Confidence levels of measurement based decisions

Jos G.M. van der Grinten

NMi Certin B.V., Dordrecht, The Netherlands

Abstract

Metrological decisions are based on measurements that have uncertainties. Examples are car velocity measurements for law enforcement, initial verifications that lead to the decision to approve or reject an instrument, and the significance of differences found during intercomparisons. The paper shows for each of these examples the relationship between acceptance criterion, tolerance, uncertainty and confidence level. From the discussion of these examples it can be concluded that uncertainties must be known in order to evaluate the risk on an erroneous decision. Confidence levels are associated with decisions for which it is impossible to achieve 100% confidence. Conformance and non-conformance are not two complementary notions. If the accepted risk on an erroneous decision is less than 50% there is a range of observations for which the instrument is not conforming and not non-conforming at the same time.

For verifications an increasing number of verification points leads to an increased risk of making an incorrect decision. In order to appreciate the extra information of more observations a curve fit procedure described by Van der Grinten and Peters [1] can be followed. If there are sufficient data, i.e. at least 6 degrees of freedom, it is best to make a curve fit with a 95% confidence envelope.

In all of the above-discussed examples the statistical distribution of the observed results is not known. So the risk analysis is based on the assumption of a Gaussian distribution of the measurement results that is the worst-case representation of our knowledge. If other distributions can be demonstrated to describe the measurement results this will certainly lead to a higher degree of confidence or acceptance criteria that are closer to the tolerances.

Introduction

In the daily practise of metrology many decisions are based on tests or measurements. Instruments may be placed on the market and put into use after it has been clearly demonstrated that the instrument meets the applicable metrological requirements, especially the accuracy requirements. And if instruments are in use for some time they may be subject to a re-verification system or an in-field inspection system that is supervised by the government (Market Surveillance). Here an instrument will be rejected after it was demonstrated that the instrument is operating outside its metrological limits. Also in law enforcement people obtain a ticket if they have exceeded the limits beyond all reasonable doubts. So one may say that in legal metrology every measurement results in a decision: good - not good, fault - not fault. In other words it is decided that the instrument is conforming or not conforming, or that the instrument is non-conforming or not non-conforming. As we will show later there is a difference between non-conforming and not conforming. The decisions based on doping tests that are carried out in modern top sports, require the same level of confidence as speeding tickets.

In scientific and industrial metrology calibrations of instruments do not result in decisions. The deviations of an instrument are simply reported on a calibration certificate without making reference to a tolerance. However, industrial production requires (statistical) process control to monitor the production quality. Adjustments are made if production is no longer within preset factory tolerances.
One of the most important decisions in intercomparison testing is if deviations between laboratories are significant or not.

In the above situations it is vital that reliable decisions are taken. The reliability of a decision is expressed by the confidence level, which is one minus the probability (risk) that an erroneous decision is taken. If the measurement value is close to the tolerance, part of the uncertainty interval is within and another part is outside the tolerances. In other words due to the measurement uncertainty four possibility arise:

a. The object is within tolerances and is approved.
b. The object is within tolerances and is not approved.
c. The object is outside tolerances and is not approved.
d. The object is outside tolerances and is approved.

Cases b and d result in incorrect decisions. In practise people want to limit the risk that an erroneous decision is taken. This risk is depending on the tolerance, the actual deviation and the uncertainty of the measurement result. A special case is where the deviation equals the tolerance, a situation displayed in Figure 1, case (2). The uncertainty band shows that 50% of the values that can reasonably attributed to the measurand, is above the tolerance, the other 50% is below the tolerance. The probability that this instrument is performing within the tolerance is 50%. The decision of approving this instrument results in a confidence level of 50%. In other words the risk associated with the approval of this result is 50%. High risks are not acceptable in cases of health, safety and custody transfer where disputes or lawsuits may involve enormous financial consequences.
The relationship between confidence level (risk), tolerance, observed deviation (error) and uncertainty will be demonstrated in the following examples: speed enforcement in traffic, initial verification in flow meters and intercomparison of laboratories.

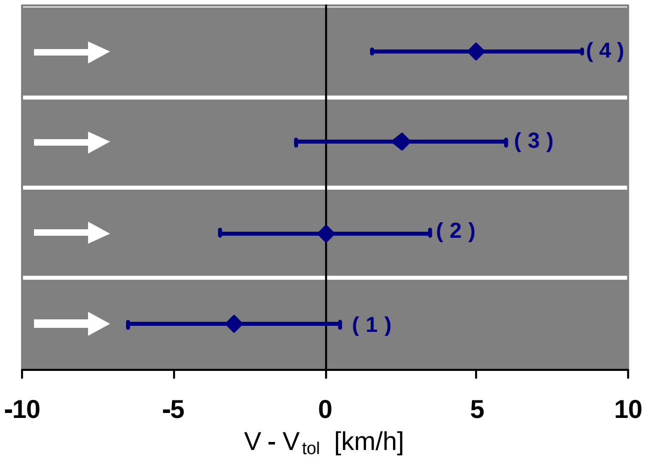

Figure 1: Four examples of excess of speed limit and uncertainty (k=2). The marks indicate the observed car velocities relative to the speed limit. The horizontal bars represent the uncertainty of the observed speed.

Driving too fast

In law enforcement the police uses speed control instrumentation like radar and laser guns. The objective of these instruments is to detect motorists that are driving to fast. The decision to give a speeding ticket to the motorist has to stand in the court of law beyond all reasonable doubt, or in other words with a high degree of confidence.

The legal tolerance or maximum permissible error MPE for in-field speed measurements is 3 km/h and above 100 km/h 3%. The meaning of this requirement is that the reading of the speed control instrumentation can actually deviate 3 km/h or 3% from the reference to which the speed meter is traceable. Based on this information the uncertainty of the speed control instrument is obtained by assuming a rectangular distribution. The relationship between the MPE and the standard uncertainty is obtained from the GUM [2] or EA-4/02 [3].

$$u_s = MPE / \sqrt{3} \qquad (1)$$

The expanded uncertainty (k=2) is

$$U_{k=2} = 2 \cdot MPE / \sqrt{3} \qquad (2)$$

The results of these calculations are tabulated in
Table I.

Table I: Relationship between MPE and measurement uncertainty.

Range	MPE	Standard uncertainty u_s	Expanded uncertainty $U_{k=2}$
0 -100 km/h	3 km/h	1,73 km/h	3,46 km/h
> 100 km/h	3%	1,73%	3,46%

In Figure 1 four examples are given of a speed control measurement as might be observed on a motorway. On the abscissa the velocity relative to the speed limit is shown. For each of the cars the observed velocity and the expanded uncertainty (k=2) are plotted. Case (1) is the motorist that is below the speed limit. Due to the uncertainty of the measurement there is a small probability that he actually driving faster than the speed limit. For the second motorist (2) this probability is actually 50%. The third motorist (3) is exceeding the speed limit, however there is still a probability that he is driving less than the speed limit. The fourth motorist (4) is clearly exceeding the speed limit. Only in this fourth case a fine is the result of a decision with sufficient confidence.

The relationship between speed excess, uncertainty and the risk of erroneously fining the motorist is a classical problem in statistical process control [4][5], which is well documented if the measurement result has a Gaussian distribution. In metrology hardly any information is available on the statistical distribution associated with the measurement uncertainty. So an assumption needs to be made that corresponds to the worst-case situation: i.e. the Gaussian distribution where the standard uncertainty equals the standard deviation. This distribution function is not a distribution in the statistical meaning but is a knowledge representation. An example is displayed in Figure 2 for a car that is driving 2,5 km/h too fast, which corresponds to car (2) in Figure 1.

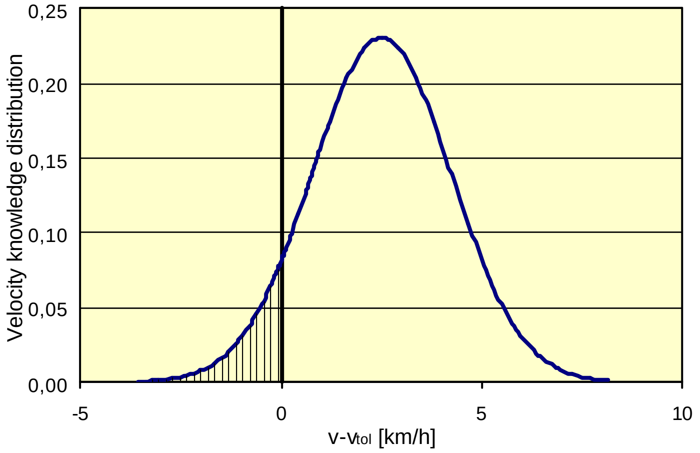

Figure 2: Velocity knowledge distribution corresponding to a car driving 2,5 km/h too fast and a standard uncertainty of 1,73 km/h. The shaded area represents the risk that the motorist gets incorrectly a speeding ticket.

Now the probability that a motorist is driving faster than the speed limit $P(v_{\text{obs}} > v_{\text{tol}})$ can be calculated from the cumulative normal distribution

$$P(v_{obs} > v_{tol}) = \int_{z_{tol}}^{\infty} \frac{1}{\sqrt{2p}} \cdot e^{-z^2/2} dz \qquad (3)$$

in which $z = (v\text{-}v_{\text{obs}})/u_{\text{s}}$, $z_{\text{tol}}= (v_{\text{tol}}\text{-}v_{\text{obs}})/u_{\text{s}}$ with v is the velocity, v_{tol} the speed limit and u_{s} is the standard uncertainty of the observed car velocity v_{obs}. The risk the car is driving less than the speed limit is 1 - P(v_{obs}>v_{tol}) and is displayed as the shaded area in Figure 2. Since only one tolerance is involved this test is called a single side test. The result for different car velocities is shown in Table II. In the first column the excess of the speed limit is expressed as a multiple of the standard uncertainty. The second column shows the probability that the motorist is exceeding the speed limit. The last column is the risk of an erroneous decision if the police give the motorist a speeding ticket. From Table II it is shown that if a motorist is exceeding the speed limit by 2 standard uncertainties or more, the risk on erroneous penalties is 2,3% or less. Due to the worst-case character of the assumed Gaussian distribution the confidence level of the decision taken is always higher than calculated.

In the Netherlands the instruction of the Prosecution Counsel is to fine the speed excess with a threshold of 7 km/h for v_{tol} < 100 km/h, for v_{tol} = 100 km/h the threshold is 8 km/h [6]. The fines are based on the observed velocities reduced with 3 km/h for v_{obs} < 100 km/h and 3% for v_{obs} = 100 km/h, respectively [7]. In combination with an in-field MPE of 3 km/h or 3% this means that the risk of erroneously getting a speeding tickets is limited to the order of 0,01%. Since the tariffs for speeding show a step-wise increment, the maximum risk of getting an incorrect amount on the ticket is 4,2%.

During type approval and initial verification the velocity meter has to stay within an MPE of 1 km/h for v_{obs} < 100 km/h or 1% for v_{obs} = 100 km/h. During a field test the velocity meter has to stay within the ±3 km/h or ±3% [7]. In addition the Prosecution Counsel has ordered all speed meters to be re-verified annually. This guarantees the motorist only a low risk of an incorrect speeding ticket.

Table II: Confidence levels $P(v_{obs}>v_{tol})$ for a single sided test depending on the observed relative speed excess $(v_{obs}-v_{tol})/u_s$. The risk is $1-P(v_{obs}>v_{tol})$.

$(v_{obs}-v_{tol})/u_s$	$P(v_{obs}>v_{tol})$	Risk
1,00	84,1%	15,9%
1,64	95,0%	5,0%
1,96	97,5%	2,5%
2,00	97,7%	2,3%
2,33	99,0%	1,0%
3,00	99,9%	0,1%

Initial and subsequent verifications, single point case

During initial verification the meter needs to stay within two tolerances or MPEs. And the starting hypothesis is that the instrument is performing within tolerances. Since two tolerances are involved this test is called a two-sided test. The confidence level of the test, i.e. the probability that the observed value is between tolerances, equals

$$P(e_{tol-} < e_{obs} < e_{tol+}) = P(z_{tol-} < z_{obs} < z_{tol+}) = \int_{z_{tol-}}^{z_{tol+}} \frac{1}{\sqrt{2p}} \cdot e^{-z^2/2} dz \tag{4}$$

in which $z = (e-e_{obs})/u_s$, $z_{tol\pm} = (e_{tol\pm}-e_{obs})/u_s$ with e is the deviation of the meter, $e_{tol\pm}$ are the + and – tolerances and u_s is the standard uncertainty of the observed meter deviation e_{obs}. Of course $z_{obs}=0$. The decision to accept the instrument has a risk of $1-P(e_{tol-} < e_{obs} < e_{tol+})$. The confidence level is also influenced by the ratio of the tolerance and uncertainty as is shown in Table III.

Table III: Confidence levels $P(v_{obs}>v_{tol})$ for a single sided test depending on the observed relative speed excess $(v_{obs}-v_{tol})/u_s$. The risk is $1-P(v_{obs}>v_{tol})$.

$\lvert e_{tol}\rvert-\lvert e_{obs}\rvert/u_s$	$\lvert e_{tol}\rvert$ » u_s	$\lvert e_{tol}\rvert = u_s$	$\lvert e_{tol}\rvert = 2\cdot u_s$	$\lvert e_{tol}\rvert = 2{,}5\cdot u_s$	$\lvert e_{tol}\rvert = 3\cdot u_s$
1	84,134%	68,269%	83,999%	84,131%	84,134%
1,64	95,000%		94,074%	94,960%	94,999%
1,96	97,500%		95,433%	97,382%	97,497%
2	97,725%		95,450%	97,590%	97,722%
2,33	99,000%			98,625%	98,988%
3	99,865%				99,730%

The alternative hypothesis is that the meter is performing outside the tolerances. The probabilities that the observed deviation is above the upper tolerance e_{tol+} or below the lower tolerance e_{tol-} are

$$P(e_{obs} < e_{tol-}) = P(z_{obs} < z_{tol-}) = \int_{-\infty}^{z_{tol-}} \frac{1}{\sqrt{2p}} \cdot e^{-z^2/2} dz \tag{5}$$

and

$$P(e_{obs} > e_{tol+}) = P(z_{obs} > z_{tol+}) = \int_{z_{tol+}}^{+\infty} \frac{1}{\sqrt{2p}} \cdot e^{-z^2/2} dz \tag{6}$$

The risks associated with the decisions based on these test are $1-P(e_{obs} < e_{tol-})$ and $1-P(e_{obs} > e_{tol+})$, respectively. A graphical display of the risks associated with the acceptance or rejection of the instrument is shown in Figure 3. This figure shows that there is a high risk associated with the acceptance of an instrument if observations are outside tolerances. Likewise, rejection of an instrument if observations are within tolerances has a high risk. If the maximum acceptable risk on an erroneous decision is 5% the width of the rectangles in Figure 3 represent the range of observations in which no decision can be taken. Between the shaded areas the instrument is said to be conforming, outside the shaded areas the instrument is non-conforming. Inside the shaded areas the instrument is not conforming and not non-conforming at the same time. The width of the rectangles is depending on the risk level accepted: the larger the risk the smaller the width of the rectangle. In case of a 50% risk there will be no rectangle at all. In the latter case acceptance or rejection of an instrument comes close to tossing coins.

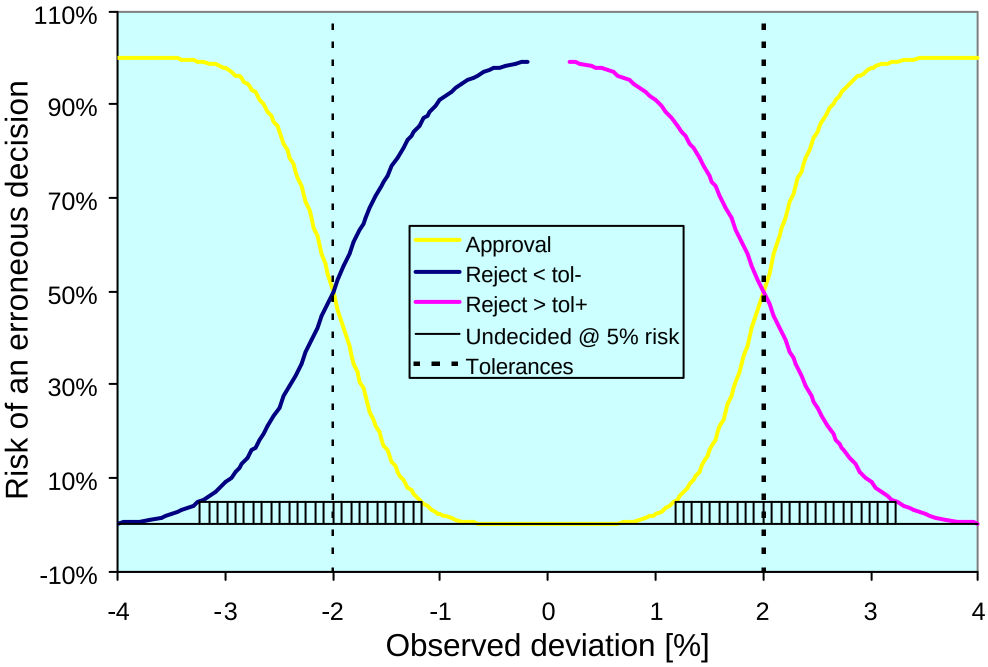

Figure 3: Risk distribution for the approval or rejection of an instrument that must perform within two tolerances ±2%. The uncertainty (*k*=2) of the approval observations is 1%. The uncertainty of the rejection observations is 1.5%. The shaded rectangles represent the range of observations where no decision with less than 5% risk is possible.

In line with previous discussion Sommer and Kochsiek [8] propose acceptance criteria obtained by reduction of the tolerance with the expanded measurement uncertainty. They also state that this results in a de-facto reduction of error limits and that common use in legal metrology seems to be unlikely due to the commercial implications of such a reduction. Moreover the acceptance criteria are variable, depending on the uncertainties that are obtained by different laboratories.

It is very clear that 100% confidence or zero risk can never be obtained when taking metrological decisions. For a given confidence level a smaller measurement uncertainty results in acceptance criteria closer to the applicable tolerances. This is an advantage for manufacturers that have their instruments verified with a low uncertainty. The confidence level or risk is also influenced by the Gaussian knowledge distribution assumed. If it can be proven that another statistical distribution gives a more adequate knowledge representation, this will most certainly result in acceptance criteria that are closer to the tolerances. The use of distributions will be an important research issue in the near future.

The current practise of initial verification and law enforcement by in-field inspection of fuel dispensers in The Netherlands is already aiming at limiting the risks of metrological decisions. Accredited organisations that perform the initial verification of a fuel dispenser use as an acceptance criterion ±0,4% where the tolerance is ±0,5%. The law enforcement officers close the fuel dispenser if it is deviating more than +0,7%. For deviations within ±0,5% no action is required. In all other cases repair by the owner is required. This practise works for more than 10 years to the satisfaction of all parties involved [9] due to a reduced risk that an already approved fuel dispenser will appear non-conforming. Another example of risk reduction concerns coriolis meters intended for metering gasses, which may be verified with water if reduced tolerances are used. The reason is that the use of gas introduces additional uncertainties. This practise has not been established yet for other metrology areas in The Netherlands. In the light of these different metrological practises the question what risks are acceptable, should be elevated to the level of the OIML technical committee on verification.

Initial and subsequent verifications, multi-point case

During the initial verification of an instrument the deviation of the instrument versus the flow rate in the range of the instrument is determined. If one of the observed deviations does not meet the acceptance criteria, the meter is not approved. Now the risk that the instrument is not conforming, is the sum of the risks of all individual observations. An example of a verification of a turbine gas meter is shown in Table IV. For a normal initial verification the deviation is measured at 6 different flow rates. For curve fitting purposes two additional verification points are added. The right-hand column shows for each observation the risk that the observation leads to an erroneous approval of the meter. The highest contribution is found at 20 m^3/h where the difference between the observed deviation and the tolerance equals the uncertainty. If the difference between tolerance and observed deviation is two uncertainties (k=2) or more, its contribution to the risk of erroneously approving the meter can be neglected. For a better impression on the performance of the meter more observations can be made. The paradox of this approach is that despite we get a better impression of the meter performance the risk attributed to erroneously approving the meter increases with an increasing number of observations.

Table IV: Example of a verification of a turbine gas flow meter. The third column represents the uncertainty of the observed deviation. The last column represents the risk that an observation leads to an erroneous approval of the meter. At the 8th row the sum of all risks is displayed. The two bottom lines are additional verification points that are used in the curve fit.

Flow rate m^3/h	Deviation e	U_e (k=2)	Tolerance	Risk
5	-1,50%	0,40%	2%	0,62%
10	0,00%	0,40%	2%	0,00%
20	0,70%	0,30%	1%	2,28%
40	0,55%	0,30%	1%	0,13%
70	0,40%	0,30%	1%	0,00%
100	0,00%	0,30%	1%	0,00%
				3,03%
55	0,50%	0,30%	1%	0,04%
85	0,30%	0,30%	1%	0,00%

A method that utilises the additional information of the extra verification points is the linear curve fit method developed by Van der Grinten and Peters [1]. As a standard curve fit methods does not count with the individual uncertainties of the data points, it is only suitable for type A evaluation of uncertainties. Verification data however contain uncertainties that are the result of both a type A and a type B evaluation. For the linear case Van der Grinten and Peters calculate the fit, which is identical to the standard regression method, and an uncertainty envelope in which the uncertainties of the individual data are included. All regression methods require at least 6 degrees of freedom. The regression line is

$$e(x) = m \cdot (x - x_0) + e_0, \quad x = Q^{-1} \tag{7}$$

with m the regression coefficient, e_0 the deviation at x_0 and Q the flow rate in m^3/h. If x_0 and e_0 are chosen to be the arithmetic averages of transformed flow rates and the deviations, respectively, the uncertainty envelope is found from [1]

$$u_e^2(x) = (x - x_0)^2 u^2(m) + u^2(e_0) \tag{8}$$

The trend line will meet the tolerances with a confidence level 1-a if the bounds $e_\pm(x)$ are within the tolerances, where

$$e_\pm(x) = e(x) \pm \tfrac{1}{2}\, t_{1-a/2}\, u_e(x) \tag{9}$$

and $t_{1\text{-a}}$ follows from the Student-t distribution for 6 degrees of freedom. The results of applying this method to all the experimental data of Table IV are displayed in Figure 4, where the results are transformed back the flow rate domain. Figure 4 shows that the 95% confidence envelope is everywhere within the tolerances except for the lowest flow rate. A better fit with smaller uncertainties will be obtained if higher order curve fits will be utilized for the trend of the meter curve. To this end the method developed in [1] has to be generalised for the multi-linear regression case.

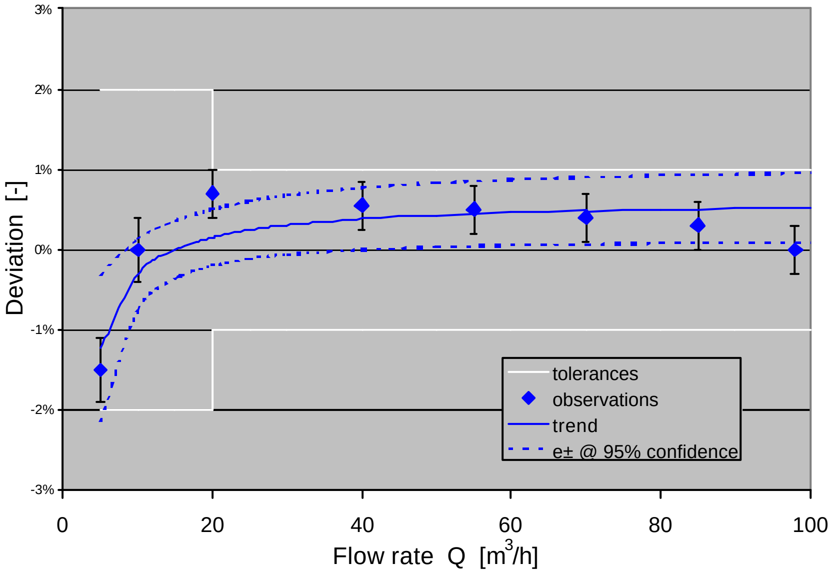

Figure 4: Linear curve fit (solid line) together with the 95% confidence envelope (dashed lines) of the verification data of Table IV. The deviation is plotted versus the indicated flow rate in m^3/h.

For the moment it seems practical to use the following strategy to determine the risk associated with approving an instrument. For low numbers of data the risk of erroneously approving an instrument is the sum of the risks that an individual data point leads to an incorrect decision. If there are sufficient data, i.e. at least 6 degrees of freedom, it is best to make a curve fit with a 95% confidence envelope.

Intercomparisons

The last example of taking decisions occurs in intercomparisons and round robins when measurement results from different laboratories are compared. If two or more laboratories measure the same specific quantity under fully comparable conditions they will get different results. According to EA recommendation EA-2/03 [10] two measurement results differ significantly if the absolute value of the difference is greater than the uncertainty of the difference.

$$| R_1 - R_2 | > \sqrt{U_1^2 + U_2^2} \qquad (10)$$

Here $R_1 \pm U_1$ is the result by laboratory 1 and $R_2 \pm U_2$ is the result by laboratory 2. U is the expanded measurement uncertainty with k=2. Another quantity used for comparing results is the normalised difference E_n

$$E_n = \frac{| R_1 - R_2 |}{\sqrt{U_1^2 + U_2^2}} \qquad (11)$$

If E_n is greater than 1 the difference is called significant. What is the degree of confidence associated with the above significance criteria? Or in other words what is the risk of the decision that the results of two laboratories are significantly different. Again the answer can be given by means of statistical testing and again the statistical distribution of the difference $|R_1$-$R_2|$ is not known. Also here the worst-case approximation is the assumption of a Gaussian distribution with an average of ? R = $|R_1$-$R_2|$ and a standard deviation s equal to s = ½v(U_1^2+U_2^2).
Applying the transformation z = (r-? R)/s the confidence level is obtained from the standard normal distribution

$$P(E_n < 1) = \int_{-2}^{2} \frac{1}{\sqrt{2p}} \cdot e^{-z^2/2} dz = 95,4\% \qquad (12)$$

So the maximum risk of the decision that two laboratories have different results is 4,6%.

Conclusions

From the previous analysis the following conclusions can be drawn.

1. From a statistical perspective confidence levels and risks are associated with decisions, not with uncertainties.
2. It is impossible to achieve 100% confidence for any decision taken, i.e. there is always a risk on an incorrect decision.
3. In order to establish a relationship between acceptance criterion, tolerance, uncertainty and confidence level a statistical distribution has to be assumed that represents our knowledge of the measurement results. For this purpose the worst case approximation is the Gaussian distribution. This means that in practice the confidence level of the decision taken will always be higher than calculated. If our knowledge can be represented by other statistical distributions this will certainly result in acceptance criteria that are closer to the tolerances.
4. With respect to the acceptable risks of metrological decisions there are different practises in different metrological areas. The juridical and commercial implications of applying lower risk levels should be discussed on OIML level.
5. Due to the thresholds used for speed enforcement the risk of incorrectly getting a speed ticket is of the order of 0,01%. Due to the step-wise tariff system for speed tickets the maximum risk of getting a wrong amount on the invoice is 4,2%.

6. There are two types of verifications; one is carried out with the objective to approve the meter; the other verification is to find instruments that are performing outside the metrological tolerances. If the accepted risk on an erroneous decision is less than 50% there is a range of observations for which the instrument can neither be approved nor rejected: the instrument is not conforming and not non-conforming at the same time.
7. For the moment it seems practical to use the following strategy to determine the risk associated with approving an instrument based on a multi-point verification. For low numbers of data the risk of erroneously approving an instrument is the sum of the risks that an individual data point leads to an incorrect decision. If there are sufficient data, i.e. at least 6 degrees of freedom, it is best to make a curve fit with a 95% confidence envelope.
8. In intercomparisons the criterion for a significant difference is if the difference of two results is bigger than the uncertainty (k=2) of the difference. The confidence level of this decision is 95,4%.

References

[1] Grinten, J.G.M. van der and Charles M.E.E Peters (1994): Trend analysis, general model and application to high-pressure gas flow standards, *in:* proceedings of the IMEKO XIII conference, September 1994, Turin, Italy, pp 1161-1164a.

[2] BIPM/IEC/IFCC/ISO/IUPAC/IUPAP/OIML (1993): Guide to the expression of uncertainty in measurement, first edition, ISO 1993.

[3] European Accreditation (1999): Expression of the uncertainty of measurement in calibration, document EA-4/02, December 1999.

[4] Grinten, P.M.E.M. van der and J.M.H. Lenoir (1973): Statistische procesbeheersing (Statistical process control), Uitgeverij Het Spectrum B.V. Utrecht / Antwerpen.

[5] Eadie, W.T., D. Drijard, F.E. James, M. Roos, B. Sadoulet (1971): Statistical methods in experimental physics, North-Holland Publishing Company, Amsterdam – London.

[6] See the website of the Dutch Openbaar Ministerie (Prosecution Council): www.om.nl/beleidsregels/dbase/verkfrm.htm document www.om.nl/beleidsregels/docs/2002a014.htm (in Dutch)

[7] Private communication with Mr. Paul Kok of NMi Certin.

[8] Sommer, K.D. and M. Kochsiek (2002): Role of measurement uncertainty in deciding conformance in legal metrology, OIML Bulletin Vol. 18, No. 2, pp. 19-24.

[9] Private communications with Mr. Aart Kooiman and Mr. Ed van Römer of Verispect.

[10] European Accreditation (1996): EA Inter-laboratory Comparison (previously EAL-P7), document EA-2/03, rev. 01, March 1996.

Contact point

Dr. ir. Jos G.M. van der Grinten
NMi Certin B.V., P.O. Box 394, 3300 AJ DORDRECHT, The Netherlands
Phone: +31 78 633 2368, Fax: +31 78 613 4647, e-mail: JvanderGrinten@nmi.nl