«Laboratórios de Engenharia Química II» (LEQ II)

T — true value
M — measured value

Error: $$e = x_{\text{measured}} - x_{\text{true}} = M - T$$

p 6 $$e_{\text{abs}} = \left|M - T\right| \qquad e_{\text{rel}} = \frac{e_{\text{abs}}}{T} = \frac{\left|M - T\right|}{T} \qquad \{1\}$$

T is estimated by $\overline{M}$.

p 6 $$T \cong \overline{M} \qquad \rightarrow \qquad T = \overline{M} \qquad \{2\}$$

p 7 Accuracy *(Exactidão)* — agreement between M and T.
Precision *(Precisão)* — agreement between several M's (in the same conditions), i.e., *repetitions* or *replicates* (statistical concept). Expresses reproducibility *(reprodutibilidade)*. [(Numerical) precision *(precisão numérica)* resolution, number of significant figures (numerical concept).])

p 9 $$m = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s = \lim_{n\to\infty} \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - m\right)^2}$$
$$d = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\left|x_i - m\right| \qquad \text{(Gauss :)} \quad d \cong 0.8s \qquad \{3\}$$

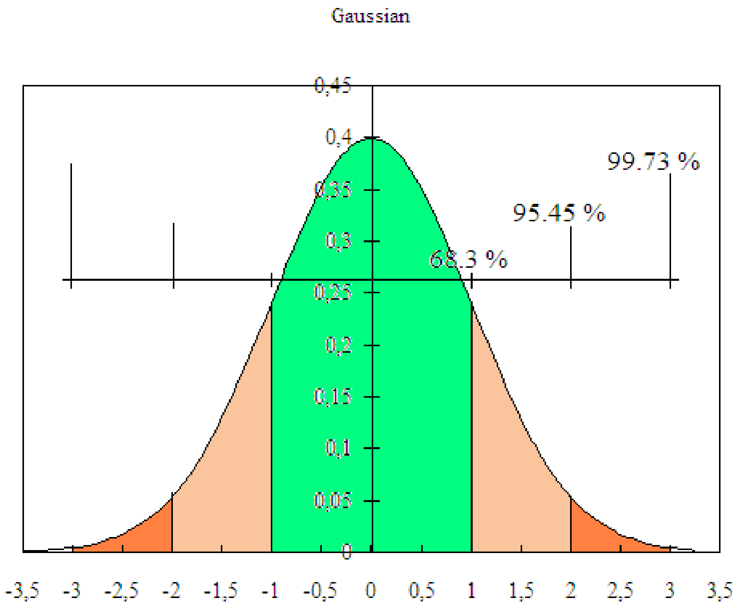
p 11 $$\begin{aligned} P_{\text{inside}}(k) &\equiv \Pr\left[x \in \left(m \pm ks\right)\right] = \Pr\left(m - ks < x < m + ks\right) = \\ &= \Pr\left(-k < z = \frac{x-m}{s} < +k\right) = \Pr\left(-k < z < +k\right) = \\ &= \Phi(k) - \Phi(-k) = \Phi(k) - \left[1 - \Phi(k)\right] = 2\Phi(k) - 1 \end{aligned} \qquad \{4\}$$

p 11 $$\begin{aligned} \Phi(k) &= \text{NORMSDIST}^{\text{Excel}}(k) = \\ &= \text{NORMDIST}^{\text{Excel}}\left(x = m + ks\,; m; s\,; \text{TRUE}\right) \end{aligned}$$
$$k = \Phi^{\text{inv}}\left(\frac{1}{2} + \frac{P_{\text{inside}}}{2}\right) = \Phi^{\text{inv}}\left(\frac{1 + P_{\text{inside}}}{2}\right) = \text{NORMINV}^{\text{Excel}}\left(\frac{1 + P_{\text{ins.}}}{2}\right) \qquad \{5\}$$

(See probabilities below.)
The classical convention is usually adopted here: lower case for *pdf* (probability *density* function), upper case for *cdf* (*cumulative* distribution function).

Gaussian distribution

x	Prob. (%)	$z\ (=\dfrac{x-m}{s})$
$(k \equiv z)$	$P = F(z)$	$z = F^{\text{inv}}(P)$
$m \pm s$	68.3	**1**
	90	1.64
	95	1.96
$m \pm 2\ s$	95.4	**2**
	98	2.33
	99	2.58
$m \pm 3\ s$	99.7	**3**
$m \pm 4\ s$	99.98	4



Sample (statistics): average *[média (amostral)]*, variance *(variância)*, standard deviation *(desvio-padrão)*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad T = \lim_{n\to\infty} \bar{x}$$

p 12 {6}

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Average deviation *(desvio médio)*:

p 12 $$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$ {7}

For a Gaussian variable ($n \to \infty$), $d \to \sim 0.80\ s$.

Coefficient of variation *(coeficiente de variação)*:

p 12 $$C_{\mathrm{v}} = \frac{s}{\bar{x}}$$ {8}

p 13 $$s\left(\bar{X}\right) = \frac{s}{\sqrt{n}}$$ $$m = \bar{x} \pm t\,\frac{s}{\sqrt{n}}$$ {9}

For large samples, i.e., if $n \to \infty$ (or $n >$ ~50), $t \to z$.

Caution: the usual notation (following), $t_{P,n}$, may be confusing.

p 13 $$t \equiv t_{P,n} = T^{\mathrm{inv}}\left(\frac{1}{2} + \frac{P_{\mathrm{inside}}}{2}; n = n-1\right) - T^{\mathrm{inv}}\left(\frac{1}{2} - \frac{P_{\mathrm{inside}}}{2}; n = n-1\right) =$$ $$= 2\,T^{\mathrm{inv}}\left(\frac{1+P_{\mathrm{ins.}}}{2}; n\right) = \mathrm{TINV}^{\mathrm{Excel}}\left(1-P_{\mathrm{ins.}}; n\right)$$ {10}

Rejection of an *outlier*, say, x_k

Calculate, *without* the (possible) outlier (so, n becomes $n - 1$):

p 14 $$\bar{x}\,,\ s$$ $$\bar{x} \pm t_{P,n}\, s \equiv \bar{x} \pm T^{\mathrm{inv}}\left(\frac{1}{2} + \frac{P_{\mathrm{inside}}}{2}; n\right) s$$ {11}

If x_k is *not* in this interval, *reject*; otherwise, *accept*.

Kurtosis *(curtose)*:

p 15 $$\mathrm{kurt} = \mathrm{KURT}^{\mathrm{Excel}}(\mathrm{vector}) =$$ $$= \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4 - 3\,\frac{(n-1)^2}{(n-2)(n-3)}$$ {12}

(See the Excel Help for kurtosis and skewness.) There are two trends for the use of kurtosis. The above definition makes kurt = 0 for a (large) *Gaussian* sample, instead of the classical value 3. Thus, relatively to the Gaussian, for kurt < 0, the distribution is flat; and for kurt > 0, it is peaked.

Skewness *(enviesamento)*:

p 15 $$\mathrm{skew} = \mathrm{SKEW}^{\mathrm{Excel}}(\mathrm{vector}) = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3$$ {13}

Indicates asymmetry (e.g., 0 for Gaussian). If *positive*, long *tail* to the *positive* side; and vice-versa.

pp 16–24 (Measurements, error propagation (upper limit of error, probable error. Error propagation. Significant figures.)

Regression analysis

Free straight line (general case)

p 25 $$y = a_0 + a_1 x \qquad \{14\}$$

Notation akin to Walpole & Myers [1989].

$$ss_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad ss_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
$$ss_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \qquad \{15\}$$

Slope (a_1), intercept (a_0) *(coeficiente angular, ordenada na origem)*:

p 28 $$\hat{a}_1 = \frac{ss_{xy}}{ss_{xx}} \qquad \hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x} \qquad \{16\}$$

$$s_{xx} = \sqrt{ss_{xx}} \qquad s_{yy} = \sqrt{ss_{yy}} \qquad \{17\}$$

Coefficient of determination *(coeficiente de determinação)*:

$$R^2 = \frac{ss_{xy}^2}{ss_{xx} ss_{yy}} \qquad \{18\}$$

Variance of the correlation:

$$e_i = \hat{y}_i - y_i$$

SSE, sum of squares of the errors (about the regression line):

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 \qquad \{19\}$$

$$\mathrm{var}_{err} = \frac{1}{n-2}\mathrm{SSE} \qquad s_{xy} = \sqrt{\mathrm{var}_{err}}$$

Standard errors:

$$\mathrm{std\_err}(a_0) = s_{xy}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ss_{xx}}} \qquad \mathrm{std\_err}(a_1) = \frac{s_{xy}}{s_{xx}} \qquad \{20\}$$

Confidence interval of the intercept and the slope, a_0 and a_1:

p 29 $$a_0 = \hat{a}_0 \pm t_{P,n}\ \mathrm{std\_err}(a_0) \qquad a_1 = \hat{a}_1 \pm t_{P,n}\ \mathrm{std\_err}(a_1)$$
$$n = n - 2 \qquad \{21\}$$

Confidence interval of (one value of) y_i:

p 30
$$\left(\hat{Y}_i\right)_1 = \hat{y}_i \pm t\, s_{xy} \sqrt{1 + \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{ss_{xx}}} \qquad \{22\}$$

Confidence interval of the *average* of (many values of) y_i:

p 31
$$\left(\hat{Y}_i\right)_{\text{ave.}} = \hat{y}_i \pm t\, s_{xy} \sqrt{\frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{ss_{xx}}} \qquad \{23\}$$

Remark that $\text{err}\left(\hat{Y}_i\right)_{\text{ave.}} < \text{err}\left(\hat{Y}_i\right)_1$, i.e., of course, the average (of predicted values) varies less than the individual predicted value.

Straight line through the origin

p 33
$$y = a_1 x \qquad \{24\}$$

Slope (a_1) *(coeficiente angular)*:

p 33
$$a_1 = \frac{\sum_{i=1}^{n} x_i y_i}{ss_{xx}} \qquad \{25\}$$

Standard error:

$$\text{std\_err}\left(a_1\right) = \frac{s_{xy}}{\sum_{i=1}^{n} x_i^2} \qquad \{26\}$$

Confidence interval of the slope, a_1: as above.

R^2 and some other statistics: not easily applicable (not recommended).

ANOVA (Analysis of variance)

p 45
$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SSE}} \qquad \{27\}$$

ANOVA (Analysis of variance)

Allows to test the hypothesis H_0: $a_1 = 0$ (null h.) against H_1: $a_1 \neq 0$, i.e.: does *y* really vary with *x* or is it just random fluctuation (chance) ?

Source of variation	Degrees of freedom	Sum of squares	Mean square	Computed F
Regression	**1**	SSR	MST = SSR / **1**	$F = \frac{\text{MST}}{\text{MSE}}$
Error	$\boldsymbol{n-2}$	SSE	$\text{MSE} = s^2 = \frac{\text{SSE}}{\boldsymbol{n-2}}$	
Total	$\boldsymbol{n-1}$	SST		

ANOVA *(Análise de variância)*

Fonte de variação	Graus de liberdade	Soma dos quadr. dos desvios	Médias quadráticas	F calculado
Desvios da regressão *vs.* *y* médio	**1**	SSR	MST = SSR / **1**	$F = \frac{\text{MST}}{\text{MSE}} = \frac{\text{SSR}}{s^2}$
Desvios entre val.s exper. e calculados	$\boldsymbol{n-2}$	SSE	$\text{MSE} = s^2 = \frac{\text{SSE}}{\boldsymbol{n-2}}$	
Total (desvios dos val.s exper. *vs.* *y* médio)	$\boldsymbol{n-1}$	SST		

If *F* is “sufficiently” large (value from software or tables), then H_0 is rejected (*y* does vary with *x*).

“*F*” (Fisher) is the ratio of two chi-square variables, each divided by its degrees of freedom: 1 for SSR, $n-2$ for s^2.

$$\textit{Reject} \text{ if } F > F_{\text{critical}} = F^{\text{inv}}(1 - P;\, n_{\text{numer.}},\, n_{\text{denom.}})$$

References:

– WALPOLE, Ronald E. and Raymond H. MYERS, 1989, “Probability and statistics for engineers and scientists”, 4.th ed., Macmillan, New York, NY (USA) (ISBN 0-02-424210-1), pp 366 ff.

❖