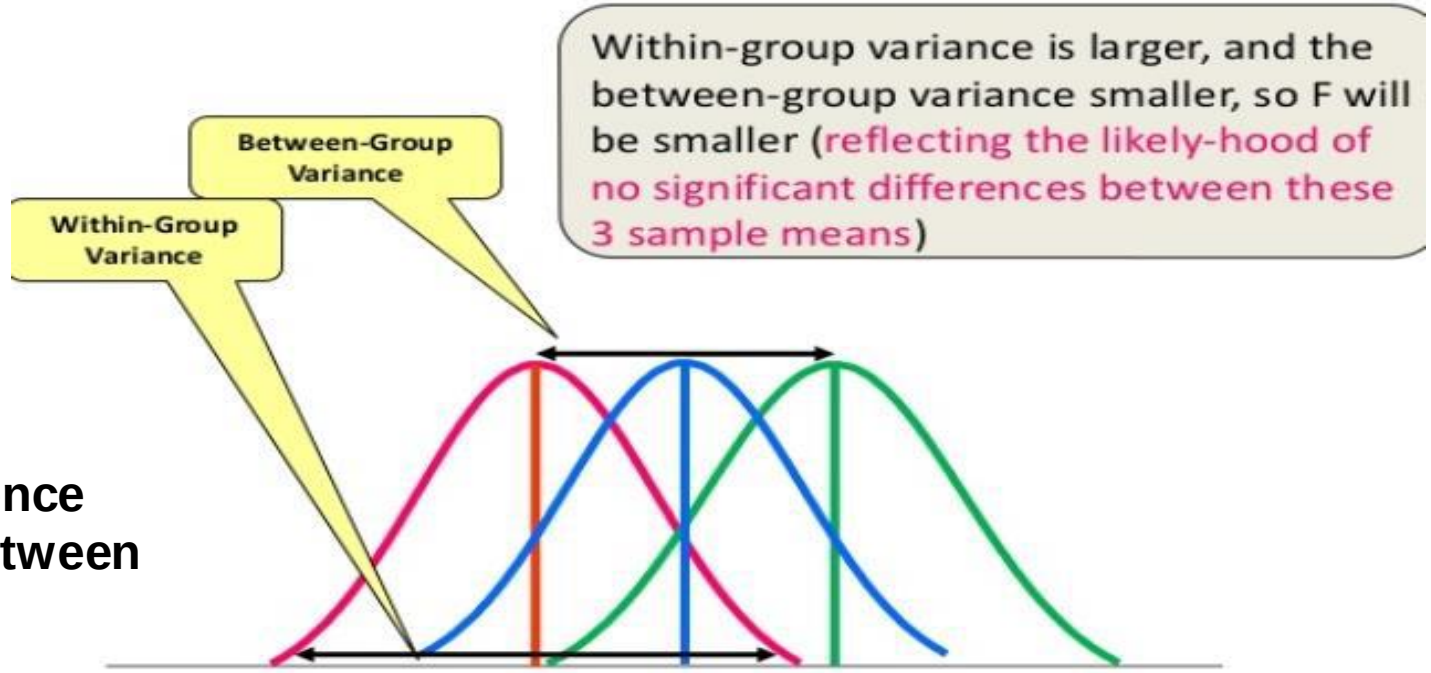
Regression analysis: The Least Squares method. Goodness of fit. Regression models.

Analysis of variance (ANOVA)

ANOVA should be applied in trying to find the difference between mean values of the measured effect (i.e. caused by the action of certain factor) in three or more groups.

MS_B – mean square between groups
MS_W – mean square within groups

$$F = MS_B / MS_W$$



ANOVA can show the presence of the difference between the groups, but it doesn't tell us, between which groups the difference occurs.

Analysis of variance (ANOVA)

$F = MS_B / MS_W$

$MS_B = SS_B / df_B$

$df_B = k - 1$;

k is the number of groups

$MS_W = SS_W / df_W$

$df_W = n - k$;

n is the total number of individuals

MS_B – mean square between groups
MS_W – mean square within groups
SS – sum of squared deviations from the mean

SS_T – sum of squared deviations of individual values from the total mean.
SS_B – sum of squared deviations of the group means from the total mean.
SS_W – sum of squared deviations of individual values from the group means.

$$F = \frac{SS_B \times (n-k)}{(SS_T - SS_B) \times (k-1)}$$

$SS_T = SS_B + SS_W$

Virtual group of the workers of radiation industry unit.
Study 2. Individuals sorted by professional code.

Nr	Individial	Cells scored	Aberrations found	Yield	Difference with mean	Diff ^{2}	Weight, f_i	Diff ^{2} * f_i
1	AAA	1000	0	0.0000	-0.0052	2.7E-05	1.9697	0.000054
2	AAB	250	0	0.0000	-0.0052	2.7E-05	0.4924	0.000014
3	AAC	513	1	0.0019	-0.0033	1.06E-05	1.0105	0.000011
4	AAD	750	0	0.0000	-0.0052	2.7E-05	1.4773	0.000041
5	AAE	500	2	0.0040	-0.0012	1.44E-06	0.9848	0.000001
6	AAF	204	0	0.0000	-0.0052	2.7E-05	0.4018	0.000011
7	GAA	450	2	0.0044	-0.0008	5.71E-07	0.8864	0.000001
8	GAB	300	2	0.0067	0.0015	2.15E-06	0.5909	0.000001
9	GAC	200	0	0.0000	-0.0052	2.7E-05	0.3939	0.000011
10	GAD	1000	2	0.0020	-0.0032	1.02E-05	1.9697	0.000021
11	GAE	600	4	0.0067	0.0015	2.15E-06	1.1818	0.000003
12	GAF	336	5	0.0149	0.0097	9.37E-05	0.6618	0.000063
13	WAA	557	1	0.0018	-0.0034	1.16E-05	1.0971	0.000013
14	WAB	500	3	0.0060	0.0008	6.4E-07	0.9848	0.000001
15	WAC	1000	3	0.0030	-0.0022	4.84E-06	1.9697	0.000010
17	WAD	1000	6	0.0060	0.0008	6.4E-07	1.9697	0.000001
18	WAE	1000	6	0.0060	0.0008	6.4E-07	1.9697	0.000001
19	RAA	127	3	0.0236	0.0184	0.000339	0.2502	0.000086
20	RAB	300	4	0.0133	0.0081	6.62E-05	0.5909	0.000040
21	RAC	400	3	0.0075	0.0023	5.29E-06	0.7879	0.000004
23	RAD	200	2	0.0100	0.0048	2.3E-05	0.3939	0.000009
24	RAE	333	1	0.0030	-0.0022	4.83E-06	0.6559	0.000003
25	RAF	200	6	0.0300	0.0248	0.000615	0.3939	0.000247
26	RAG	80	2	0.0250	0.0198	0.000392	0.1576	0.000063
27	RTA	1200	10	0.0083	0.0031	9.82E-06	2.3636	0.000024
28	RTB	200	0	0.0000	-0.0052	2.7E-05	0.3939	0.000011
	Total	13200	68	0.00515			n = 26	Σ (Diff ^{2} * f_i) = 0.000730
			Weighted mean yield					Dispersion = 0.0000292

SS_T

$\sigma = 0.00540484$

The group was split into 4 subgroups: A (“Administrators”), G (“Guards”), W (“Workers”) and R (“Radiation technologists”). Groups are not equal by the number of individuals and cells scored.

Nr	Individual	Cells scored, N_j	Aberrations found	Yield
1	AAA	1000	0	0.0000
2	AAB	250	0	0.0000
3	AAC	513	1	0.0019
4	AAD	750	0	0.0000
5	AAE	500	2	0.0040
6	AAF	204	0	0.0000
	$n_j = 6$	3217	3	0.00093

Nr	Individual	Cells scored, N_j	Aberrations found	Yield
7	GAA	450	2	0.0044
8	GAB	300	2	0.0067
9	GAC	200	0	0.0000
10	GAD	1000	2	0.0020
11	GAE	600	4	0.0067
12	GAF	336	5	0.0149
	$n_j = 6$	2886	15	0.00520

Nr	Individual	Cells scored, N_j	Aberrations found	Yield
13	WAA	557	1	0.0018
14	WAB	500	3	0.0060
15	WAC	1000	3	0.0030
17	WAD	1000	6	0.0060
18	WAE	1000	6	0.0060
	$n_j = 5$	4057	19	0.00468

Nr	Individial	Cells scored, N_j	Aberrations found	Yield
19	RAA	127	3	0.0236
20	RAB	300	4	0.0133
21	RAC	400	3	0.0075
23	RAD	200	2	0.0100
24	RAE	333	1	0.0030
25	RAF	200	6	0.0300
26	RAG	80	2	0.0250
27	RTA	1200	10	0.0083
28	RTB	200	0	0.0000
	$n_j = 9$	3040	31	0.01020

Group	A	G	W	R	Sum	
Yield Y_j	0.00093	0.0052	0.00468	0.0102		
Weight $f_j = N_j * n / N$	6.33652	5.68455	7.99106	5.98788	26	Mean yield = $\Sigma (Y_j * f_j) / n$
$Y_j * f_j$	0.00589	0.02956	0.03740	0.06108	0.133927	0.00515
Difference with mean	-0.00422	4.89568E-05	-0.00047	0.005049		
Difference^2	1.78172E-05	2.39677E-09	2.22E-07	2.55E-05		
Diff $^2 * f_j$	0.0001129	1.36245E-08	1.77E-06	0.000153	0.000267	= SS_B

$MS_B = SS_B / df_B = 0.000267 / (4 - 1) = 8.911E\text{-}05$

$SS_W = SS_T - SS_B = 0.000730 - 0.000267 = 0.000463$

$MS_W = SS_W / df_w = 0.000463 / (26 - 4) = 2.103E\text{-}05$

$MS_B > MS_W$!!! $\rightarrow F = MS_B / MS_W = 8.911E\text{-}05 / 2.103E\text{-}05 = 4.237$; $p = 0.01657$

F.DIST.RT(4.237;3;22)

Kruskal–Wallis *H* test

This test is a non-parametric analogue of ANOVA, thus should be applied for ranks of the measurements, without calculations of the mean and/or dispersion. It is similar to Mann – Whitney *U* test, but is used for the comparison of 3 or more groups in the study. In each group $n_i \geq 5$.

Algorithm

1. Put the observations (e.g., aberration yields) from all groups into one set and sort them in ascending order.

2. Assign numeric ranks to all the observations (e.g., aberration yields), beginning with 1 for the smallest value. Where there are groups of tied values, assign a rank equal to the midpoint of unadjusted rankings (e.g., the ranks of (3, 5, 5, 5, 5, 8) are (1, 3.5, 3.5, 3.5, 3.5, 6), whereas the unadjusted ranks would be (1, 2, 3, 4, 5, 6)). Total number of ranks, N, must be equal to $n_1 + n_2 + \ldots + n_i$.

3. Add up the ranks for the observations which came from group 1 (= R_1), from group 2 (= R_2), … from i^{th} group (= R_i).

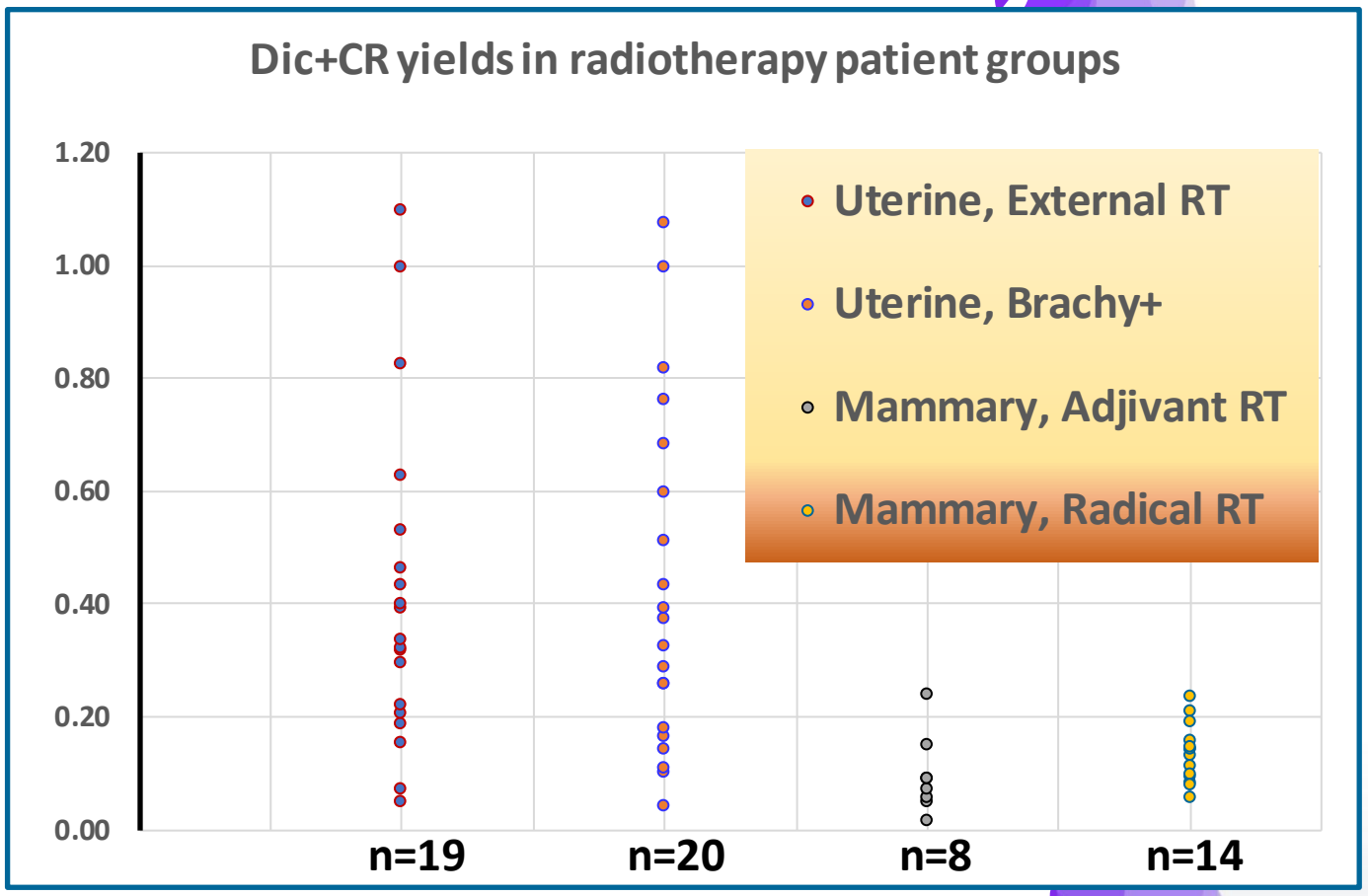
4. Calculate $(\Sigma R_i)^2$ and divide by n_i.

5. Calculate *H*. Use it as $\chi 2$ for $df = k - 1$, where k is the number of groups.

$$H = \frac{12}{N(N+1)} \sum \frac{(\Sigma R_l)^2}{n_l} - 3(N+1)$$

Radiotherapy patients. Groups are not equal by the number of individuals and cells scored.

n=19	n=20	n=8	n=14
Uterine cancer, External RT	Uterine cancer, Brachy+	Mammary cancer, Adjivant RT	Mammary cancer, Radical RT
0.0526	0.0438	0.0166	0.0601
0.0732	0.1020	0.0526	0.0813
0.1551	0.1100	0.0571	0.0857
0.1889	0.1449	0.0732	0.0945
0.2075	0.1667	0.0907	0.1012
0.2244	0.1806	0.0935	0.1131
0.2976	0.2584	0.1500	0.1324
0.3177	0.2600	0.2418	0.1429
0.3220	0.2899		0.1438
0.3396	0.3252		0.1480
0.3929	0.3739		0.1600
0.4030	0.3924		0.1921
0.4348	0.4364		0.2110
0.4667	0.5143		0.2386
0.5333	0.6000		
0.6282	0.6846		
0.8261	0.7647		
1.0000	0.8182		
1.1011	1.0000		
	1.0769		



Individual aberration yields in the groups are heterogeneous; no modal classes (values) are apparent.

$k = 4$; $n_i \geq 5$; $N = 61$

Aberration yields from all groups were pooled into one set and sorted in ascending order. The ranks were assigned from 1.0 to 61.0.

Group	Dic+CR yield	Rank
3	0.0166	1.0
2	0.0438	2.0
1	0.0526	3.5
3	0.0526	3.5
3	0.0571	5.0
4	0.0601	6.0
1	0.0732	7.5
3	0.0732	7.5
4	0.0813	9.0
4	0.0857	10.0
3	0.0906	11.0
3	0.0935	12.0
4	0.0945	13.0
4	0.1012	14.0
2	0.1020	15.0
2	0.1100	16.0
4	0.1131	17.0
4	0.1324	18.0
4	0.1429	19.0
4	0.1438	20.0
2	0.1449	21.0
4	0.1480	22.0
3	0.1500	23.0
1	0.1551	24.0
4	0.1600	25.0
2	0.1667	26.0
2	0.1806	27.0
1	0.1889	28.0
4	0.1921	29.0
1	0.2075	30.0
4	0.2110	31.0

Group	Dic+CR yield	Rank
1	0.2244	32.0
4	0.2386	33.0
3	0.2418	34.0
2	0.2584	35.0
2	0.2600	36.0
2	0.2899	37.0
1	0.2976	38.0
1	0.3177	39.0
1	0.3220	40.0
2	0.3252	41.0
1	0.3396	42.0
2	0.3739	43.0
2	0.3924	44.0
1	0.3929	45.0
1	0.4030	46.0
1	0.4348	47.0
2	0.4364	48.0
1	0.4667	49.0
2	0.5143	50.0
1	0.5333	51.0
2	0.6000	52.0
1	0.6282	53.0
2	0.6846	54.0
2	0.7647	55.0
2	0.8182	56.0
1	0.8261	57.0
1	1.0000	58.5
2	1.0000	58.5
2	1.0769	60.0
1	1.1011	61.0

Group 1; n=19

Dic+CR yield	Rank
0.0526	3.5
0.0732	7.5
0.1551	24.0
0.1889	28.0
0.2075	30.0
0.2244	32.0
0.2976	38.0
0.3177	39.0
0.3220	40.0
0.3396	42.0
0.3929	45.0
0.4030	46.0
0.4348	47.0
0.4667	49.0
0.5333	51.0
0.6282	53.0
0.8261	57.0
1.0000	58.5
1.1011	61.0
Total R_1	Σ=751.5

Group 2; n=20

Dic+CR yield	Rank
0.0438	2.0
0.1020	15.0
0.1100	16.0
0.1449	21.0
0.1667	26.0
0.1806	27.0
0.2584	35.0
0.2600	36.0
0.2899	37.0
0.3252	41.0
0.3739	43.0
0.3924	44.0
0.4364	48.0
0.5143	50.0
0.6000	52.0
0.6846	54.0
0.7647	55.0
0.8182	56.0
1.0000	58.5
1.0769	60.0
Total R_2	Σ=776.5

Group 3; n=8

Dic+CR yield	Rank
0.0166	1.0
0.0526	3.5
0.0571	5.0
0.0732	7.5
0.0906	11.0
0.0935	12.0
0.1500	23.0
0.2418	34.0
Total R_3	Σ=97.0

Group 4; n=14

Dic+CR yield	Rank
0.0601	6.0
0.0813	9.0
0.0857	10.0
0.0945	13.0
0.1012	14.0
0.1131	17.0
0.1324	18.0
0.1429	19.0
0.1438	20.0
0.1480	22.0
0.1600	25.0
0.1921	29.0
0.2110	31.0
0.2386	33.0
Total R_4	Σ=266.0

$$H = \frac{12}{61\times(61+1)} \times \left(\frac{751.5^{\,2}}{19} + \frac{776.5^{\,2}}{20} + \frac{97.0^{\,2}}{8} + \frac{266.0^{\,2}}{14}\right) - 3 \times (61+1) = 23.74$$

df = *k* - 1 = 3; p = 2.837E-05 by CHISQ.DIST.RT(23.74,3)

Statistical evaluation of the sample size for the detection of the 2-fold increase of biomarker's yield above the background level : *a single measurement, Poisson distribution, t-test*

Given: **The numbers of cells scored in the control and exposed samples are equal (*N*).**

The aberration distributions in both samples are in agreement with the Poisson statistics, i.e. $\sigma^2 = Y_{mean}$**. A statistical significance of the increase above the background can be concluded if the difference of the mean yields between exposed and control samples is at least twice higher, than the error of this difference (thus, *t* = 2.0 > 1.96).**

Question: **How many aberrations should be found in the exposed (X) and control samples?**

Answer: $t = Y_{diff} \,/\, SE_{diff} = (Y_{exposed} - Y_{control}) \,/\, \sqrt{(SE^2_{exposed} + SE^2_{control})} =$

$= (Y_{exposed} - Y_{control}) \,/\, \sqrt{(\sigma^2_{exposed} / N + \sigma^2_{control} / N)} = (Y_{exposed} - Y_{control}) \,/\, \sqrt{((Y_{exposed} + Y_{control}) / N)} =$

$= (2 \times Y_{control} - Y_{control}) \,/\, \sqrt{((2 \times Y_{control} + Y_{control}) / N)} = (X_{control} / N) \,/\, \sqrt{((3 \times X_{control} / N) / N)} =$

$= (X_{control} / N) \,/\, \sqrt{((3 \times X_{control}) / N^2)} = X_{control} \,/\, (\sqrt{(3 \times X_{control})} = \sqrt{(X_{control} / 3)} = t \geq 2.0.$

Solution: $X_{control} = 12$; $X_{exposed} = 24$**. These values define the lower limit of the sample size.**

Detection and measurement of the induced cytogenetic effect at the level of doubling the background yield

Given: $X_{control}$ = 12; $X_{exposed}$ = 24. The numbers of cells scored in the control and exposed samples are equal (*N*). Historical background level of Dic+CR in control populations $Y_{control}$ = 0.001 per cell.

If the samples are 12 Dic+CR in 12 000 cells in the control group ($Y_{control}$ = 0.00100±0.00029) and 24 Dic+CR in 12 000 cells in the exposed group ($Y_{exposed}$ = 0.00200±0.00041), then the Poisson 95% Confidence limits:

For X = 12: CL_{lw} = 6.686 & CL_{up} = 20.335; for Yield in 12000 cells CL_{lw} = 0.000557 & CL_{up} = 0.001695

For X = 24: CL_{lw} = 14.921 & CL_{up} = 34.665; for Yield in 12000 cells CL_{lw} = 0.001243 & CL_{up} = 0.002889

$CL_{up\ control}$ = 0.001695 > $CL_{lw\ exposed}$ = 0.001243. This is a detection of the induced effect.

If the samples are 24 Dic+CR in 24 000 cells in the control group ($Y_{control}$ = 0.00100±0.00020) and 48 Dic+CR in 24 000 cells in the exposed group ($Y_{exposed}$ = 0.00200±0.00029), then the Poisson 95% Confidence limits:

For X = 24: CL_{lw} = 14.921 & CL_{up} = 34.665; for Yield in 24 000 cells CL_{lw} = 0.000622 & CL_{up} = 0.001444

For X = 48: CL_{lw} = 34.665 & CL_{up} = 62.810; for Yield in 24000 cells CL_{lw} = 0.001444 & CL_{up} = 0.002617

$CL_{up\ control}$ = $CL_{lw\ exposed}$ = 0.001444. This is a borderline for the measurement of the induced effect.

Correlation analysis

In statistics, correlation is any statistical relationship, whether causal or not, between two random variables or bivariate data. Although in the broadest sense, "correlation" may indicate any type of association, in statistics it normally refers to the degree to which a pair of variables are linearly related.

Correlations can indicate a predictive relationship that can be exploited in practice. However, the presence of a correlation is not sufficient to infer the presence of a causal relationship.

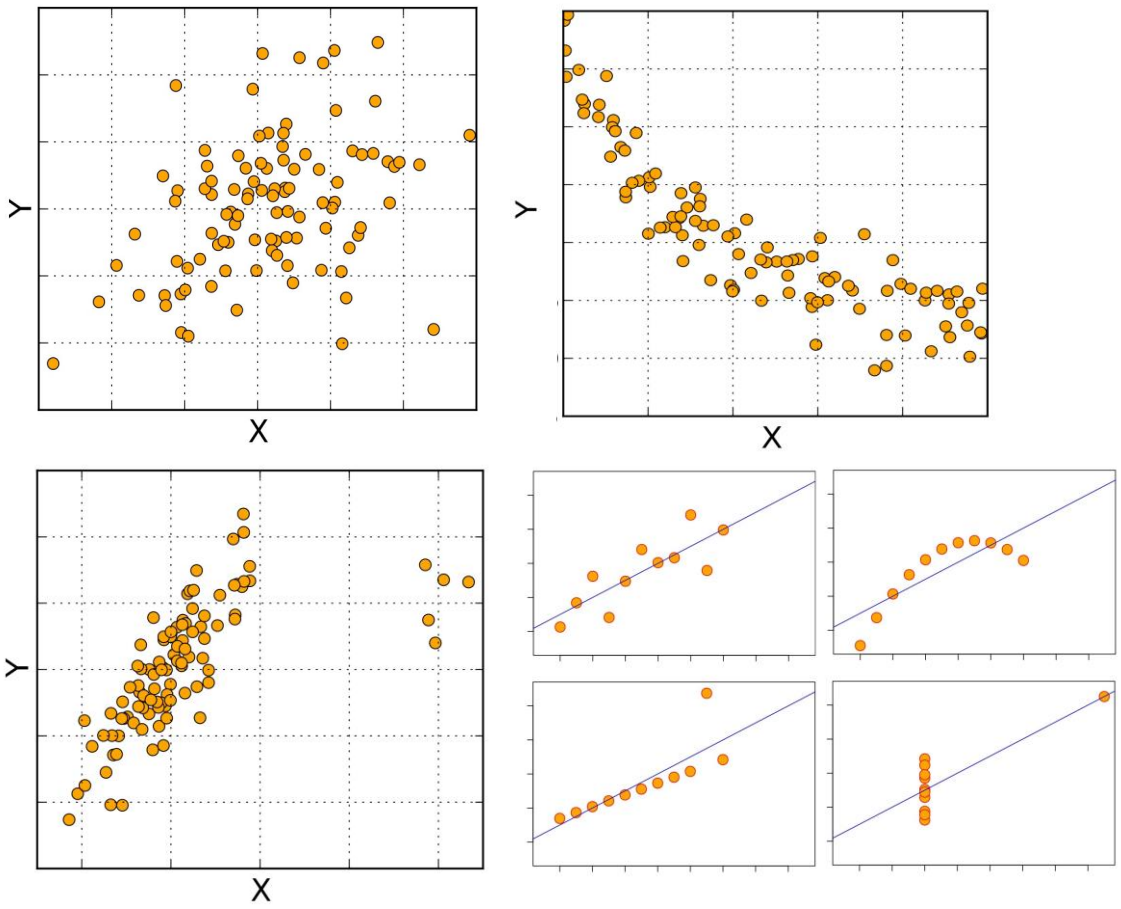
Correlation is the measure of how two or more variables are related to one another. Correlation reflects the strength and direction of a linear relationship, but not the slope of that relationship, nor the character of nonlinear relationships.

The most common is the *Pearson correlation coefficient*, which should be used only to test a linear relationship between two variables, e.g. between a factor and an effect (even when one variable is a nonlinear function of the other). It is applicable for quantitative data, which are expressed by numerical scales (discrete, continuous, interval or ratio data).

Non-parametric correlation coefficients – such as *Spearman's rank correlation* – are more robust than Pearson's, that is, more sensitive to nonlinear relationships. These coefficients should be applied if at least one of the variables has an ordinal scale or is not normally distributed. *Kendall's coefficient τ* (tau) should be used, if both variables are expressed as ranks, and/or abnormal, marginal values (outliers) are present in the original data.

Correlation analysis: Formalism for Pearson's coefficient (*r*)

Starting step in any correlation analysis:
Making a scatter plot and visual assessment



$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (\bar{y}_i - y)^2}} \quad \text{or} \quad r_{xy} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

For weighted data

$$r_{xy} = \frac{n \sum_{i=1}^{n} x_i y_i f_i - \sum_{i=1}^{n} x_i f_i \sum_{i=1}^{n} y_i f_i}{\sqrt{n \sum x_i^2 f_i - (\sum x_i f_i)^2} \sqrt{n \sum y_i^2 f_i - (\sum y_i f_i)^2}}$$

For n < 30

$$r_{Corrected} = r \left[1 + \frac{1 - r^2}{2(n-3)} \right]$$

Error:

$$SE_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Significance by t-test; $df = n - 2$

$$t = \frac{r}{SE_r} = r \sqrt{\frac{n-2}{1-r^2}} \geqslant t_{st}$$

Determination coefficient (r^2 or R^2) shows, which proportion of the variation of the effect depends on the variations of the factor. For $r \geq 0.70$ more than 50% of the changes of the effect is determined by the changes of the acting factor ($R^2 \geq 0.50$).

Correlation analysis: Example – Virtual group; Study 2; Aberration yield versus Age

Nr	Individial	Age	Cells scored	Dic+CR found	Dic+CR Yield	Weight, f_i	Yield * f_i	Age * f_i	Age * Y * f_i	Age^2 * f_i	Yield ^{2} * f_i
1	AAA	54	1000	0	0.0000	1.9697	0.00000	106.3636	0.00000	5743.63636	0.00000
2	AAB	34	250	0	0.0000	0.4924	0.00000	16.7424	0.00000	569.24242	0.00000
3	AAC	44	513	1	0.0019	1.0105	0.00197	44.4600	0.08667	1956.24000	0.00000
4	AAD	24	750	0	0.0000	1.4773	0.00000	35.4545	0.00000	850.90909	0.00000
5	AAE	33	500	2	0.0040	0.9848	0.00394	32.5000	0.13000	1072.50000	0.00002
6	AAF	33	204	0	0.0000	0.4018	0.00000	13.2600	0.00000	437.58000	0.00000
7	GAA	25	450	2	0.0044	0.8864	0.00394	22.1591	0.09848	553.97727	0.00002
8	GAB	60	300	2	0.0067	0.5909	0.00394	35.4545	0.23636	2127.27273	0.00003
9	GAC	22	200	0	0.0000	0.3939	0.00000	8.6667	0.00000	190.66667	0.00000
10	GAD	56	1000	2	0.0020	1.9697	0.00394	110.3030	0.22061	6176.96970	0.00001
11	GAE	38	600	4	0.0067	1.1818	0.00788	44.9091	0.29939	1706.54545	0.00005
12	GAF	43	336	5	0.0149	0.6618	0.00985	28.4582	0.42348	1223.70182	0.00015
13	WAA	36	557	1	0.0018	1.0971	0.00197	39.4964	0.07091	1421.86909	0.00000
14	WAB	55	500	3	0.0060	0.9848	0.00591	54.1667	0.32500	2979.16667	0.00004
15	WAC	54	1000	3	0.0030	1.9697	0.00591	106.3636	0.31909	5743.63636	0.00002
17	WAD	61	1000	6	0.0060	1.9697	0.01182	120.1515	0.72091	7329.24242	0.00007
18	WAE	57	1000	6	0.0060	1.9697	0.01182	112.2727	0.67364	6399.54545	0.00007
19	RAA	44	127	3	0.0236	0.2502	0.00591	11.0067	0.26000	484.29333	0.00014
20	RAB	27	300	4	0.0133	0.5909	0.00788	15.9545	0.21273	430.77273	0.00011
21	RAC	28	400	3	0.0075	0.7879	0.00591	22.0606	0.16545	617.69697	0.00004
23	RAD	30	200	2	0.0100	0.3939	0.00394	11.8182	0.11818	354.54545	0.00004
24	RAE	27	333	1	0.0030	0.6559	0.00197	17.7095	0.05318	478.15773	0.00001
25	RAF	26	200	6	0.0300	0.3939	0.01182	10.2424	0.30727	266.30303	0.00035
26	RAG	40	80	2	0.0250	0.1576	0.00394	6.3030	0.15758	252.12121	0.00010
27	RTA	42	1200	10	0.0083	2.3636	0.01970	99.2727	0.82727	4169.45455	0.00016
28	RTB	53	200	0	0.0000	0.3939	0.00000	20.8788	0.00000	1106.57576	0.00000
	Total	1046	13200	68		26	0.13394 = $= \Sigma y \times f_i$	1146.4286 = $= \Sigma x \times f_i$	5.706212 = $= \Sigma (x \times y \times f_i)$	54642.622273 = $= \Sigma (x^2 \times f_i)$	0.001420 = $= \Sigma (y^2 \times f_i)$
Mean age, unweighted		1046 / 26 = 40.2308	Mean aberration yield 68 / 13200 = 0.00515				0.13394 / 26 =0.00515	1146 / 26= 44.0934			
						Mean aberration yield, weighted		Mean age, weighted			

Correlation analysis: calculations

$$r = \frac{(26 * 5.706212 - 1146.4286 * 0.13394)}{\sqrt{(26 * 54642.62273 - 1146.4286^2)} * \sqrt{(26 * 0.00142 - 0.13394^2)}} = -0.11547$$

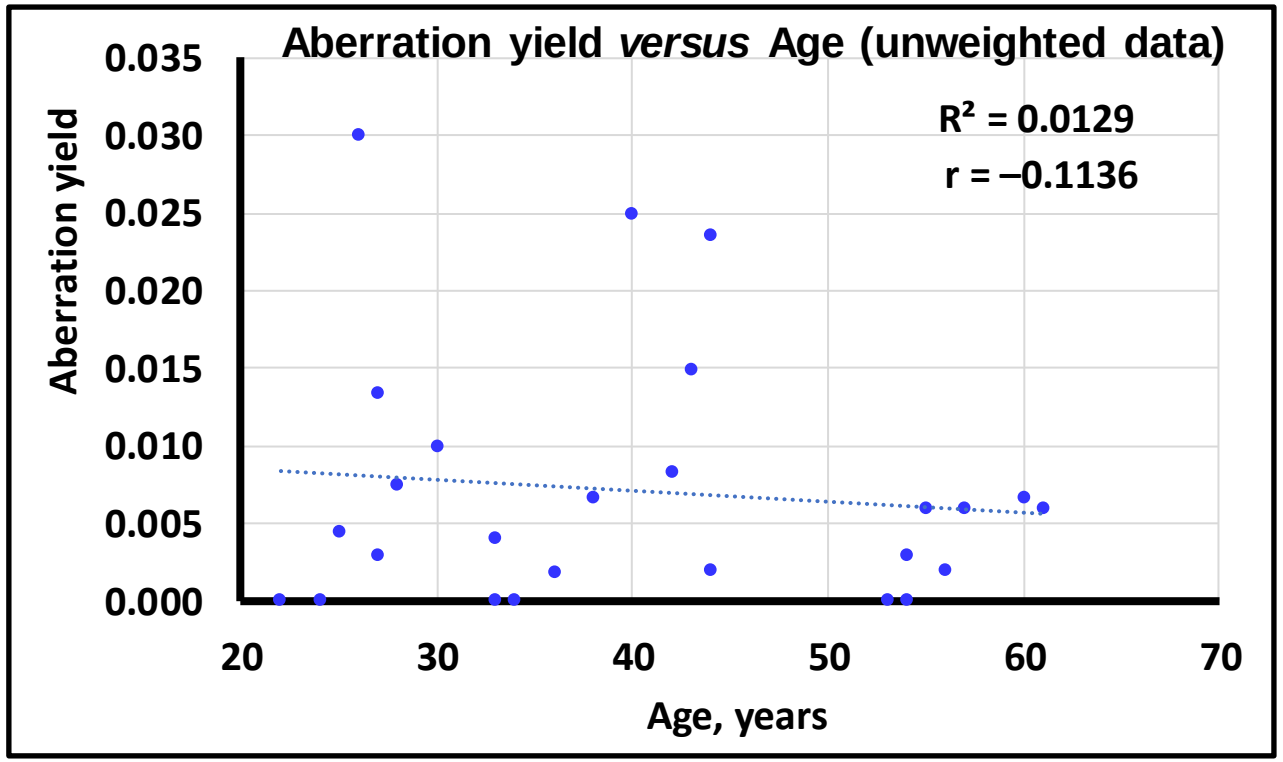
$$r_{Corrected} = -0.11547 * [\,1 + \frac{(1 - 0.11547^2)}{2 * (26 - 3)}\,] = -0.11795 \qquad R^2_{Corrected} = 0.01391$$

$$\text{Error} = \sqrt{[(1 - 0.11795^2)/(26 - 2)]} = 0.20270$$

$t = 0.11795 / 0.20270 = 0.58189;$
$df = 26 - 2 = 24$
$p = 0.71697$ by T.DIST.RT(0.58189,24)

Conclusion: The correlation between the age and individual aberration yield is low, negative, statistically insignificant.

Pearson's correlation coefficient can be obtained by the in-built function in MS Excel (Scatter Plot; "Add Trendline" & "Display R-squared value").



Correlation analysis: Formalism for *Spearman's rank correlation*

The test should be applied to the datasets containing from 5 to 40 pairs of measurements. The normality of the distribution of the individual values within a group is not necessary.

1. **Put the variable *X* (acting factor) in ascending order. Assign ranks (R_i) as a serial number, beginning with 1 for the smallest value. If two identical values occur, they are assigned the same rank value equal to the arithmetic mean of the ranks of these values.**
2. **Perform the ranking procedure for the variable *Y* (effect).**
3. **Compute the difference in ranks of each pair of the linked values of *X* and *Y*, $d_i = d_x - d_y$.**
4. **Put the difference d_i into square and find the sum, $\sum d^2$.**
5. **Calculate the correlation coefficient of ranks using the formula: $r_S = 1 - \frac{6\sum d^2}{n(n^2-1)}$**

   where *n* is the number of pairs (i.e. the size of the group).

6. **If there are identical values of the compared variables, a correction should be made for the same ranks.**

$$r_S^* = 1 - \frac{6\sum d^2 + T_x + T_y}{n(n^2-1)}$$

, where $T_x = (A_x^3 - A_x):12$ and $T_y = (A_y^3 - A_y):12$

A_x и A_y, respectively, are the numbers of identical ranks in the sets of variable *X* and variable *Y*.

With many identical ranks in the group, the coefficient gives approximate values, thus better, try Pearson's linear regression coefficient.

The statistical significance of the Spearman's rank correlation coefficient

Method 1. A comparison with a critical ('statistical') value of r_s for p<0.05 or p<0.01:

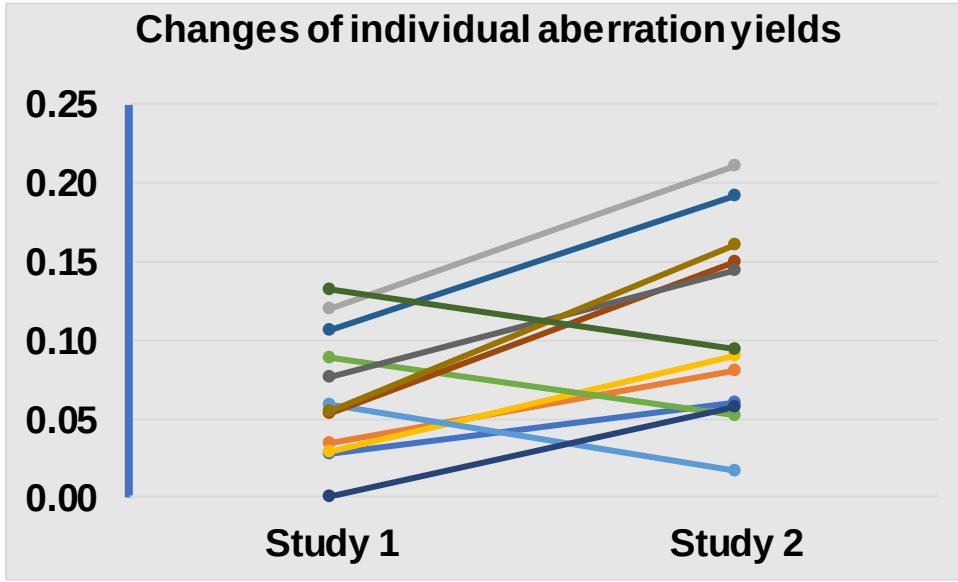
$$r_{st} = \frac{t}{\sqrt{n-1}}\left(1 - \frac{m}{n-1}\right) \quad ;$$

t=1.96 and m=0.16 for p<0.05
t=2.58 and m=0.69 for p<0.01

Method 2. Student's *t*-test (for n ≥ 10): $t_s = r_s\sqrt{\frac{n-2}{1-r_s^2}} \geqslant t_{st}$ for $df = n - 2$

Example – Radiotherapy patients; Aberration yields measured in the mid- and end of the RT course

Patients	Cells	DR	Yield Study 1	Cells	DR	Yield Study 2
IV-8-2	214	6	0.0280	183	11	0.0601
IV-7-2	58	2	0.0345	320	26	0.0813
IV-3-2	25	3	0.1200	109	23	0.2110
I-2-4	69	2	0.0290	695	63	0.0906
I-1-4	137	8	0.0584	603	10	0.0166
I-3-2	224	20	0.0893	133	7	0.0526
III-1-2	65	0	0.0000	140	8	0.0571
II-2-3	75	4	0.0533	100	15	0.1500
IV-4-2	65	5	0.0769	598	86	0.1438
IV-5-2	55	3	0.0545	125	20	0.1600
IV-6-2	600	64	0.1067	229	44	0.1921
IV-2-2	370	49	0.1324	254	24	0.0945



Step 1. Sort by "Yield Study 1" and assign ranks for X

Patients	Cells	DR	Yield Study 1	Rank X	Cells	DR	Yield Study 2
III-1-2	65	0	0.0000	1	140	8	0.0571
IV-8-2	214	6	0.0280	2	183	11	0.0601
I-2-4	69	2	0.0290	3	695	63	0.0906
IV-7-2	58	2	0.0345	4	320	26	0.0813
II-2-3	75	4	0.0533	5	100	15	0.1500
IV-5-2	55	3	0.0545	6	125	20	0.1600
I-1-4	137	8	0.0584	7	603	10	0.0166
IV-4-2	65	5	0.0769	8	598	86	0.1438
I-3-2	224	20	0.0893	9	133	7	0.0526
IV-6-2	600	64	0.1067	10	229	44	0.1921
IV-3-2	25	3	0.1200	11	109	23	0.2110
IV-2-2	370	49	0.1324	12	254	24	0.0945

Step 2. Sort by "Yield Study 2" and assign ranks for Y

Patients	Cells	DR	Yield Study 1	Rank X	Cells	DR	Yield Study 2	Rank Y
I-1-4	137	8	0.0584	7	603	10	0.0166	1
I-3-2	224	20	0.0893	9	133	7	0.0526	2
III-1-2	65	0	0.0000	1	140	8	0.0571	3
IV-8-2	214	6	0.0280	2	183	11	0.0601	4
IV-7-2	58	2	0.0345	4	320	26	0.0813	5
I-2-4	69	2	0.0290	3	695	63	0.0906	6
IV-2-2	370	49	0.1324	12	254	24	0.0945	7
IV-4-2	65	5	0.0769	8	598	86	0.1438	8
II-2-3	75	4	0.0533	5	100	15	0.1500	9
IV-5-2	55	3	0.0545	6	125	20	0.1600	10
IV-6-2	600	64	0.1067	10	229	44	0.1921	11
IV-3-2	25	3	0.1200	11	109	23	0.2110	12

Step 3. Calculate d_i and d_i^2

Patients	Cells	DR	Yield Study 1	Rank X	Cells	DR	Yield Study 2	Rank Y	d_i = Rank Y – Rank X	d_i^2
I-1-4	137	8	0.0584	7	603	10	0.0166	1	-6	36
I-3-2	224	20	0.0893	9	133	7	0.0526	2	-7	49
III-1-2	65	0	0.0000	1	140	8	0.0571	3	2	4
IV-8-2	214	6	0.0280	2	183	11	0.0601	4	2	4
IV-7-2	58	2	0.0345	4	320	26	0.0813	5	1	1
I-2-4	69	2	0.0290	3	695	63	0.0906	6	3	9
IV-2-2	370	49	0.1324	12	254	24	0.0945	7	-5	25
IV-4-2	65	5	0.0769	8	598	86	0.1438	8	0	0
II-2-3	75	4	0.0533	5	100	15	0.1500	9	4	16
IV-5-2	55	3	0.0545	6	125	20	0.1600	10	4	16
IV-6-2	600	64	0.1067	10	229	44	0.1921	11	1	1
IV-3-2	25	3	0.1200	11	109	23	0.2110	12	1	1
										$\Sigma d_i^2 = 162$

Step 4. Calculate r_s, t and p

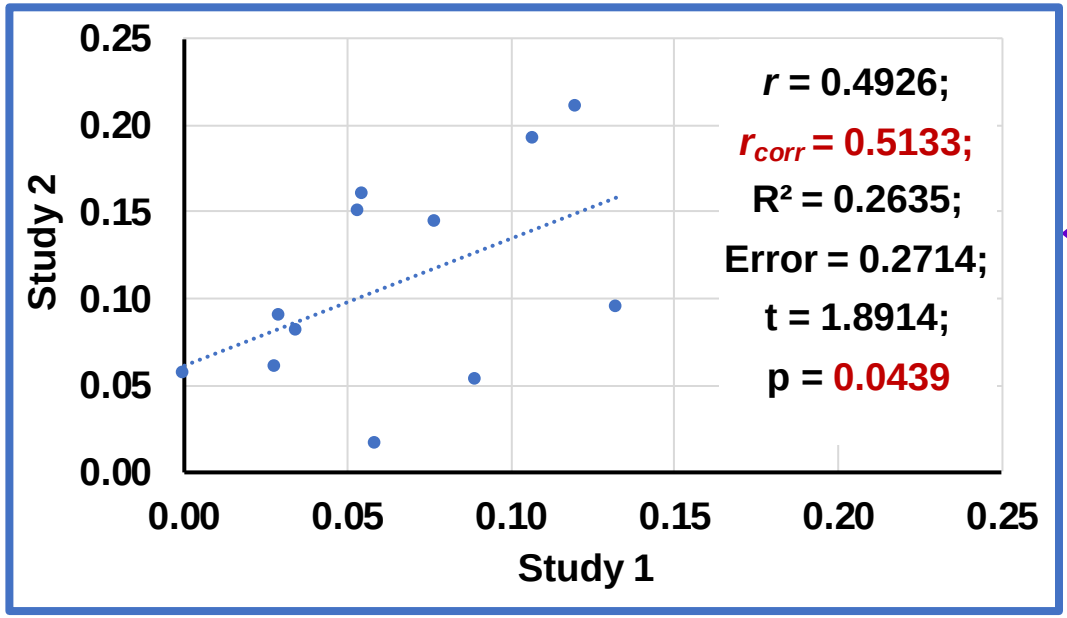
$r_s = 1 - 6x162 / (12 \times (12^2 - 1)) =$
$= 0.4336$

$t = 0.4336 \times \sqrt{[(12 - 2) / (1 - 0.4336^2)]}$
$= 1.5216$

For $df = 12 - 2 = 10$
$p = 0.0795$ by T.DIST.RT (1.5216, 10)

Correlation analysis: Pearson's coefficient for the same example

Pearson's coefficient; unweighted data



Actually, the analysis of data must be started from making a scatter plot like this.

We can use this Figure also to illustrate Spearman's correlation test.

Pearson's coefficient, adapted for weighted data for X and Y

$$r_{xy} = \frac{n\sum_{l=1}^{n} x_l y_l f_{i_x} f_{i_y} - \sum_{l=1}^{n} x_l f_{i_x} f_{i_y} \sum_{l=1}^{n} y_l f_{i_x} f_{i_y}}{\sqrt{n\sum x_l^2 f_{i_x} f_{i_y} - (\sum x_l f_{i_x} f_{i_y})^2}\ \sqrt{n\sum y_l^2 f_{i_x} f_{i_y} - (\sum y_l f_{i_x} f_{i_y})^2}}$$

Correlation analysis: Pearson's coefficient; weighted data for X and Y for the same example

	Study 1				Study 2										
Patients	Cells	DR	Y_1	f_1	Cells	DR	Y_2	f_2	$Y_1 * f_1 * f_2$	Y_1^2	$Y_1^2 * f_1 * f_2$	$Y_2 * f_1 * f_2$	Y_2^2	$Y_2^2 * f_1 * f_2$	$Y_1 * Y_2 * f_1 * f_2$
IV-8-2	214	6	0.02804	1.31221	183	11	0.06011	0.62941	0.02316	0.00079	0.00065	0.04965	0.00361	0.00298	0.00139
I-3-2	224	20	0.08929	1.37353	133	7	0.05263	0.45744	0.05610	0.00797	0.00501	0.03307	0.00277	0.00174	0.00295
IV-2-2	370	49	0.13243	2.26878	254	24	0.09449	0.87360	0.26248	0.01754	0.03476	0.18728	0.00893	0.01770	0.02480
I-1-4	137	8	0.05839	0.84006	603	10	0.01658	2.07395	0.10174	0.00341	0.00594	0.02889	0.00028	0.00048	0.00169
IV-7-2	58	2	0.03448	0.35565	320	26	0.08125	1.10060	0.01350	0.00119	0.00047	0.03180	0.00660	0.00258	0.00110
III-1-2	65	0	0.00000	0.39857	140	8	0.05714	0.48151	0.00000	0.00000	0.00000	0.01097	0.00327	0.00063	0.00000
I-2-4	69	2	0.02899	0.42310	695	63	0.09065	2.39037	0.02932	0.00084	0.00085	0.09168	0.00822	0.00831	0.00266
IV-4-2	65	5	0.07692	0.39857	598	86	0.14381	2.05675	0.06306	0.00592	0.00485	0.11789	0.02068	0.01695	0.00907
IV-6-2	600	64	0.10667	3.67910	229	44	0.19214	0.78762	0.30909	0.01138	0.03297	0.55677	0.03692	0.10698	0.05939
IV-3-2	25	3	0.12000	0.15330	109	23	0.21101	0.37489	0.00690	0.01440	0.00083	0.01213	0.04452	0.00256	0.00146
II-2-3	75	4	0.05333	0.45989	100	15	0.15000	0.34394	0.00844	0.00284	0.00045	0.02373	0.02250	0.00356	0.00127
IV-5-2	55	3	0.05455	0.33725	125	20	0.16000	0.42992	0.00791	0.00298	0.00043	0.02320	0.02560	0.00371	0.00127
Total	1957	166		12	3489	337		12	0.88168	0.06925	0.08720	1.16704	0.18389	0.16818	0.10703

$$r = \frac{12 \times 0.10703 - 0.88168 \times 1.16704}{\sqrt{(12 \times 0.08720 - 0.88168^2)} \times \sqrt{(12 \times 0.16818 - 1.16704^2)}} = 0.60681$$

$r_{corrected} = 0.62910$; $R^2_{corrected} = 0.39577$; Error = 0.24581;

$t = 1.61006$; $df = 12 - 2 = 10$; $p = 0.06923$

To compare with Spearman's correlation:

$r_s = 0.4336$; t = 1.5216; p = 0.0795

Correlation analysis: other coefficients

Apart from Pearson's and Spearman's tests, correlation can be estimated by other coefficients, depending on the type of data for one or both variables (X and Y, e.g., "Factor" and "Effect").

Type of data		Correlation coefficient
Variable X	Variable Y	
Numerical (discrete, (continuous, interval or ratio)	Numerical (discrete, (continuous, interval or ratio)	**Pearson's (*r*)**
Numerical or Rank (ordinal)	Numerical or Rank (ordinal)	**Spearman's (r_s or ρ)**
Rank (ordinal)	Rank (ordinal)	**Kendall's τ (tau)**
Biserial (dichotomic)	Biserial (dichotomic)	**φ (phyta, close to χ^2)**
Biserial (dichotomic)	Numerical (discrete, (continuous, interval or ratio)	**Biserial coefficient**
Biserial (dichotomic)	Rank (ordinal)	**Rank-biserial coefficient**

$$\varphi = \frac{(\,|\,ad - bc\,|\,) - 0{,}5n}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

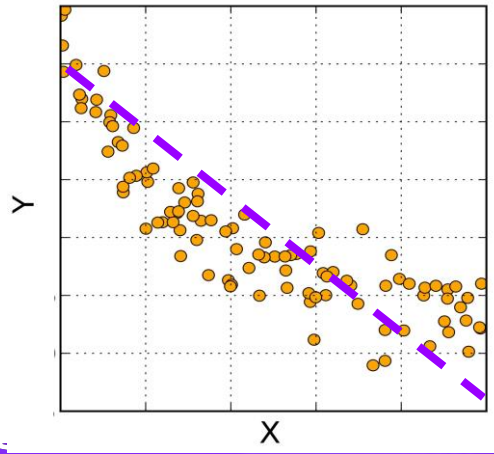
$$r_{bs} = \frac{\bar{y}_1 - \bar{y}_2}{\sigma_y} \sqrt{\frac{n_1 n_2}{N(N-1)}}$$

$$r_{rbs} = \frac{2 \times (\bar{R}_1 - \bar{R}_2)}{N}$$

Regression analysis: least squares methods

Regression analysis is a set of statistical processes for estimating the quantitative relationships between a dependent variable (effect) and one or more independent variables (often called ‘factors’, 'predictors’, 'covariates', 'explanatory variables’ or 'features’).
Regression analysis is used, in some situations, for inferring causal relationships between the independent variable (factor) and dependent variable (effect), but generally for prediction of quantitative changes of the effect associated with the changes of acting factor.

The most common form of regression analysis is **linear regression**, in which one finds the line that most closely fits the data according to a specific mathematical criterion. The method of ordinary **least squares** computes the unique line that minimizes the sum of squared differences between the true data and that line. σ_x



$$y_x = a_{yx} + b_{yx}x \quad \text{and} \quad x_y = a_{xy} + b_{xy}y$$

$$Q = \sum (y_i - y_x)^2 = \sum (y_i - f(x))^2 = Q_{\min}$$

Regression analysis: Basic formalism

$$\begin{cases} an + b\sum x = \sum y; \\ a\sum x + b\sum x^2 = \sum xy \end{cases},$$

for weighted data

$$\begin{cases} an + b\sum x f_i = \sum y f_i; \\ a\sum x f_i + b\sum x^2 f_i = \sum xy f_i \end{cases},$$

where

$$b_{yx} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \quad \text{or} \quad b_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x}$$

for weighted data

$$b_{yx} = \frac{\sum xy f_i - n\bar{x}\bar{y}}{\sum x^2 f_i - n\bar{x}^2}$$

$$a_{yx} = \frac{\sum y \sum x^2 - \sum x \sum yx}{n\sum x^2 - (\sum x)^2} \quad \text{or} \quad a_{yx} = \bar{y} - b_{yx}\bar{x}$$

for weighted data

$$a_{yx} = \frac{\sum y f_i \sum x^2 f_i - \sum x f_i \sum yx f_i}{n\sum x^2 f_i - (\sum x f_i)^2}$$

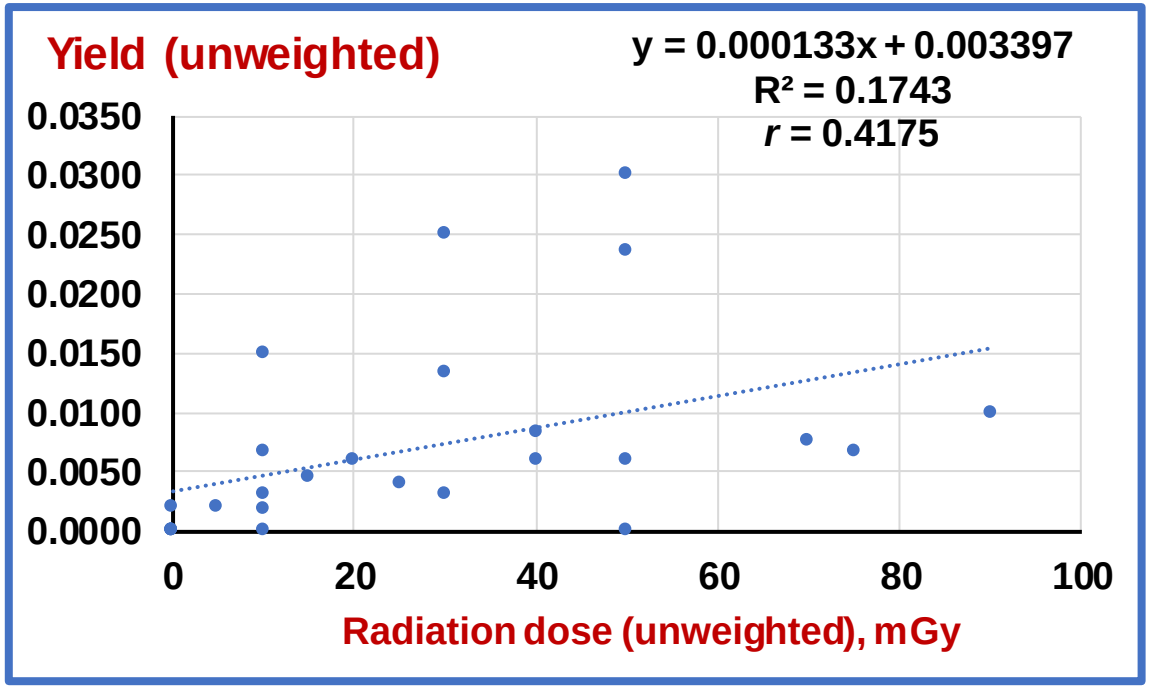
The goodness of fit of a regression model describes how well it fits a set of data.

Estimates are r^2 (or R^2) or χ^2.

Example of the regression fitting by the Least Squares method

Nr	Individial	Dose, mGy	Cells scored	Dic+CR found	Yield
1	AAA	0	1000	0	0.0000
2	AAB	0	250	0	0.0000
3	AAC	0	513	1	0.0019
4	AAD	0	750	0	0.0000
5	AAE	25	500	2	0.0040
6	AAF	10	204	0	0.0000
7	GAA	15	450	2	0.0044
8	GAB	75	300	2	0.0067
9	GAC	0	200	0	0.0000
10	GAD	5	1000	2	0.0020
11	GAE	10	600	4	0.0067
12	GAF	10	336	5	0.0149
13	WAA	10	557	1	0.0018
14	WAB	20	500	3	0.0060
15	WAC	30	1000	3	0.0030
17	WAD	40	1000	6	0.0060
18	WAE	50	1000	6	0.0060
19	RAA	50	127	3	0.0236
20	RAB	30	300	4	0.0133
21	RAC	70	400	3	0.0075
23	RAD	90	200	2	0.0100
24	RAE	10	333	1	0.0030
25	RAF	50	200	6	0.0300
26	RAG	30	80	2	0.0250
27	RTA	40	1200	10	0.0083
28	RTB	50	200	0	0.0000

– Virtual group; Study 2; Aberration yield versus recorded radiation dose (mGy)



Example of the regression fitting by the Least Squares method

Nr	Individial	Dose, mGy	Cells scored	Dic+CR found	Yield	Dose ^{2}	Dose x Yield
1	AAA	0	1000	0	0.0000	0.00	0.0000
2	AAB	0	250	0	0.0000	0.00	0.0000
3	AAC	0	513	1	0.0019	0.00	0.0000
4	AAD	0	750	0	0.0000	0.00	0.0000
5	AAE	25	500	2	0.0040	625.00	0.1000
6	AAF	10	204	0	0.0000	100.00	0.0000
7	GAA	15	450	2	0.0044	225.00	0.0667
8	GAB	75	300	2	0.0067	5625.00	0.5000
9	GAC	0	200	0	0.0000	0.00	0.0000
10	GAD	5	1000	2	0.0020	25.00	0.0100
11	GAE	10	600	4	0.0067	100.00	0.0667
12	GAF	10	336	5	0.0149	100.00	0.1488
13	WAA	10	557	1	0.0018	100.00	0.0180
14	WAB	20	500	3	0.0060	400.00	0.1200
15	WAC	30	1000	3	0.0030	900.00	0.0900
17	WAD	40	1000	6	0.0060	1600.00	0.2400
18	WAE	50	1000	6	0.0060	2500.00	0.3000
19	RAA	50	127	3	0.0236	2500.00	1.1811
20	RAB	30	300	4	0.0133	900.00	0.4000
21	RAC	70	400	3	0.0075	4900.00	0.5250
23	RAD	90	200	2	0.0100	8100.00	0.9000
24	RAE	10	333	1	0.0030	100.00	0.0300
25	RAF	50	200	6	0.0300	2500.00	1.5000
26	RAG	30	80	2	0.0250	900.00	0.7500
27	RTA	40	1200	10	0.0083	1600.00	0.3333
28	RTB	50	200	0	0.0000	2500.00	0.0000
	Total	720	13200	68	0.1842	36300.00	7.2796
		Σ x	n=26		Σ y	Σ x^2	Σ xy
	Mean	27.6923			0.0071		

Solution 1:

$$\begin{cases} an + b\sum x = \sum y; \\ a\sum x + b\sum x^2 = \sum xy \end{cases}$$

$$\begin{cases} a \times 26 + b \times 720 = 0.1842 \\ a \times 720 + b \times 36300 = 7.2796 \end{cases}$$

$a = (0.1842 - b \times 720) / 26$

$720 \times (0.1842 - b \times 720) / 26 + b \times 36300 = 7.2796$

$5.10092 - b \times 19\,938.46154 + b \times 36300 = 7.2796$

$b \times 16\,361.53846 = 2.17868$

$b = 0.00013316$

$a = (0.1842 - b \times 720) / 26 = 0.003397$

Solution 2: $b_{yx} = \dfrac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$ $\quad a_{yx} = \bar{y} - b_{yx}\bar{x}$

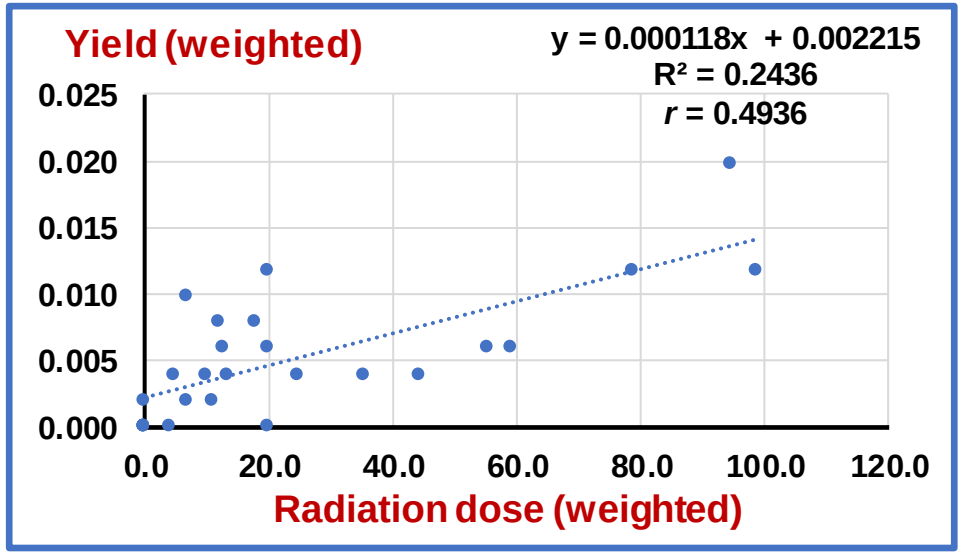
$$b = \frac{(7.2796 - 26 \times 27.6923 \times 0.00708)}{(36300 - 26 \times 27.6923^2)} = 0.00013316$$

$a = 0.00708 - b \times 27.6923 = 0.00708 - 0.00369 = 0.003397$

Example of the regression fitting by the Least Squares method

– Virtual group; Study 2; Aberration yield versus recorded radiation dose (mGy); weighted data

Nr	Individial	Dose, mGy	Cells scored	Dic+CR found	Yield	Weight, f_i	Dose*f_i	Yield*f_i	Dose^2*f_i	Dose * Yield * f_i
1	AAA	0	1000	0	0.0000	1.9697	0.0000	0.00000	0	0
2	AAB	0	250	0	0.0000	0.4924	0.0000	0.00000	0	0
3	AAC	0	513	1	0.0019	1.0105	0.0000	0.00197	0	0
4	AAD	0	750	0	0.0000	1.4773	0.0000	0.00000	0	0
5	AAE	25	500	2	0.0040	0.9848	24.6212	0.00394	615.5303	0.0985
6	AAF	10	204	0	0.0000	0.4018	4.0182	0.00000	40.1818	0
7	GAA	15	450	2	0.0044	0.8864	13.2955	0.00394	199.4318	0.0591
8	GAB	75	300	2	0.0067	0.5909	44.3182	0.00394	3323.8636	0.2955
9	GAC	0	200	0	0.0000	0.3939	0.0000	0.00000	0	0
10	GAD	5	1000	2	0.0020	1.9697	9.8485	0.00394	49.2424	0.0197
11	GAE	10	600	4	0.0067	1.1818	11.8182	0.00788	118.1818	0.0788
12	GAF	10	336	5	0.0149	0.6618	6.6182	0.00985	66.1818	0.0985
13	WAA	10	557	1	0.0018	1.0971	10.9712	0.00197	109.7121	0.0197
14	WAB	20	500	3	0.0060	0.9848	19.6970	0.00591	393.9394	0.1182
15	WAC	30	1000	3	0.0030	1.9697	59.0909	0.00591	1772.72723	0.1773
17	WAD	40	1000	6	0.0060	1.9697	78.7879	0.01182	3151.5152	0.4727
18	WAE	50	1000	6	0.0060	1.9697	98.4848	0.01182	4924.2424	0.5909
19	RAA	50	127	3	0.0236	0.2502	12.5076	0.00591	625.3788	0.2955
20	RAB	30	300	4	0.0133	0.5909	17.7273	0.00788	531.8182	0.2364
21	RAC	70	400	3	0.0075	0.7879	55.1515	0.00591	3860.6061	0.4136
23	RAD	90	200	2	0.0100	0.3939	35.4545	0.00394	3190.9091	0.3545
24	RAE	10	333	1	0.0030	0.6559	6.5591	0.00197	65.5909	0.0197
25	RAF	50	200	6	0.0300	0.3939	19.6970	0.01182	984.8485	0.5909
26	RAG	30	80	2	0.0250	0.1576	4.7273	0.00394	141.8182	0.1182
27	RTA	40	1200	10	0.0083	2.3636	94.5455	0.01970	3781.8182	0.7879
28	RTB	50	200	0	0.0000	0.3939	19.6970	0.00000	984.8485	0
	Total		13200	68		26	648	0.13394	28932.3864	4.84545
						n=26	Σ (x*f_i)	Σ (y*f_i)	Σ (x^2*f_i)	Σ (x*y*f_i)
			Y = 68 / 13200 =	0.00515		Σ x / n =	24.9091	0.00515 = Σ y / n		
			Mean Yield, true			Mean dose, weighted			Mean Yield, weighted	



This plot is a mathematical abstraction!!!

Solution: $b_{yx} = \frac{\sum x y f_i - n\bar{x}\bar{y}}{\sum x^2 f_i - n\bar{x}^2}$ $\quad a_{yx} = \bar{y} - b_{yx}\bar{x}$

$$b = \frac{(4.84545 - 26 \times 24.9091 \times 0.00515)}{(28932.3864 - 26 \times 24.9091^2)} =$$

$$= 0.000118$$

$$a = 0.00515 - b \times 24.9091 = 0.002215$$

Testing regressions for the goodness of fit: unweighted *versus* weighted

Unweighted yield: $Y = 0.003397 + 0.000133 \times Dose$; $r = 0.4175$; $r_{corrected} = 0.4250$; $R^2_{corrected} = 0.1806$

Nr	Individial	Dose, mGy	Cells scored	Dic+CR found	Yield	Expected Yield	Expected Dic+CR	(Obs - Exp)	(Obs - Exp)^2	/ Expected
1	AAA	0	1000	0	0.0000	0.003397	3.397	-3.3970	11.539609	3.397
2	AAB	0	250	0	0.0000	0.003397	0.84925	-0.8493	0.721226	0.84925
3	AAC	0	513	1	0.0019	0.003397	1.742661	-0.7427	0.551545	0.316496
4	AAD	0	750	0	0.0000	0.003397	2.54775	-2.5478	6.491030	2.54775
5	AAE	25	500	2	0.0040	0.006722	3.361	-1.3610	1.852321	0.551122
6	AAF	10	204	0	0.0000	0.004727	0.964308	-0.9643	0.929890	0.964308
7	GAA	15	450	2	0.0044	0.005392	2.4264	-0.4264	0.181817	0.074933
8	GAB	75	300	2	0.0067	0.013372	4.0116	-2.0116	4.046535	1.008708
9	GAC	0	200	0	0.0000	0.003397	0.6794	-0.6794	0.461584	0.6794
10	GAD	5	1000	2	0.0020	0.004062	4.062	-2.0620	4.251844	1.046737
11	GAE	10	600	4	0.0067	0.004727	2.8362	1.1638	1.354430	0.477551
12	GAF	10	336	5	0.0149	0.004727	1.588272	3.4117	11.639888	7.328649
13	WAA	10	557	1	0.0018	0.004727	2.632939	-1.6329	2.666490	1.012743
14	WAB	20	500	3	0.0060	0.006057	3.0285	-0.0285	0.000812	0.000268
15	WAC	30	1000	3	0.0030	0.007387	7.387	-4.3870	19.245769	2.605357
17	WAD	40	1000	6	0.0060	0.008717	8.717	-2.7170	7.382089	0.846861
18	WAE	50	1000	6	0.0060	0.010047	10.047	-4.0470	16.378209	1.630159
19	RAA	50	127	3	0.0236	0.010047	1.275969	1.7240	2.972283	2.329432
20	RAB	30	300	4	0.0133	0.007387	2.2161	1.7839	3.182299	1.435991
21	RAC	70	400	3	0.0075	0.012707	5.0828	-2.0828	4.338056	0.853478
23	RAD	90	200	2	0.0100	0.015367	3.0734	-1.0734	1.152188	0.37489
24	RAE	10	333	1	0.0030	0.004727	1.574091	-0.5741	0.329580	0.209378
25	RAF	50	200	6	0.0300	0.010047	2.0094	3.9906	15.924888	7.925196
26	RAG	30	80	2	0.0250	0.007387	0.59096	1.4090	1.985394	3.359608
27	RTA	40	1200	10	0.0083	0.008717	10.4604	-0.4604	0.211968	0.020264
28	RTB	50	200	0	0.0000	0.010047	2.0094	-2.0094	4.037688	2.0094
	Total	720	13200	68					$\chi^2 =$	43.8549
								Degrees of freedom	$df = n - 1 =$	25
									Chi-squared critical =	37.65
									Exact p =	0.011258

The linear regression doesn't fit the data, because $p < 0.05$

Weighted yield: $Y = 0.002215 + 0.000118 \times Dose$; $r = 0.4936$; $r_{corrected} = 0.5017$; $R^2_{corrected} = 0.2517$

Weight, f_i	Expected Yield	Expected Dic+CR	(Obs - Exp)	(Obs - Exp)^2	/ Expected	x f_i
1.9697	0.00221476	2.21476	-2.2148	4.905162	2.21476	4.362406
0.4924	0.00221476	0.55369	-0.5537	0.306573	0.55369	0.272650
1.0105	0.00221476	1.13617188	-0.1362	0.018543	0.01632	0.016491
1.4773	0.00221476	1.66107	-1.6611	2.759154	1.66107	2.453853
0.9848	0.00516226	2.58113	-0.5811	0.337712	0.130839	0.128856
0.4018	0.00339376	0.69232704	-0.6923	0.479317	0.692327	0.278189
0.8864	0.00398326	1.792467	0.2075	0.043070	0.024028	0.021298
0.5909	0.01105726	3.317178	-1.3172	1.734958	0.523022	0.309059
0.3939	0.00221476	0.442952	-0.4430	0.196206	0.442952	0.174496
1.9697	0.00280426	2.80426	-0.8043	0.646834	0.230661	0.454333
1.1818	0.00339376	2.036256	1.9637	3.856290	1.893814	2.238144
0.6618	0.00339376	1.14030336	3.8597	14.897258	13.06429	8.646187
1.0971	0.00339376	1.89032432	-0.8903	0.792677	0.419334	0.460060
0.9848	0.00457276	2.28638	0.7136	0.509254	0.222734	0.219359
1.9697	0.00575176	5.75176	-2.7518	7.572183	1.316498	2.593103
1.9697	0.00693076	6.93076	-0.9308	0.866314	0.124996	0.246203
1.9697	0.00810976	8.10976	-2.1098	4.451087	0.548856	1.081079
0.2502	0.00810976	1.02993952	1.9701	3.881138	3.768317	0.942650
0.5909	0.00575176	1.725528	2.2745	5.173223	2.998052	1.771576
0.7879	0.01046776	4.187104	-1.1871	1.409216	0.336561	0.265169
0.3939	0.01282576	2.565152	-0.5652	0.319397	0.124514	0.049051
0.6559	0.00339376	1.13012208	-0.1301	0.016932	0.014982	0.009827
0.3939	0.00810976	1.621952	4.3780	19.167304	11.81743	4.655512
0.1576	0.00575176	0.4601408	1.5399	2.371166	5.153132	0.812009
2.3636	0.00693076	8.316912	1.6831	2.832785	0.340605	0.805067
0.3939	0.00810976	1.621952	-1.6220	2.630728	1.621952	0.638951
26					$\chi^2 =$	33.90542
				Degrees of freedom	$df = n - 1 =$	25
					Chi-squared critical =	37.65
					Exact p =	0.109945

The linear regression does fit the data, because $p > 0.05$

Regression analysis: errors of the regression coefficient, its significance and confidence limits

Error of the regression coefficient

$$S_b = \sqrt{\frac{\sum (y_i - \bar{y})^2 - \frac{[\sum (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum (x_i - \bar{x})^2}}{(n-2)\sum (x_i - \bar{x})^2}}$$

or

$$SE_{b_{yx}} = \sqrt{\frac{(1 - r^2)\sum (y_i - \bar{y})^2 f_i}{(n-2)\sum (x_i - \bar{x})^2 f_i}}$$

Significance by *t*-test

$t = b_{yx} / SE_{b_{yx}}$

$df = n - 2$

Residual standard deviation of the entire regression

$$\sigma_{yx} = \sqrt{\frac{\sum (y_i - \hat{y}_x)^2}{n-2}}$$

, where $\hat{y}_x$ – yield, expected by the regression for given x.

or $$\sigma_{yx} = \sigma_y \sqrt{1 - r_{xy}^2} \frac{n-1}{n-2}$$

Error of the regression "central line" $$SE_{y_x} = \sigma_{yx} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{n\sigma_x^2}}$$

Error of the point prognosis by the regression $$SE_{\hat{y}_x} = \sigma_{yx} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{n\sigma_x^2}}$$

Confidence Limits $= y_x \pm t\, SE_{y_x}$ $df = n - 2$; $t = 1.96$ for $p<0.05$ (CL=95%)

In the given example "Radiation Workers Group; Study 2":

Weighted yield versus recorded dose: Y = 0.002215 + 0.000118 x Dose; *r* = 0.4936

$$SE_{b_{yx}} = \sqrt{\frac{(1-r^2)\sum (y_i-\bar{y})^2 f_i}{(n-2)\sum (x_i-\bar{x})^2 f_i}} = \sqrt{\frac{(1-0.49359^2) \times 0.000730}{(26-2) \times 12800.35331}} = 0.000042$$

***t*-test:** $t = b_{yx} / Se_b = 0.000118 / 0.000042 = 2.78278$

$df = n - 2 = 24$; p = 0.005167 by MS Excel T.DIST.RT(2.78278, 24)

Residual standard deviation of the entire regression:

σ_y taken from ANOVA calculations, Slide 5

$$\sigma_{y\,x} = \sigma_y \sqrt{1-r^2_{xy}} \frac{n-1}{n-2} = 0.005405 \times \sqrt{(1-0.49359^2)} \frac{(26-1)}{(26-2)} = 0.004897$$

Error of the regression "central line":

$$SE_{y_x} = \sigma_{yx} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{n\sigma_x^2}} = 0.004897 \sqrt{(1/26) + \frac{(\text{Dose} - 24.90909)^2}{26 \times 512.01413}}$$

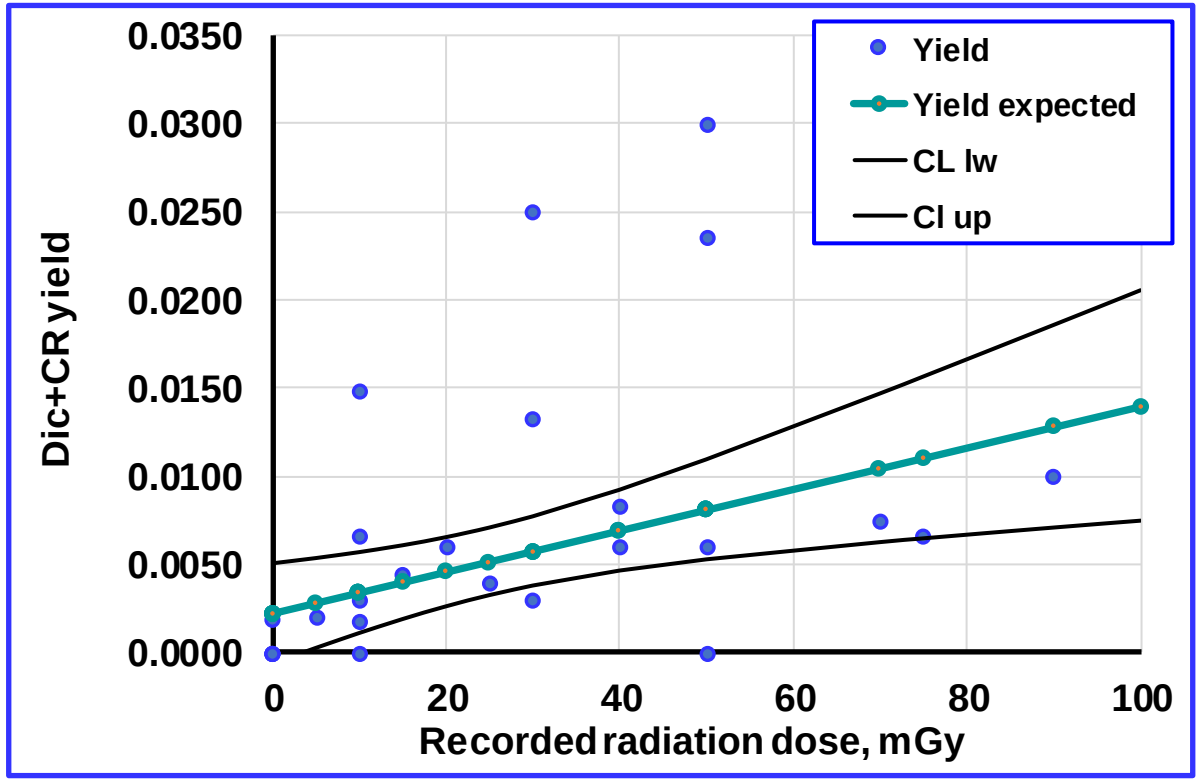
σ_x^2 estimated:

$$\sigma_x^2 = \frac{\sum (x_i - x_{mean})^2 \times f_i}{(n-1)}$$

In the given example “Radiation Workers Group; Study 2”:

Nr	Individial	Dose, mGy	Cells scored	Dic+CR found	Yield	Yield expected	SE	Cl_{lw}= = Y – t*SE	Cl_{up}= = Y + t*SE
1	AAA	0	1000	0	0.0000	0.002215	0.001428	-0.000584	0.005015
2	AAB	0	250	0	0.0000				
3	AAC	0	513	1	0.0019				
4	AAD	0	750	0	0.0000				
9	GAC	0	200	0	0.0000				
10	GAD	5	1000	2	0.0020	0.002805	0.001279	0.000298	0.005312
6	AAF	10	204	0	0.0000	0.003395	0.001150	0.001141	0.005649
11	GAE	10	600	4	0.0067				
12	GAF	10	336	5	0.0149				
13	WAA	10	557	1	0.0018				
24	RAE	10	333	1	0.0030				
7	GAA	15	450	2	0.0044	0.003985	0.001048	0.001930	0.006040
14	WAB	20	500	3	0.0060	0.004575	0.000982	0.002649	0.006501
5	AAE	25	500	2	0.0040	0.005165	0.000960	0.003283	0.007047
15	WAC	30	1000	3	0.0030	0.005755	0.000984	0.003826	0.007684
20	RAB	30	300	4	0.0133				
26	RAG	30	80	2	0.0250				
17	WAD	40	1000	6	0.0060	0.006935	0.001154	0.004672	0.009198
27	RTA	40	1200	10	0.0083				
18	WAE	50	1000	6	0.0060	0.008115	0.001434	0.005304	0.010926
19	RAA	50	127	3	0.0236				
25	RAF	50	200	6	0.0300				
28	RTB	50	200	0	0.0000				
21	RAC	70	400	3	0.0075	0.010475	0.002141	0.006278	0.014672
8	GAB	75	300	2	0.0067	0.011065	0.002332	0.006493	0.015638
23	RAD	90	200	2	0.0100	0.012835	0.002925	0.007102	0.018568
		100				0.014015	0.003329	0.007491	0.020539

Individual data were sorted by recorded radiation dose. Expected aberration yield was calculated for particular dose values using the regression equation; virtual point 100 mGy was added. Standard errors and Confidence Limits (CL) for the regression “central line” were calculated using formulas given on the previous slide. The true regression line with a given probability (95 %) must lie within this confidence zone.



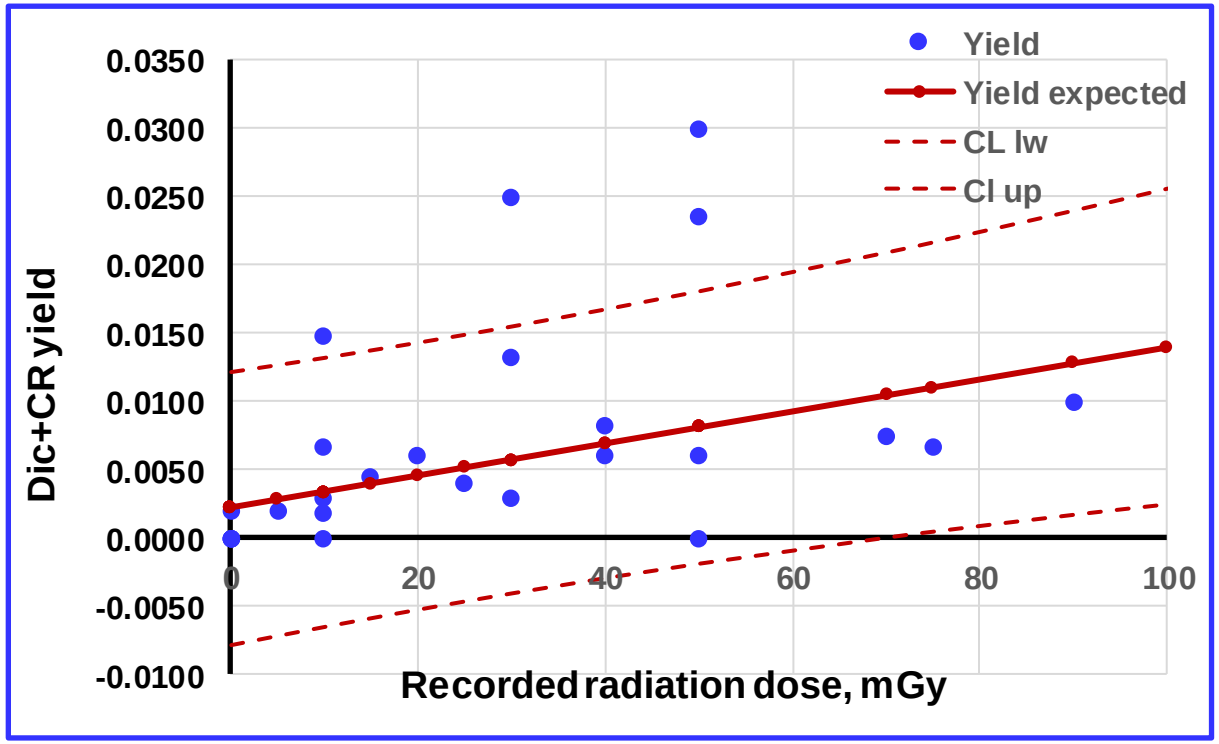
In the given example “Radiation Workers Group; Study 2”:

Nr	Individial	Dose, mGy	Cells scored	Dic+CR found	Yield	Yield expected	SE	CI_{lw}= = Y – t*SE	CI_{up}= = Y + t*SE
1	AAA	0	1000	0	0.0000	0.002215	0.005101	-0.007783	0.012213
2	AAB	0	250	0	0.0000				
3	AAC	0	513	1	0.0019				
4	AAD	0	750	0	0.0000				
9	GAC	0	200	0	0.0000				
10	GAD	5	1000	2	0.0020	0.002805	0.005061	-0.007115	0.012725
6	AAF	10	204	0	0.0000	0.003395	0.005030	-0.006464	0.013254
11	GAE	10	600	4	0.0067				
12	GAF	10	336	5	0.0149				
13	WAA	10	557	1	0.0018				
24	RAE	10	333	1	0.0030				
7	GAA	15	450	2	0.0044	0.003985	0.005008	-0.005831	0.013801
14	WAB	20	500	3	0.0060	0.004575	0.004995	-0.005214	0.014364
5	AAE	25	500	2	0.0040	0.005165	0.004990	-0.004616	0.014946
15	WAC	30	1000	3	0.0030	0.005755	0.004995	-0.004035	0.015545
20	RAB	30	300	4	0.0133				
26	RAG	30	80	2	0.0250				
17	WAD	40	1000	6	0.0060	0.006935	0.005031	-0.002926	0.016796
27	RTA	40	1200	10	0.0083				
18	WAE	50	1000	6	0.0060	0.008115	0.005103	-0.001886	0.018116
19	RAA	50	127	3	0.0236				
25	RAF	50	200	6	0.0300				
28	RTB	50	200	0	0.0000				
21	RAC	70	400	3	0.0075	0.010475	0.005345	-0.000001	0.020951
8	GAB	75	300	2	0.0067	0.011065	0.005424	0.000433	0.021697
23	RAD	90	200	2	0.0100	0.012835	0.005704	0.001655	0.024015
		100				0.014015	0.005921	0.002410	0.025620

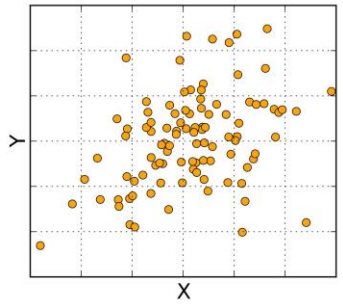
Standard errors and Confidence Limits (CL) <u>for the point prognoses</u> were calculated

$$SE = 0.004897 \sqrt{1 + (1/26) + \frac{(\text{Dose} - 24.90909)^2}{26 \times 512.01413}}$$

One can expect that 95 % values of aberration Yield predicted by this regression must lie within this confidence zone.



Other regressions: when linear regression doesn't fit well.

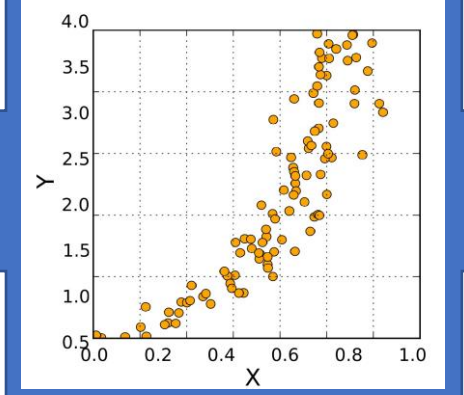


Multiply Linear Regression?

$$y = a + bx + cz$$

Exponential ?

$$y = ab^x \text{ or } y = ae^{bx}$$



Power ?

$$y_x = ax^b$$

Logarithmic ?

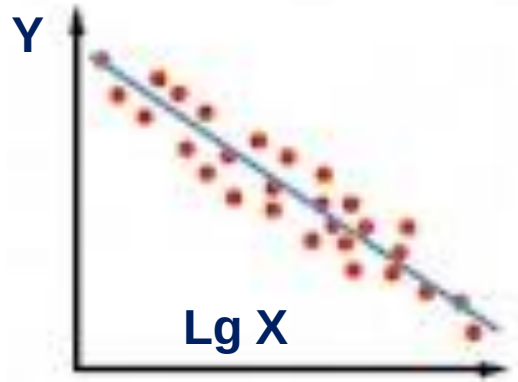
$$y = a + b\,Lg\,(x)$$

$$y = a + bx + cx^2$$

Linear-Quadratic ?

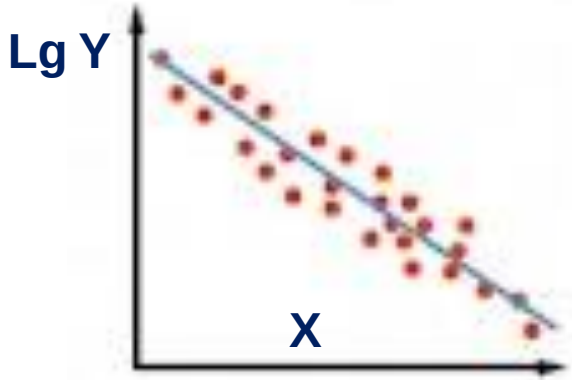
How to choose between non-linear regressions to fit the data:

Make scatter plots Y = *f* (Lg X); Lg Y = *f* (X) and Lg Y = *f* (Lg X). Check for the linear correlation.



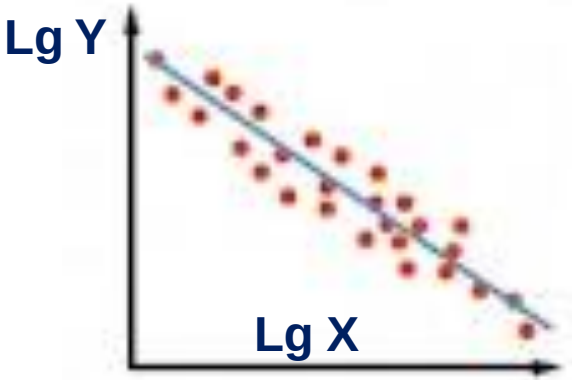
Lg X and Y ⟶ Logarithmic

$$y = a + b\,Lg\,(x)$$



X and Lg Y ⟶ Exponential

$$y = ab^x \text{ or } y = ae^{bx}$$



Lg X and Lg Y ⟶ Power

$$y_x = ax^b$$

The Linear-Quadratic model can be applied, when X ≤ 0 or Y ≤ 0, i.e. when taking logarithm is impossible.

If logarithmic transformation doesn't lead to arranging data points in a straight line, then a higher degree polynomic model should be tested.

Linear fit of Lg X and Lg Y is also indicative for Linear-Quadratic – the most universal model

$$y = a + bx + cx^2$$

Fitting of non-linear regressions by the Least Squares method

$$y = ae^{bx}$$

$$\begin{cases} n \ln a + \ln b \sum x = \sum \ln y; \\ \ln a \sum x + \ln b \sum x^2 = \sum (x \ln y) \end{cases}$$

$$\begin{cases} an + b \sum x + c \sum x^2 = \sum y; \\ a \sum x + b \sum x^2 + c \sum x^3 = \sum xy; \\ a \sum x^2 + b \sum x^3 + c \sum x^4 = \sum yx^2. \end{cases}$$

$$y = a + bx + cx^2$$

$$\begin{cases} an + b \sum x = \sum y; \\ a \sum x + b \sum x^2 = \sum xy \end{cases}$$

$$y = a + b\,Lg\,(x)$$

$$\begin{cases} na + b \sum \lg x = \sum y; \\ a \sum (\lg x) + b \sum (\lg x)^2 = \sum (\lg x)\, y \end{cases}$$

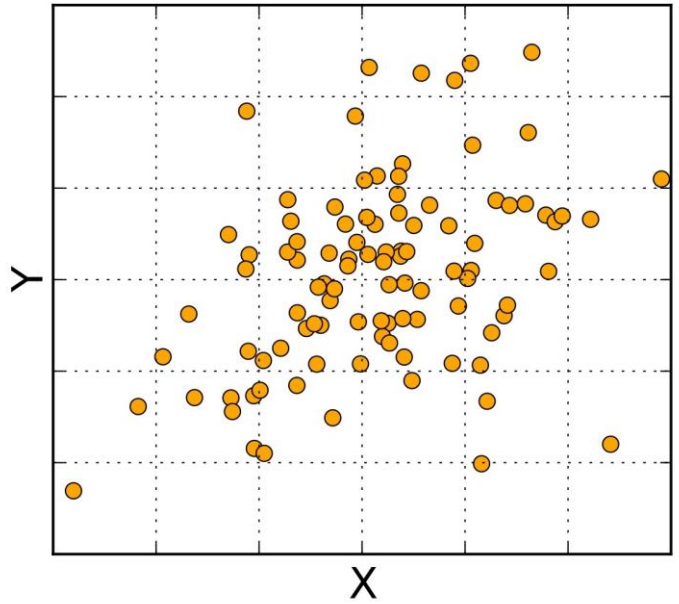
$$\begin{cases} n \lg a + b \sum \lg x = \sum \lg y; \\ \lg a \sum \lg x + b \sum (\lg x)^2 = \sum (\lg x \lg y) \end{cases}$$

$$y_x = ax^b$$

Multiple linear regression by the Least Squares method

If data show a weak linear inter-relationship with high heterogeneity of points, and possible combined action of two or more factors influencing the measured effect can be suggested, then a multiply linear regression can be tried.

Example: Stable translocation yield as a function of individual's age, life habits and the rate of accumulation of radiation dose.



$$y = a + bx + cz$$

$$\begin{cases} an + b\sum x + c\sum z = \Sigma y; \\ a\sum x + b\sum x^2 + c\sum xz = \sum xy; \\ a\sum z + b\sum xz + c\sum z^2 = \sum yz. \end{cases}$$

In case of any regression model: The number of experimental points (n) must be sufficient to maintain $df = (n - k) \geq 3$, where k is the number of coefficients in the model.

Important thing to remember:

To fit a regression for the dose response in radiation cytogenetics it is recommended to use

the Iteratively Reweighted Least Squares method

or

the Least Squares method,

but not the Weighted Least Squares method!

THANK YOU FOR YOUR ATTENTION!!!