

physical sciences data 39

**statistical methods
in applied chemistry**



physical sciences data 39

statistical methods in applied chemistry

jurand czermiński

University of Gdańsk, Poland

andrzej iwasiewicz

The Cracow Academy of Economics, Poland

zbigniew paszek

The Cracow Academy of Economics, Poland

andrzej sikorski

The Cracow Academy of Economics, Poland



ELSEVIER

Amsterdam—Oxford—New York—Tokyo

PWN-Polish Scientific Publishers

Warszawa

1990

Translated by Elżbieta Krasodomska from the Polish revised and enlarged edition *Metody statystyczne dla chemików*, published by Państwowe Wydawnictwo Naukowe, Warszawa 1986

Distribution of this book is being handled by the following publishers:

For the USA and Canada

ELSEVIER SCIENCE PUBLISHING CO., INC.

655 Avenue of the Americas

New York, N. Y. 10010

For Albania, Bulgaria, Cuba, Czechoslovakia, German Democratic Republic, Hungary, Korean People's Democratic Republic, Mongolia, People's Republic of China, Poland, Romania, the USSR, Vietnam and Yugoslavia

ARS POLONA

Krakowskie Przedmieście 7, 00-068 Warszawa, Poland

For all remaining areas

ELSEVIER SCIENCE PUBLISHERS B. V.

Sara Burgerhartstraat 25

P. O. Box 211, 1000 AE Amsterdam, The Netherlands

ISBN 0-444-98862-9 (vol.39)

ISBN 0-444-41689-7 (series)

Copyright © by PWN—Polish Scientific Publishers—Warszawa 1990

All rights reserved

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise without the prior written permission of the copyright owner

Printed in Poland by D.R.P.

"I can see it but I do not believe it"

GEORG CANTOR about a result

Instead of a foreword

A few words on the need and use of probabilistic thinking

Mathematical statistics is known to examine variation using calculus of probability methods in order to find some regularities which occur among random events. Random events are, in turn, characterized by random variables. It is their realizations that we encounter in experimentation, hence the possibility or even necessity (we do not hesitate to use this emphatic expression) to apply statistical methods.

This is particularly the case where a researcher wishes to assign to his decisions a numerically defined probability of their being accurate (or inaccurate, for that matter). It cannot be done without using calculus of probability.

Although such is the principal use of statistical methods, it is by no means the only one. We shall not proceed to enumerate the benefits which are derived from applying the methods or to discuss them.¹ We shall only mention some of them, i.e. the possibility of detecting the significance of differences between sets of observations, of answering questions concerning the real effect of variables, their correlation or interaction, etc.

We share the opinion of W. Hamilton and J. Hadamard² who state that expressing a thought by means of words makes it stable. In our opinion, the application of probability as a measure of the researcher's risk makes the results of his experiments stable. It can be easily seen that such evaluation of experi-

¹ In our book, we do not shun discussion of a number of advanced problems even though we know that many people regard discussion as a way of disagreement which only makes them more confirmed in their errors.

² J. Hadamard, *An Essay on the Psychology of Invention in the Mathematical Field*, Dover Publications, 1954.

ments is not only a necessary part of a research method but it also facilitates realization of the aims of science which are assumed to be the basic ones, i.e., explaining and predicting.

To detect a regularity in variation is our main task as experimenters. This is the "integrating spark" which enables us to see the subject of chemical research in probabilistic perspective.

Acknowledgements

We owe special gratitude to Professor A. K. Jonscher, University of London, who has continuously supported and encouraged this work. We are indebted to Professor T. M. Krygowski, A. G. Dickson, PhD, who made many valuable suggestions concerning Chapter 5. We are obliged to the staff of the Computer Centre at Gdańsk University for their kindness and assistance in preparing the material which comprises this chapter. It is our pleasure to acknowledge the contribution that Mrs J. Kosmulska, M. Sc. made to this text at its final stage of preparation.

The Authors

List of symbols

Chapters 1 and 2

Ω	—set of elementary events
$A, B, C, \dots$	—random events
$\bar{A}$	—event opposite to A
$\mathfrak{B}$	—Borel's body of events
$P(A)$	—probability of random event A
$P(A B)$	—conditional probability of random event A
A_n	—sequence of random events
$V_n(A)$	—relative frequency of random event A
$X, Y, Z, \dots$	—random variables
$x, y, z, \dots$	—realizations of random variables
$F(x)$	—distribution function of random variable
$f(x)$	—probability density function
$E(x), \mu_x$	—expected value of general population
D^2X, σ_x^2	—general population variance
$D(X), \sigma_x$	—standard deviation of general population
$\varphi(u)$	—probability density of standardized variable
$\Phi(u)$	—distribution function of standardized random variable
$\theta(u)$	—Laplace's function
$H(x)$	—error function
$\Gamma(p)$	—Euler's function
$(X, Y), (Y, T), (X_1, X_2) \dots$	—two-dimensional random variables
$F(x, y)$	—distribution function of two-dimensional random variable
$f(x, y)$	—probability density function of two-dimensional random variable
$f_1(x), f_2(y)$	—marginal probabilities of random variables (X, Y)
$F_1(x), F_2(x)$	—distribution functions of marginal distributions of random variables (X, Y)
ρ	—linear correlation coefficient of probability distribution of two-dimensional random variable (X, Y)
$\bar{x}$	—sample arithmetic mean
$\tilde{x}$	—median
$\hat{x}$	—modal value
r_n	—sample range
s_n^2	—sample variance
s_x	—standard deviation
v_x	—variation coefficient

n	—sample size
N	—size of general population
$\text{cov}(X, Y)$	—covariance of variables X, Y
χ^2	—chi-square random variable
t	—random variable with Student's t distribution
F	—random variable with F -distribution
Q	—value of general population parameter being estimated
α	—level of significance
γ_2	—confidence coefficient
σ_x^2	—mean error of arithmetic mean

Chapter 3

H_0	—null hypothesis
H_1, H_{-1}	—alternative hypotheses
α	—probability of Type I error; level of significance of test
β	—probability of Type II error
D	—statistic of H_0 (general denotation)
D_c	—critical region (general denotation)
u	—statistic of H_0 in $N(0; 1)$ distribution test
t	—statistic of H_0 in Student's test
χ	—statistic of H_0 in chi-square test
F	—statistic of H_0 in F -test
$F_{\max}$	—statistic of H_0 in Hartley's test
v	—statistic of H_0 in Aspin-Welch test
U^2	—statistic of H_0 in Bartlett's test
r_n	—statistic of H_0 in sign test of differences
T_n	—statistic of H_0 in rank sign test of differences
$\bar{x}$	—sample arithmetic mean; also control chart used for controlling the expected value of random variable X with $N(\mu_x, \sigma_x)$ distribution
x_s, x_d	—upper and lower control limits of chart $\bar{x}$
$\bar{x}-s_d$	—two-line control chart used for controlling the expected value and standard deviation of random variable X with $N(\mu_x, \sigma_x)$ distribution
s_s, s_d	—upper and lower control limits for lines s of control chart $\bar{x}-s$
$\bar{x}-r$	—two-line control line for controlling the expected value and range of random variable $X-N(\mu_x, \sigma_x)$ by means of arithmetic mean $\bar{x}$ and sample range (r_n)
$r_n, r_{n,d}$	—upper and lower control limits on line r of control chart $\bar{x}-r$
np	—control chart used for controlling the number of defectives in a sample with determined size
z_s, z_d	—upper and lower control limits of chart np
p	—control chart for controlling sample fraction defectives
w_s, w_d	—upper and lower control limits of chart p
X, Y, Z, U	—random variables
U	—standardized random variable with $N(0; 1)$ distribution

Chapter 4

η_i	—effect of the i th object on the value of observation x_{ij} or x_{ipj}
$\hat{\eta}_i$	—estimate (e.g. $\hat{\eta}_i$ estimate of effect η_i)
e_{ij}, e_{ipj}	—random error (deviation)
k	—number of columns
r	—number of rows
n_i	—number of observations in the i th object
n	—number of observations when $n_i = n$
S	—total sum of squares
S_{ij}	—sum of squares within objects
S_i	—sum of squares among objects (columns)
σ_η^2	—denotes contribution of systematic variation of means to variation among objects
μ_i	—mean (real) value of the i th object
μ	—total real value
$\bar{x}_i$	—arithmetic mean of the i th object (column)
$\bar{x}$	—total arithmetic mean
σ_i	—variance of the i th general population
s_i^2	—estimate of variance σ_i^2
F_0	—computed value of F -test
F_θ	—tabulated value of F (boundary value)
$\bar{s}_{\bar{x}_i}^2$	—estimate of mean variance
$\bar{x}_p$	—group mean
d_L	—difference $\bar{x}_i - \bar{x}_p$
μ_0	—determined value
θ_p	—effect of the p th object on the value of observation x_{ipj}
λ_{ip}	—measure of interaction of factors A and B , λ_{ip} effect
$\bar{x}_p$	—mean of the p th object (row); group mean in Tukey's method
$\bar{x}_{ip}$	—mean of subclass a_ib_p
a_i	—levels of factor A
b_p	—levels of factor B
σ_v^2	—represents contribution of systematic variation of means μ_p to variation among objects (rows)
σ_λ	—interaction variance
S_p	—sum of squares among objects (rows)
S_{ip}	—sum of squares for interaction
S_{ipj}	—sum of squares for subclasses a_ib_p
w_i	—weight of the i th object
$\bar{x}_w$	—weighted total mean
f_1, f_2	—degrees of freedom

Chapter 5

B	—matrix constructed from table of values of explanatory variable expanded by vector composed of ones only
C	—matrix of system of normal equations
D	—dielectric permeability
E	—extinction
E_s	—steric component of free energy excess
$E(X)$	—expected value of random variable X
F_D	—free energy of ion transmission between environments with different permittivity
$F(X, \beta)$	—function defining exact model of experiment
H_0	—null hypothesis
H	—Hessian matrix
J	—Jacobian matrix
K	—dissociation constant
K_a	—association constant
L	—likelihood function
L	—matrix of sample central moments of II order (estimate of matrix)
N	—number of measuring points
P	—correlation matrix
Q	—non-negative measure which is a superposition of several r types errors
Q'_i	—first derivative of function Q in relation to b_i
Q''_{ij}	—second derivative of function Q in relation to b_i and b_j
R	—matrix of estimates of sample correlation coefficients
R	—positive definite matrix generating direction vector in gradient methods
V	—covariance matrix
b	—vector of parameter estimates β_i
$f(x), g(x), h(x)$	—functions of one real variable
$f_+(x)$	—upper confidence curve
$f_-(x)$	—lower confidence curve
e	—vector of residuals
k	—reaction-velocity constant
h	—hyperconjugation constant of free energy excess
h_{ij}	—elements of Hessian matrix
l	—vector of right sides of system of normal equations
l_{ij}	—elements of L matrix
n	—number of parameters under estimation
Δn	—difference between numbers of hydrogen atoms involved in hyperconjugation
n_i	—mean ligand number
q_i	—gradient calculated at the i th iterative step
r	—sample correlation coefficient
t	—Student's test statistic
w	—vector of statistical weights
v_i	—direction vector in gradient methods at the i th iterative step
x	—vector of explanatory variables of a model

$\bar{x}$	— sample mean value (arithmetic mean)
v	—vector of explanatory variables
g_{ik}	—elements of matrix Λ^{-1}
α	—level of significance
β	—vector of model parameters under estimation
β_i	—complex constants in the Rosottis equation
δ	—proportionality coefficient of steric component
$\delta(X, Y)$	—measure of dependence of random variables X and Y
ε	—error vector
λ	—multiplier controlling convergence of Marquardt's method
λ	—equivalent conductivity
λ_{ij}	—elements of matrix Λ
Λ	—matrix of II order central moments
$\varphi_i(x)$	—orthogonal polynomial of the i th degree
Ψ	—component of free energy excess due to π -electron substructure coupling with reaction centre
θ_{yx}	—correlation ratio (formula (5.3))
Σ	—diagonal matrix composed of standard deviations
ξ	—random variable whose realizations are errors e_i
$\dim(x)$	—dimension of space, number of components of vector x
$\text{sign}(x)$	—signum function (sign)
$[L]$	—equilibrium concentration of ligand in the Rosotti's equation

There are no new symbols in Chapters 6 and 7.

Contents

Instead of a foreword	XI
Acknowledgements	XIII
List of symbols	XIV
Chapter 1. Random variables	1
1.1. Properties and variables	1
1.2. Random events and elementary event space	4
1.3. Probability	4
1.4. Basic properties of probability	6
1.5. One-dimensional random variable	9
1.5.1. Probability function and probability density function	11
1.5.2. Distribution function	16
1.6. Measures of position and dispersion of one-dimensional random variable	18
1.7. Moments of one-dimensional random variable	20
1.8. Selected distributions of one-dimensional random variables	22
1.8.1. Distributions of discrete random variables	22
1.8.1.1. Two-point distribution	22
1.8.1.2. Binomial distribution	22
1.8.1.3. Pascal's distribution	26
1.8.1.4. The Poisson distribution	27
1.8.2. Continuous random variable distributions	32
1.8.2.1. Uniform distribution	32
1.8.2.2. Normal distribution	33
1.8.2.3. Truncated normal distribution	37
1.8.2.4. The logarithmic-normal distribution	40
1.8.2.5. Exponential distribution	42
1.9. Multi-dimensional random variable	44
1.9.1. Probability function, probability density function and distribution function of a two-dimensional random variable	45
1.9.2. Marginal and conditional distributions	47

1.9.3. Moments of two-dimensional random variable	49
1.9.4. Linear correlation coefficient	50
1.9.5. Two-dimensional normal distribution	51
Chapter 2. Estimation of parameters of random variable distribution	56
2.1. Population and sample	56
2.2. Empirical distributions	58
2.3. Point estimation	62
2.3.1. Estimation methods	62
2.3.2. Criteria for estimators evaluation	66
2.3.3. Estimation of expected value	68
2.3.3.1. Sample arithmetic mean	68
2.3.3.2. Order statistics	75
2.3.4. Estimation of variance and standard deviation	80
2.3.4.1. Sampling variance and standard deviation	80
2.3.4.2. Estimation of standard deviation from sample range	88
2.3.5. Variation coefficient	92
2.4. Probability distribution of selected sample characteristics	94
2.4.1. Distribution of sample mean	94
2.4.2. Distribution of sample variance and standard deviation	98
2.4.3. Distribution of sample median	102
2.5. Interval estimation	102
2.5.1. General principles of determination of confidence intervals	102
2.5.2. Confidence interval for the expected value	103
2.5.3. Confidence interval for variance and standard deviation	105
2.5.4. Confidence interval for median	107
Chapter 3. Testing statistical hypotheses	109
3.1. Introduction	109
3.2. Verification of parametric hypotheses	115
3.2.1. Verification of the significance of the difference between the expected random variable μ_x and the determined value μ_0	115
3.2.1.1. The u test	116
3.2.1.2. Control chart $\bar{x}$	118
3.2.1.3. The t test	124
3.2.2. Verification of the significance of the difference between the expected values of two random variables	126

3.2.2.1. The u test	127
3.2.2.2. The t test	127
3.2.2.3. The Aspin-Welch test	131
3.2.3. Verification of the significance of the difference between the variance of the random variable σ_x^2 and the determined value σ_0^2	133
3.2.3.1. The chi-square test	134
3.2.3.2. The control chart $\bar{x}-s$	134
3.2.3.3. The control chart $\bar{x}-r$	137
3.2.4. Testing of the significance of the difference between the variances of two random variables	138
3.2.5. Testing homogeneity of variance of several ($k \geq 2$) random variables	141
3.2.5.1. Hartley's F_{max} test	141
3.2.5.2. Bartlett's test	143
3.2.6. Testing hypotheses about fractions	144
3.2.6.1. The u test	145
3.2.6.2. The control chart np and the control chart p	147
3.3. Non-parametric hypothesis testing	149
3.3.1. An analysis of contingency tables by the chi-square test	149
3.3.2. A comparison of the empirical and theoretical distributions by means of the chi-square test	153
3.3.3. Graphical method for testing of normal distribution	161
3.3.4. Testing hypotheses concerning the goodness of fit of two empi- rical distributions	165
3.3.4.1. The sign test for differences	165
3.3.4.2. The ranked signs test for differences	166
3.3.5. Testing the hypothesis concerning randomness of a sample	168
3.4. Sequential tests	172
3.4.1. Basic concepts and definitions	172
3.4.2. Testing the hypotheses concerning the expected value μ_x of the normal random variable	174
3.4.2.1. The sequential probability ratio test (SPRT)	174
3.4.2.2. The cumulative sum control chart (CSCC)	177
3.4.3. Testing hypotheses concerning the standard deviation σ_x of the normal random variable	179
3.4.3.1. The sequential probability ratio test (SPRT)	179
3.4.3.2. The cumulative sum control chart (CSCC)	181
3.4.4. Testing hypotheses on fractions	181

3.4.4.1. The sequential probability ratio test (SPRT)	182
3.4.4.2. The cumulative sum control chart (CSCC)	184
Chapter 4. Analysis of variance	186
4.1. Theoretical model of analysis of variance for one-way classification	187
4.2. Computation of sums of squares	194
4.3. Models and assumptions of the analysis of variance	198
4.4. Grouping of object means. The $\mu = \mu_0$ hypothesis testing	201
4.5. Approximate test in the case of non-homogeneity of variance	208
4.6. Analysis of variance for two-way classification	209
Chapter 5. Correlation and regression	223
5.1. Correlation	223
5.1.1. Model and its verification	223
5.1.2. Measures of dependence	229
5.1.3. Correlation coefficient assessment	232
5.1.4. Study of the significance of correlation	233
5.2. Regression	234
5.2.1. Variable, function, regression	234
5.2.2. The least squares method	235
5.2.2.1. Approximation and its error	235
5.2.2.2. The minimum of the multivariate function	238
5.2.3. Several linear regression models by the least squares method	240
5.2.3.1. Defining of approximation error	240
5.2.3.2. Regression of y with regard to x (y/x)	242
5.2.3.3. Orthogonal regression ($y \perp x$)	244
5.2.3.4. Regression of x with regard to y (x/y)	246
5.2.3.5. Weighted regression	246
5.2.3.6. Summary	249
5.2.4. Residual and its distribution	253
5.2.4.1. The adequacy of the linear model	253
5.2.4.2. Residual variance	257
5.2.4.3. Stochastic properties of solution vector. Mean standard deviation of intercept and regression coefficient	259
5.2.4.4. The propagation of error	263
5.2.4.5. Standard deviation x , given y	263
5.2.5. Testing significance of regression coefficient (slope) and intercept	264

5.2.6. Confidence intervals determination in regression analysis . . .	265
5.2.6.1. Confidence intervals of regression coefficient and intercept . . .	265
5.2.6.2. Confidence interval for the expected value $E(Y)$, given x . . .	266
5.2.7. Tolerance region for values off the regression line (outliers) . . .	267
5.2.8. Multiple linear regression	279
5.2.8.1. Introduction	279
5.2.8.2. Regression coefficients	281
5.2.8.3. Residual variance	284
5.2.8.4. Simple correlation coefficients	285
5.2.8.5. Partial correlation coefficients	286
5.2.8.6. Multiple correlation coefficient	287
5.2.9. Non-linear regression	296
5.2.9.1. Linearized non-linear regression	296
5.2.9.2. Curvilinear (polynomial) regression	297
5.2.10. Non-linearized non-linear regression	307
5.2.10.1. Iterative schemes	307
5.2.10.2. Covariance matrix of vector b	317
5.2.10.3. Transformation of non-linear problems. Job's plot of the second kind	317
 Chapter 6. Methodical guidelines	 324
6.1. Criteria of choice and evaluation of research methods	324
6.1.1. Sensitivity threshold and sensitivity of a method	324
6.1.2. Precision and accuracy of a method	330
6.2. Operation on approximate numbers. Calculation of errors	337
6.2.1. Preliminary definitions	337
6.2.2. Errors of the elementary operations. Basic theorems	342
6.2.3. Error of arbitrary function	346
6.2.4. Calculus of errors and statistical methods	351
6.2.5. Numerical stability of algorithms. Numerically ill-conditioned problems	354
6.3. Formation of sample population	361
6.3.1. Simple random sample and other types of samples	362
6.3.2. Sampling	364
6.3.3. Sample size	367
6.4. Registration and analysis of results	371
6.4.1. Experiment record	371

6.4.2. Construction of the interval series	372
6.4.3. Graphical presentation of results. Histogram	378
Chapter 7. Examples of a complete analysis of experiment results.	381
7.1. A technological example	381
7.1.1. A general outline of the problem	381
7.1.2. Experiment results and hypotheses	383
7.1.3. An analysis of the results	385
7.1.3.1. Testing randomness: distributions	385
7.1.3.2. Hypotheses testing	388
7.1.4. Discussion of results	398
7.2. An analytical example.	401
7.2.1. A general outline of the problem and results of the experiment	401
7.2.2. An analysis of the results	404
7.2.2.1. Testing randomness	404
7.2.2.2. Testing normality of distributions	406
7.2.2.3. Testing homogeneity of variance	409
7.2.3. Discussion of results	410
Appendix A. Statistical tables	414
I. The normal distribution function	414
II. Student's t distribution function	417
III. Student's t distribution	419
IV. The F distribution	420
V. Upper bounds for the ratio $F_{\max} = s_{\max}^2/s_{\min}^2$	425
VI. χ^2 distribution function	426
VII. The χ^2 distribution	428
VIII. The series distribution	429
IX. The signs number distribution	432
X. The distribution of sum of rows	433
XI. The Poisson distribution function	434
XII. The v distribution	444
XIII. Parameters of the range distribution	446
XIV. Random numbers	447
Appendix B. Computer programmes and procedures	448
Bibliography	485
Index	487