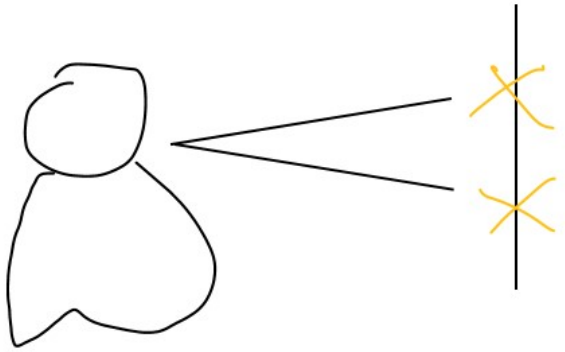
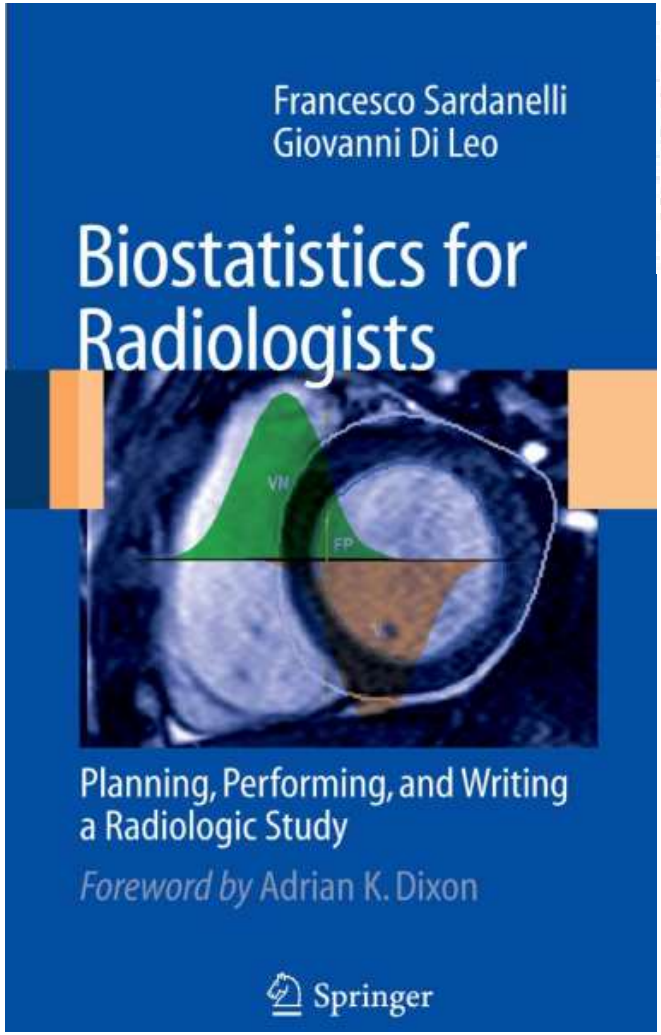


Delineamento intraparticipantes

Testes estatísticos de concordância

Prof. José Siqueira
Prof. Paulo Silveira
2020

1



Originally published as:
Biostatistica in Radiologia
Progettare, realizzare e scrivere un lavoro scientifico radiologico
Francesco Sardanelli, Giovanni Di Leo
© Springer-Verlag Italia 2008
All rights reserved

2

Delineamento intraparticipantes

Teste de concordância entre dois métodos de avaliação

Tabela bidimensional

- Duas variáveis nominais dicotômicas
  - Teste qui-quadrado de mudança de McNemar: `coin::mh_test, mcnemar.test`
  - Teste kappa de Cohen: `epiR::epi.kappa`
  - Intervalo de confiança para diferença de proporção em dados pareados: `diffdepprop::diffpci`
- Duas variáveis ordinais
  - Teste de homogeneidade marginal: `coin::mh_test`
  - Teste kappa de Cohen: `psych::cohen.kappa`

Tabela tridimensional

- Duas variáveis ordinais com variável de estratificação
  - Teste de homogeneidade marginal: `coin::mh_test`

Duas variáveis intervalares com erros de mensuração

- Regressão de Deming: `mcr::mcreg`

Teste qui-quadrado de mudança de McNemar
Siegel & Castellan (1988), p. 75-7

- O teste de McNemar para a significância das mudanças é particularmente aplicável aos projetos "antes-depois“ de intervenção, nos quais cada sujeito é usado como seu próprio controle e em que as medidas são feitas em escala nominal ou ordinária.
- As condições podem ser usadas para testar a eficácia de um tratamento específico (reuniões, editoriais de jornais, discursos, visitas pessoais etc.) sobre as preferências dos eleitores sobre candidatos a cargos públicos ou para testar o efeito de migração do campo para a cidade pela afiliação política de pessoas.
- Observe que nesses estudos as pessoas podem servir como seu próprio controle e que a escala nominal (ou categorização) é usada adequadamente para avaliar a mudança "antes-depois".

Teste qui-quadrado de mudança de McNemar
Siegel & Castellan (1988), p. 75-7

- Assim, b + c é o total de pessoas cujas respostas foram alteradas.
- A hipótese nula é que o número de mudanças em cada direção é o mesmo na população.
- Portanto, dos indivíduos com b + c que mudaram, esperamos que (b + c) / 2 indivíduos mudem de + para - e (b + c) / 2 pessoas mudem de - para +.
- Em outras palavras, quando H_0 é verdadeira, a frequência esperada em cada uma das duas caselas é (b + c) / 2.

		Pós	
		+	-
Pré	+	a	b
	-	c	d

Teste qui-quadrado de mudança de McNemar
Siegel & Castellan (1988), p. 75-7

- No teste de McNemar para a significância das mudanças, estamos interessados apenas nas caselas nas quais as mudanças podem ocorrer.
- Assim, se b é o número de casos observados cujas respostas mudaram de + para -, c é o número observado de casos que mudaram de - para + e (b + c) / 2 é o número esperado de casos nas caselas b e c.

- $X^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{O_i} = \frac{\left(b - \frac{b+c}{2}\right)^2}{\frac{b+c}{2}} + \frac{\left(c - \frac{b+c}{2}\right)^2}{\frac{b+c}{2}} = \frac{(b-c)^2}{b+c}$
- A distribuição amostral de X^2 quando H_1 é verdadeira, é distribuída assintoticamente como qui-quadrado com graus de liberdade iguais a um.
- H_0: P(+ → –) = P(– → +) vs. H_1: P(+ → –) ≠ P(– → +)

6

Teste de concordância para tabela 2x2
Teste_McNemar_&_MHOrdinal_&_kappa.R

```
library(coin)
# Siegel & Castellan (1988), p. 77
Tabela <- ("
 TVDebate Carter Reagan
  Carter  28     13
  Reagan  7      27
")
print(TC <- as.matrix(read.table(textConnection(Tabela),
                                 header=TRUE, row.names=1)))
print(mcnemar.test(TC,correct=FALSE)) # classico
# McNemar's Chi-squared test
# data:  TC
# McNemar's chi-squared = 1.8, df = 1, p-value = 0.1797
print(coin::mh_test(as.table(TC), distribution = "exact")) # robusto
# Exact Marginal Homogeneity Test
# data:  response by
# conditions (Var1, Var2)
# stratified by block
# chi-squared = 1.8, p-value = 0.2632
```

Análise de concordância em estudos clínicos e experimentais

Agreement analysis in clinical and experimental trials

Hélio Amante Miot[1]

J Vasc Bras. 2016 Abr.-Jun.; 15(2):89-92

8

Medidas de concordância entre dois métodos de avaliação

- “Análise de concordância se refere à capacidade de aferir resultados idênticos (mesma unidade de medida), aplicados ao mesmo sujeito/fenômeno, quer por instrumentos diferentes, pelo mesmo instrumento em tempos diferentes, por avaliadores diferentes, ou por alguma combinação dessas situações.”
- “Exemplos triviais são calibragem de instrumentos, fidedignidade de escala/medida, avaliação de equivalência entre ferramentas de mensuração, julgamento em provas de habilidades, avaliação de repetitividade ou reprodutibilidade, e análise diagnóstica (concordância interpessoal e intrapessoal) e psicométrica (estabilidade temporal).”

Kappa de Cohen

- “A situação mais simples ocorre quando a variável de interesse é dicotômica (por exemplo, doente × saudável, indicação cirúrgica × clínica, aprovado × reprovado), e a estimativa ocorre por dois avaliadores ou dois instrumentos; nesse caso, classicamente se emprega a estatística kappa de Cohen.
- O valor, o intervalo de confiança e a significância estatística de kappa devem ser interpretados como a dimensão da concordância que ultrapassa a coincidência de avaliações que ocorrem ao acaso.”

10

Teste de concordância para tabela 2x2
TC2x2_intrapartic.R

```
library(epiR)
library(rcompanion)
library(diffdepprop)
sink("TC2x2_intrapartic.txt")
# Foram analisadas 315 amostras usando os métodos Bell e
# Kato-Katz para detecção ovos de Schistosoma mansoni nas fezes.
# Sleigh et al. (1982) Transactions of the Royal Society of
# Tropical Medicine and Hygiene 76: 403-6.
Tabela <- ("
  BellxKK P   N
  P       184 54
  N       14  63
")
print(TC <- as.matrix(read.table(textConnection(Tabela),
                                 header=TRUE, row.names=1)))
```

20

Teste de concordância para tabela 2x2
TC2x2_intrapartic.R

```
cat("\nTeste de concordancia: delineamento intraparticipantes\n")
cat("\nSoftware de referencia: IBM SPSS Statistics 24\n")
cat("Teste kappa de Cohen\n")
cat("H0: kappa = 0 vs H1: kappa > 0\n",sep="")
res <- epiR::epi.kappa(TC,alternative="greater")
print(res$kappa)
print(res$z)
z <- as.numeric(res$z[1])
cat("z^2 = ", z^2,"\n",sep="")
out <- chisq.test(TC, correct=FALSE)
a <- TC[1,1]; b <- TC[1,2]; c <- TC[2,1]; d <- TC[2,2]; n <- sum(TC)
pae <- out$expected[1,1]/n
pbe <- out$expected[1,2]/n
pce <- out$expected[2,1]/n
pde <- out$expected[2,2]/n
kmax <- (min(pae+pbe,pae+pce)+min(pce+pde,pbe+pde)-pae-pde)/(1-pae-pde)
cat("kappa de Cohen maximo = ",kmax,"\n",sep="")
k <- as.numeric(res$kappa[1])
if(k>0){
  k <- k/kmax
  cat("kappa de Cohen corrigido pelo kappa maximo = ",k,"\n",sep="")
}
if(k<=0){
  cat("\nkappa de Cohen não-positivo\n")
  cat("kappa de Cohen nao corrigido pelo kappa maximo = ",k,"\n",sep="")
}
```

Teste de concordância para tabela 2x2
TC2x2_intrapartic.R

```
# Landis & Koch (1997)
if (0 <= k & k < 0.1) {gkl <- "Poor"}
if (0.1 <= k & k < 0.2) {gkl <- "Slight"}
if (0.2 <= k & k < 0.4) {gkl <- "Fair"}
if (0.4 <= k & k < 0.6) {gkl <- "Moderate"}
if (0.6 <= k & k < 0.8) {gkl <- "Substancial"}
if (0.8 <= k & k <= 1.0) {gkl <- "Almost perfect"}
cat("Grau de concordancia entre dois metodos/avaliadores = ",gkl,"\n",sep="")
```

Teste de concordância para tabela 2x2
TC2x2_intrapartic.R

```
# HRIPCSAK, G & ROTHSCHILD, AS (2005) Agreement, the F-measure, and reliabilty
# in information retrieval. J Am Med Inform Assoc 12:296-8.
F <- 2*a/(2*a+b+c)
cat("F-measure = ",F,"\n",sep="")
V <- rcompanion::cramerV(TC,ci=TRUE,R=1e4)
cat("fi = V de Cramer = ",V$Cramer.V,"\n",sep="")
if(!is.na(V$lower.ci) & !is.na(V$lower.ci)){
if (0 <= V$lower.ci & V$lower.ci < 0.1) {gVl <- "minimo"}
if (0.1 <= V$lower.ci & V$lower.ci < 0.3) {gVl <- "pequeno"}
if (0.3 <= V$lower.ci & V$lower.ci < 0.5) {gVl <- "intermediario"}
if (0.5 <= V$lower.ci & V$lower.ci <= 1.0) {gVl <- "grande"}
if (0 <= V$upper.ci & V$upper.ci < 0.1) {gVu <- "minimo"}
if (0.1 <= V$upper.ci & V$upper.ci < 0.3) {gVu <- "pequeno"}
if (0.3 <= V$upper.ci & V$upper.ci < 0.5) {gVu <- "intermediario"}
if (0.5 <= V$upper.ci & V$upper.ci <= 1.0) {gVu <- "grande"}
cat("IC95(V de Cramer) = [", V$lower.ci, ", ", V$upper.ci,"]\n",sep="")
cat("Grau de dependencia entre as duas variaveis nominais =
    [", gVl, ", ", gVu, "]\n",sep="")
}
```

Teste de concordância para tabela 2x2
TC2x2_intrapartic.R

```
alfa <- .05
# Bonnet, DG & Price, RM (2012) Adjusted Wald confidence interval for a
# difference of binomial proportion based ond paired data
# Journal of Edicational and Behavioral Statistics 37(4): 479-88.
cat("\nH0: pi_+1 - pi_1+ = 0 vs. H0: pi_+1 - pi_1+ != 0\n")
diffdepprop::diffpci(a=a, b=b, c=c, d=d, n=n, alpha=alfa)
#        Method Estimator lower limit upper limit
# 1        Wald 0.1269841  0.07762879   0.1763395
# BellxKK: A probabilidade média do teste Bell ser positivo é 12.7% maior
# que a probabilidade do teste KK ser positivo.
sink()
```

Teste de concordância para tabela 2x2
TC2x2_intrapartic.R

```
    P  N
P 184 54
N  14 63

Teste de concordancia: delineamento intraparticipantes

Software de referencia: IBM SPSS Statistics 24
Teste kappa de Cohen
H0: kappa = 0 vs H1: kappa > 0
       est         se     lower     upper
1 0.502924 0.05388122 0.3973187 0.6085292
  test.statistic      p.value
1       9.333938 5.100442e-21
z^2 = 87.1224
kappa de Cohen maximo = 0.7076023
kappa de Cohen corrigido pelo kappa maximo = 0.7107438
Grau de concordancia entre dois metodos/avaliadores = Substancial

F-measure = 0.8440367
fi = V de Cramer = 0.5259
IC95(V de Cramer) = [0.4247, 0.619]
Grau de dependencia entre as duas variaveis nominais = [intermediario, grande]
```

Teste de concordância para tabela 2x2
TC2x2_intrapartic.R

```
    P  N
P 184 54
N  14 63

Teste da nulidade da diferenca das probabilidades de mudanca
       Method Estimator lower limit upper limit
1        Wald 0.1269841  0.07762879   0.1763395
2     Wald.cc 0.1269841  0.07445419   0.1795141
3     Agresti 0.1261830  0.07673844   0.1756275
4       Tango 0.1269841  0.07834078   0.1783027
5  Exact.cond 0.1269841  0.07717984   0.1651947
6  Exact.midp 0.1269841  0.08025921   0.1630940
7      Uncond 0.1269841  0.07850000   0.1775000
8      Wilson 0.1269841  0.07742303   0.1759014
9   Wilson.cc 0.1269841  0.07583324   0.1774363
10 Wilson.phi 0.1269841  0.07702527   0.1762761
11      np.nv 0.1269841  0.07755027   0.1764180
12       np.t 0.1269841  0.07745351   0.1765147
```

Teste de concordância
Teste_McNemar_&_MHOrdinal_&_kappa.R

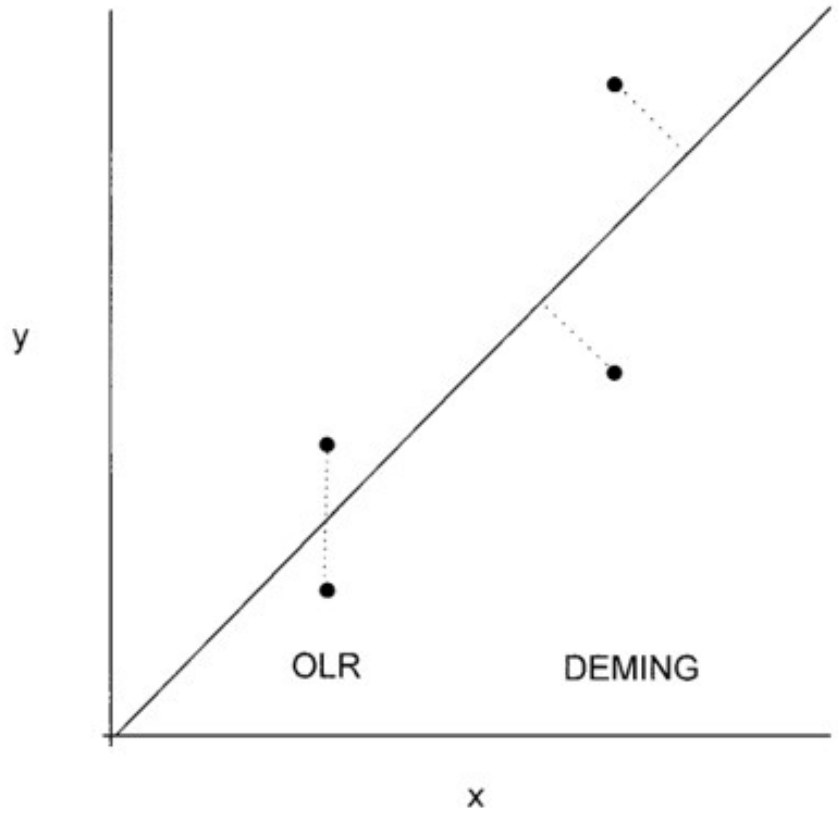
```
# Agresti (1990), apud Mehta, C. R. &
# Patel, N. R. (1996) SPSS Exact Tests 7 for  Windows. IL: SPSS, p. 72.
# O objetivo é analisar a concordância de diagnóstico entre 2 patologistas
# que  classificaram conforme a sereridade de uma determinada
# lesão uterina de 118 slides de diferentes mulheres.
# N=Negativo, HEA=Hiperplasia escamosa atipica, CIS=carcinoma in situ
# CE=Carcinoma escamoso, CI=Carcinoma invasivo
Tabela <- ("
 Patologistas N  HEA CIS CE CI
          N   22 2   2   0  0
          HEA 5  7   14  0  0
          CIS 0  2   36  0  0
          CE  0  1   14  7  0
          CI  0  0   3   0  3
")
print(TC <- as.matrix(read.table(textConnection(Tabela),
                                 header=TRUE, row.names=1)))
cat("\nTeste de concordancia: delineamento intraparticipantes\n")
```

Teste de concordância
Teste_McNemar_&_MHOrdinal_&_kappa.R

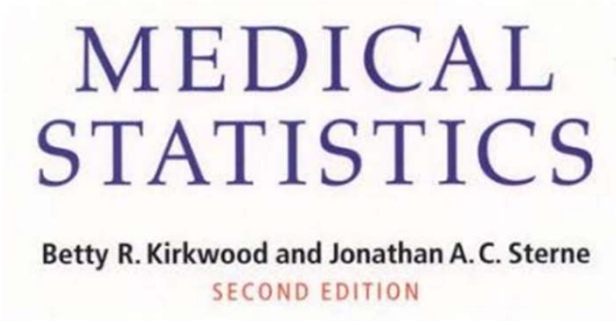
```
# Exact Marginal Homogeneity Test for Ordered Data
print(coin::mh_test(as.table(TC), distribution = "exact",
                    scores = list(response = 1:nrow(TC)))) # robusto
# data:  response (ordered) by
# conditions (Var1, Var2)
# stratified by block
# Z = 1.1523, p-value = 0.3073
# alternative hypothesis: two.sided
cat("Teste kappa de Cohen\n")
cat("H0: kappa = 0 vs H1: kappa != 0\n",sep="")
print(res <- psych::cohen.kappa(TC))
# Cohen Kappa and Weighted Kappa correlation coefficients and
# confidence boundaries
#                  lower estimate upper
# unweighted kappa  0.39     0.50  0.61
# weighted kappa    0.78     0.78  0.78

k <- res$kappa
if (0 <= k & k < 0.1) {gkl <- "Poor"}
if (0.1 <= k & k < 0.2) {gkl <- "Slight"}
if (0.2 <= k & k < 0.4) {gkl <- "Fair"}
if (0.4 <= k & k < 0.6) {gkl <- "Moderate"}
if (0.6 <= k & k < 0.8) {gkl <- "Substancial"}
if (0.8 <= k & k <= 1.0) {gkl <- "Almost perfect"}
cat("Grau de concordancia entre dois metodos/avaliadores = ",gkl,"\n",sep="")
```

Regressão de Deming



29



36.4 NUMERICAL VARIABLES: METHOD COMPARISON STUDIES

We will now consider analyses appropriate to **method comparison studies**, in which two different methods of measuring the same underlying (true) value are compared. For example, lung function might be measured using a spirometer, which is expensive but relatively accurate, or with a peak flow meter, which is cheap (and can therefore be used by asthma patients at home) but relatively inaccurate. The appropriate analysis of such studies was described, in an influential paper, by Bland and Altman (1986).

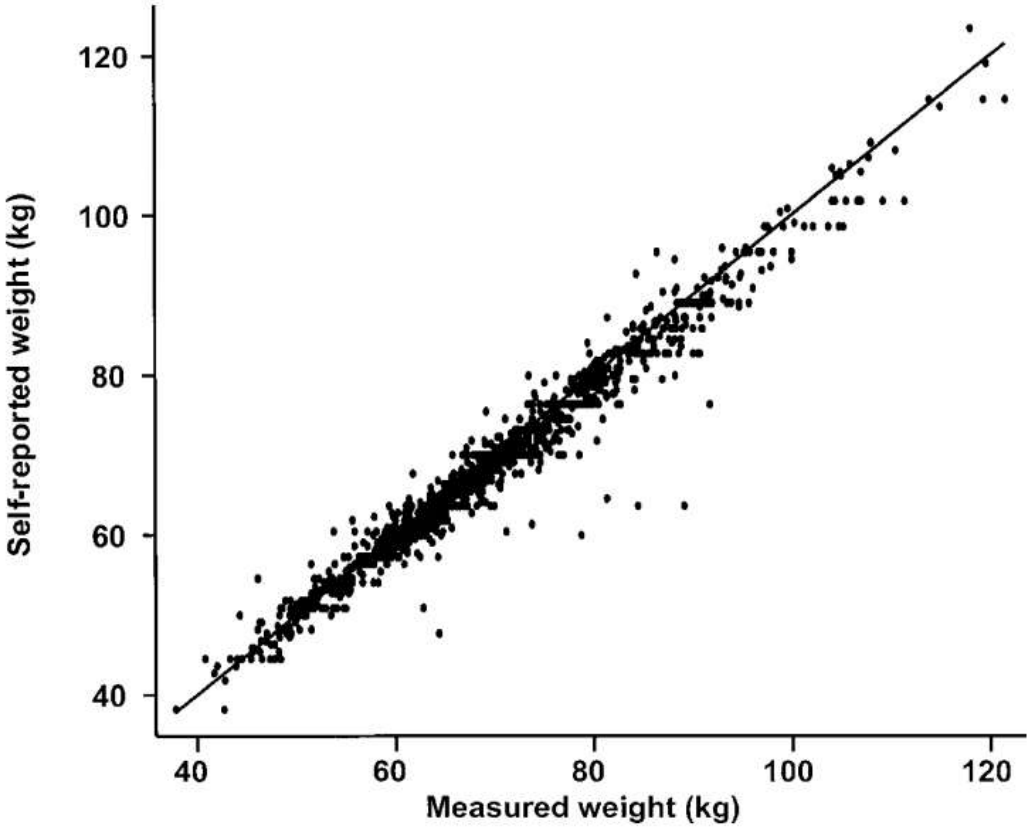


Fig. 36.3 Scatter plot of self-reported versus measured weight (kg) in 1236 women who participated in the British Regional Women's Heart Study. The solid line is the *line of equality*. Data displays and analyses by kind permission of Dr Debbie Lawlor and Professor Shah Ebrahim.

31

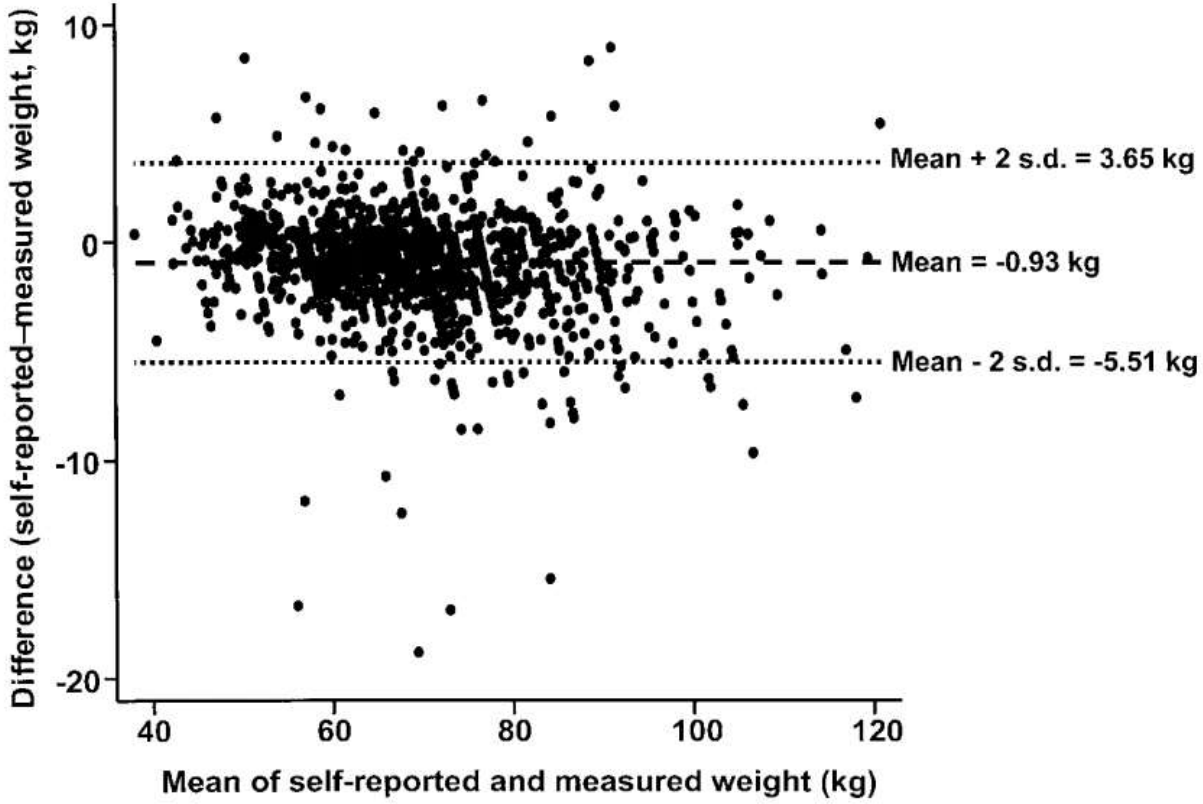


Fig. 36.4 Scatter plot (Bland–Altman plot) of self-reported minus measured weight (vertical axis) against mean of self-reported and measured weight (horizontal axis) in 1236 women who participated in the British Regional Women's Heart Study. The dashed horizontal line corresponds to the mean difference (−0.93 kg) while the dotted horizontal lines correspond to the 95% limits of agreement.

32

J. chron. Dis. Vol. 15, pp. 969–977. Pergamon Press Ltd. Printed in Great Britain

1962

A NOTE ON THE ANALYSIS OF REPEATED MEASUREMENTS OF THE SAME SUBJECTS

P. D. OLDHAM, M.A.(Oxon.), F.S.S.

Pneumoconiosis Research Unit, Llandough Hospital, Penarth, Glamorgan

33

SUMMARY

When repeated measurements have been made on the same subject, care must be taken in choosing the indices which are to be used to summarize the measurements, or erroneous conclusions may be reached. For example, when two measurements have been made, it is natural to think of analysing the first, and the difference between second and first. These indices are not independent, however, and to treat them as independent is to introduce spurious correlations into the results. The best solution is to use as indices the functions of the repeated measurements given by orthogonal polynomials. In the case of two measurements these are the mean of the two and their difference.

Misuses of correlation and regression analyses in orthodontic research: The problem of mathematical coupling

Yu-Kang Tu,[a] Zararna L. Nelson-Moon,[b] and Mark S. Gilthorpe[c]
Leeds, United Kingdom

Introduction: The aim of this article was to encourage good practice in the statistical analyses of orthodontic research data. Our objective was to highlight the statistical problems caused by mathematical coupling (MC) in correlation and regression analyses. These statistical problems are among the most common pitfalls in orthodontic research when exploring associations among clinical variables. This article will show why these problems arise and how they can be avoided and overcome. **Methods:** Four orthodontic journals were electronically and manually searched for articles that used correlation and regression analyses. Studies that seemed to suffer from MC in their statistical analyses were identified and carefully examined. **Results:** Several examples from our search illustrate that MC in correlation and regression analyses can potentially cause misleading results. More appropriate statistical methods are available and should be used to eliminate confusing results and improve any subsequent interpretations. Because many clinical and radiographic variables used in orthodontic research are correlated due to direct or indirect MC, interpretation of studies in the literature needs to be cautious. **Conclusions:** Correlation and regression analyses are useful tools in orthodontic research when their assumptions and limitations are recognized. However, greater care is required in formulating research questions and experimental designs. It is prudent to seek statistical advice when orthodontic research involves complex data analyses. (Am J Orthod Dentofacial Orthop 2006;130:62-8)

35

Oldham's method

In 1962, Oldham noticed the problem of MC in correlation/regression when seeking whether there was a relationship between treatment effect and baseline disease severity in patients with hypertension.[30] He warned against the common practice of regressing (or correlating) change $(x - y)$ with baseline (x), as the null hypothesis—that correlation coefficient or regression slope is zero—is no longer valid. He suggested that the change $(x - y)$ should be regressed on the arithmetic mean of the pretreatment and posttreatment values. The Pearson correlation between change and the mean is: [30]

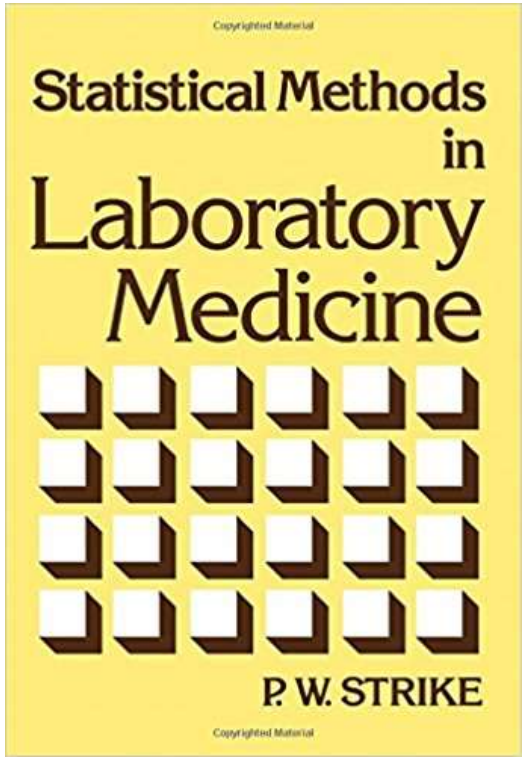
$$Corr[x - y,\ (x + y)/2] = \frac{s_x^2 - s_y^2}{\sqrt{(s_x^2 + s_y^2)^2 - 4r_{xy}^2 s_x^2 s_y^2}},$$

where s_x^2, s_y^2, and r_{xy} are as defined previously.

The rationale behind Oldham's method is that x and y are 2 repeated measurements made on the same subject on successive occasions, so that their variances should be almost identical if there were no intervention and no other time-related biologic variation between measurement occasions.[30]

Delineamento

- A e B são os métodos de referência e o outro, respectivamente, que produzem observações da mesma variável quantitativa com erros de mensuração realizadas na mesma unidade experimental por meio de dois métodos de mensuração.

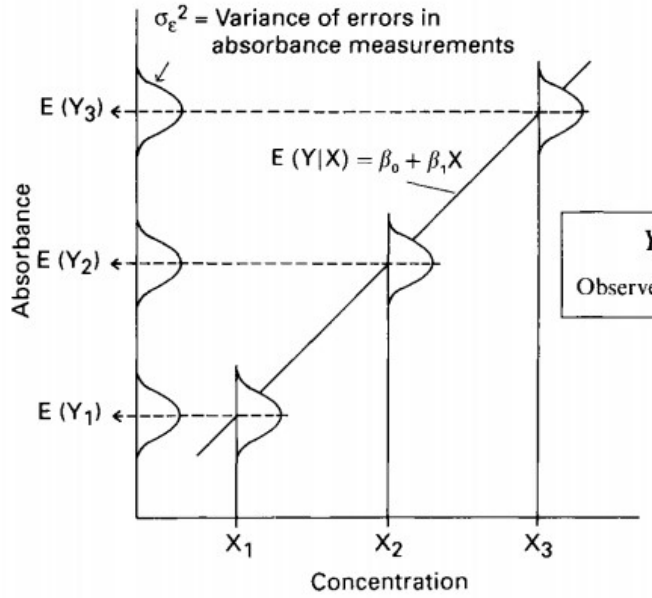


Statistical Methods in Laboratory Medicine

Butterworth-Heinemann Ltd
Linacre House, Jordan Hill, Oxford OX2 8DP

P. W. Strike MPhil, PG DipBiom, FSS
Clinical Statistician/Principal Scientific Officer
Royal Air Force Medical Branch
Institute of Pathology and Tropical Medicine,
RAF Halton, Aylesbury, UK

Strike, Paul W.
Statistical methods in laboratory medicine/P.W. Strike.
p. cm.
Includes bibliographical references and index.
ISBN 0 7506 1345 9
1. Medicine—Research—Statistical methods. I. Title.
R853.S7S83 1991



$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

Observed value — Expected value — Disturbance

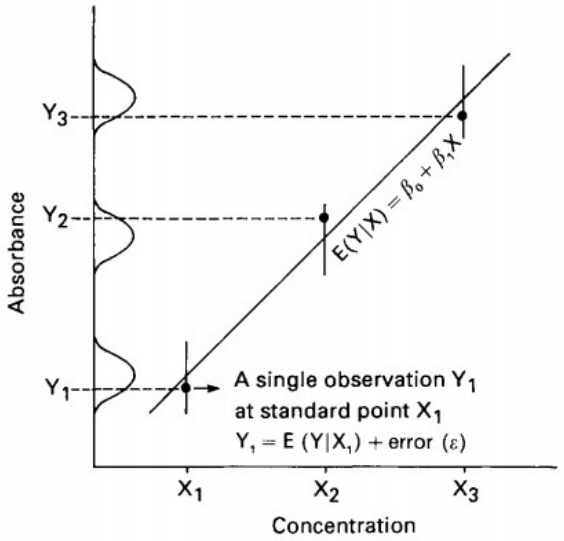


Fig. 7.2 The straight line model with X known and Y subject to observation error (ε)

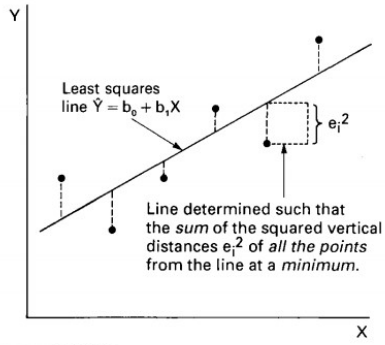
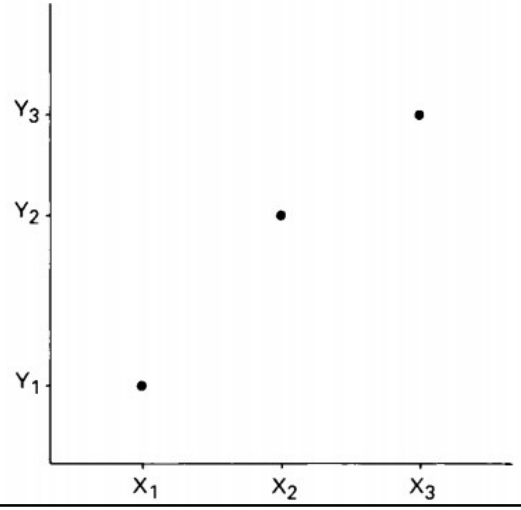


Fig. 7.4 The least squares principle



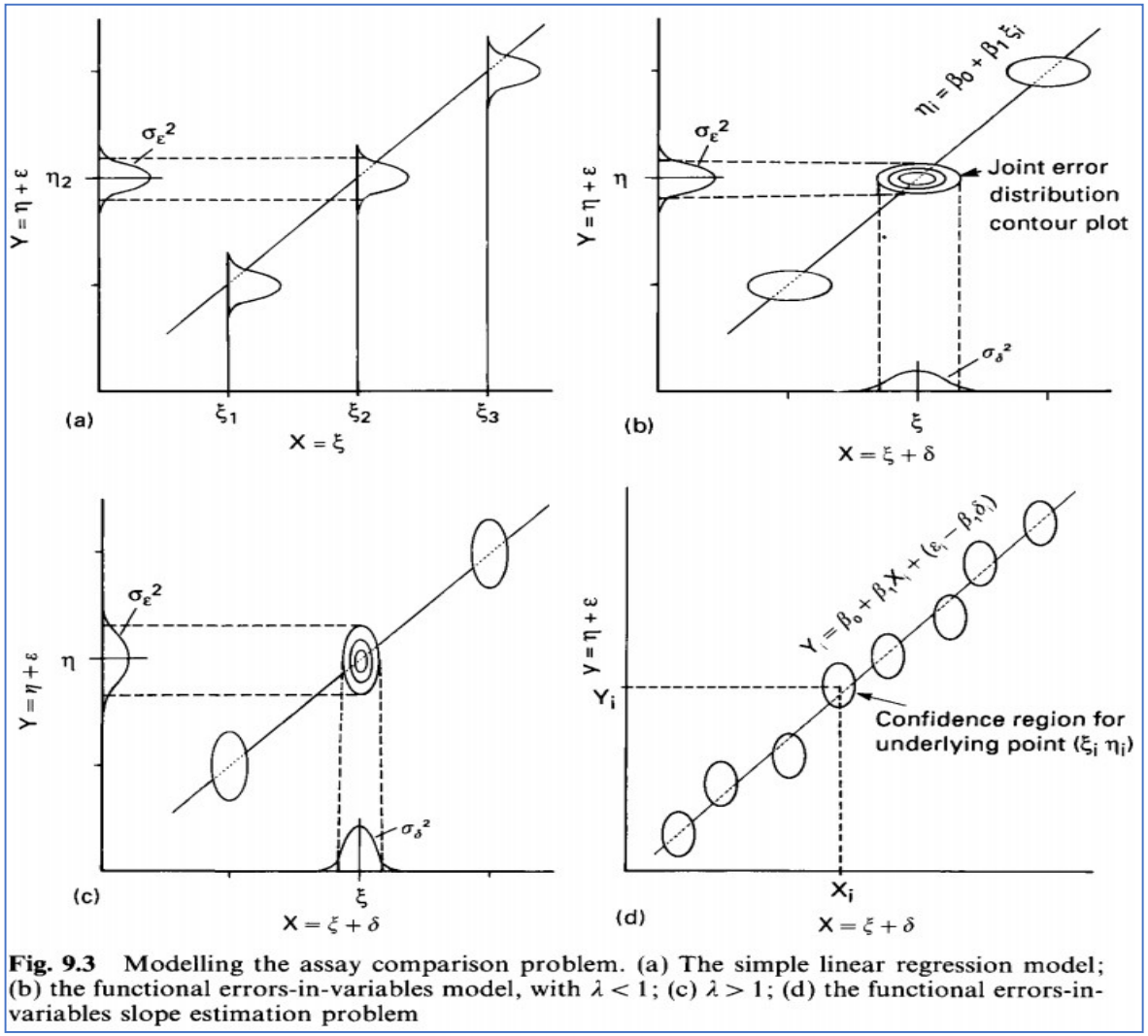


Fig. 9.3 Modelling the assay comparison problem. (a) The simple linear regression model; (b) the functional errors-in-variables model, with $\lambda < 1$; (c) $\lambda > 1$; (d) the functional errors-in-variables slope estimation problem

(6) Simultaneous 95% c.i. on $E(Y|X_i)$

A 95% c.i. on the expected or mean values of Y for the entire regression line was derived by Working and Hotelling (1929) as follows:

$$\underset{\text{multiple inference}}{95\%\text{ c.i. on } E(Y|X_i)} = \hat{Y}_i \pm (2F_{2/n-2})^{1/2}\left[S_e^2\left(\frac{1}{n}+\frac{(X_i-\bar{X})^2}{S_{XX}}\right)\right]^{1/2} \qquad (7.31)$$

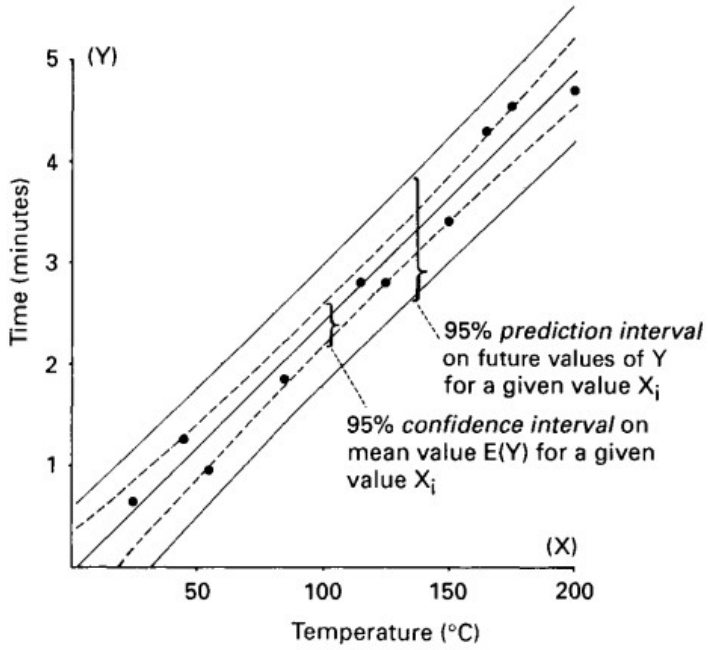


Fig. 7.15 Interval estimates on the Y variable from eq. 7.29 and eq. 7.30. (Note: these equations strictly admit inference on Y for one value of X_i only (see text).)

4.5 Effects of Measurement Errors

In our discussion of regression models up to this point, we have not explicitly considered the presence of measurement errors in the observations on either the response variable Y or the predictor variable X. We now examine briefly the effects of measurement errors in the observations on the response and predictor variables.

Applied linear statistical model - 5e - Kutner et al - 2004

Measurement Errors in *Y*

When random measurement errors are present in the observations on the response variable Y, no new problems are created when these errors are uncorrelated and not biased (positive and negative measurement errors tend to cancel out). Consider, for example, a study of the relation between the time required to complete a task (Y) and the complexity of the task (X). The time to complete the task may not be measured accurately because the person operating the stopwatch may not do so at the precise instants called for. As long as such measurement errors are of a random nature, uncorrelated, and not biased, these measurement errors are simply absorbed in the model error term ε. The model error term always reflects the composite effects of a large number of factors not considered in the model, one of which now would be the random variation due to inaccuracy in the process of measuring Y.

Measurement Errors in *X*

Unfortunately, a different situation holds when the observations on the predictor variable X are subject to measurement errors. Frequently, to be sure, the observations on X are accurate, with no measurement errors, as when the predictor variable is the price of a product in different stores, the number of variables in different optimization problems, or the wage rate for different classes of employees. At other times, however, measurement errors may enter the value observed for the predictor variable, for instance, when the predictor variable is pressure in a tank, temperature in an oven, speed of a production line, or reported age of a person.

Clinical Chemistry 44:5
1024–1031 (1998)

Laboratory Management

Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies

KRISTIAN LINNET

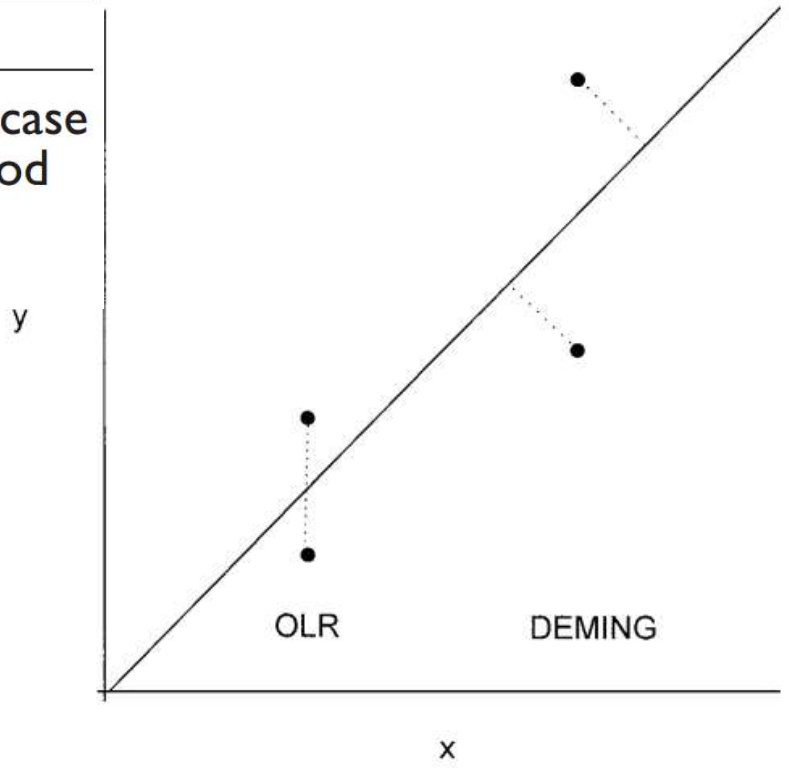


Fig. 1. Comparison of analysis methods.

OLR analysis: projection of measurement sets (x, y) onto the line in the vertical direction so that the sum of squared deviations is minimized; Deming regression analysis: projection of measurement sets (x, y) onto the line at an angle determined by λ so that the sum of squared deviations is minimized. Here λ = 1 and the angle is 90°.

Métodos de mensuração equivalentes

- Os dois métodos de mensuração são equivalentes se:
  - A reta estrutural é a bissetriz.

Suposições e Condições

- APENAS uma população definida por critérios de inclusão e de exclusão
- Método A: Referência ou Gold Standard
- Método B: Novo (mais econômico/ rápido/ menos invasivo)
- Concordância estrita
  - Relação linear passando pela origem entre os métodos A e B: bissetriz do primeiro quadrante do plano cartesiano
- Testes de "aceitação" de H_0
  - Os testes principais são de "aceitação" e não de rejeição de hipótese nula: o tamanho da amostra para obter poder prospectivo de 90% deve ser determinado.
  - Usando método *bootstrapping* não é necessário supor normalidade e homocedasticidade.

Regressão de Deming

- $x_i = X_i + \varepsilon_i$
- $y_i = Y_i + \delta_i$
- $i = 1,2, \ldots, n$ unidades observacionais distintas (independentes)
- X: novo e Y: referência são os verdadeiros valores dos valores observados x e y, respectivamente.
- ε e δ são os erros de mensuração com média nula e independentes entre ele e independentes de X e Y, respectivamente.
- As variâncias de V(ε) e V(δ) não são necessariamente iguais.
- $\lambda = \frac{V(\varepsilon)}{V(\delta)}$: razão das precisões

Regressão linear simples funcional e estrutural

- Observações dos métodos A: referência (y) e B: novo (x) com erros de mensuração
  - $y = Y + \delta$
    - y é variável aleatória observada do valor observado do método A: média das observações intraparticipantes
    - Y é variável aleatória do verdadeiro valor da medida
    - δ é variável aleatória do erro de mensuração com média nula
    - Y e v são independentes
  - $x = X + \varepsilon$
    - x é variável aleatória do valor observado do método B: média das observações intraparticipantes
    - X é variável aleatória do verdadeiro valor da medida
    - ε é variável aleatória do erro de mensuração com média nula
    - X e u são independentes
  - δ e ε são independentes
- Reta de regressão estrutural
  - $Y = \alpha + \beta X$
- Reta de regressão funcional
  - $y = \alpha + \beta x + (\delta - \beta \varepsilon)$

47

Comparação de métodos por regressões lineares simples estruturais e funcionais

Teste de concordância

- H_0 estrutural:
  - Y = X ou
    Y = α + β X , α = 0 e β = 1 ou
    banda de confiança de 95% contém a bissetriz

THE LANCET, FEBRUARY 8, 1986

Measurement

STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT

J. MARTIN BLAND DOUGLAS G. ALTMAN

Department of Clinical Epidemiology and Social Medicine, St George's Hospital Medical School, London SW17; and Division of Medical Statistics, MRC Clinical Research Centre, Northwick Park Hospital, Harrow, Middlesex

Summary In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for the new to replace the old. Such investigations are often analysed inappropriately, notably by using correlation coefficients. The use of correlation is misleading. An alternative approach, based on graphical techniques and simple calculations, is described, together with the relation between this analysis and the assessment of repeatability.

PEFR MEASURED WITH WRIGHT PEAK FLOW AND MINI WRIGHT PEAK FLOW METER

Subject	Wright peak flow meter		Mini Wright peak flow meter	
	First PEFR (l/min)	Second PEFR (l/min)	First PEFR (l/min)	Second PEFR (l/min)
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
4	434	401	428	444
5	476	470	500	500
6	557	611	600	625
7	413	415	364	460
8	442	431	380	390
9	650	638	658	642
10	433	429	445	432
11	417	420	432	420
12	656	633	626	605
13	267	275	260	227
14	478	492	477	467
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443

A miniature Wright peak-flow meter

B M WRIGHT

British Medical Journal, 1978, **2**, 1627-1628

Summary and conclusions

A new miniature Wright peak-flow meter has been designed and produced. The meter is tubular with a spring-loaded piston and a longitudinal slot through which air escapes. Its dynamic characteristics have been carefully designed to make it respond only to peak flow and not to rate of rise. Performance tests on early instruments showed fairly close correlation with the Wright peak-flow meter but with a constant error of ±38 l/min. On later models the correlation was increased to 0·990 and the error reduced to ±3%.

The mini-meter correlates as well with the standard instrument as two standard instruments correlate with each other and should prove useful clinically.

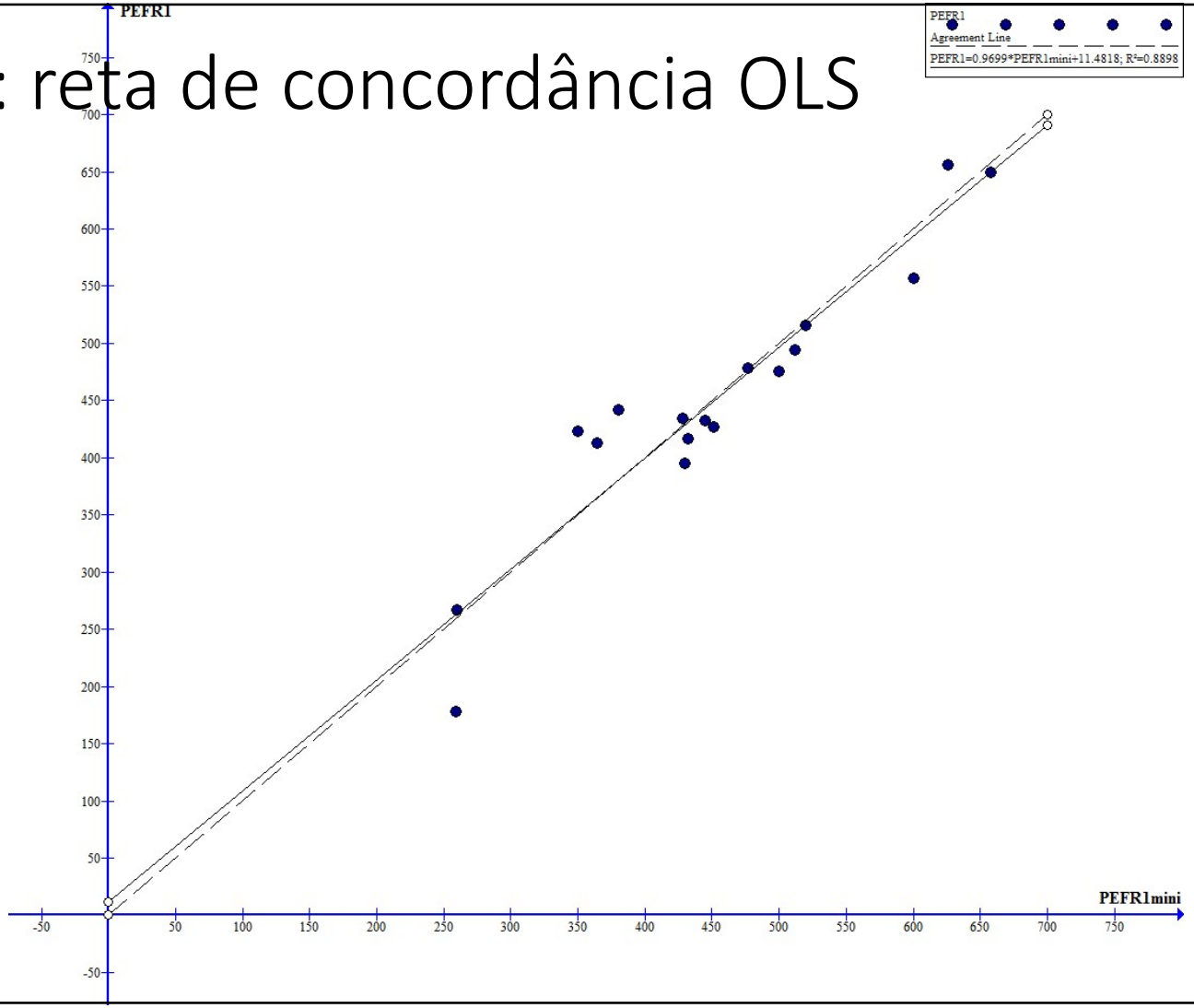
Performance

The instrument has been tested by several workers. Perks *et al*[10] found a correlation coefficient of 0·970 with the standard instrument on a hundred pairs of PEF measurements ranging from 100 to 700 l/min, but with a constant error of +38 l/min. They noted that this error was being corrected in later instruments. Pride (N R Pride, personal communication) in a study of 14 paired observations over a similar range found a correlation coefficient of 0·990. McDermot and Oldham (M McDermot and H G Oldham, personal communication) studied the mini-meter at various times during its production. With early models they found the same high readings as Perks *et al* but with later models the error was reduced to +3%, and the correlation coefficient was 0·990 on a group of 44 comparisons.

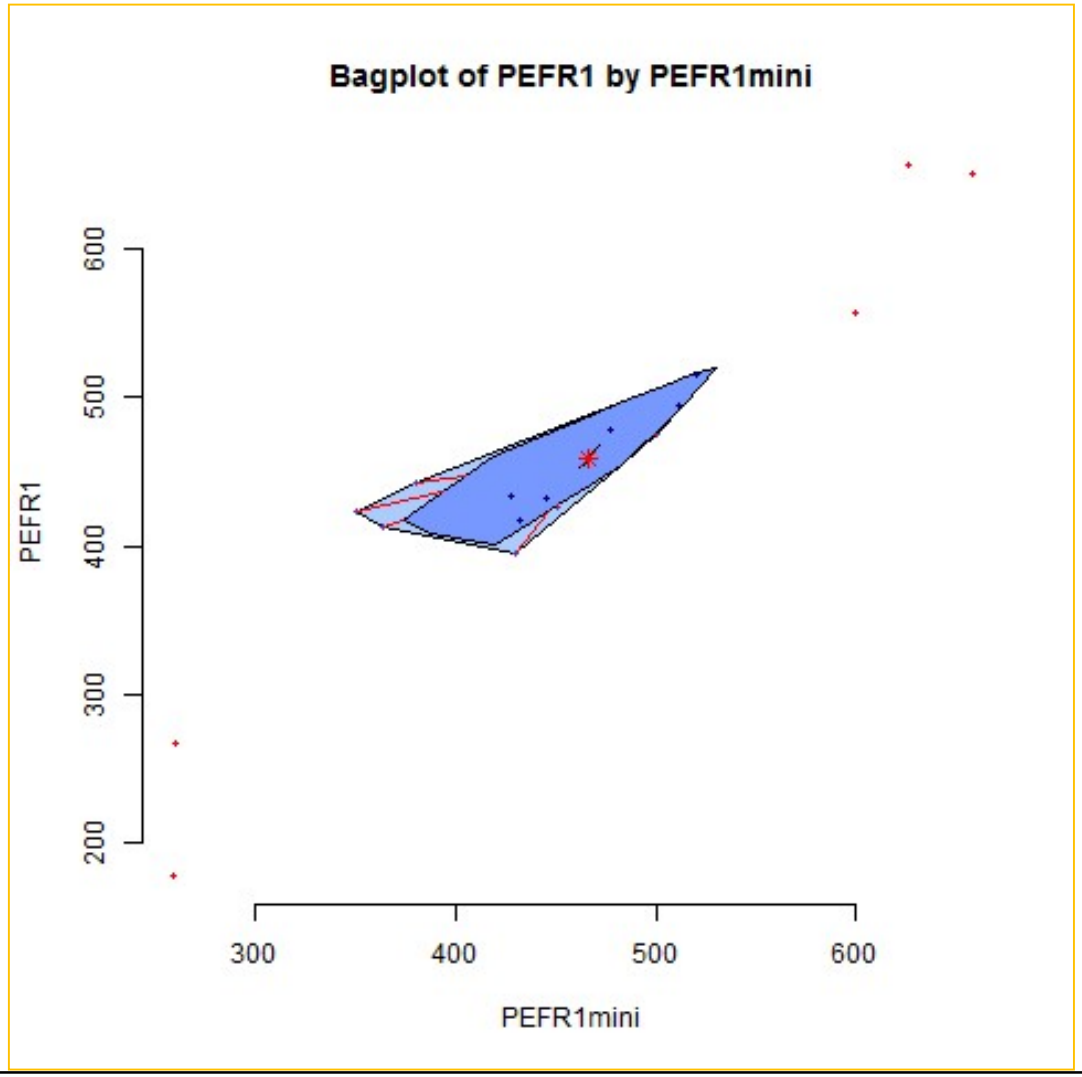
In all these studies subjects blew alternately through the two instruments in a randomised fashion. This was the procedure followed by Wright and McKerrow and was unavoidable because of the nature of the standard peak-flow meter and the pneumotachograph with which it was compared. It does, however, inevitably increase the scatter, as the breaths measured are not the same.

The mini-meter, however, being small and tubular, lends itself to being enclosed in a case and connected in series with a standard peak-flow meter. When this was done the correlation coefficient rose to 0·995, which was a significant improvement. For practical purposes therefore the performance of the mini-meter is identical with that of the standard instrument.

Bissetriz: reta de concordância OLS



52


Bagplot of PEFR1 by PEFR1mini
PEFR1
200
300
400
500
600
300
400
500
600
PEFR1mini

53

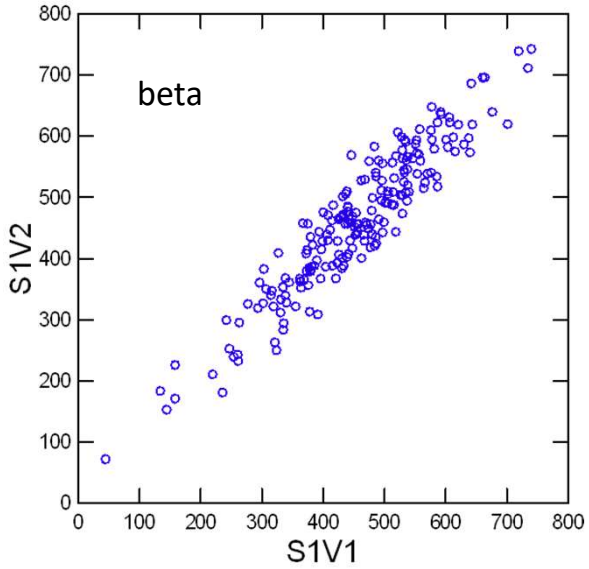
PEFR1	PEFR2	PEFR.media	PEFRmini1	PEFRmini2	PEFRmini.media
494	490	492,0	512	525	518,5
395	397	396,0	430	415	422,5
516	512	514,0	520	508	514,0
434	401	417,5	428	444	436,0
476	470	473,0	500	500	500,0
557	611	584,0	600	625	612,5
413	415	414,0	364	460	412,0
442	431	436,5	380	390	385,0
650	638	644,0	658	642	650,0
433	429	431,0	445	432	438,5
417	420	418,5	432	420	426,0
656	633	644,5	626	605	615,5
267	275	271,0	260	227	243,5
478	492	485,0	477	467	472,0
178	165	171,5	259	268	263,5
423	372	397,5	350	370	360,0
427	421	424,0	451	443	447,0
PEFR.mediag =		447,9	PEFRmini.mediag =		453,9

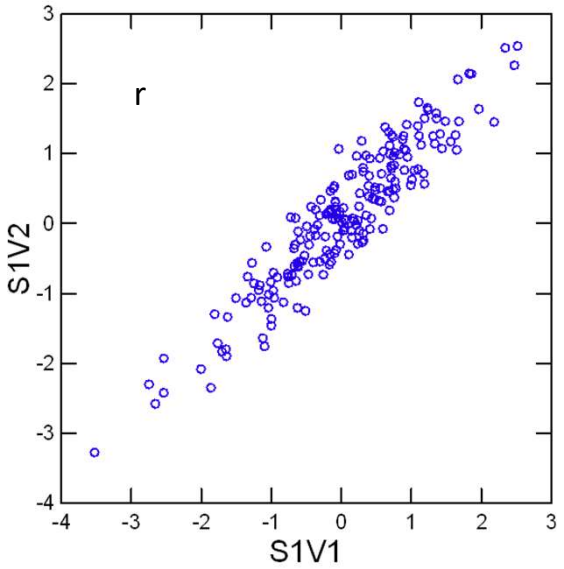
55

Normal bivariada (SYSTAT 13)

Seed = 1, N = 200

Métodos equivalentes



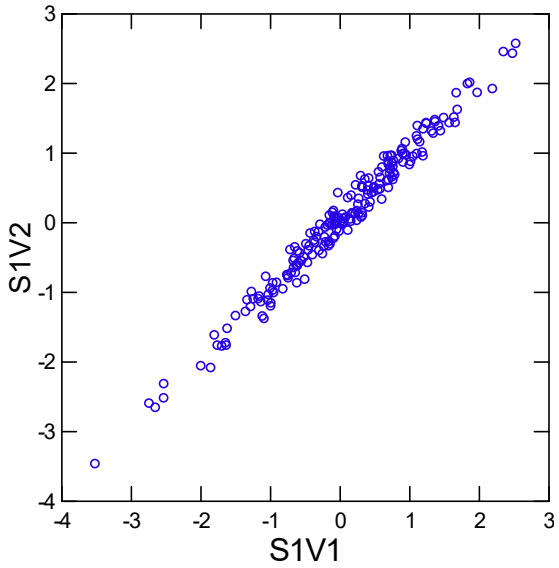


450	450
13225	12475
12475	13225

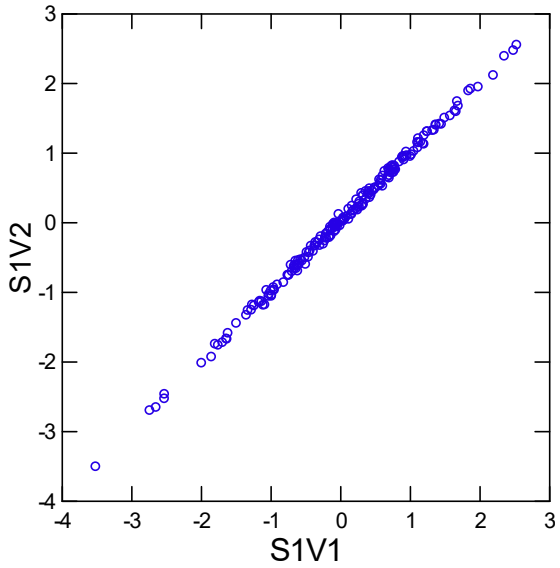
0	0
1	0.94
0.94	1

Normal bivariada (SYSTAT 13)

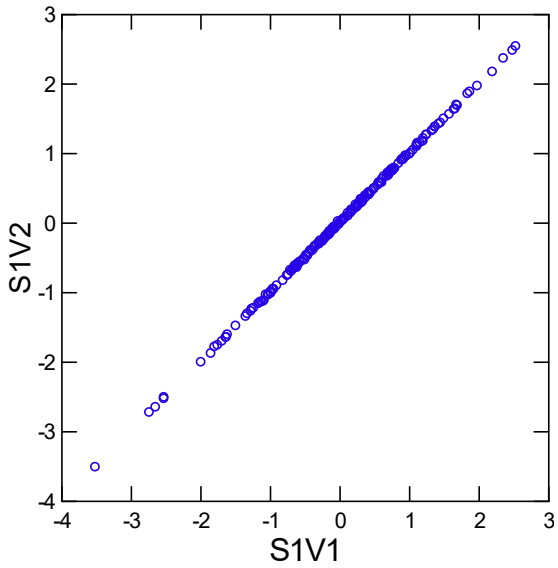
Seed = 1, N = 200



0	0
1	0.99
0.99	1



0	0
1	0.999
0.999	1



0	0
1	0.9999
0.9999	1

57

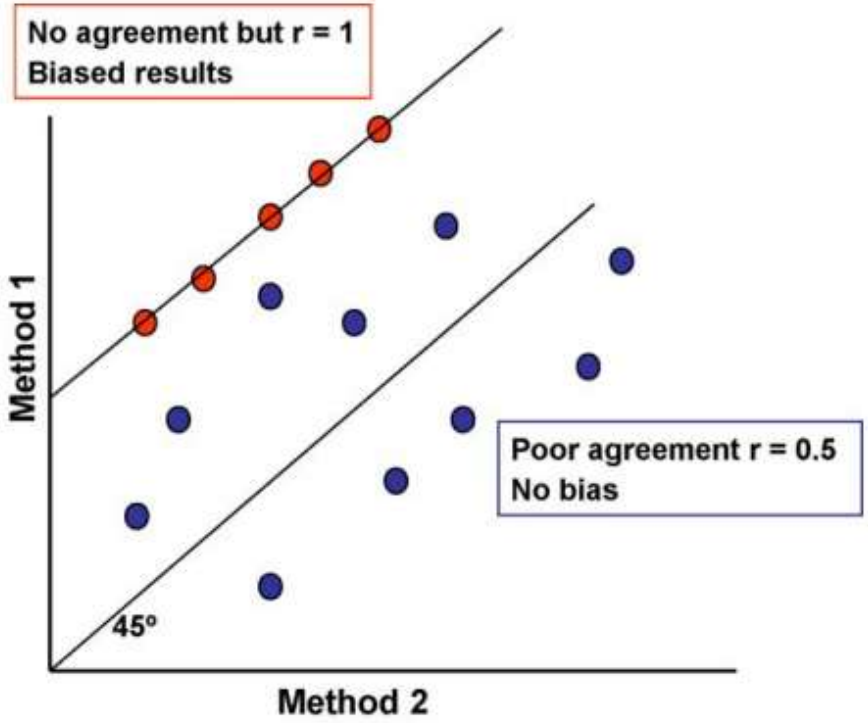


Fig. 5. Diagram showing two kinds of association between the results of Method 1 and those of Method 2. The red circles on the upper line demonstrate perfect correlation but no agreement. The blue circles around the lower line demonstrate poor correlation but no systematic difference between the two methods.

58

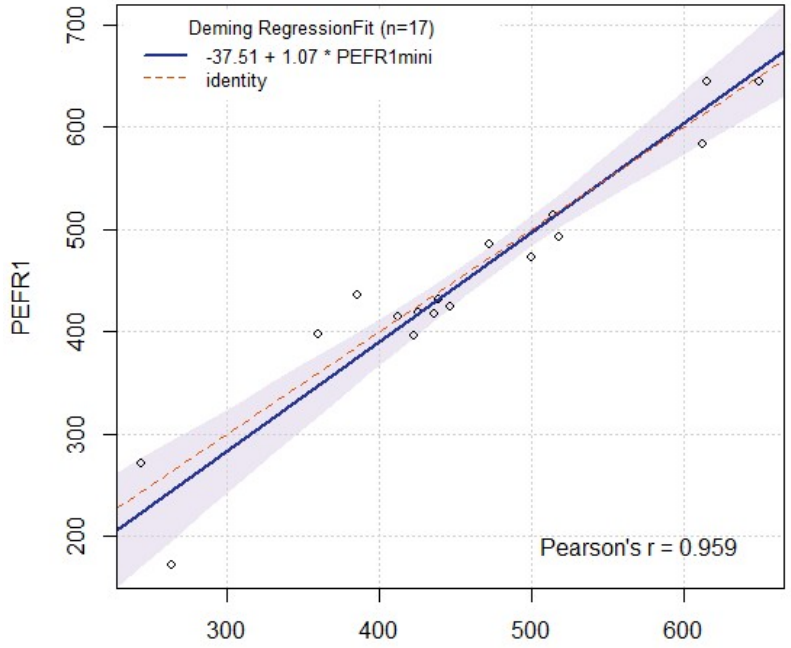
A reta estrutural é bissetriz? Sim
DemingRegression.R

```
library(readxl)
library(mcr)
# x = PEFRmini
# y = PEFR
# Estimacao pontual do lambda para medidas repetidas usada na Regressao de Deming
# Chapter 303 do NCSS 11 (2016): Deming regression
PEFRDATA <- readxl::read_excel("PEFRDATA.xls")
PEFR.media <- (PEFRDATA$PEFR1 + PEFRDATA$PEFR2)/2
PEFR.mediag <- mean(PEFR.media, na.rm=TRUE)
PEFRmini.media <- (PEFRDATA$PEFRmini1 + PEFRDATA$PEFRmini2)/2
PEFRmini.mediag <- mean(PEFRmini.media, na.rm=TRUE)
n <- nrow(PEFRDATA)
VARdelta <- (sum((PEFRDATA$PEFR1 - PEFR.media)^2, na.rm=TRUE) +
               sum((PEFRDATA$PEFR2 - PEFR.media)^2, na.rm=TRUE))/n
VARepsilon <- (sum((PEFRDATA$PEFRmini1 - PEFRmini.media)^2, na.rm=TRUE) +
               sum((PEFRDATA$PEFRmini2 - PEFRmini.media)^2, na.rm=TRUE))/n
lambda_rm <- VARepsilon/VARdelta # lambda = 1.692
out_boot_rm <- mcr::mcreg(x=PEFRmini.media,y=PEFR.media,error.ratio=lambda_rm,
             method.reg="Deming",
             method.ci="bootstrap", nsamples = 1e4, rng.seed = 123,
             mref.name = "PEFR1mini", mtest.name = "PEFR1", na.rm=TRUE)
printSummary(out_boot_rm)
plot(out_boot_rm)
```

DEMING REGRESSION FIT:

	EST	SE	LCI	UCI
Intercept	-37.509068	NA	-139.5273223	69.534865
slope	1.069352	NA	0.8433128	1.280572

Deming Regression Fit



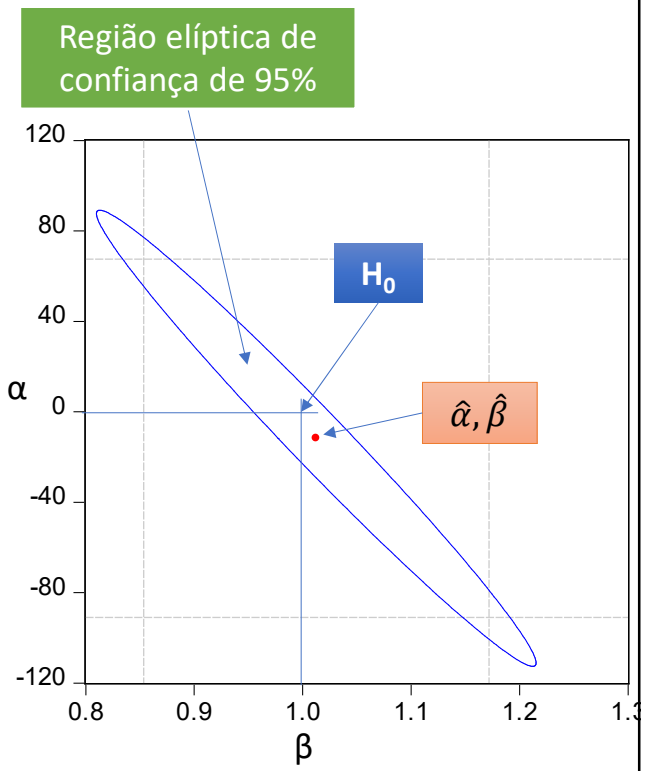
The 0.95-confidence bounds are calculated with the bootstrap(quantile) method

The confidence intervals are calculated with bootstrap (quantile) method.
Confidence level: 95%
Error ratio: 1.692066

60

Região de confiança elíptica de 95% para intercepto e inclinação

- $H_0: \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $vs\, H_a: \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \neq \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
- $F = \left(\frac{n-2}{2}\right) \frac{n\hat{\alpha}^2 + 2\hat{\alpha}(\hat{\beta}-1) \sum_{i=1}^{n} \hat{X}_i + (\hat{\beta}_1 - 1)^2 \sum_{i=1}^{n} \hat{X}_i{}^2}{\sum_{i=1}^{n} d_i^2} \leq F_{2,n-2}^{95\%}$
- $\hat{\alpha} = -37.5 \quad \hat{\beta} = 1.069$



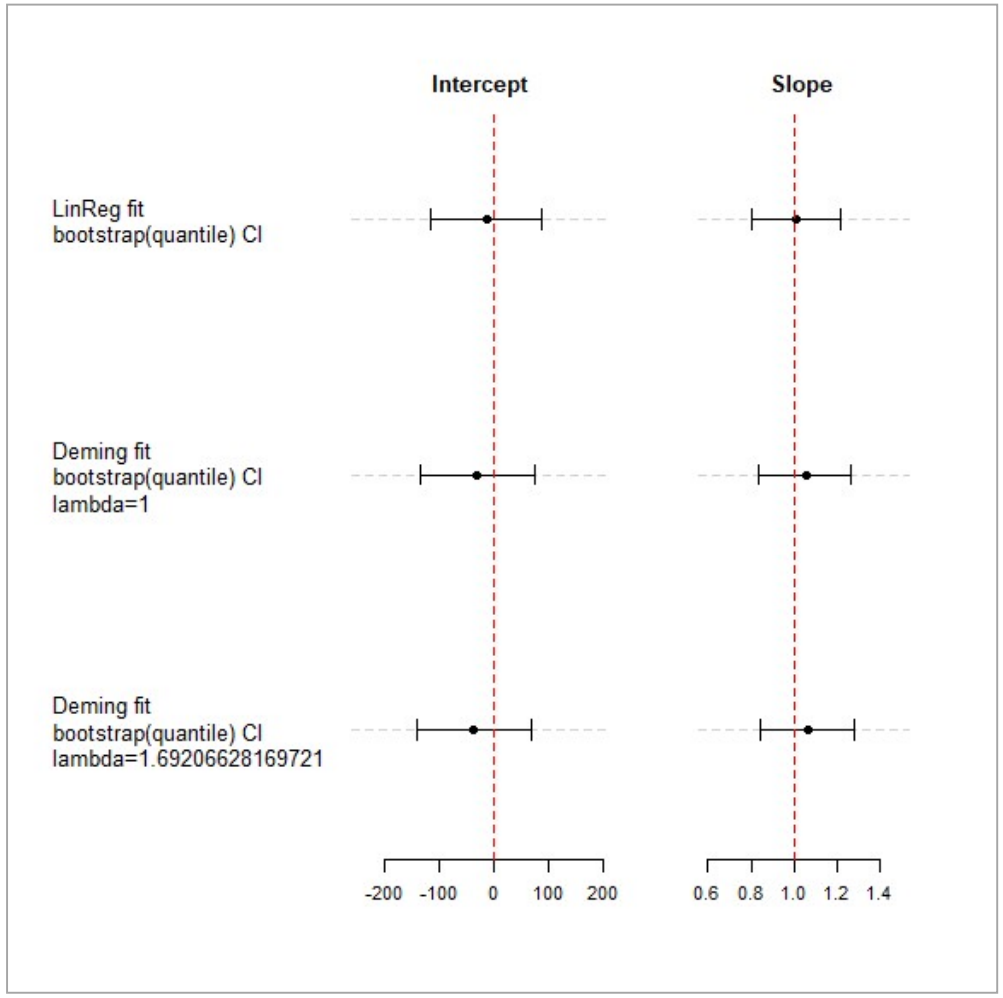
Regressão N-W lag=0
EVIEWS 10

PEFRMINI	1.012584
C	-11.74124

61

A reta estrutural é bissetriz? Sim
DemingRegression.R

```
out_boot_rm <- mcr::mcreg(x=PEFRmini.media,y=PEFR.media,error.ratio=lambda_rm,
             method.reg="Deming",
             method.ci="bootstrap", nsamples = 1e4, rng.seed = 123,
             mref.name = "PEFR1mini", mtest.name = "PEFR1", na.rm=TRUE)
mcr::printSummary(out_boot_rm)
mcr::plot(out_boot_rm)
out_boot_rm1 <- mcr::mcreg(x=PEFRmini.media,y=PEFR.media,error.ratio=1,
                           method.reg="Deming",
                           method.ci="bootstrap", nsamples = 1e4, rng.seed = 123,
                           mref.name = "PEFR1mini", mtest.name = "PEFR1", na.rm=TRUE)
mcr::printSummary(out_boot_rm1)
mcr::plot(out_boot_rm1)
out_lr_rm <- mcr::mcreg(x=PEFRmini.media,y=PEFR.media,error.ratio=1,
                        method.reg="LinReg",
                        method.ci="bootstrap", nsamples = 1e4, rng.seed = 123,
                        mref.name = "PEFR1mini", mtest.name = "PEFR1", na.rm=TRUE)
mcr::printSummary(out_lr_rm)
mcr::plot(out_lr_rm)
mcr::compareFit(out_boot_rm, out_boot_rm1, out_lr_rm)
```


Intercept
Slope
LinReg fit
bootstrap(quantile) CI
Deming fit
bootstrap(quantile) CI
lambda=1
Deming fit
bootstrap(quantile) CI
lambda=1.69206628169721
-200
-100
0
100
200
0.6
0.8
1.0
1.2
1.4

63

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Theriogenology 73 (2010) 1167–1179

Theriogenology

www.theriojournal.com

Invited Review

Method agreement analysis: A review of correct methodology

P.F. Watson [a,*], A. Petrie [b]

[a] *The Royal Veterinary College, London, United Kingdom*
[b] *The UCL Eastman Dental Institute, London, United Kingdom*

Abstract

The correct approach to analyzing method agreement is discussed. Whether we are considering agreement between two measurements on the same samples (repeatability) or two individuals using identical methodology on identical samples (reproducibility) or comparing two methods, appropriate procedures are described, and worked examples are shown. The correct approaches for both categorical and numerical variables are explained. More complex analyses involving a comparison of more than two pairs of data are mentioned and guidance for these analyses given. Simple formulae for calculating the approximate sample size needed for agreement analysis are also given. Examples of good practice from the reproduction literature are cited, and common errors of methodology are indicated.
© 2010 Elsevier Inc. All rights reserved.

Keywords: Agreement analysis; Reliability; Repeatability; Reproducibility; Sample size calculation

2.2.2. The Bland and Altman diagram

A display of the differences between the pairs of readings may offer an insight into the pattern (and extent) of the agreement. The *Bland and Altman diagram* [12] is such a display; the difference between a pair is plotted on the vertical axis of the diagram against the mean of the pair on the horizontal axis. Fig. 3 shows the Bland and Altman plot of the follicular diameter data obtained from 20 mares in two repeated cycles. If a random scatter of points is observed, a single measure of repeatability is acceptable. To determine such a measure, we first estimate the standard deviation of the differences (s_d). Assuming a Normal distribution of differences, approximately 95% of the differences in the population are expected to lie between $d \pm 2s_d$, where d is the mean of the observed differences. The upper and lower limits of this interval, usually displayed on the Bland and Altman diagram, provide the *limits of agreement*; from them, we can decide (subjectively) whether the agreement between pairs of readings in a given situation is acceptable (see Fig. 3). For the mare data, the standard deviation of the differences is estimated as 2.97 mm and the 95% limits of agreement by –6.12 mm and 5.52 mm. The limits of agreement are shown as red lines in Fig. 3. The purple line is the line corresponding with the mean difference of –0.30 mm (it is negative in the diagram, indicating that on average the diameter measurements from the second cycle are greater than those of the first cycle).

Furthermore, the *British Standards Institution repeatability/reproducibility coefficient* ($2\ s_d$) may be used as a single measure of agreement. It indicates the maximum likely difference between a pair of readings. The British Standards repeatability coefficient for the mare data is $2 \times 2.97 = 5.94$ mm, which the investigators found represented acceptable repeatability.

65

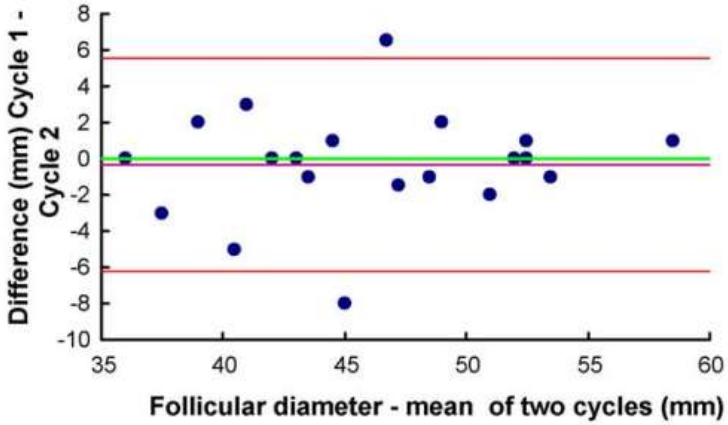


Fig. 3. Bland and Altman diagram showing the plot of the difference between the diameters (mm) of the equine follicle just prior to ovulation in two consecutive cycles of the mare against the mean of the pair (n = 20). Red lines show limits of agreement, and the purple line shows the mean value of the differences. The green line is the zero line used to assess the discrepancy of the observed mean difference from zero. (Data from Ref. 10, courtesy of Dr. Cuervo-Arango.)

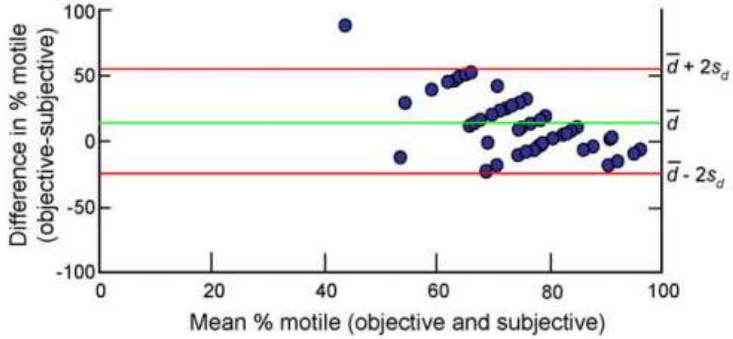


Fig. 4. Bland and Altman plot showing limits of agreement between two methods of measuring sperm motility, an objective Hamilton-Thorne computer-based semen analyzer and a subjective visual assessment, of samples of boar semen. $\bar{d}$ is the mean difference, and s_d is the standard deviation of the differences between pairs of measurements. (Redrawn and modified from Vyt P,Maes D, Rijsselaere T, Dejonkheere E, Castryck F, Van Soom A. Motility assessment of porcine spermatozoa: a comparison of methods. Reprod Dom Anim 2004;39:447.)

It should be noted that if the extent of agreement between the pairs depends on the magnitude of the measurement, a single measure of agreement is inappropriate. This would be evident on inspecting the Bland and Altman diagram if a funnel effect were observed. In such a situation, the variation in the differences is larger (say) for smaller mean values and decreases as the mean values become larger.

No funnel effect is observed in Fig. 3, but an example of its occurrence is shown in Fig. 4 (e.g., Vyt et al. [13]). These authors compared boar semen motility scores using a Hamilton-Thorne computer-based semen analyzer (HTR) with subjective microscope scoring from two experienced individuals. Fig. 4 shows the Bland and Altman diagram comparing the HTR with results from the first of the two individuals, in which the differences get smaller with the higher percentages. Note also that the mean difference departs substantially from zero indicating that the automated system gives systematically higher values for percentage motility.

In this situation, where a funnel effect is observed, the problem must be reassessed. An appropriate transformation of the raw data may resolve the issue, so that when the process is repeated on the transformed observations, the required conditions are satisfied. Otherwise, we should not calculate a *single* measure of reproducibility.

The Bland and Altman diagram can also be used to detect outliers. Outliers are occasional extreme readings departing from the main body of the data, possibly caused by errors of measurement.

Ou outras populações!!!

Statistical Methods in Medical Research 1999; **8**: 135–160

Measuring agreement in method comparison studies

J Martin Bland Department of Public Health Sciences, St George's Hospital Medical School, London, UK and **Douglas G Altman** ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK

Agreement between two methods of clinical measurement can be quantified using the differences between observations made using the two methods on the same subjects. The 95% limits of agreement, estimated by mean difference ± 1.96 standard deviation of the differences, provide an interval within which 95% of differences between measurements by the two methods are expected to lie. We describe how graphical methods can be used to investigate the assumptions of the method and we also give confidence intervals. We extend the basic approach to data where there is a relationship between difference and magnitude, both with a simple logarithmic transformation approach and a new, more general, regression approach. We discuss the importance of the repeatability of each method separately and compare an estimate of this to the limits of agreement. We extend the limits of agreement approach to data with repeated measurements, proposing new estimates for equal numbers of replicates by each method on each subject, for unequal numbers of replicates, and for replicated data collected in pairs, where the underlying value of the quantity being measured is changing. Finally, we describe a nonparametric approach to comparing methods.

Referências

- ALTMAN DG, BLAND JM. (1983) Measurement in medicine: the analysis of method comparison studies. **The Statistician**, 32(3): 307-17.
- ALTMAN DG, BLAND JM. (1999) Measuring agreement in method comparison studies. **Statistical Methods in Medical Research**, 8: 135-60.
- BLAND JM, ALTMAN DG. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. **Lancet**, i: 307-10.
- DRAPER, NR, SMITH, H (1998) **Applied regression analysis**. 3rd ed. NJ: Wiley.
- LINNET, K (1998) Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. **Clinical Chemistry** 44(5): 1024–1031.
- THORESEN, M, LAAKE, P (2007) On the simple linear regression model with correlated measurement errors. **Journal of Statistical Planning and Inference**, 137: 68-78.
- WATSON, PF, PETRIE, A (2010) Method agreement analysis: A review of correct methodology. **Theriogenology**, 73: 1167-79.
- SHOUKRI, MM (2011) **Measures of interobserver agreement and reliability**. 2nd ed. NY: Chapman & Hall/CRC.
- KUTNER, MH et al. (2005) **Applied linear statistical models**. 5th ed. CA: McGraw-Hill/Irwin.
- NCSS Data Analysis 12 (2018). Chapter 303: Deming Regression. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Deming_Regression.pdf .

70