E

TWC/25/12 Rev.
ORIGINAL: English
DATE: October 9, 2007

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Fifth Session
Sibiu, Romania, September 3 to 6, 2007

REVIEW OF TEST DESIGN: CHECKING LEVELS OF QUALITY (REVISED)

Document prepared by an expert from France

REVIEW OF TEST DESIGN: CHECKING LEVELS OF QUALITY

Dr. Sylvain Grégoire
GEVES
France
sylvain.gregoire@geves.fr

INTRODUCTION:

1. Checking quality is a task that many organizations deal with. It covers almost the entire range of human activity, from manufactured products to intellectual production such as writing software. It also covers for instance checking of the quality of raw material, of the quality of work provided by a team or persons, of the quality of a quality assurance system, etc.

2. A check of quality rarely relies only on human expertise; most of the time a decision is taken on the basis of measurements, or assessments, on a number of characteristics which define what quality is.

3. As an outcome of a quality check, decisions are taken which are generally of three kinds:

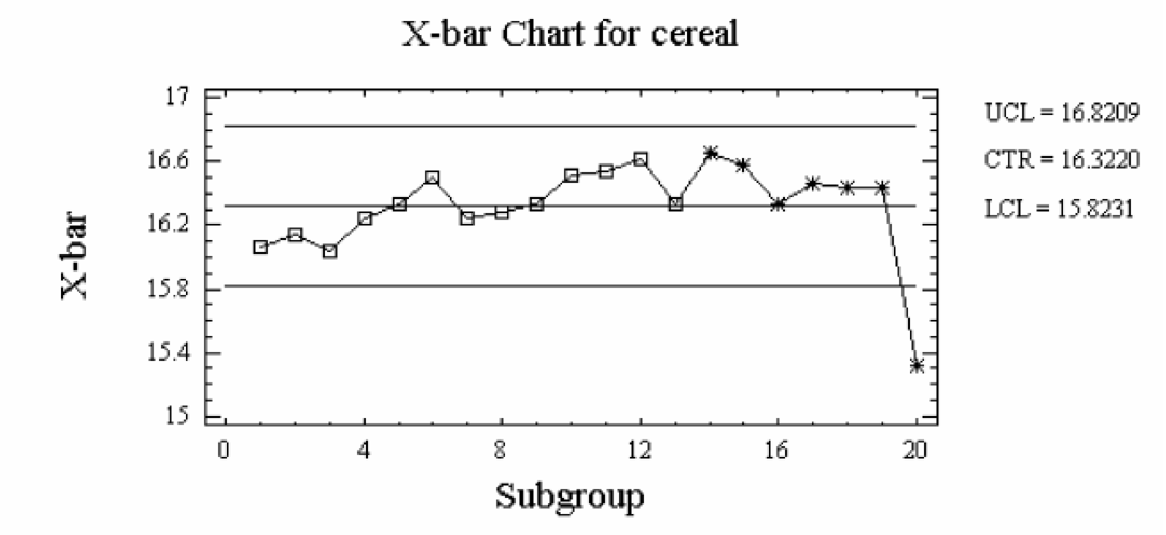
- We can go on with the product (to use it, to sell it, to buy it…).
- We do not use the product (not use it, not sell it, not buy it…).
- We try to improve the product, in order that it meets the quality required.

There are circumstances where the labelling, destination, etc., of a product are a consequence of its quality level. Top quality product is sold at a higher price or to more selective markets, while average quality product is sold on easier markets.

I Example of quality control charts on a set of measures that are supposed to stay between limits

4. When results are quantitative, the use of charts such as below (figure 1) is common for processes where successive (through time on the horizontal axis) data points are obtained on a production line where the value of the production is assumed to be constant. It can be applied to sets of data where time is not involved; in that case, the horizontal axis simply identifies the different lots.

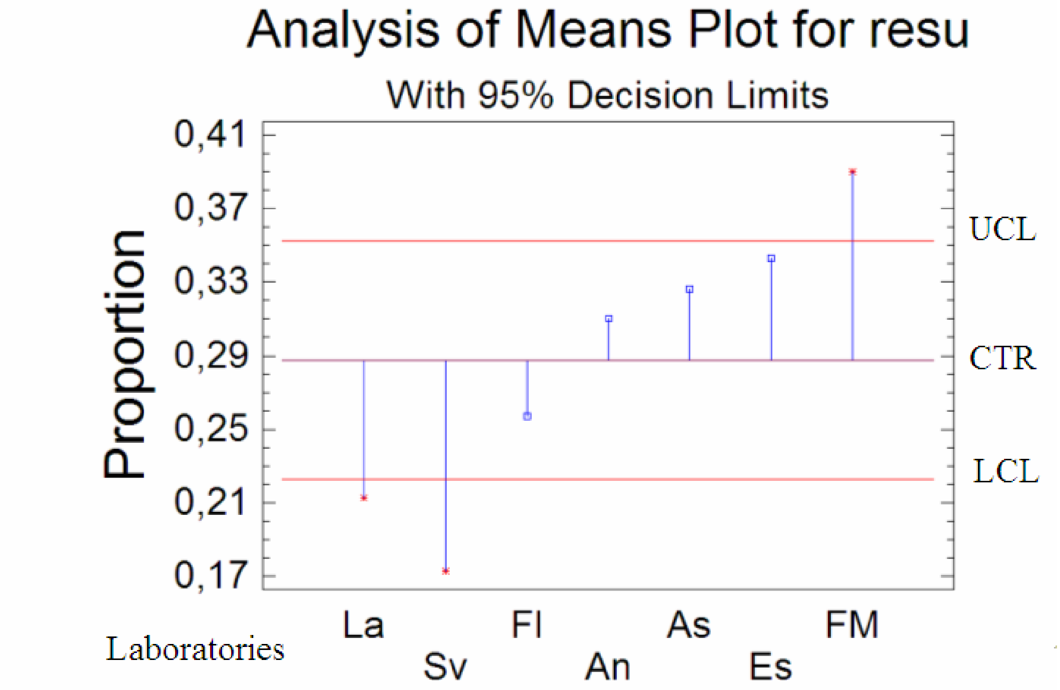
Figure 1: Control chart of a quantitative measure using 3 times the standard deviation to compute control limits (from Statgraphics)



5. The reference value (CTR) can be a pre-determined value which corresponds to the target level of quality. The reference value can also be the average of all data points available, or the moving average using the most recent set of data points. If some points are above the Upper Control Limit (UCL), or below the Lower Control Limit (LCL), action is needed, such as to withdraw lots with unexpected values, or to check the production line. The two lines (UCL and LCL) can be set as *a priori* quality limits (need to stay within LCL and UCL). The two lines (UCL and LCL) can also be computed on the basis of the observed variability of the process. If the standard deviation is used, 3 times the standard deviation from the reference quality level is a common way to compute UCL and LCL limits. In the example in figure 1, only the last point is below the limit, and is an alert for action.

6. The same kind of analysis can be applied when results are percentages or proportions. In the following example (see figure 2) the average proportion (p) of the 7 results obtained by different laboratories on samples from the same lot is used as the reference quality level (CTR), and the confidence interval of this percentage (binomial, with p=0,287 and n=300 plants observed) is drawn as lines to define the upper and lower control limits (UCL, LCL).

Figure 2: Control chart of proportions where the confidence interval of average proportions is used to define control limits of a characteristic, "resu" in the title of the graph stands as a short name for results on different observed variables (from Statgraphics)



7. In the above example, two data points are below the LCL (underestimation) and one data point is above the UCL (overestimation).

8. <u>The same type of chart can apply when results are counts</u>. Depending on the nature of the counts, quantitative or percentage methods as above may apply

9. In case of counts we can also be in a situation where we want to check a % of defects (see sections IV and V below).

II Example of quality check by assessing repeatability and reproducibility of laboratories

10. It is common practice for laboratories that use a given method to check repeatability and reproducibility. This is done prior to the use of a new method, or to check consistency at regular time intervals.

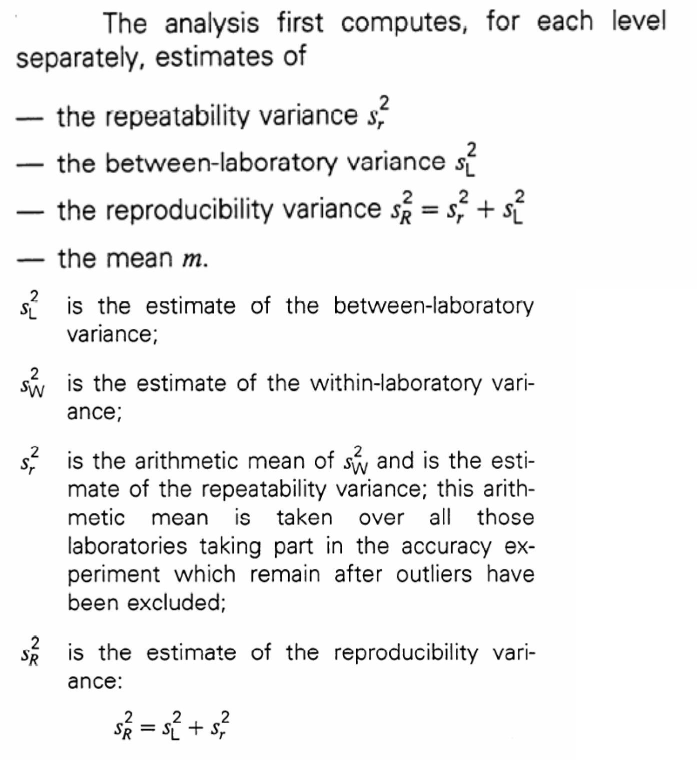
11. Samples from a small number of lots (usually 3 to 6) are distributed to participating laboratories, and on each sample laboratories perform an analysis, with some repeats. By repeats it is meant the laboratory will perform the technical check a number of times on each sample received, either on the whole sample or on sub-samples.

12. ISO 5725-2 describes a way to compute repeatability and reproducibility of measures as well as *h* and *k* values. Free software is available on the ISTA website to obtain values and graphs according to this ISO norm.

13. Repeatability is a measure of variability which measures within laboratory consistency.

14. Reproducibility is a measure of variability which sums repeatability and between laboratory variability.

15. We need to be aware that calculation of repeatability and reproducibility may vary from one reference document to another. The model in ISO5725-2 is y = m + B + e, where m is the general mean, B is the laboratory component of bias under repeatable conditions; e is the random error occurring in repeatable conditions.



The analysis first computes, for each level separately, estimates of

— the repeatability variance s_r^2

— the between-laboratory variance s_L^2

— the reproducibility variance $s_R^2 = s_r^2 + s_L^2$

— the mean m.

s_L^2 is the estimate of the between-laboratory variance;

s_W^2 is the estimate of the within-laboratory variance;

s_r^2 is the arithmetic mean of s_W^2 and is the estimate of the repeatability variance; this arithmetic mean is taken over all those laboratories taking part in the accuracy experiment which remain after outliers have been excluded;

s_R^2 is the estimate of the reproducibility variance:

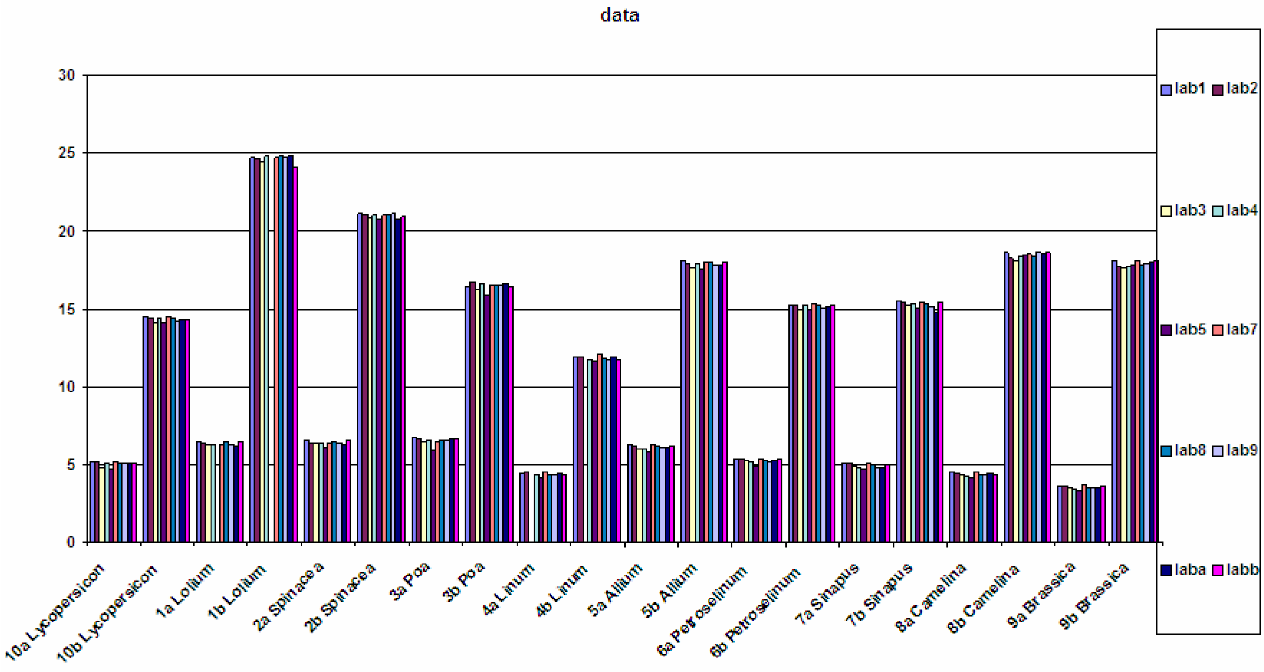
$$s_R^2 = s_L^2 + s_r^2$$

16. The ISO norm indicates which formulae to use, including when missing values occur, and gives examples of computation on small datasets.

17. Sometimes repeatability is considered to measure within laboratory variability (which is right), and reproducibility to measure between laboratories variability (which is wrong if we accept ISO 5725-2 definitions). Both being calculated on a set of samples that are supposed have the same measured value.

18. It can be seen from the definitions above that reproducibility is the sum of repeatability and a between laboratory variance. In other words, repeatability is a part of reproducibility.

Figure 3: Simple visualisation of concordance of results on 20 samples (horizontal axis) for the 10 laboratories (10 color bars per sample) average seed moisture on the vertical axis (From ISO5725-2 software free on ISTA website)



19. It is usually difficult to compare differences between laboratories without comparing each result to the average of the results on the same lot.

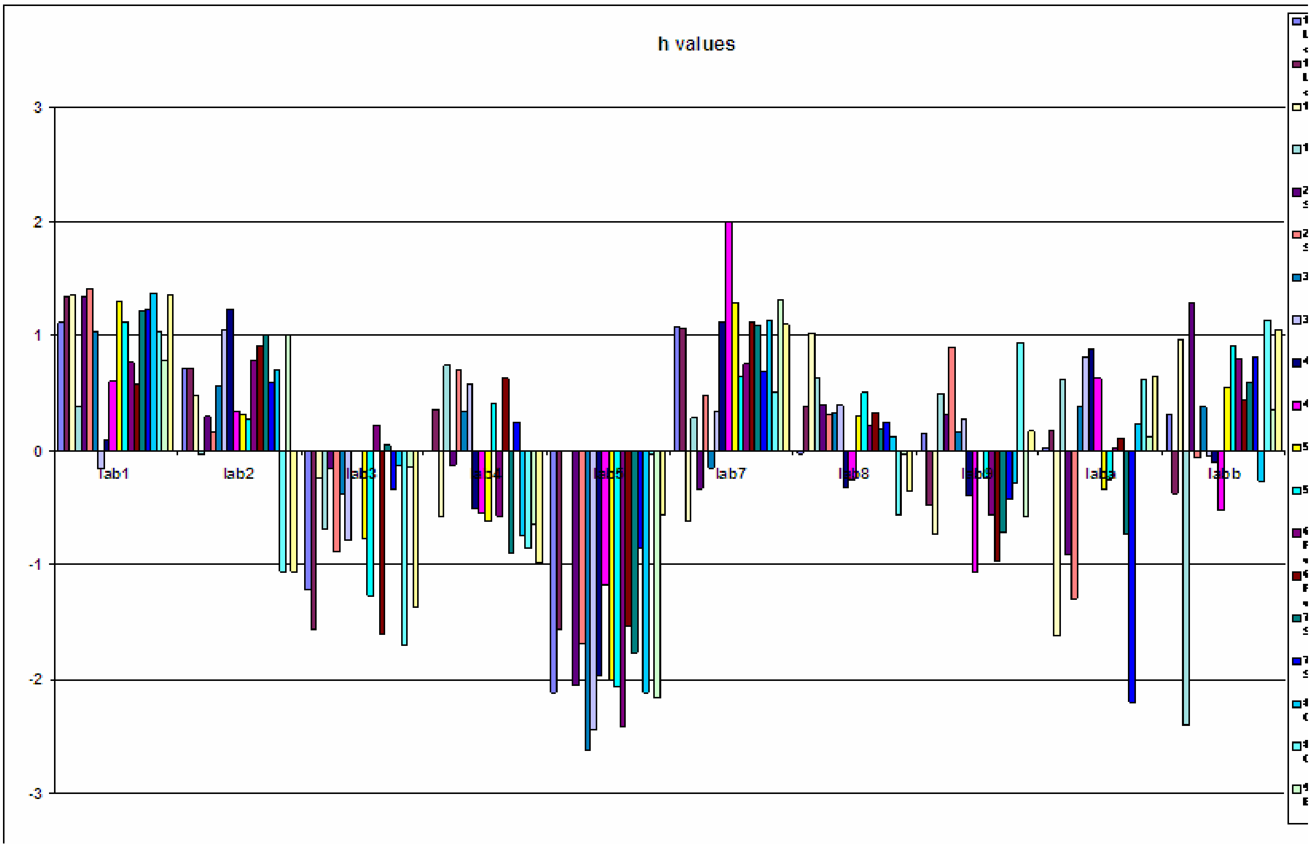
20. So, ISO 5725-2 defines *h* values which allow a better viewing of differences between laboratories, and k values which allow a better viewing of consistency of repeated measures.

21. *h* are standardised values which computes the difference of a result compared to the average of all laboratories on the samples form a given lot. *h* is equal to zero if the result obtained by a laboratory is equal to the average of all participating laboratories.

22. The *h* values are positive when a result is over the average result of all labs on the samples sent from the same lot, *h* values are negative if the result is below the average.

23. In this example (see Figure 3 above) 10 laboratories worked on 20 different samples. In the above graph, for each sample the values of the 10 labs are placed together. In the graph below (see Figure 4) the 20 *h* values of a given laboratory are plotted. “lab1” and “lab2” on the left of the graph have positive *h* values, indicating a tendency to over estimate, compared to the average from all laboratories.

Figure 4: Plot of h values (vertical axis) for the 10 laboratories (horizontal axis), for each lab. 20 color bars per laboratory, one per sample (From ISO5725-2 free on ISTA website)

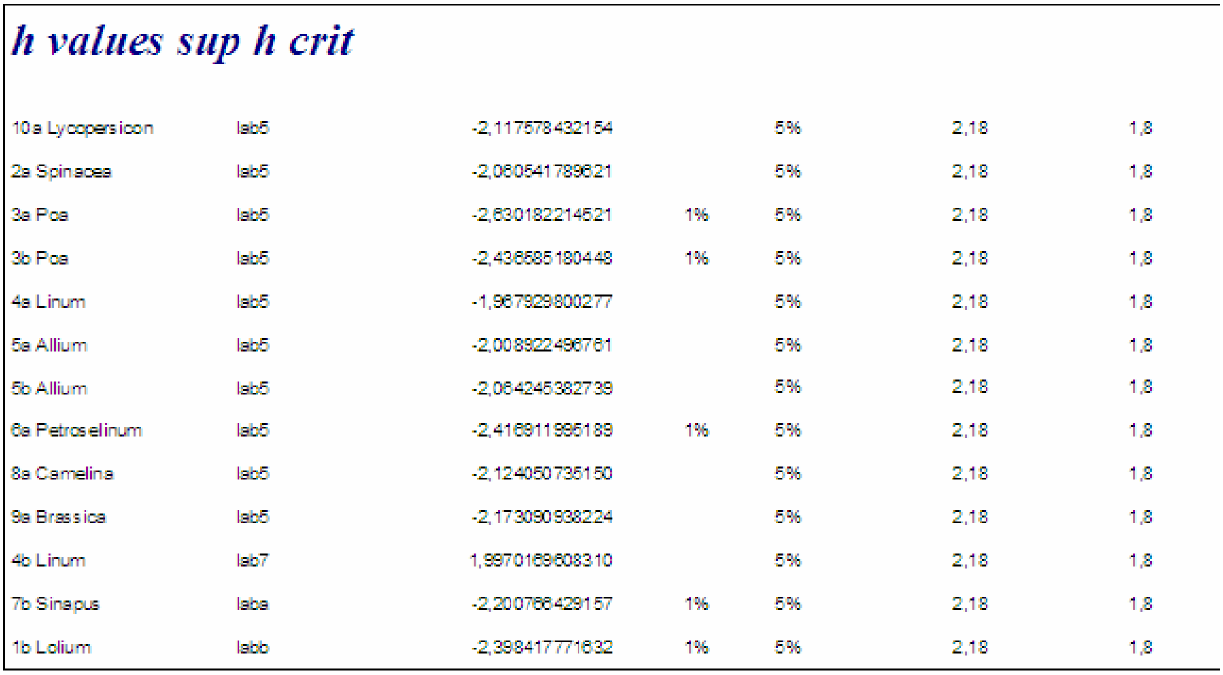


24. In the table below (see figure 5), the statistical analysis of *h* values indicates that “lab5” under estimates results. Thirteen *h* values are statistically significant, ten of them belonging to “lab5”. An ideal situation for high consistency of results is having small *h* values (with about equal numbers of positive and negative), indicating small departures from the average value which stands as reference.

25. A classical situation is a systematic over estimation (only positive *h* values) as “lab1” and “lab2”, or under estimation (only negative *h* values) as “lab3” and “lab5”, giving indication of necessary work on harmonization. Small, and either positive or negative values, as for “lab8”; indicates a laboratory has results which are consistent with the average results. The presence of high values, some positive some negative, as for “laba” and “labb” (right part of figure4) indicates that, depending of the samples a laboratory, either over-or under-estimate, which is more problematic than a systematic departure from the average.

26. If a laboratory is very different from others, a new computation is usually made after discarding the results from this laboratory. The average of reference (or experienced) laboratories is sometimes preferred to the average of all results, but ISO5725-2 explicitly defines the average of all results as the reference value.

Figure 5: Statistical test which identifies among h values, those which are significant (From ISO5725-2 free on ISTA website) 13 values at alpha level =5%, 5 values at alpha level =1%.

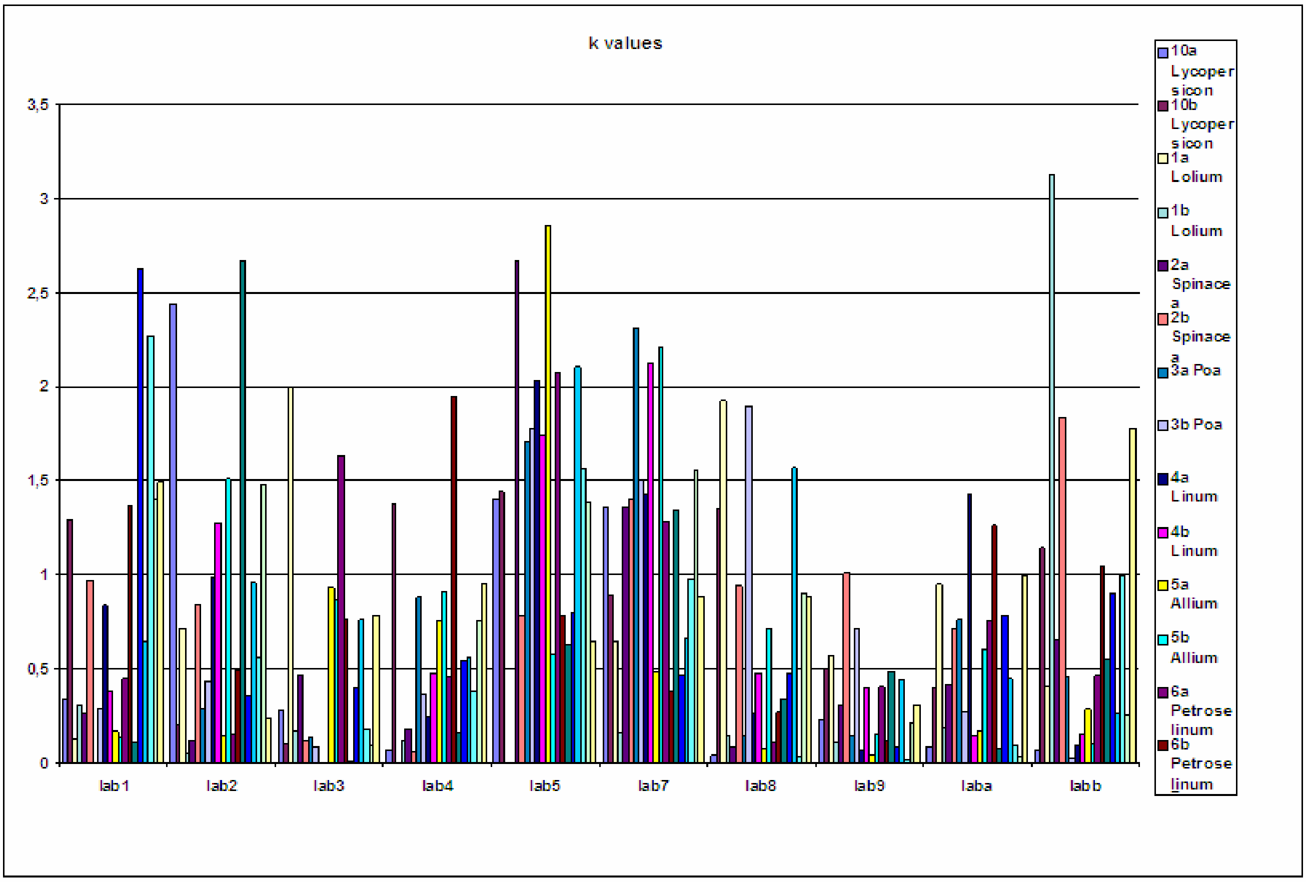


h values sup h crit

10a Lycopersicon	lab5	-2,11757843215 4		5%	2,18	1,8
2a Spinacea	lab5	-2,060541789621		5%	2,18	1,8
3a Poa	lab5	-2,630182214521	1%	5%	2,18	1,8
3b Poa	lab5	-2,436585180448	1%	5%	2,18	1,8
4a Linum	lab5	-1,967929800277		5%	2,18	1,8
5a Allium	lab5	-2,008922496761		5%	2,18	1,8
5b Allium	lab5	-2,064245382739		5%	2,18	1,8
6a Petroselinum	lab5	-2,416911995189	1%	5%	2,18	1,8
8a Camelina	lab5	-2,124050735150		5%	2,18	1,8
9a Brassica	lab5	-2,173090938224		5%	2,18	1,8
4b Linum	lab7	1,997016960831 0		5%	2,18	1,8
7b Sinapus	laba	-2,200766429157	1%	5%	2,18	1,8
1b Lolium	labb	-2,398417771632	1%	5%	2,18	1,8

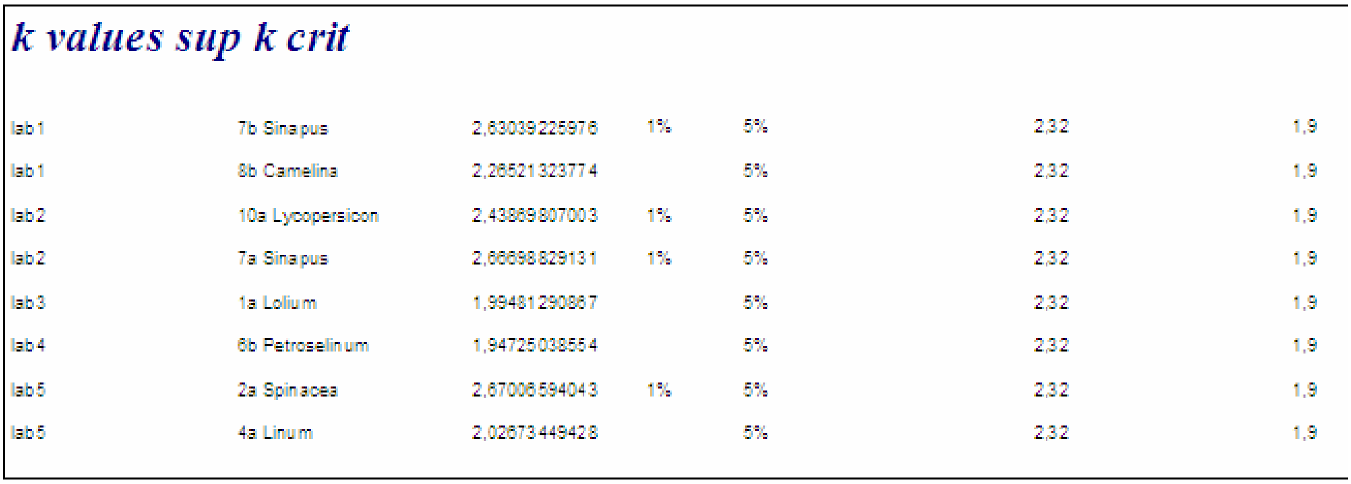
27. *k* values indicate the variability of repeated results. They are always positive. A zero value corresponds to a laboratory having obtained exactly the same value on all repeats from a sample.

Figure 5a: Plot of k values (vertical axis) for the 10 laboratories (horizontal axis), for each lab. 20 color bars per laboratory, one per sample (From ISO5725-2 free on ISTA website)



28. A statistical test on *k* values allows the spotting of unexpectedly high *k* values, but is not able to detect unexpectedly low *k* values.

Figure 5b: Statistical test which identifies among k values, those which are significant (From ISO5725-2 free on ISTA website) 8 values at alpha level =5%, 4 values at alpha level =1%.



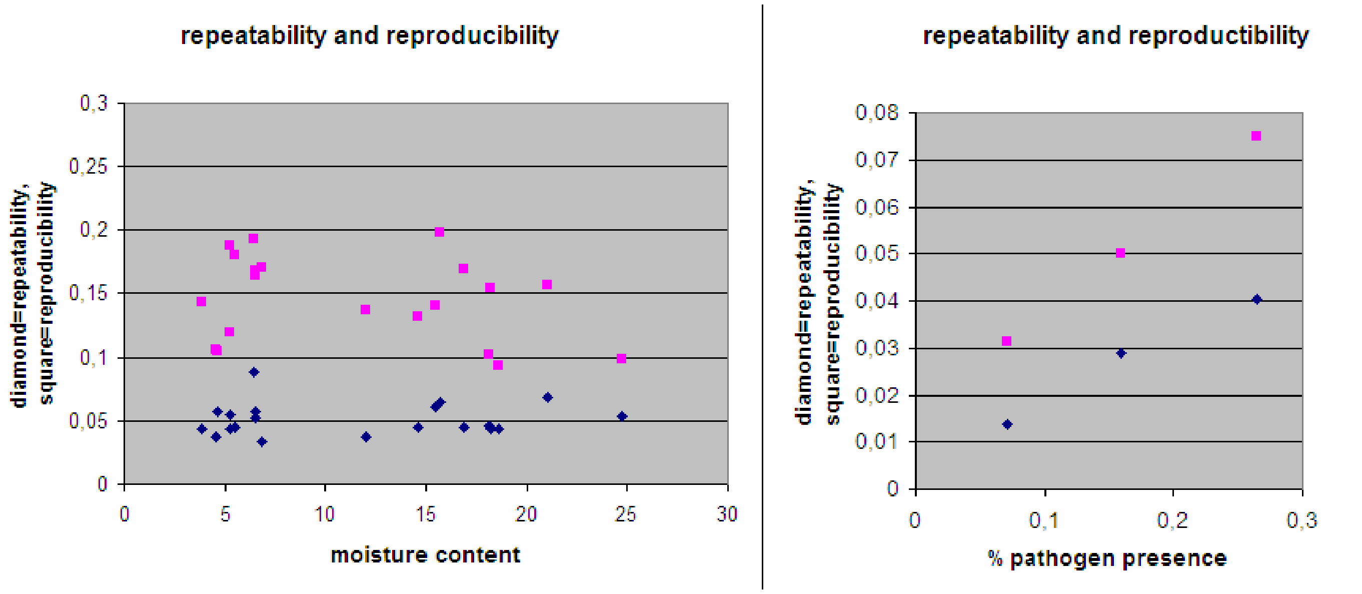
k values sup k crit

lab1	7b Sinapus	2,63039225976	1%	5%	2,32	1,9
lab1	8b Camelina	2,26521323774		5%	2,32	1,9
lab2	10a Lycopersicon	2,43869807003	1%	5%	2,32	1,9
lab2	7a Sinapus	2,66698829131	1%	5%	2,32	1,9
lab3	1a Lolium	1,99481290867		5%	2,32	1,9
lab4	6b Petroselinum	1,94725038554		5%	2,32	1,9
lab5	2a Spinacea	2,67006594043	1%	5%	2,32	1,9
lab5	4a Linum	2,02673449428		5%	2,32	1,9

29. The *h* and *k* values are independent of the unit and scale, which allows easy comparisons between studies. The *h* and *k* graphs are usually very useful for crop experts to visualise differences in estimation of the result (*h* values) and intra-laboratory variability (*k* values). Repeatability and reproducibility values are unit and scale dependant, which makes them difficult to use and to compare. It is not possible to know *a priori* if repeatability and reproducibility values are good or acceptable. Only when a set of values in conditions that can be compared are available, one can judge if repeatability and reproducibility meet the quality standards that are corresponding to the status of the art.

30. Depending on the nature of the characteristic, there are cases where repeatability and reproducibility are of the same magnitude over the range of measurement as shown below (figure 6) on the left graph (moisture level in different species of seeds), while in other cases we observe a link between the variability observed and the level of the measure as on the right graph (presence of pathogens in seeds).

Figure 6: values or repeatability and reproducibility (vertical axis) along quality criterion (horizontal axis) (values copied from ISO5725-2 free on ISTA website)



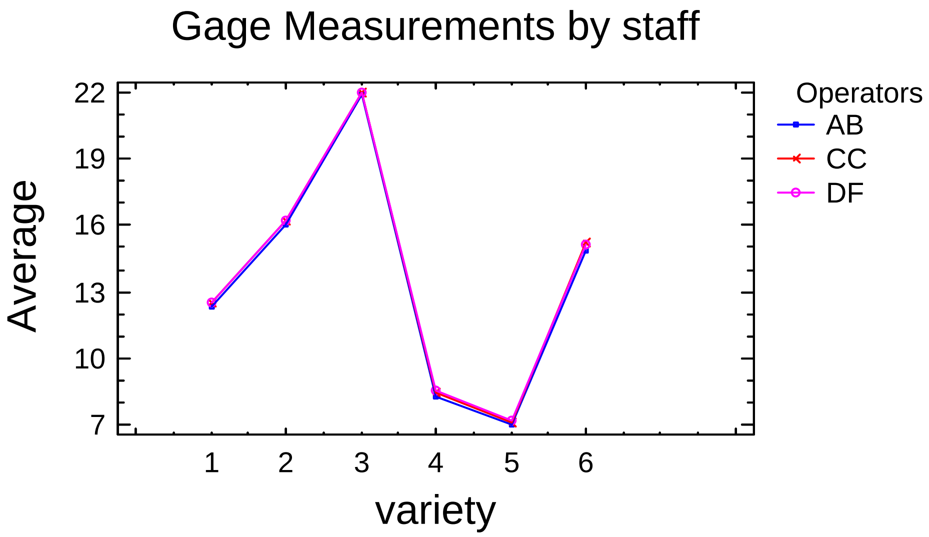
III Example of a check of the repeatability/reproducibility of staff making measurements:

31. Machines can be checked using reference samples in order to check repeatability/reproducibility of their results.

32. In variety testing, not only machines perform measurements. Measures also rely on the expertise and knowledge of persons who observe or measure the plants. Staff can also be involved in quality checks, to see if the same results are obtained by a given person when they measure the same objects, and see if members of a team obtain the same results on a given set of objects.

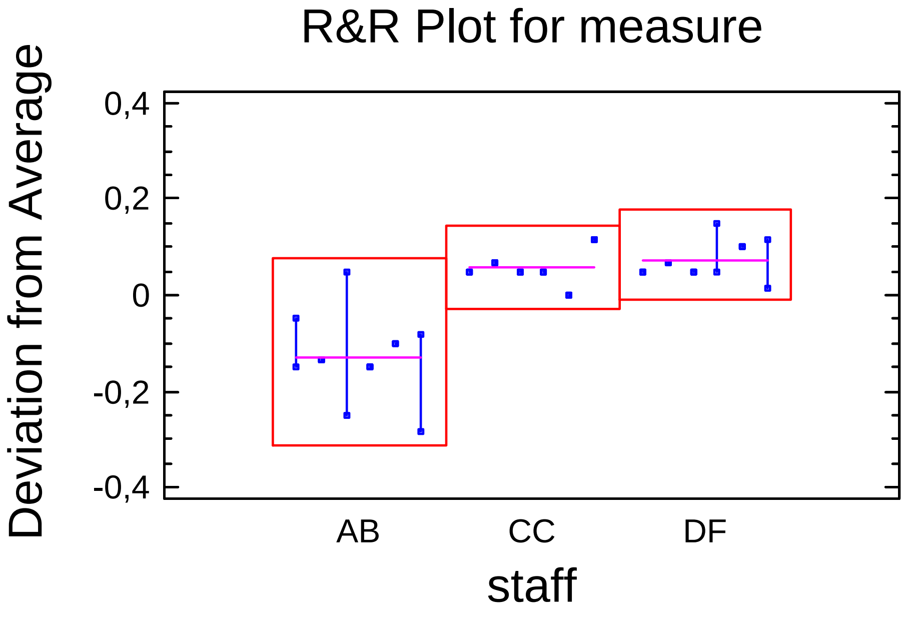
33. On the graph below (see figure 7), a measurement of the height of 6 plants made twice by 3 persons is shown. The visualization of the measures as shown in figure 7 does not show discrepancies between operators.

Figure 7: Plot of average measurement (vertical axis) on 6 varieties (horizontal axis) by 3 different persons identified by their initials (From Statgraphics)



34. With the same fictitious data set, the analysis of repeatability and reproducibility as shown below in figure 8 indicates clearly that AB did not produce the same measure twice on 3 of the 6 plants, and made measures that are lower to those delivered by CC and DF.

Figure 8: repeatability and reproducibility plots with individual data points per person (box) on 6 varieties twice (From Statgraphics)



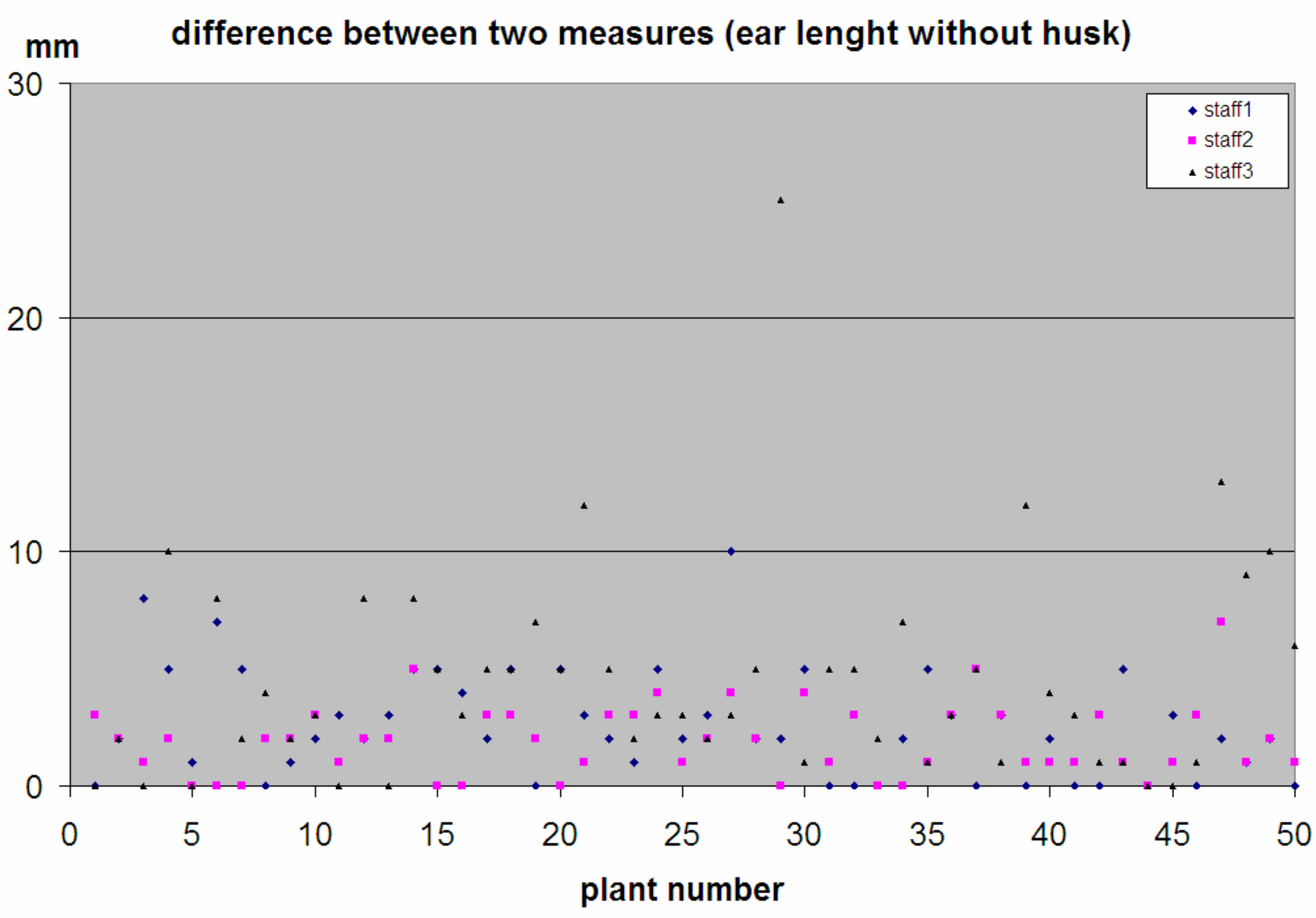
35. When 2 measures on the same plant are different, they are linked by a line (variety 1, 3 and 6 for AB, variety 4and 6 for DF). In other cases the two measures are identical, shown by a unique data point on the graph. On two plants, DF did not produce the same measure, but there is a global consistency between CC and DF, CC and DF obtained very similar measures, see data set shown on figure 8 below.

staff	repeat	variety 1	variety 2	variety 3	variety 4	variety 5	variety 6
AB	1	12,4	16	22	8,3	7	15
AB	2	12,3	16	21,7	8,3	7	14,8
CC	1	12,5	16,2	22	8,5	7,1	15,2
CC	2	12,5	16,2	22	8,5	7,1	15,2
DF	1	12,5	16,2	22	8,5	7,2	15,2
DF	2	12,5	16,2	22	8,6	7,2	15,1

36. This data is fictitious in order to be easier to read; in practice many more plants per variety would be checked, as in the following example.

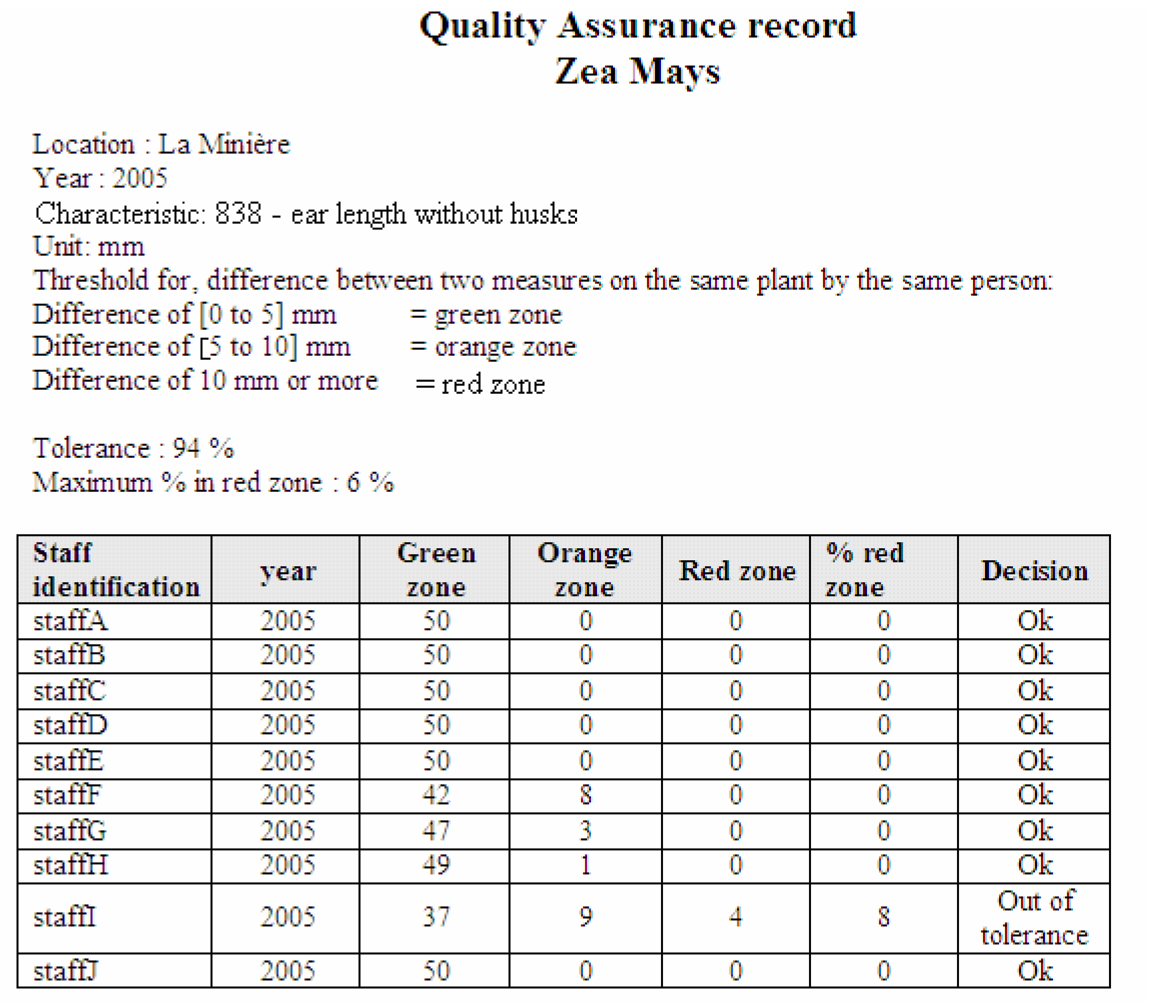
37. This kind of check is performed in GEVES, as part of quality assurance system, to check staff will give the same values when measuring plants in the field. Some of the persons are crop experts while others are temporary staff. The check is performed after temporary staff have learnt how to take the measurement.

Figure 9: Difference between two measures on the same plant by the same person, 3 staff members - 50 plants (from GEVES Quality assurance check characteristic ear length without husk)



38. In figure 9, staff3 is less repeatable than staff1 and staff2, and has a number of large deviations. The decision rule in that situation will be that staff3 will not perform the measurements.

Figure 10: Quality assurance record (from GEVES Quality assurance check)



Quality Assurance record
Zea Mays

Location : La Minière
Year : 2005
Characteristic: 838 - ear length without husks
Unit: mm
Threshold for, difference between two measures on the same plant by the same person:
Difference of [0 to 5] mm = green zone
Difference of]5 to 10] mm = orange zone
Difference of 10 mm or more = red zone

Tolerance : 94 %
Maximum % in red zone : 6 %

Staff identification	year	Green zone	Orange zone	Red zone	% red zone	Decision
staffA	2005	50	0	0	0	Ok
staffB	2005	50	0	0	0	Ok
staffC	2005	50	0	0	0	Ok
staffD	2005	50	0	0	0	Ok
staffE	2005	50	0	0	0	Ok
staffF	2005	42	8	0	0	Ok
staffG	2005	47	3	0	0	Ok
staffH	2005	49	1	0	0	Ok
staffI	2005	37	9	4	8	Out of tolerance
staffJ	2005	50	0	0	0	Ok

39. In figure 10 above, the quality check is Ok (i.e. within quality assurance limits) for nine of the staff.

IV <u>A general framework for test design when checking levels of defects:</u>

40. A good way to design tests when the aim is to check for percentage levels of defects is to use acceptance probability curves, with two types of quality levels: AQL and LQL.

41. **AQL** is a good quality level that we want to accept often. “Often” is defined by the acceptance probability at the AQL level; for instance 95%, which corresponds to an alpha risk of 5% (reject a good lot).

42. **LQL** is a poor quality level that we want to reject often. “Often” is defined by the acceptance probability at the LQL level; for instance 95%, which corresponds to a beta risk of 5% (accept a bad lot).

43. An **acceptance probability curve** can be drawn for a given sample size, and a given decision rule. For instance look at 200 individual plants (sample size) and accept up to 5 off-types (decision rule).

44. The horizontal axis is the true level of defects in a lot, and the vertical axis is the probability to accept the lot using the test design (sample size + decision rule).

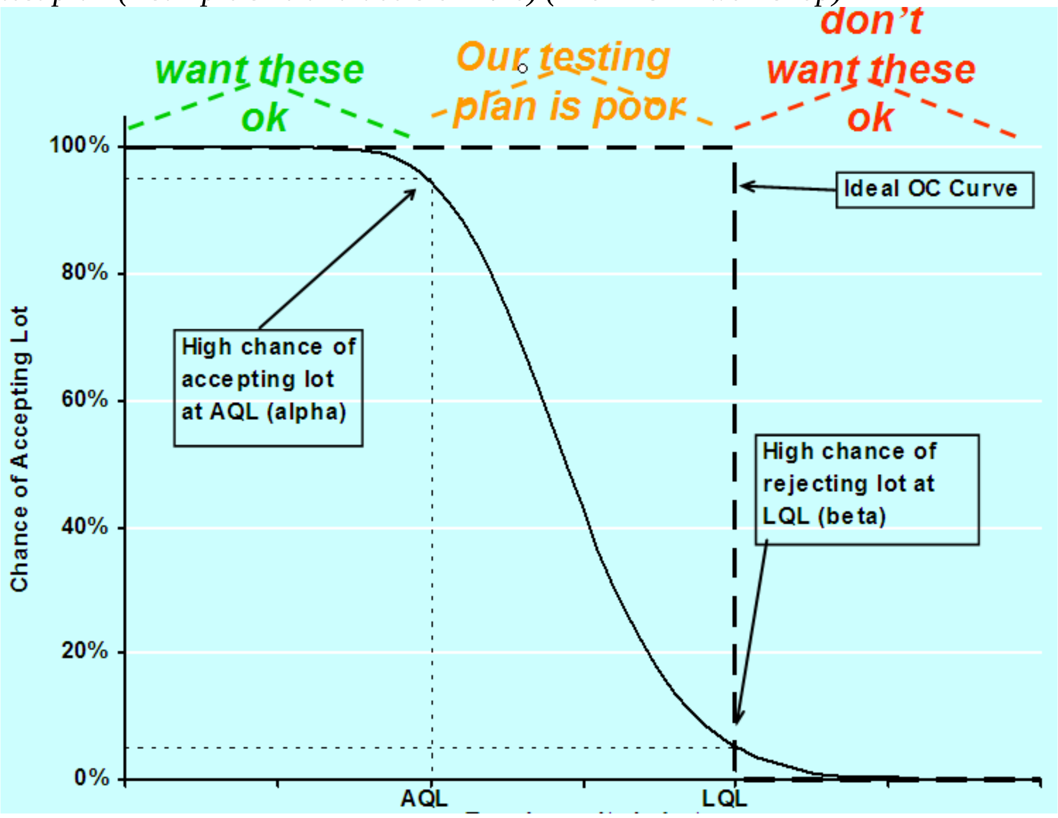
45. The acceptance probability curve comes from 100% at the upper left, where if we have no defects we will always accept the lot. On the right of the curve, when the level of defects becomes great, the decision rule will always lead to rejection of the lots (acceptance probability=0.)

46. Between those 2 extreme situations the curve goes down as shown below. From 0 defects to AQL level we very often accept the lots; above LQL level, we reject them very often. Between AQL and LQL are defect levels for which our testing plan is not very good, because we will sometimes accept, sometimes reject lots of this quality.

47. To reduce the non precise zone between AQL and LQL it is necessary to increase the number of checked objects.

48. If practical or economic reasons do not permit an increase in the sample size, i.e. in the number of objects to be checked, then the impact of the sample size will show on the acceptance probability curve 'quality zone' (between the AQL and the LQL)where we will not be accurate in our decisions.

Figure 11: General appearance of an acceptance probability curve. A curve is specific for a test plan (=sample size and decision rule) (From ISTA workshop)



49. The black descending curve is the acceptance probability curve. The dashed line with a vertical descent at LQL level indicates an ideal situation where we would always accept until a given level of quality and always reject above. The only test plan that would correspond to

that ideal situation is a check of each individual object produced, which is impossible in variety testing. The drawing of the acceptance probability curves is based on the fact that at some stages we look at samples, this variability can not be avoided. Other sources of variability can be added such as for instance false positive, and/or false negatives. The variability introduced by analytical processes can also be computed and taken into account.

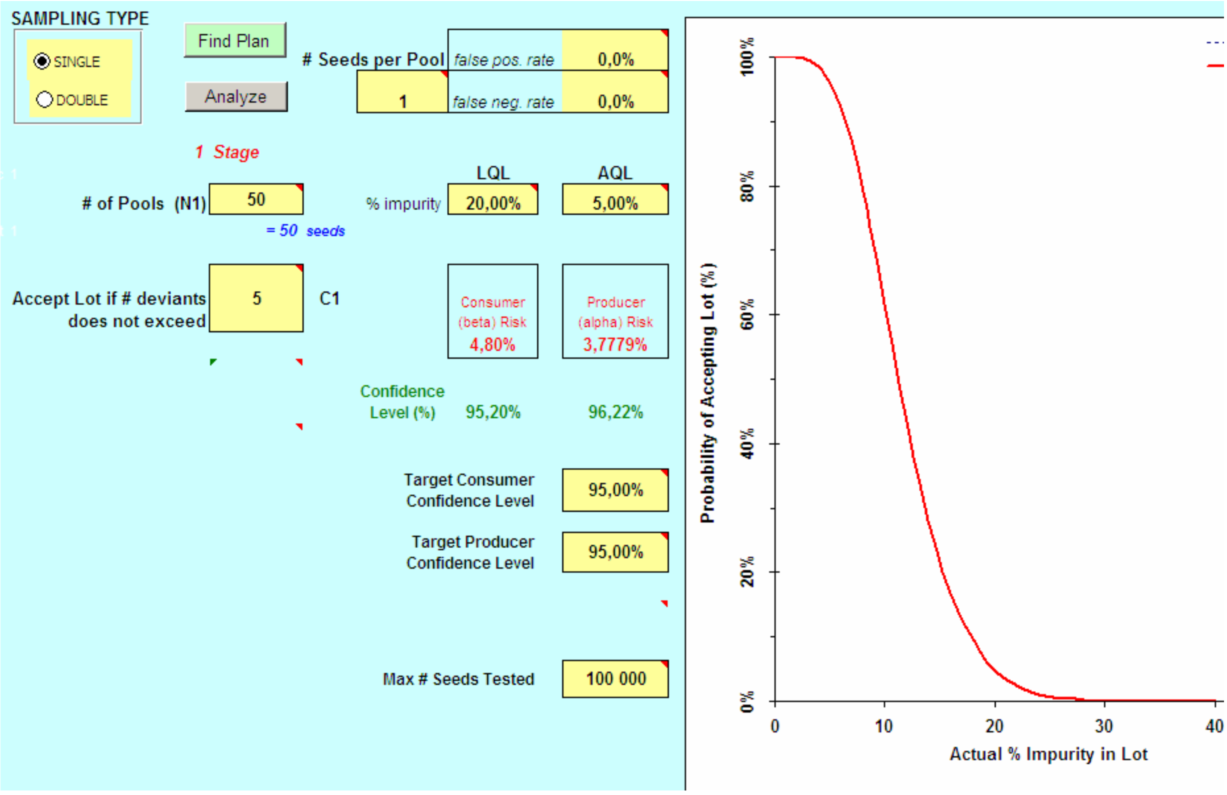
50. Sometimes the focus is on AQL (as in UPOV off-types check for instance), sometimes the focus is on LQL (for instance to avoid the presence of seeds with specified traits in conventional seeds).

51. It is always advisable to consider both AQL and LQL levels with appropriate risks. In doing so we are able to select testing plans which allow us to meet both objectives (accept often lots with good quality and reject often lots of poor quality).

52. Sometimes economic or practical considerations do not permit us to use an efficient testing plan. In such situations, it is still of value to know the efficiency of our testing plan (see section 5 for examples on UPOV off-types check).

53. In many circumstances the use of simple human logic to define a testing plan results in the choice of inappropriate test plans and decision rules. For instance, to check 90% purity, a common sense approach consists of using the level as a decision rule, for instance accept 5 impurities from 50 plants, with the belief that doing so we will reject lots with 10% of impurities. The acceptance probability curve (figure 12) shows that in fact it is a 20% impurity level that we will reject often, instead of our intended level (10%).

Figure 12: Acceptance curve intuitive testing plan, accept 5 impurities from 50 plants to check a 10% impurity level, this level to be rejected often, i.e. the LQL (From Seed Calc on ISTA website)



SAMPLING TYPE
● SINGLE
○ DOUBLE

1 Stage

# Seeds per Pool: 1
false pos. rate: 0,0%
false neg. rate: 0,0%

# of Pools (N1): 50
= 50 seeds

	LQL	AQL
% impurity	20,00%	5,00%
	Consumer (beta) Risk 4,80%	Producer (alpha) Risk 3,7779%
Confidence Level (%)	95,20%	96,22%

Accept Lot if # deviants does not exceed: 5 C1

Target Consumer Confidence Level: 95,00%
Target Producer Confidence Level: 95,00%

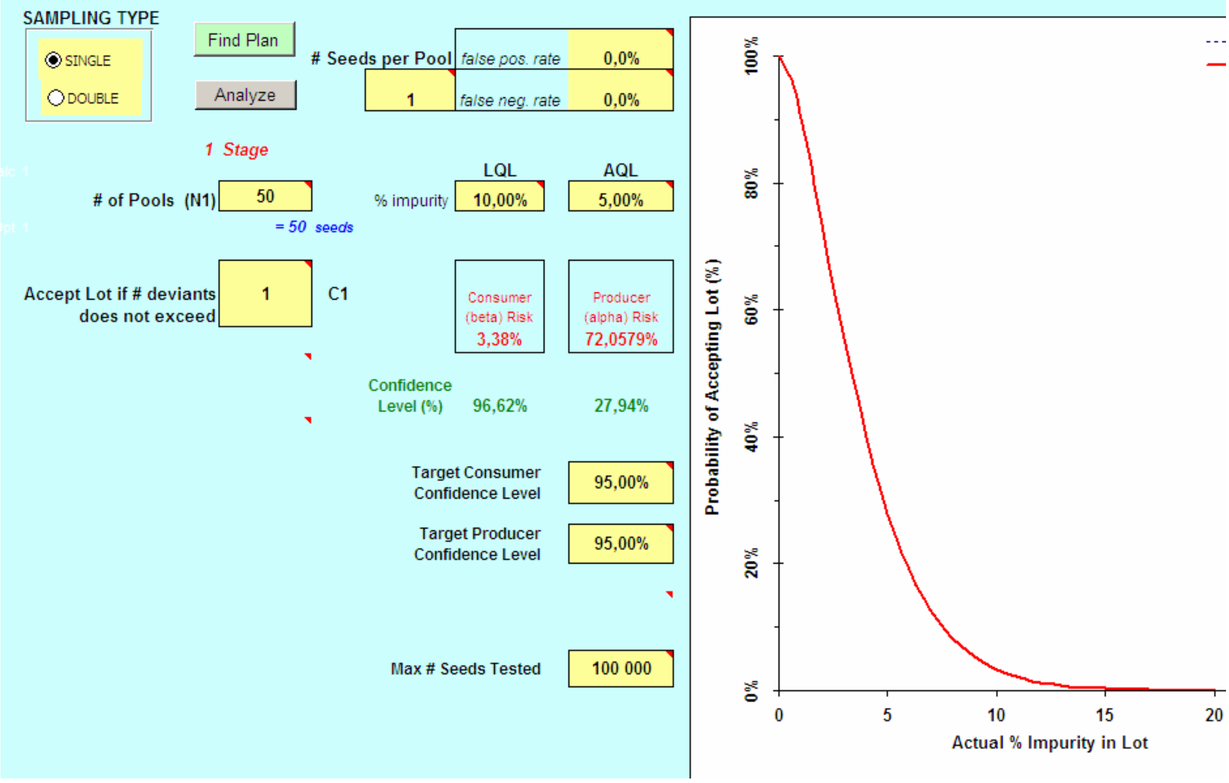
Max # Seeds Tested: 100 000

54. The intuitive testing plan shown above in figure 12 is intended to check 90% of purity (reject when more than 10% =5/50). This testing plan in fact corresponds to checking a AQL level of 5%, and a LQL level of 20% (both acceptance and rejection probabilities for AQL and LQL respectively are above 95%). The acceptance probability curve shown in figure 13 also shows that if we look only at 50 plants, to reject often (>95%) a 10% level of impurity we need to accept only 1 impurity from 50 plants, which is not intuitive.

55. Figure 13 shows a testing plan which corresponds to the intention to avoid accepting lots with 10% impurities or more, with the same number of plants (50) assuming 50 plants is what is possible in routine (time and money spent) and an *ad hoc* decision rule (=accept 0 or 1 defects). This decision rule is adapted because we will reject lots having 10% impurities in more than 95% of the checks; the beta risk is below 5%.

56. A possible problem for this testing plan is that good levels of purity will be often rejected (i.e. 95% purity will have an acceptance rate of 28%).

Figure 13: Correct decision rule to often reject 10% (LQL) impurity rate, if 50 plants sample Size is kept (From Seedcalc on ISTA website)



SAMPLING TYPE

● SINGLE

○ DOUBLE

Find Plan

Analyze

# Seeds per Pool	false pos. rate	0,0%
1	false neg. rate	0,0%

1 Stage

# of Pools (N1) 50

= 50 seeds

	LQL	AQL
% impurity	10,00%	5,00%

Accept Lot if # deviants does not exceed 1 C1

	Consumer (beta) Risk	Producer (alpha) Risk
	3,38%	72,0579%
Confidence Level (%)	96,62%	27,94%

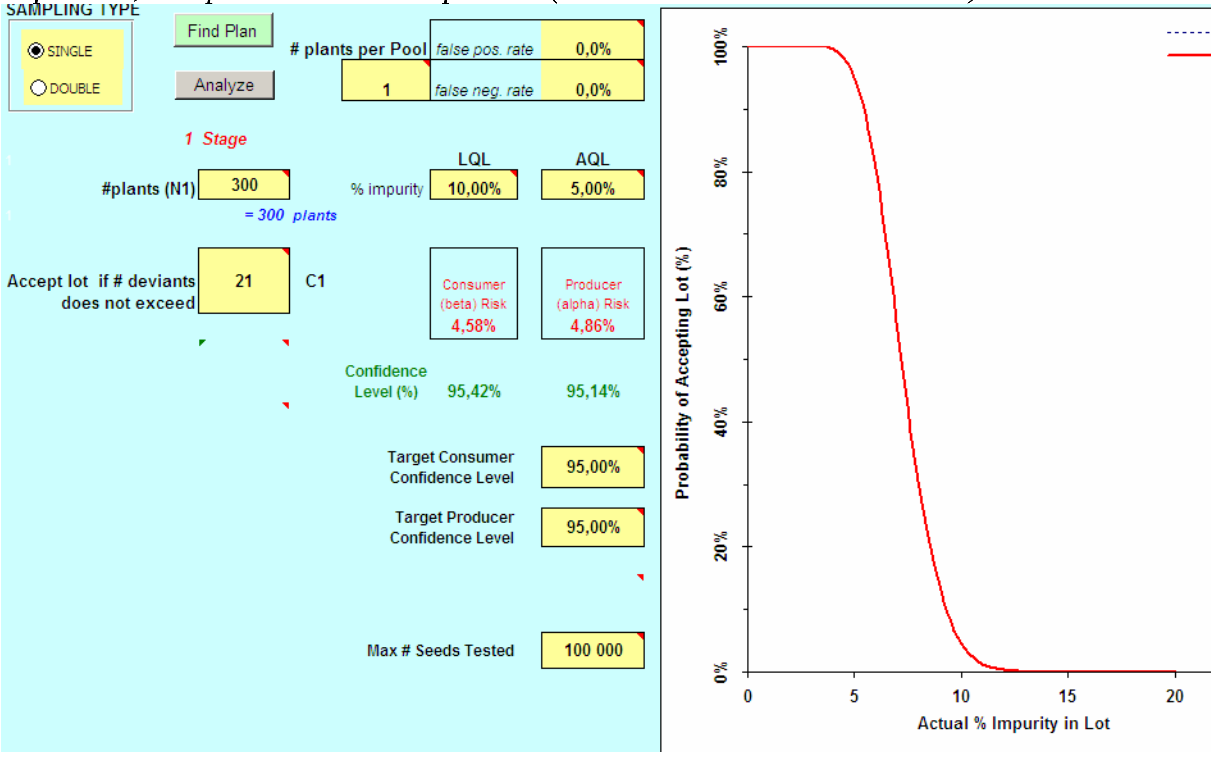
Target Consumer Confidence Level 95,00%

Target Producer Confidence Level 95,00%

Max # Seeds Tested 100 000

57. If the intention is to keep a high level of rejection of lots with 10% impurities, and accept at least 95% of lots having 95% purity, about 300 seeds need to be checked, and up to 21 impurities can be accepted (see below in figure 14)

Figure 14: A testing plan that allow both objectives to be met, reject >95% of lots at 10% Impurities, accept >95% at 5% impurities (From Seedcalc on ISTA website)



SAMPLING TYPE

● SINGLE

○ DOUBLE

Find Plan

Analyze

# plants per Pool	false pos. rate	0,0%
1	false neg. rate	0,0%

1 Stage

#plants (N1) 300

= 300 plants

	LQL	AQL
% impurity	10,00%	5,00%

Accept lot if # deviants does not exceed 21 C1

	Consumer (beta) Risk	Producer (alpha) Risk
	4,58%	4,86%
Confidence Level (%)	95,42%	95,14%

Target Consumer Confidence Level 95,00%

Target Producer Confidence Level 95,00%

Max # Seeds Tested 100 000

58. Seedcalc which is freely available on the ISTA website, allows the user to find appropriate testing plans, or to check the efficiency of testing plans, for different situations:

- Checks where individual objects are checked (as in the off-types check in UPOV)
- Checks where the presence of defects in made by looking for the presence/absence of defects in pools of objects (presence of pathogens in seeds for instance)
- Checks where a check is made on objects through an analytical process (check %GM in seed lots for instance)

59. A benefit of Seedcalc is that it allows the efficiency of the testing plan to be checked in the presence of false positives and/or false negatives. For given quality objectives, it can be shown that from a set of testing plans that meet AQL and LQL with alpha and beta risks settled, some are robust to false observations, while others are not robust. The testing plans that are not robust to false observation should be avoided.

V Examples in UPOV work, the check of uniformity:

60. There are two major ways to assess Uniformity within UPOV

1. Compute intra-varietal variability, and compare it to an acceptable variability level
2. Count off-types in varieties, and compare it to an acceptable number of off-types

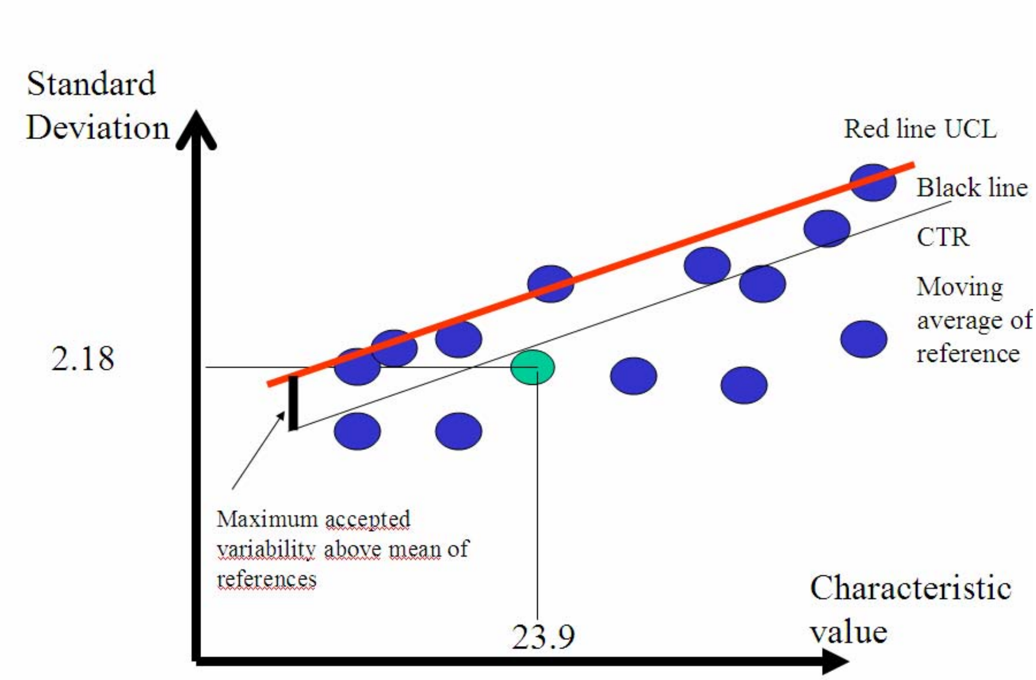
61. COYU is an example of implementation of the first type of computation. In this approach the intra-varietal variability is calculated. UPOV checks that the variability is not greater than the variability usually observed in reference varieties.

62. UPOV tables for the off-types method are an example of implementation of the second type of check. In this approach all plants from a given variety are expected to be very much alike. A lack of uniformity is checked by looking if some plants are not true to the variety type of expression, a plant which differs from the others is called an off-type.

V(a) Intra-varietal uniformity check with UPOV COYU:

63. Uniformity in COYU is assessed from the standard deviation (SD) of a set of plants per variety.

Figure 15: COYU computes an upper confidence limit (UCL) and a reference value (CTR) for the uniformity quality criterion



64. In figure 15, the standard deviation (SD) is plotted on Vertical axis, and the mean of the measures from the set of plants per variety is plotted on Horizontal axis. The green circle is the value obtained for a candidate variety, the blue circles indicates the values obtained on reference varieties. The slope indicates a tendency for an increase in standard deviation when measures on the horizontal axis increase. In COYU, the CTR value is computed from the variability observed on the 9 reference varieties that have the most similar values on the horizontal axis to the candidate. The upper limit of tolerance (UCL) indicates the maximum tolerated value for variability, significantly exceeding the variability of the reference varieties. In the UPOV check for uniformity, varieties that are more uniform than the reference value are not a concern, in that situation no lower control limit (LCL) is used. For computational reasons, COYU uses Log(SD+1) instead of SD (standard deviation) to test the significance to the UCL. The Alpha risk is usually set at 1 per 1000 in a two year test.

V(b) UPOV check of uniformity via off-types count:

65. This approach is used for crops in which plants from a given variety are expected to be very much alike. When a plant, or part of a plant, is not the true-to-type in a variety, it is counted as an off-type.

66. UPOV defines the Population Standard as the % of off-types that can be accepted in a variety. A population standard of 1% as in figure 16, correspond to 1% of the plants being off-types.

67. UPOV defines Acceptance Probability as the probability to accept a variety when the variety shows a % of off-types equal to the population standard. >= 95% as in figure 16 corresponds to acceptance of varieties having 1% of off-types at least in 95% of the controls.

Figure 16: UPOV table to select a sample size and a maximum number of off-types if population is 1% and acceptance probability >= 95% (From UPOV document TGP/10)

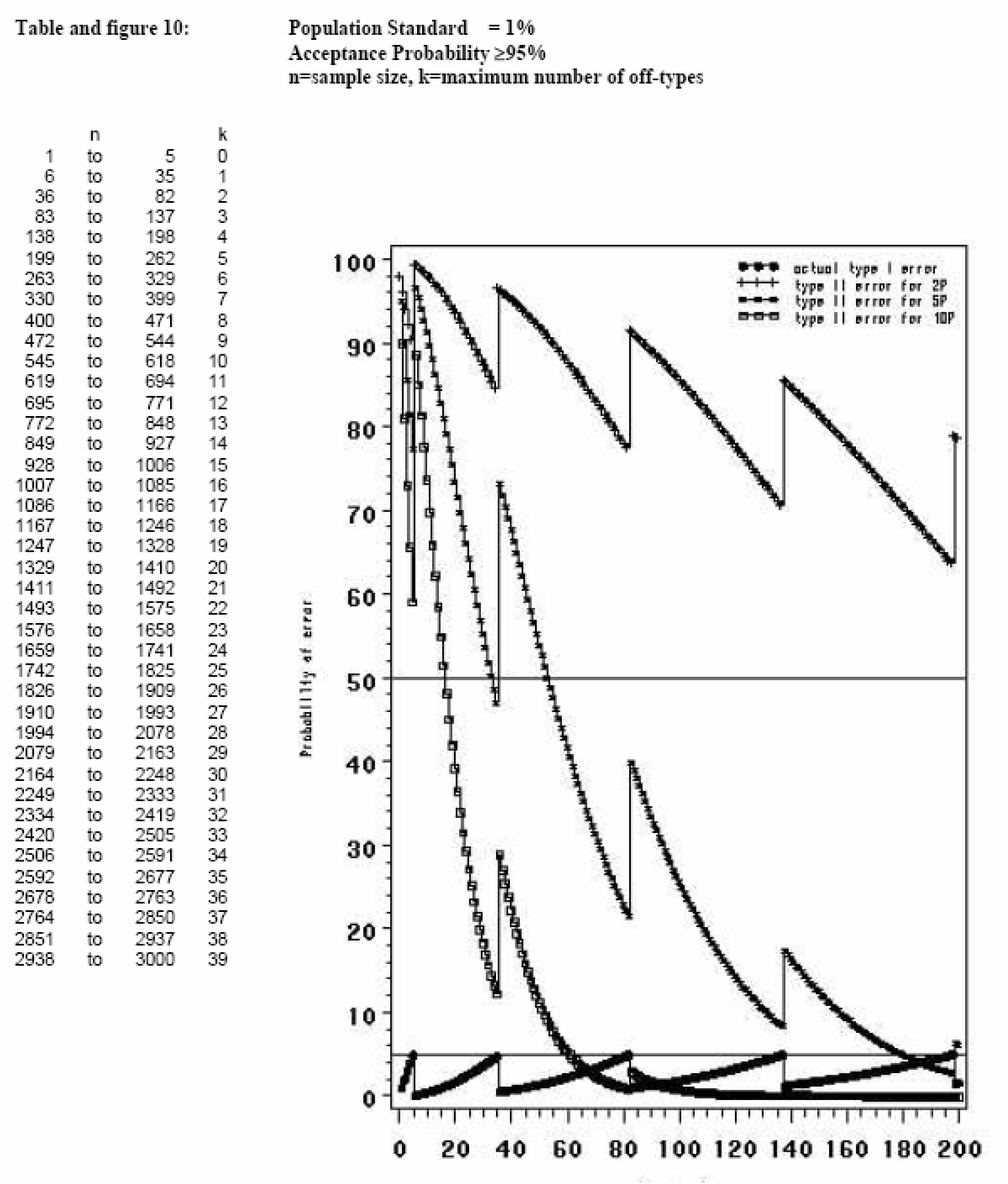


Table and figure 10:

Population Standard = 1%
Acceptance Probability ≥95%
n=sample size, k=maximum number of off-types

n			k
1	to	5	0
6	to	35	1
36	to	82	2
83	to	137	3
138	to	198	4
199	to	262	5
263	to	329	6
330	to	399	7
400	to	471	8
472	to	544	9
545	to	618	10
619	to	694	11
695	to	771	12
772	to	848	13
849	to	927	14
928	to	1006	15
1007	to	1085	16
1086	to	1166	17
1167	to	1246	18
1247	to	1328	19
1329	to	1410	20
1411	to	1492	21
1493	to	1575	22
1576	to	1658	23
1659	to	1741	24
1742	to	1825	25
1826	to	1909	26
1910	to	1993	27
1994	to	2078	28
2079	to	2163	29
2164	to	2248	30
2249	to	2333	31
2334	to	2419	32
2420	to	2505	33
2506	to	2591	34
2592	to	2677	35
2678	to	2763	36
2764	to	2850	37
2851	to	2937	38
2938	to	3000	39

68. The original sample sizes and limits of tolerance were first established for the certification of seed lots in European countries.

69. C. Hutin, co-founder of the UPOV TWC, showed how these decision rules were linked to statistical tests in the very early TWC meetings.

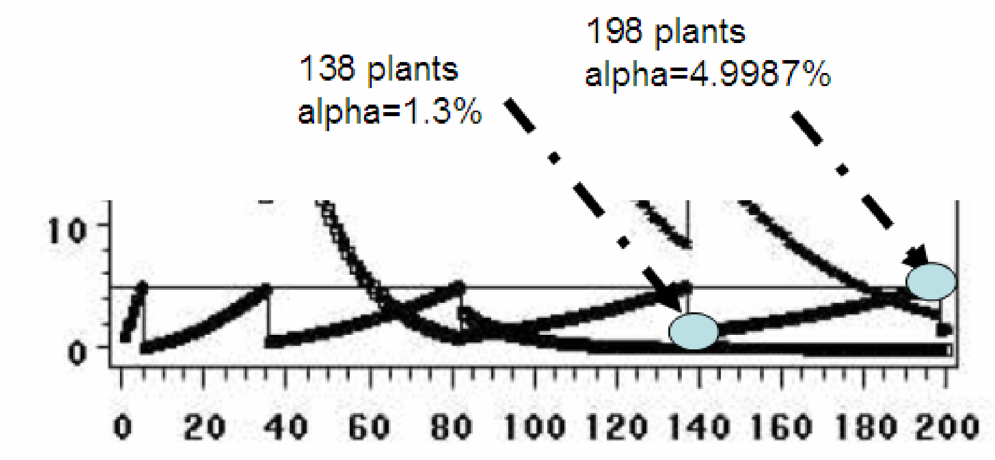
70. When the TWC revised the principles of the off-types method, K. Kristensen provided a set of tables showing the relationship between sample-size, population standards, decision

rule, alpha and beta risks. UPOV Test Guidelines now refer to these notions explicitly, as well as CPVO guidelines.

71. The focus has been given to the population standard, which is in fact an AQL (see general framework section above), with different alpha risk levels. The alpha risk is described as the type I error in figure 16. A population standard at alpha risk=5% has at least 95% probability to be accepted using UPOV tables (see figure 16 above).

72. The alpha risk level is equal to (1 – acceptance probability) and can be seen on the lower part of the graphs. Alpha risk values increase regularly, and step down at each change of sample size, as seen in figure 17 below.

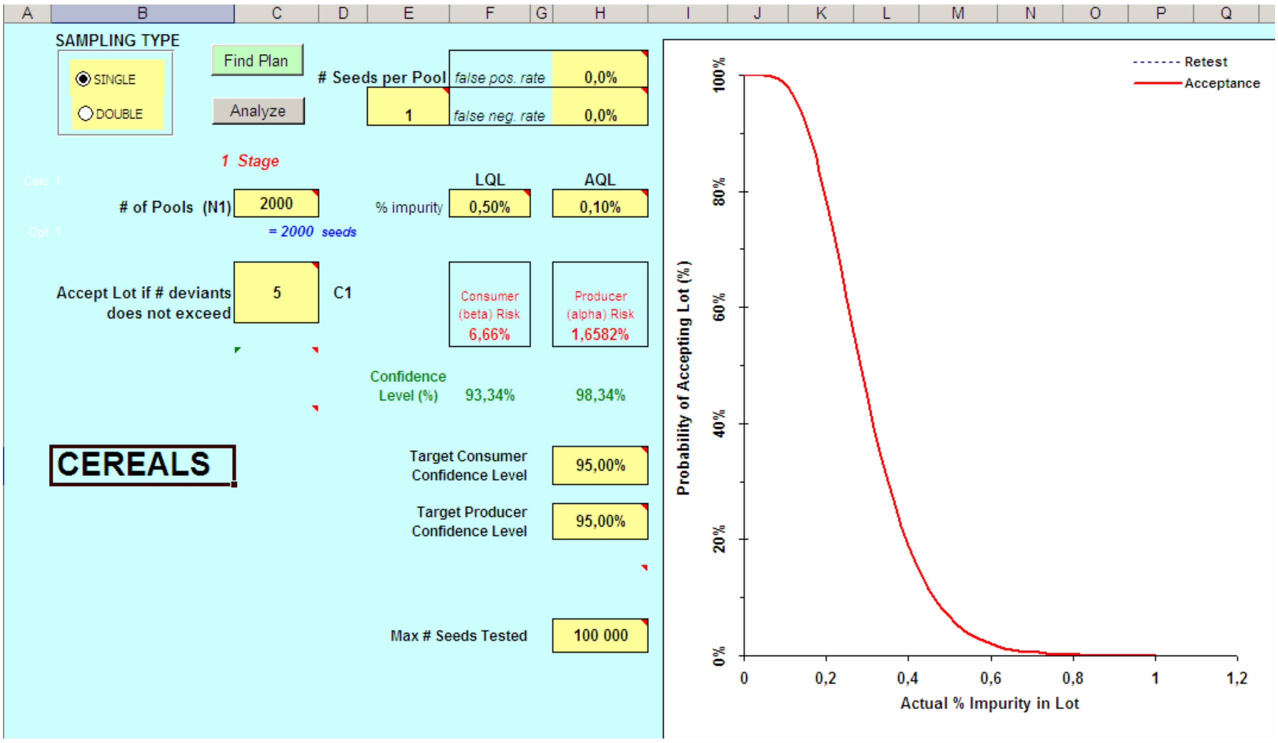
Figure 17: Extract from figure 16 showing alpha risks (below a 5% line on vertical axis)



73. If we look at the part of the graph with a sample size from 138 to 198 plants; accept until a maximum *k*=4 off-types, we see with 138 plants alpha risk=1.3%, while with 198 plants alpha risk =4.9987% (i.e. less but very near to 5%).

74. There are big differences between crops in the level of off-types checked and in the precision of the check. For some characteristics in cereals, 2000 plants are checked with *k*=5 (population standard 0.1%). This sample size allows a good check for actual off-type levels from approximately 0.1% to 0.5% (see figure 18 below).

Figure 18: Efficiency of UPOV testing plan in some cereal guidelines (From Seedcalc on ISTA website)



SAMPLING TYPE

● SINGLE

○ DOUBLE

Seeds per Pool: 1

false pos. rate: 0,0%

false neg. rate: 0,0%

1 Stage

# of Pools (N1): 2000

= 2000 seeds

	LQL	AQL
% impurity	0,50%	0,10%
	Consumer (beta) Risk 6,66%	Producer (alpha) Risk 1,6582%
Confidence Level (%)	93,34%	98,34%

Accept Lot if # deviants does not exceed: 5 C1

CEREALS

Target Consumer Confidence Level: 95,00%

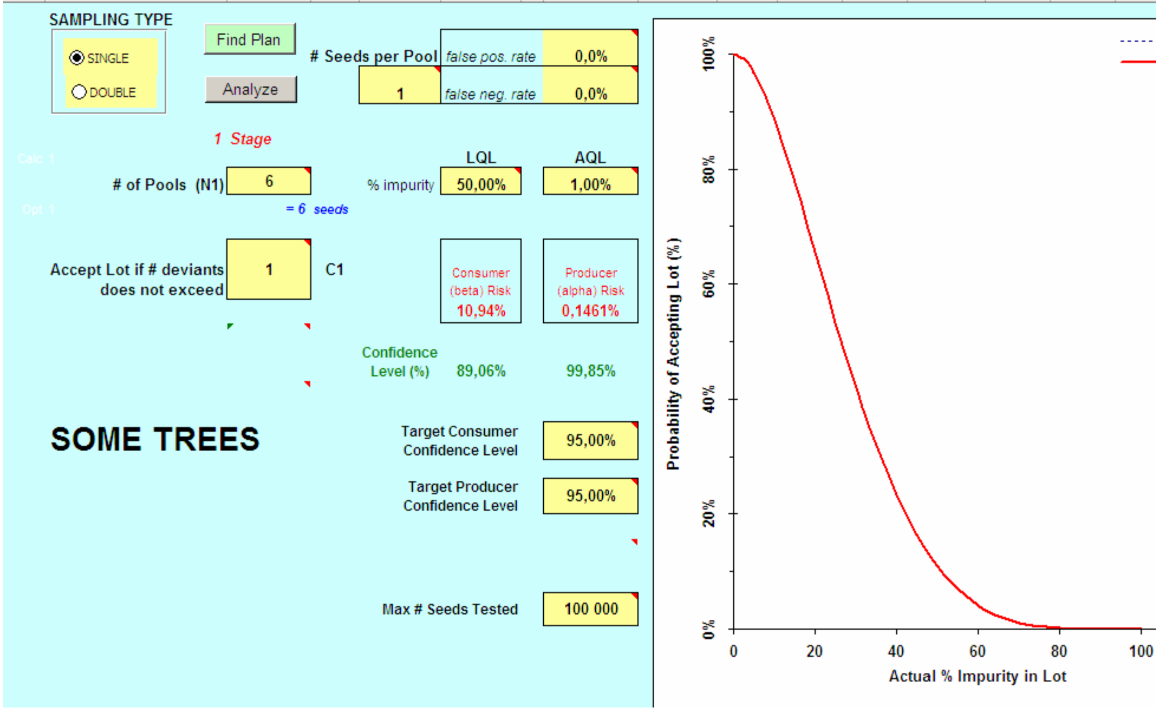
Target Producer Confidence Level: 95,00%

Max # Seeds Tested: 100 000

75. For some crops, in UPOV the sample size is 6 and *k*=1. In that situation, 1% (population standard) is very often accepted (100-0.1461=99.85%), while varieties having 50% off-types would be accepted at a 10.94% rate (see figure 19 below).

76. The decision rule to accept 0 off-types from 6-plant samples was changed to the rule to accept 0 or 1 off-types from 6 plant samples. This was done in order to ensure the acceptance probability was at least 95%. With the previous rule (accept 0 off-types from 6 plant samples) the alpha risk was equal to 5.8%.

Figure 19: Efficiency of UPOV testing plan in some cereal guidelines (From Seedcalc on ISTA website)



SAMPLING TYPE

● SINGLE

○ DOUBLE

Find Plan

Analyze

# Seeds per Pool		false pos. rate	0,0%
	1	false neg. rate	0,0%

1 Stage

# of Pools (N1) 6

= 6 seeds

	LQL	AQL
% impurity	50,00%	1,00%

Accept Lot if # deviants does not exceed 1 C1

	Consumer (beta) Risk	Producer (alpha) Risk
	10,94%	0,1461%
Confidence Level (%)	89,06%	99,85%

SOME TREES

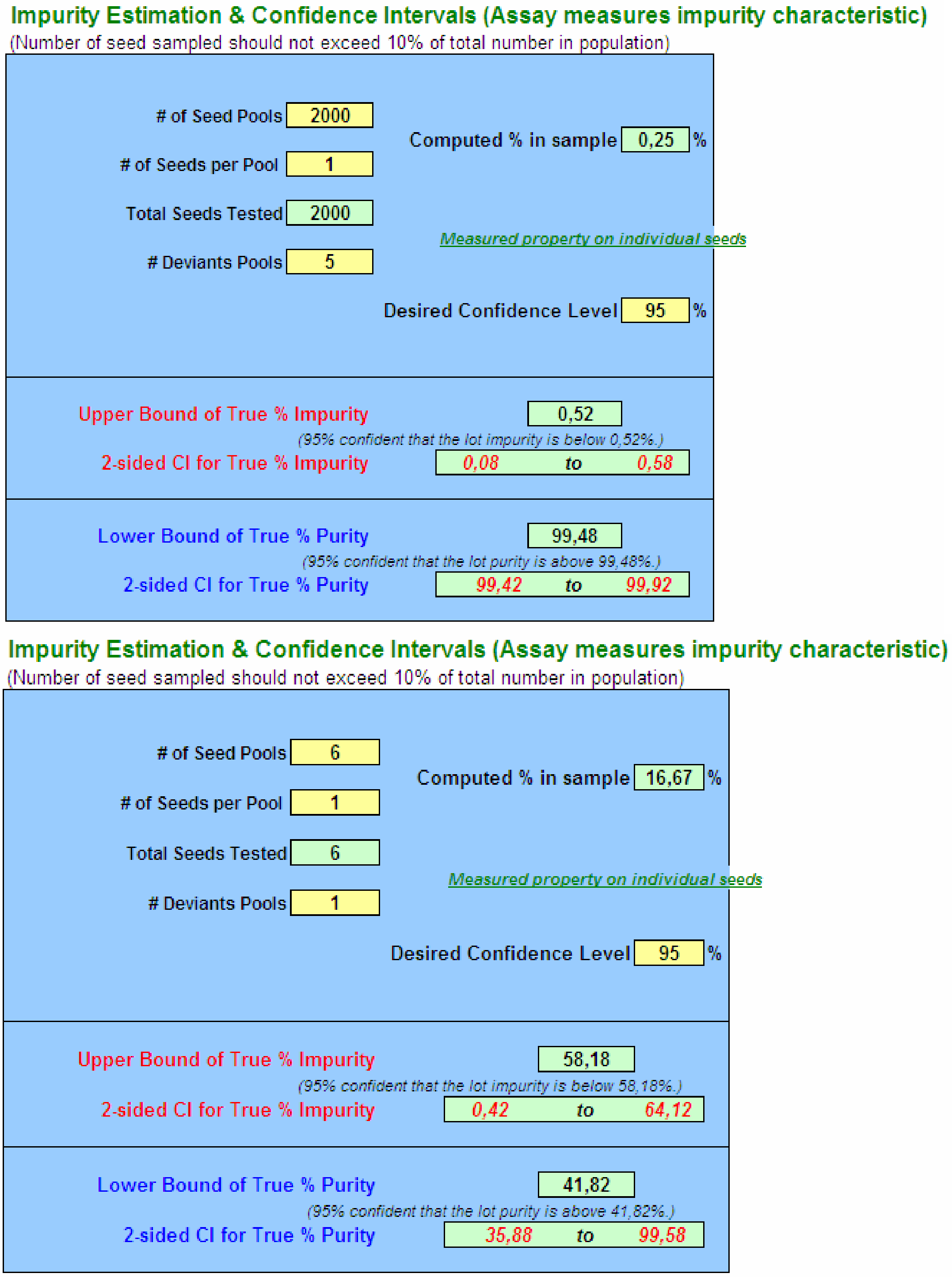
Target Consumer Confidence Level 95,00%

Target Producer Confidence Level 95,00%

Max # Seeds Tested 100 000

77. Another way to look at those decision rules is to look at estimates of off-types and the confidence interval when *k* (max off-types tolerated) off-types are found. With *k*=5 from 2000 plants, the confidence interval is from 0.08% to 0.58%. With *k*=1 from 6 plants, the confidence interval is from 0.42% to 64% (see figure 20 below)

Figure 20: Estimate and confidence interval of off-types in 2000 plants 5 off-types maximum and 6 plants 1 off-type maximum(From Seedcalc on ISTA website)



Impurity Estimation & Confidence Intervals (Assay measures impurity characteristic)

(Number of seed sampled should not exceed 10% of total number in population)

# of Seed Pools	2000	Computed % in sample	0,25	%
# of Seeds per Pool	1			
Total Seeds Tested	2000			
# Deviants Pools	5			
		Desired Confidence Level	95	%

Measured property on individual seeds

Upper Bound of True % Impurity 0,52

(95% confident that the lot impurity is below 0,52%.)

2-sided CI for True % Impurity 0,08 to 0,58

Lower Bound of True % Purity 99,48

(95% confident that the lot purity is above 99,48%.)

2-sided CI for True % Purity 99,42 to 99,92

Impurity Estimation & Confidence Intervals (Assay measures impurity characteristic)

(Number of seed sampled should not exceed 10% of total number in population)

# of Seed Pools	6	Computed % in sample	16,67	%
# of Seeds per Pool	1			
Total Seeds Tested	6			
# Deviants Pools	1			
		Desired Confidence Level	95	%

Measured property on individual seeds

Upper Bound of True % Impurity 58,18

(95% confident that the lot impurity is below 58,18%.)

2-sided CI for True % Impurity 0,42 to 64,12

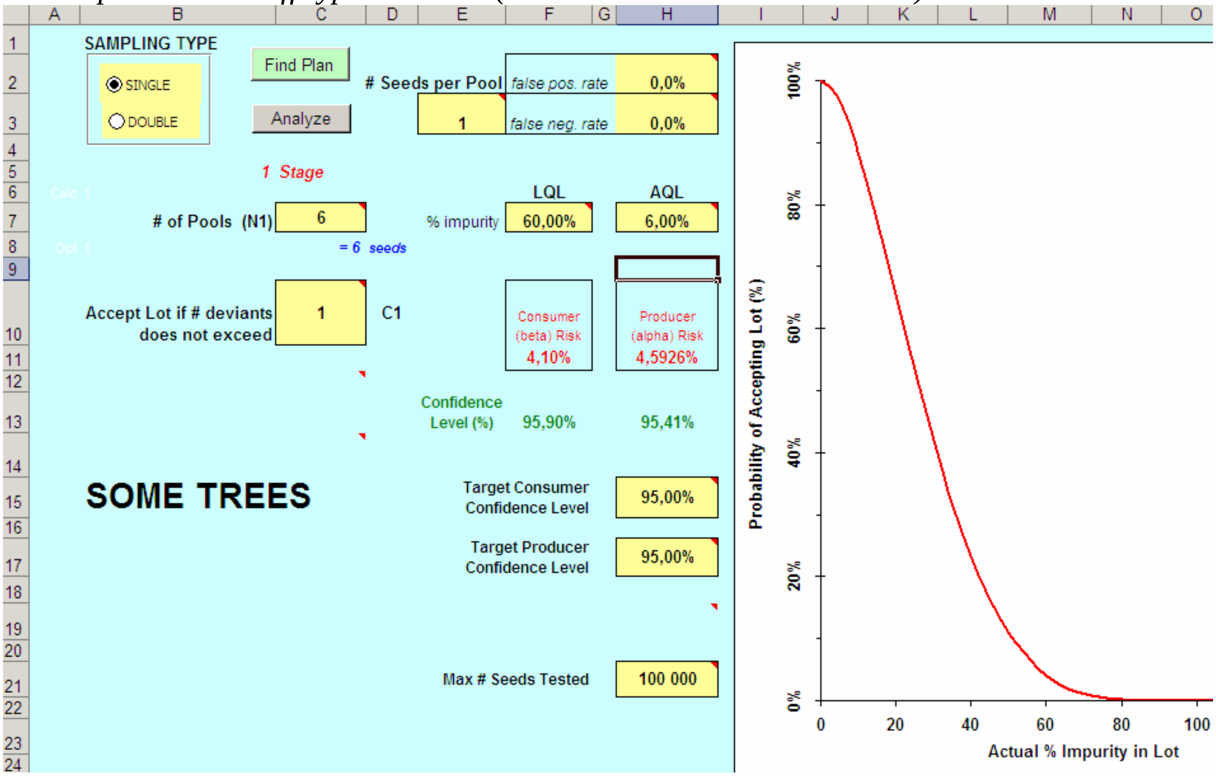
Lower Bound of True % Purity 41,82

(95% confident that the lot purity is above 41,82%.)

2-sided CI for True % Purity 35,88 to 99,58

78. Accepting 1 off-type per 6 plants is a very lenient decision rule. If this testing plan was looked at considering balanced alpha and beta risks (both below 5% for instance), it would correspond to accepting correctly good levels from 0 to 6% and rejecting correctly bad levels of more than 60% of off-types (see figure 21 below).

Figure 21 : AQL and LQL levels of off-types corresponding to both risks being less than 5% with 6 plants and 1 off-type maximum(From Seedcalc on ISTA website)



SAMPLING TYPE
● SINGLE
○ DOUBLE

# Seeds per Pool: 1
false pos. rate: 0,0%
false neg. rate: 0,0%

1 Stage

# of Pools (N1): 6
= 6 seeds

	LQL	AQL
% impurity	60,00%	6,00%

Accept Lot if # deviants does not exceed: 1 C1

	Consumer (beta) Risk	Producer (alpha) Risk
	4,10%	4,5926%
Confidence Level (%)	95,90%	95,41%

SOME TREES

Target Consumer Confidence Level: 95,00%

Target Producer Confidence Level: 95,00%

Max # Seeds Tested: 100 000

79. Nevertheless, in practice the crop experts do not seem to be concerned and it seems difficult to ask for more than 6 plants for some crops. It is better to have a check, knowing it is very lenient, rather than to have no check. In some countries the number of plants to check has been reduced to 5, with acceptance if there are no off-types in 5 plants.

80. Statistical tools are available in order to define appropriate testing plans, so the situation may seem easy to manage. In fact, there are a number of difficulties and questions for checks using the off-types method.

81. The intention in the following part in italics is not to challenge the UPOV system, which is scientifically based, backed by a long experience, and has already proved its efficiency. The purpose is to recall practical difficulties that need to be known in order to perform and use the results of checks.

> *The basis of the check relies on the ability to detect correctly off-types on a characteristic-by- characteristic basis.*
>
> *This is not an easy task, as some off-types will be obvious to any crop expert, while others would be judged as off-type in one country and not in another.*
>
> *The layout of the experiment can reveal different types, and number, of off-types. An overview of plants on a dense plot will show fewer off-types compared to a close look at individual plants.*
>
> *Some types of off-types may be revealed under some environmental conditions, and not in others.*

> *The number of off-types is checked on a characteristic-by- characteristic basis, but the number of off-types is summed over all characters when compared to the k value. This is complicated by the fact that, for some crops, some characteristics will be looked at on 100 ear-rows plots and other characteristics on 2000 plants in a separate plot; the two types of characteristics having different population standards.*
> …
> *The layout, sample size, and decision rule have evolved very little since their first establishment. It is as if the system has been very conservative, rather than questioning itself about the quality criteria to be met.*
>
> *Despite the efforts in making the guide lines clear and unambiguous, there are probably still different understandings and applications among countries.*

Perspectives

82. The statistical framework for quality control is well established, and has been spread with the increasing use of quality assurance systems.

83. Nevertheless there is a wide range of methods, and for each method different levels of risks depending on the context.

84. Furthermore, the choice of an appropriate method and level of check needs to be established in cooperation with different persons in order to meet quality objectives and precision requirements.

85. There are still opportunities for improvement in defining appropriate test designs and correct decision rules in each practical situation.

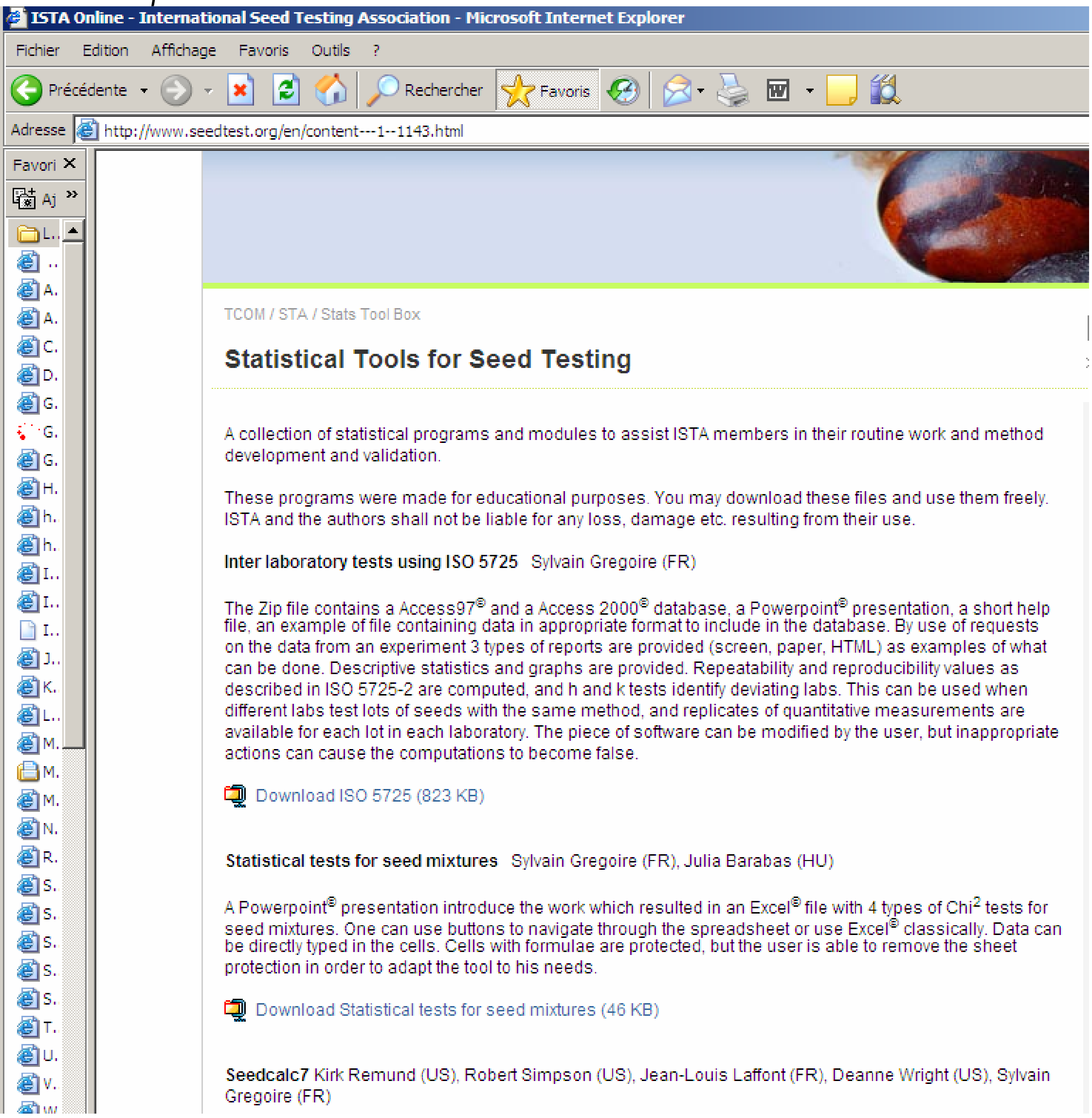
86. In order to obtain a common understanding, and an easy check, the use of graphical outputs such as charts or acceptance probability curves is of great benefit.

[Annex follows]

ANNEX

Download page from ISTA web site where ISO5725-2 and Seedcalc programs can be obtained.
http://www.seedtest.org/en/content---1--1143.html

Screen shot from ISTA Web site



ISTA Online - International Seed Testing Association - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Adresse
http://www.seedtest.org/en/content---1--1143.html
Favori
Aj

TCOM / STA / Stats Tool Box

Statistical Tools for Seed Testing

A collection of statistical programs and modules to assist ISTA members in their routine work and method development and validation.

These programs were made for educational purposes. You may download these files and use them freely. ISTA and the authors shall not be liable for any loss, damage etc. resulting from their use.

Inter laboratory tests using ISO 5725 Sylvain Gregoire (FR)

The Zip file contains a Access97$^{\circledR}$ and a Access 2000$^{\circledR}$ database, a Powerpoint$^{\circledR}$ presentation, a short help file, an example of file containing data in appropriate format to include in the database. By use of requests on the data from an experiment 3 types of reports are provided (screen, paper, HTML) as examples of what can be done. Descriptive statistics and graphs are provided. Repeatability and reproducibility values as described in ISO 5725-2 are computed, and h and k tests identify deviating labs. This can be used when different labs test lots of seeds with the same method, and replicates of quantitative measurements are available for each lot in each laboratory. The piece of software can be modified by the user, but inappropriate actions can cause the computations to become false.

Download ISO 5725 (823 KB)

Statistical tests for seed mixtures Sylvain Gregoire (FR), Julia Barabas (HU)

A Powerpoint$^{\circledR}$ presentation introduce the work which resulted in an Excel$^{\circledR}$ file with 4 types of Chi^{2} tests for seed mixtures. One can use buttons to navigate through the spreadsheet or use Excel$^{\circledR}$ classically. Data can be directly typed in the cells. Cells with formulae are protected, but the user is able to remove the sheet protection in order to adapt the tool to his needs.

Download Statistical tests for seed mixtures (46 KB)

Seedcalc7 Kirk Remund (US), Robert Simpson (US), Jean-Louis Laffont (FR), Deanne Wright (US), Sylvain Gregoire (FR)

Acknowledgments:

The author thanks the reviewers from United Kingdom, Germany and France for their comments on this document.

[End of Annex and of document]