Statistical Methods in Chemistry

Dr.V.A.Kalantre

Evaluation of analytical data

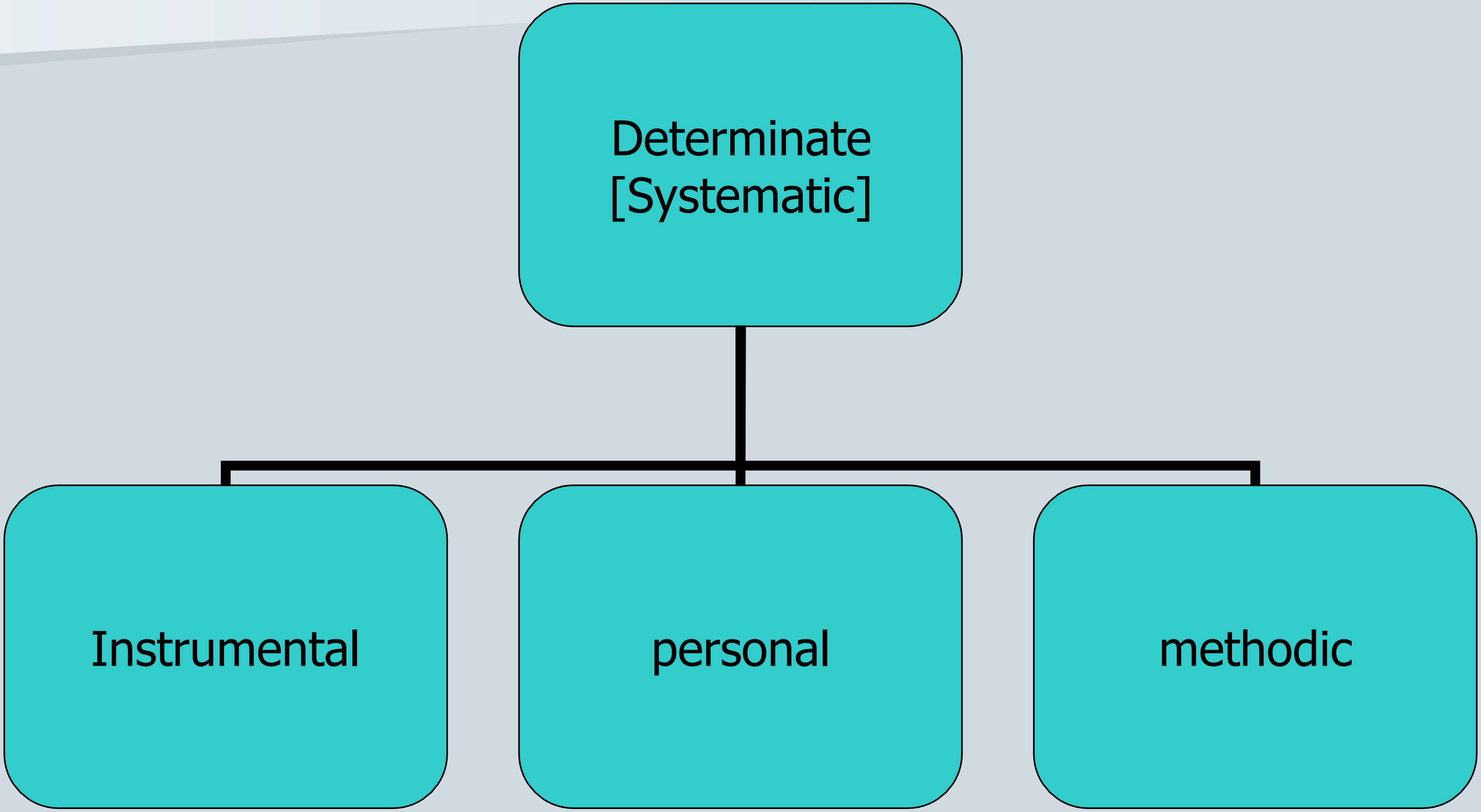
- Precision and Accuracy

Absolute error and relative error.

$$E = \overline{x} - x_i$$

$$E_r = \frac{\overline{x} - x_i}{x_i} \times 100\%$$

Types of error



Indeterminate error [Random error]

- Sample mean

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Population mean

$$\mu = \lim_{N \to \infty} \frac{\sum_{i=1}^{N} x_i}{N}$$

Standard Deviations

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})}{N}}$$

$$\sigma = \sqrt{\lim_{N\to\infty}\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}}$$

$$s_{pool} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x}_1) + \sum_{j=1}^{N}(x_j - \overline{x}_2) + \sum_{k=1}^{N}(x_k - \overline{x}_3)}{N_1 + N_2 + N_3}}$$

Confidence intervals

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{N}}$$

$$\mu = \bar{x} \pm \frac{z\sigma}{\sqrt{N}}$$

Confidence Limit

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{N}}$$

v	95%	99%
5	2.571	4.032
6	2.447	3.707
7	2.365	3.500
8	2.306	3.335

Error propagation in Arithmetic Calculation

- Addition or subtraction
- X= p+q-r

$$s_x = \sqrt{s_p^2 + s_q^2 + s_r^2}$$

- Multiplication or division
- X=p x q/r

$$\frac{s_x}{x} = \sqrt{\left(\frac{s_p}{p}\right)^2 + \left(\frac{s_q}{q}\right)^2 + \left(\frac{s_r}{r}\right)^2}$$

Rejection of Data

- 2.5 d Rule
- 4 d Rule
- Q Test

$$Q = \frac{difference}{Range}$$

n	q99
3	0.994
4	0.926
5	0.821

Least-Squares Line

$$SS_{resid} = \sum_{i=1}^{N} \left[y_i - \left(b + mx_i \right) \right]^2$$

$$S_{xx} = \sum \left(x_i - \bar{x} \right)^2 = \sum x_i^2 - \frac{\left(\sum x_i \right)^2}{N}$$

$$S_{yy} = \sum \left(y_i - \bar{y} \right)^2 = \sum y_i^2 - \frac{\left(\sum y_i \right)^2}{N}$$

$$S_{xy} = \sum \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{N}$$

- 1.The slope of the line m.

$$m = \frac{S_{xy}}{S_{xx}}$$

- 2.The intercept,b:

$$b = \bar{y} - m\bar{x}$$

- 3.The standard deviation about regression, s_r

$$S_r = \sqrt{\frac{S_{yy} - m^2 S_{xx}}{N-2}}$$

- 4.The standard deviation of the slop, S_m

$$S_m = \sqrt{\frac{S_r^2}{S_{xx}}}$$

- 5.The standard deviation of the intercept, S_b

$$S_b = S_r \sqrt{\frac{\sum x_i^2}{N\sum x_i^2 - \left(\sum x_i\right)^2}}$$

- 6.The standard deviation of results

$$S_c = \frac{S_r}{m}\sqrt{\frac{1}{M} + \frac{1}{N} + \frac{\left(\bar{y}_c - \bar{y}\right)^2}{m^2 S_{xx}}}$$

uncertainty

- Parameter, associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand

Standard uncertainty

- Uncertainty of measurement that is expressed in terms of standard deviation
- Type A: method of evaluation of uncertainty by statistical analysis of observations
- Type B : method of evaluation of uncertainty other than statistical method

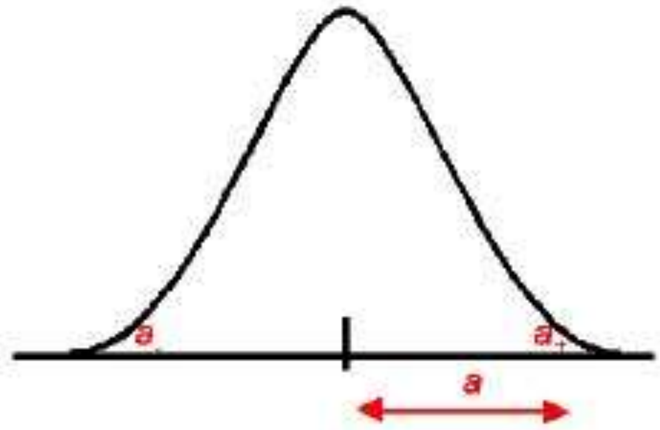
Linear Equation

- Least Square Method
- Method of Average
- Co-relation Co-efficient

$$r = \frac{\sum x_i y_i - n\overline{xy}}{\sqrt{\left(\sum x_i^{\,2} - n\bar{x}^2\right)\left(\sum y_i^{\,2} - n\bar{y}^2\right)}}$$

The Process of Measurement Uncertainty Estimation

- Specify measurand
- Identify uncertainty sources
- Quantify uncertainty components
- Calculate combined uncertainty



mean value $= (a_+ + a_-)/2$

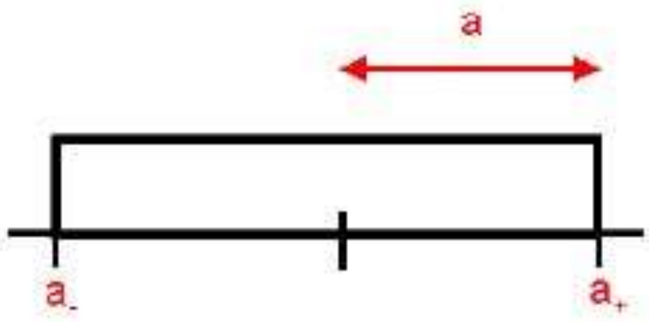
Confidence interval = 2a
Uncertainty estimate is for ± a

Standard uncertainty =

a/1.96	for 95 % confidence interval
a/2.576	for 99 % confidence interval
a/3	for 99.73 % confidence interval (3 standard deviations)

Use when evaluated from:
- limits of random replication
- standard deviation
- confidence interval

Normal distribution



$$\text{mean value} = (a_+ + a_-)/2$$

$a = (a_+ - a_-)/2$

Interval = 2a

Standard uncertaint y = $a/\sqrt{3}$

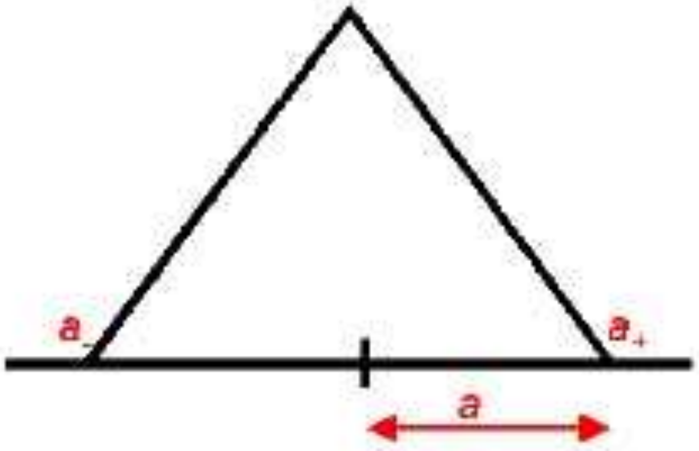
Default model if information is limited

Use when evaluated from:

•specification with no confidence level

•maximum range with unknown shape of distribution

Uniform distribution



$$\text{mean value} = (a_+ + a_-)/2$$

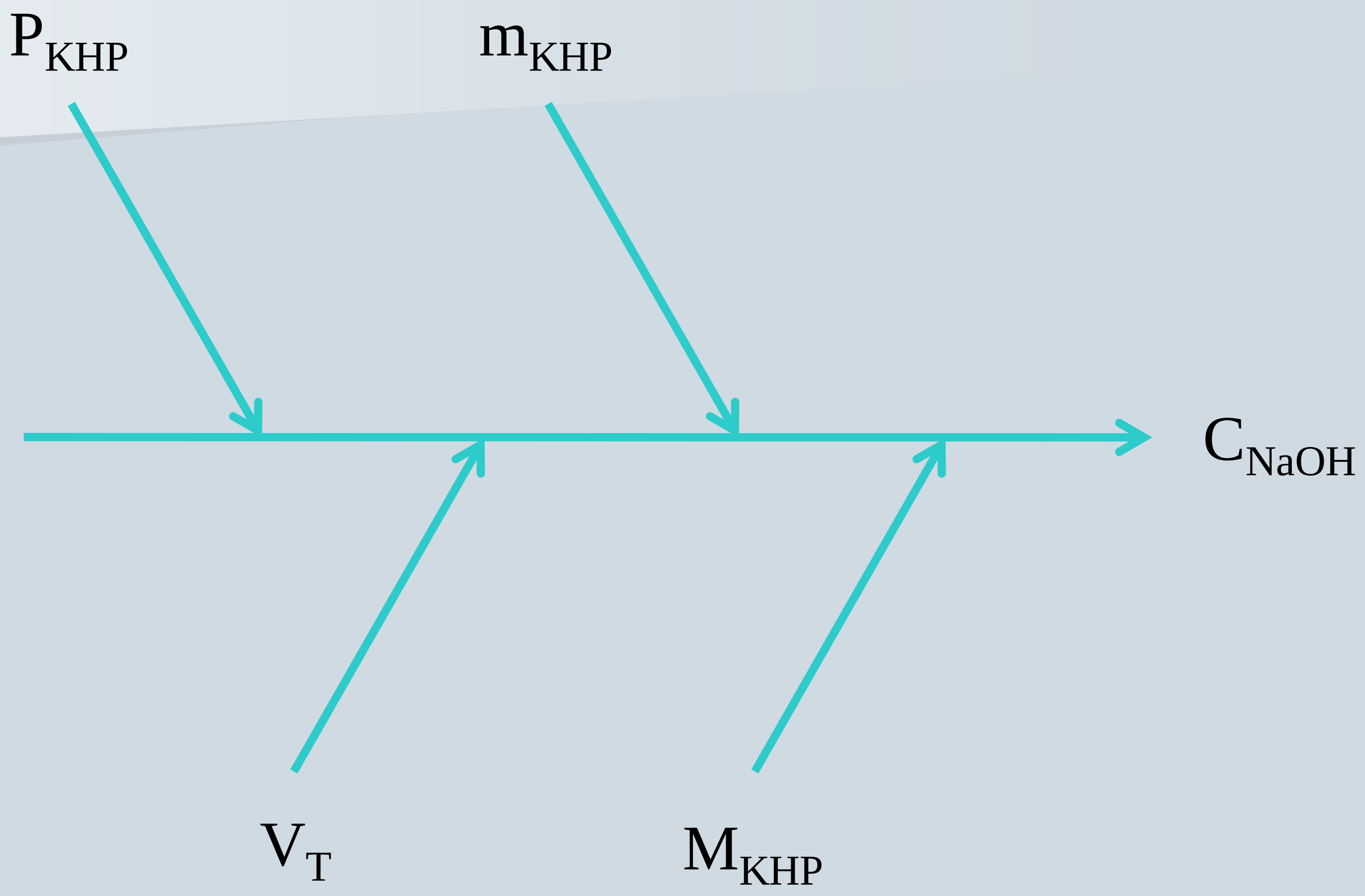
Interval = 2a

Standard uncertainty = $a/\sqrt{6}$

Use when evaluated from:

- maximum range with central tendency
- maximum range with symmetric distribution

Triangular distribution



Simple cause and effect diagram for standardization of a NaOH solution with KHP

Step 1) Specify measurand, express mathematically the equation relating measurand and input quantities. Identify all uncertainty sources.

$$c_{Cd} = \frac{1000 \cdot m \cdot P}{V} [mg/l]$$

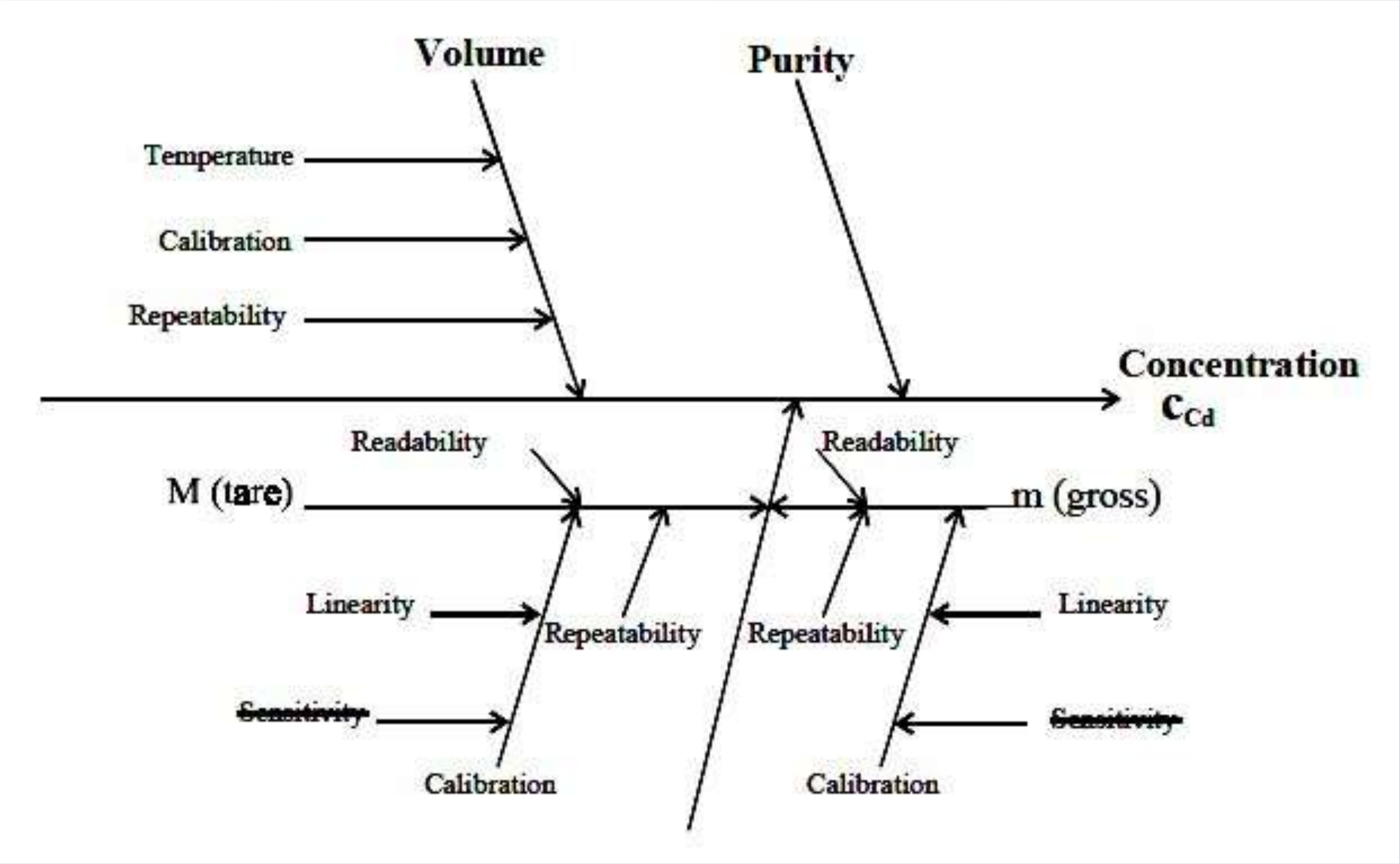
c_{Cd}: concentration of the calibration standard obtained
m: mass of the clean high purity Cd piece [mg]
P: Purity of the metal
V: volume of the flask [ml]
1000: conversion factor from ml to l

Listing the components of uncertainty

1. Purity of the Cd: supplier's certificate: 99.99 ± 0.01 %.
2. Mass of the metal from weighing in the flask. The piece weighed 0.10028g. The manufacturer's literature identifies 3 uncertainty sources for tare weighing:
   - 2.1 repeatability,
   - 2.2 readability of the balance scale,
   - 2.3 calibration (involving sensitivity of the balance and linearity).

   Sensitivity can be neglected because weighing was done on the same balance over a narrow range. Buoyancy correction can be neglected [3] being very small.
3. Volume of the solution: 100ml. Uncertainty sources are:
   - 3.1 uncertainty in the certified internal volume of the flask
   - 3.2 Filling the flask to the mark
   - 3.3 Temperature influences

“Cause and Effect” diagram:



Step 2) Determine the input quantities

1. Purity of the Cd: 99.99 ± 0.01 % i.e. 0.9999 ± 0.0001.
2. Mass of the Cd: 0.10028 g.
3. Volume of the solution: 100ml

Step 3) Quantifying the single uncertainty components

1. Purity of the Cd: Type B evaluation: 99.99 ± 0.01 % i.e. 0.9999 ± 0.0001.
A rectangular distribution is assumed, because there is no further information. Therefore the standard uncertainty of the purity is:
$$u(P) = \frac{0.0001}{\sqrt{3}} = 0.000058$$

2. Mass of the Cd: 0.10028 g. The manufacturer of the balance recommends 0.05 mg as uncertainty estimation, this value can be taken directly as
$u(m) = 0.05mg$ (Type B)

3. Volume of the solution: 100ml

3.1 uncertainty in the certified internal volume of the flask: the manufacturer quotes a volume for the flask of 100 ± 0.1 ml at 20°C.
No confidence level is given, so a triangular distribution was chosen, because in an effective production process the nominal value is more likely than extremes.
Therefore
$u(V_1) = \frac{0.1ml}{\sqrt{6}} = 0.04ml$ (Type B)

3.2 Filling the flask to the mark: An experiment of 10 fill and weigh experiments gave a standard deviation of 0.02 ml. This can be used directly as
$u(V_2) = 0.02ml$ (Type A)

3.3 Temperature influences: The laboratory temperature varies between the limits of 20°C ± 4. The volume expansion of water is large compared to flask material, which is therefore neglected. The volume expansion of water is $2.1 \cdot 10^{-4} / °C$, leading to volume variation of
$$\pm(100 \cdot 4 \cdot 2.1 \cdot 10^{-4})ml = \pm 0.084ml$$
Assuming rectangular distribution gives
$$u(V_3) = \frac{0.084}{\sqrt{3}} = 0.05ml$$

The three volume effects add to each other and are treated like a sum. The combined uncertainty from volume effects is then
$$u(V_{total}) = \sqrt{0.04^2 + 0.02^2 + 0.05^2} = 0.07ml$$

Step 4) Identify the covariances (of correlated input quantities)

Correlation effects are not known and the approximation is made that there is no correlation.

Step 5) Calculate the result of the measurement from the input quantities

Determination of the concentration

$$c_{Cd} = \frac{1000 \cdot m \cdot P}{V}[mg/l] = \frac{1000 \cdot 100.28 \cdot 0.9999}{100} mg/l = 1002.7 mg/l$$

The concentration of the calibration standard is 1002.7 mg/l.

Step 6) Calculate the combined uncertainty

Because the above equation is a multiplicative expression, the uncertainties are combined by:

$$\frac{u_{combined}(c_{Cd})}{c_{Cd}} = \sqrt{\frac{u(P)^2}{P^2} + \frac{u(m)^2}{m^2} + \frac{u(V_{total})^2}{V_{total}{}^2}} = \sqrt{\frac{0.000058^2}{0.9999^2} + \frac{0.05^2}{100.28^2} + \frac{0.07^2}{100^2}} = 0.0009$$

$$u_{combined}(c_{Cd}) = 0.9 mg/l$$

Comparing the uncertainties from the components shows that volume and mass uncertainties contribute in a similar way to the overall uncertainty, while the purity has almost no influence on it.

Step 7) Calculate the expanded uncertainty

The expanded uncertainty is

$$U = k \cdot u_{combined}(c_{Cd}) = 2 \cdot 0.9 mg/l = 1.8 mg/l$$

The coverage factor k is chosen to be 2 as recommended by the GUM [1].

Step 8) Give the result together with the uncertainty as estimated

The concentration of the Cd standard is 1002.7 ± 1.8 mg/l. The reported expanded uncertainty is based on a standard uncertainty multiplied by a coverage factor k = 2, providing a level of confidence of approximately 95%.