J Vet Clin 28(1) : 108-112 (2011)

Journal of Veterinary Clinics
www.ksvc.or.kr

검사법의 일치도 평가를 위한 분석기법

박선일[1] · 오태호*
강원대학교 수의과대학 및 동물의학종합연구소, *경북대학교 수의과대학

(게재승인: 2010년 12월 8일)

Statistical Test of Agreement between Measurements in Method-comparison Study

Son-Il Pak[1] and Tae-Ho Oh*
College of Veterinary Medicine and Institute of Veterinary Science, Kangwon National University, Chuncheon 200-701, Korea
**College of Veterinary Medicine, Kyungpook National University, Daegu, 702-701, Korea*

Abstract : In clinical settings, researchers often want to assess agreement between two measurements (or tests) of the same continuous variable. For example, when new point-of-care analyzer for testing blood glucose level were introduced clinicians need to compare results from standard or established laboratory method of measurement to those of new or point-of-care analyzer. The question in a method-comparison study would either of two different methods be used to measure the same variable equivalently. In this paper common misuse of statistical methodologies seen in the medical literatures such as correlation coefficient and paired t-test are discussed. The Bland-Altman technique has been widely used for this purpose and provides a graphic in presentation of the findings from a method-comparison study, with a mean value of measurement, this bias and the limits of agreement. For ease of application and interpretation of this technique we discussed the analysis procedure and illustrated with two worked examples. Finally, a number of alternative ways in which data can be analysed and reported in such studies were reviewed.

Key words **:** bias, Bland-Altman plot, method-comparison.

서 론

임상연구에서 진단방법이나 측정 장비를 비교하는 상황은 매우 흔하다(4,9,11,12,14,17-19). 이를테면 혈당(blood glucose)을 실험실적인 방법과 휴대용 장비를 이용하여 측정한 결과를 비교하는 연구(10), 승모판 역류량(amount of mitral regurgitation)을 Doppler echocardiography와 심장카테터(cardiac catheterization)로 측정한 연구(13), 돼지의 체온을 폐동맥과 직장에서 측정한 연구(7), 사람에서 혈압을 pulse-contour analysis와 폐동맥 카테터를 이용하여 측정한 연구(6), 두 종류의 면역분석 기법을 이용하여 procalcitonin을 평가한 연구(8), 개에서 휴대용 bioimpedance 모니터링 장비와 dual-energy x-ray absorptiometry를 이용하여 체지방율(percentage of body fat)을 비교한 연구(5), 돼지의 대퇴동맥(femoral artery)과 이동맥(auricular artery)에서 혈압을 측정하여 비교한 연구(1), 개에서 두 종류의 장비를 이용하여 혈청 cortisol을 평가한 연구(16) 등은 전형적인 예다.

이러한 연구에서는 연구자는 흔히 기존의 장비 (표준검사)와 새로운 장비가 동일한 결과를 보이는지에 관심을 갖는다. 흔히 표준검사는 비용과 시간 측면에서 임상적으로 적용하기 어렵기 때문에 보다 간편하고 저렴한 검사법을 선호하게 되는데 이를 위해서는 신속한 검사법 혹은 간편한 측정 장비가 표준검사와 동일한 결과를 보이는 다는 것이 입증된 경우에만 가능하다. 예를 들어 동일한 개체를 대상으로 두 가지 방법 (a와 b)으로 특정한 효소의 농도를 측정하는 경우 연구자는 두 방법에 의한 검사결과가 일치하는지에 관심을 갖는다. 이러한 문제는 두 측정 방법 간의 일치도(agreement)를 평가하는 것으로 다양한 기법이 개발되어 있다. 본 연구에서는 연속형으로 측정된 자료(continuous data)의 일치도를 평가하는 수단으로 널리 사용되는 Bland-Altman method에 대하여 설명한다.

[1]Corresponding author.
E-mail : paksi@kangwon.ac.kr

결 론

분석기법의 오류

일치도 평가를 위한 분석기법으로 흔히 범하는 오류는 t 검정이나 상관분석을 사용하는 경우이다. 예를 들어 Hemocytometer(C)를 이용하여 정자수를 측정하는 경우 시간이 많이 소요되므로 비교적 간단한 Colorimeter(H)로 측정하는 방

법을 제시하는 상황을 가정하자. Table 1은 22두의 양을 대상으로 C와 H 두 방법을 이용하여 정자 수 (단위: $10^9/ml$)를 측정한 결과이다(15).

이 자료에 대하여 일치도를 평가하기 위하여 t 검정이나 상관계수를 분석하는 방법은 적절하지 못하다. 그 이유는 첫째, 동일한 개체를 측정도구를 달리하여 2회 측정한 자료이므로 짝지은(paired) t 검정을 사용하는 것인데 이 방법은 "평균 차이가 0 이다"는 귀무가설을 검정한다. 이 분석에서 얻는 결과는 두 측정 방법 간에 차이가 없다는 것이지 두 측정 결과가 일치한다는 것을 의미하는 것이 아니다. 둘째, 흔히 상관계수(r)를 사용하는 오류를 범하는데 이 방법은 "두 방법에 의한 측정치는 선형 연관성(linear correlation, linear association)이 없다"는 귀무가설을 평가하는 것이다. 예를 들어 분석결과가 유의한 경우 귀무가설을 기각하고 연관성이 높다고 해석한다. 그러나 이러한 결과가 두 방법 간의 일치도를 의미하는 것은 아니다. 그 이유는 상관계수는 두 변수 간의 선형적 상관성의 강도(strength of linear relationship)를 측정하는 수단이기 때문이다. 이를테면 위의 자료를 산점도(scatter plot)로 표현할 때 두 측정치가 모두 대각선상에 위치하는 경우에 완벽한 일치도(perfect agreement)가 있다고 표현하지만 대각선 이외의 다른 어떤 직선상에 위치한다면 완벽한 상관성(perfect correlation)이 있지만 일치성이 있다고는 할 수 없다. 또한 상관계수에 대한 유의성 검정에서 귀무가설은 두 측정치 간에 연관성이 없다는 것이다. 동일한 대상을 측정도구를 달리하여 2회 측정하는 경우 (당연히 연관성이 있지만) 귀무가설을 연관성이 없다고 설정하는 것은 논리적으로 맞지 않다. 상관성과 일치도에 영향을 미치는 요인은 많다. 자료의 척도(scale)가 변하는 경우 상관성에는 영향을 미치지 못하지만 일치도에는 영향을 미칠 수 있다. 예를 들어 체중을 kg 단위로 측정하는 경우와 파운드 단위로 측정하는 경우 전자는 후자에 비하여 약 2배의 값으로 측정되기 때문에 상관성은 동일하지만 일치도는 다를 수 있다(2). 상관계수는 측정 자료의 범위에 영향을 받는다. 즉 다른 모든 조건이 동일하다고 할 때 자료의 범위가 넓어질수록 상관성은 높아지고, 특히 이상값(outlier)이 있을 경우 나머지 자료들이 넓게 퍼져있다고 하더라도 높은 상관성을 보일 수 있다. 따라서 일치도 분석으로 t 검정이나 상관성을 분석하는 방법은 적절하지 못하다.

Table 1. Results of sperm counts ($10^9/ml$) of 22 sheep ejaculates using two methods: Colorimeter (C) and Hemocytometer (H)

No.	C	H	No	C	H
1	0.82	1.01	12	1.73	1.52
2	2.34	2.46	13	3.11	3.37
3	2.34	2.20	14	3.76	3.60
4	4.13	4.29	15	1.12	1.09
5	4.03	3.82	16	3.28	3.43
6	4.70	4.59	17	3.25	3.16
7	4.78	4.66	18	1.28	1.41
8	5.00	4.75	19	3.82	3.77
9	5.04	4.97	20	3.48	3.49
10	5.17	5.24	21	1.43	1.23
11	5.27	5.35	22	3.14	3.33

Minimum of difference = –0.26
Maximum of difference = 0.25
Mean difference = 0.0127
Standard deviation of difference = 0.155
Source: Petrie (1999).

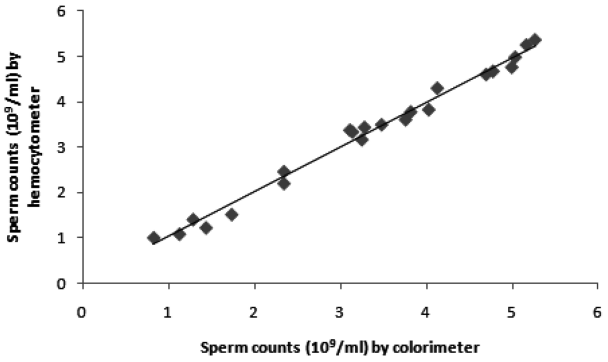

Fig 1. Scatter plot of the sperm counts ($10^9/ml$) measured with colorimeter and hemocytometer, with linear trend line.

Bland-Altman method

생물학적 현상 (변수)을 측정하는 서로 다른 두 방법이 완벽하게 동일한 결과를 보일 것으로 기대하기는 어렵기 때문에 두 방법에 의한 측정 결과가 어느 정도의 차이를 보이는지에 관심을 두는 것이 합리적이다. 표준검사에 의한 결과와 새로운 검사법에 의한 절대적 차이가 이를테면 20 이하로 매우 작다면 (임상적으로 20 이하의 차이는 무시할 정도로 작음을 의미함) 새로운 검사법을 표준검사의 대용으로 사용할 수 있다고 판단하는 것이다. 그렇다면 두 측정치 간 어느 정도의 차이가 적절한 것인가?

먼저 짝지은 두 측정치(A, B)에 대하여 산점도를 작성하면 두 방법에 의한 측정치 간 연관성을 파악할 수 있다. 그러나 대부분의 관찰치들이 선형 추세선(trend line)에 근접하여 있기 때문에 측정 방법 간 차이의 정도를 구분하기 쉽지 않다(Fig 1). 다른 방법으로 짝지은 측정치(A, B)간의 차이(difference, $d = A - B$)를 Y축, 평균 $[(A + B)/2]$ 을 X축으로 하여 그래프로 표현하면 유용한 결과를 얻을 수 있다(Fig 2). 이 그림을 Bland-Altman plot (1986)이라고 하며 동일한 변수를 두 방법으로 측정한 자료의 일치도를 평가하는 수단으로 널리 사용되며, 관찰치간 불일치(bias) 정도, 이상값 유무, 자료의 추세 등을 파악하는데 매우 유용하며, bias의 신뢰구간을 계산할 수 있다(2,15). 이러한 통계량에 근거하여 측정 장비의 정확도(accuracy)와 정밀도(precision)를 평가할 수 있다. 예를 들어 두 측정치 중 어느 하나가 표준검사일 경우 X축을 표준검사 결과로 하고 Y축을 두 측정치의 차이

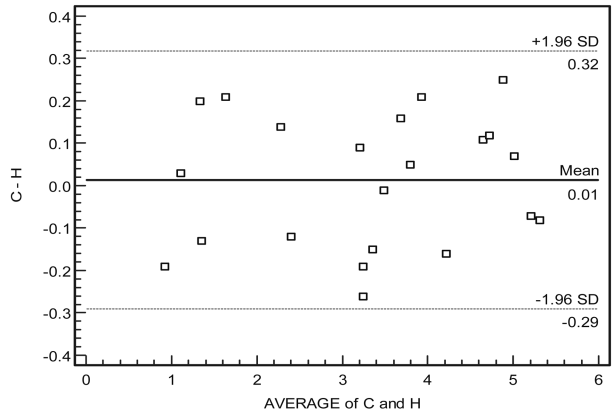

Fig 2. A Bland-Altman plot of the difference against mean for sperm counts ($10^9/ml$) measured with colorimeter (C) and hemocytometer (H) in 22 sheeps.

로 하여 그래프로 작성할 때 다양한 직선이 가능하며 표준검사에 대하여 다른 한 검사가 어느 정도 편견되어 있는지 그 크기를 파악할 수 있다.

Bland-Altman plot (Fig 2)에서 새로운 검사법 C와 표준검사법 H로 측정한 정자 수는 일치도가 상당히 높다는 것을 육안적으로 확인할 수 있다. 그러나 Fig 1과 같은 그림에서는 두 측정치 간 무수히 많은 선형 추세선이 가능하기 때문에 두 방법 간 연관성이 높은 것이지 일치도가 높다고 해석할 수 없다는 점이다. 또한 Bland-Altman plot에서 관찰치의 차이와 평균 간에 어떤 관련성이 없고 random하게 분포하고 있음을 알 수 있다. 이는 두 방법 간의 불일치 (bias를 의미함)의 크기 (size)가 정자 수와 관련이 없다는 것을 의미한다. 만일 불일치의 크기와 정자 수에 관련이 있고 bias가 일정하다면 새로운 검사법 C로 측정한 정자 수에 $\bar{d}$를 빼주어 측정값을 보정해야 한다. 본 자료에서는 random하게 분포하므로 두 방법 간의 일치도 정도를 평가하는 것이 적절하며 평균차이($\bar{d}$)와 차이의 표준편차(s)를 이용하여 bias의 정도를 평가할 수 있다. 즉 관찰치 간 차이(d)가 정규분포를 따르면 95%는 $\bar{d} \pm 2s$(보다 엄밀히 정의하면 $\bar{d} \pm 1.96s$)에 위치하는 것을 95% 신뢰할 수 있다. 본 자료에서 $\bar{d} = 0.0127$, $s = 0.155$이므로 일차도 한계(limit of agreement)는 [−0.2976, 0.3231]로 계산된다.

일치도 한계: $\bar{d} \pm 2s \Leftrightarrow 0.0127 \pm 0.310$
$\bar{d} - 2s$: $0.0127 - 2 \times 0.155 = -0.2976$
$\bar{d} + 2s$: $0.0127 + 2 \times 0.155 = 0.3231$

계산결과 두 방법 간의 차이가 한계 범위 이내에 위치하면 적절한 것으로 판정한다. 본 자료의 경우 정자수를 측정하기 위한 목적으로 Hemocytometer의 대안으로 Colorimeter를 사용하는 것이 인정할만한 수준이라는 결론을 얻는다. 일치도 한계는 모집단에 대한 추정치이므로 평균 차이에 대한 추정치의 정확도가 어느 정도인지 신뢰구간을 계산할 필요가 있다.

평균 차이의 신뢰구간

두 측정치 간의 차이(difference, d)의 표준오차는 $\sqrt{s^2/n}$ 이므로(2) 95% 신뢰구간은 자유도가 $n-1$인 t 분포로 계산할 수 있다(20). 본 예제의 경우 $s = 0.155$이므로 $\bar{d}$의 표준오차는 0.033 ($10^9/ml$)이다. 유의수준이 0.05이고 자유도가 21일 때 t = 2.08 이므로 C와 H 측정치 간의 평균차이(bias, $\bar{d} =$ 0.0127)에 대한 95% 신뢰구간은 [–0.056, 0.081]로 계산된다.

신뢰구간: $\bar{d} \pm t_{1-\alpha/2,df=n-1}SE$ [단, $SE = \sqrt{s^2/n}$, n = 표본크기]
$SE = \sqrt{s^2/n} = \sqrt{0.155^2/22} = 0.033$
평균 차이에 대한 95% 신뢰구간: $0.0127 \pm (2.08 \times 0.033) \leftrightarrow [-0.056, 0.081]$

일치도 한계의 신뢰구간

한편 일치도 한계($\bar{d} \pm 2s$)의 표준오차는 $\sqrt{3s^2/n}$ 이므로(2) 95% 신뢰구간은 자유도가 $n-1$인 t 분포로 계산할 수 있다 (20).

신뢰구간: $\bar{d} \pm t_{1-\alpha/2,df=n-1}SE$ [단, $SE = \sqrt{3s^2/n}$, n = 표본크기]
$SE = \sqrt{3s^2/n} = \sqrt{3 \times 0.155^2/22} = 0.058$
$\bar{d} - 2s = -0.2976$
$\bar{d} + 2s = 0.3231$
일치도 하한값의 95% 신뢰구간: $-0.2976 \pm (2.08 \times 0.058) \leftrightarrow [-0.418, 0.177]$
일치도 상한값의 95% 신뢰구간: $0.3231 \pm (2.08 \times 0.058) \leftrightarrow [0.202, 0.444]$

예 제

Table 2. Peak expiratory flow rate (PEER) with wright peak flow meter (Wright) and mini wright peak flow meter (Mini)

No.	Wright	Mini	No.	Wright	Mini
1	494	512	10	433	445
2	395	430	11	417	432
3	516	520	12	656	626
4	434	428	13	267	260
5	476	500	14	478	477
6	557	600	15	178	259
7	413	364	16	423	350
8	442	380	17	427	451
9	650	658			

Minimum of difference = –81
Maximum of difference = 73
Mean difference = –2.117
Standard deviation of difference = 38.765
Source: Bland and Altman (1986).

Table 2는 17명을 대상으로 분당 최대 호기율(peak expiratory flow rate, Liter)을 Wright peak flow meter와 Mini Wright meter를 이용하여 측정한 결과이다(2). 전술한 절차를 이용하여 Mini Wright meter를 Wright peak flow meter의 대안으로 사용할 수 있는지 평가하고, 추정치의 신뢰구간을 계산하여 보자.

이 자료에서 PEER의 차이(Large-Mini 측정치)의 평균과 표준편차(s)는 각각 –2.117과 38.765이다. 측정치 간 차이의 분포가 정규분포에 근사한다고 가정하면 일치도 한계는 다음과 같이 계산된다(Fig 3). 계산결과 Mini meter의 측정값이 분당 –79.6 L 보다 작거나 75.4 L 이상이라면 임상적으로 수용할 수 없는 것으로 간주할 수 있다.

$$\bar{d} - 2s = -2.117 - (1.96 \times 38.765) = -78.096\ L/\text{min}$$
$$\bar{d} + 2s = -2.117 + (1.96 \times 38.765) = 73.862\ L/\text{min}$$

한편, 계산된 일치도에 대한 추정치의 신뢰구간은 [–22.0 ~17.8] 로 계산되며, $\bar{d} \pm 2s$ 의 표준오차는 $16.3(\sqrt{3 \times 38.8^2}/\sqrt{17})$ 이므로 일치도에 대한 95% 신뢰구간은 다음과 같다.

평균 차이에 대한 95% 신뢰구간: $-2.1 \pm (2.12 \times 9.4) \leftrightarrow$ [–22.0~17.8]

일치도 하한값의 95% 신뢰구간: $-79.6 \pm (2.12 \times 16.3) \leftrightarrow$ [–114.3 ~ –45.1]

일치도 상한값의 95% 신뢰구간: $75.4 \pm (2.12 \times 16.3) \leftrightarrow$ [40.9~110.1]

기타 분석법

본 연구에서는 측정 방법의 일치도를 평가하는 기법으로 Bland-Altman 방법을 소개하였다. 변수들 간의 관계를 이용하여 측정도구나 검사법의 정확도를 평가하거나 비교하는 목적으로 사용할 수 있는 통계기법으로는 receiver-operating characteristic (ROC) curve, 범주형 자료에 대한 일치도 분석으로 사용되는 Kappa 통계량, Passing-Bablok regression analysis, Deming regression, 상관분석, Intra-class correlation coefficient (ICC), scatter plot, absolute difference 등 매우 많다(1,3,5,8,12-19). 중요한 것은 연구자가 관심을 갖고 있는 궁극적인 목적이 무엇인지에 따라 적절한 분석기법을 선택해야 한다는 점이다. 모든 상황에 적용할 수 있는 유일한 분석방법이 있는 것은 아니지만 기법을 선택할 때 다음과 같은 사항을 고려할 필요가 있다. 첫째, 연속형 자료를 범주형 자료(categorical data)로 변환하여 분석하는 것은 자료에 내재된 정보를 희생하기 때문에 연속형 자료에 부합하는 분석기법을 사용하는 것이 좋다. 둘째, 범주형 분석을 이용한다면 KAPPA 통계량을 사용하여 우연(chance)에 의한 일치도를 보정하는 것이 적절하다. 셋째, 연속형 자료는 분석하기 이전에 산점도나 Bland-Altman plot을 작성하여 개별 관찰치의 변동성을 요약하는 것이 좋다. 넷째, 연속형 자료에 대하여 일치도를 분석할 때 회귀분석이나 상관계수로 분석하는 것은 바람직하지 못하다. 다섯째, 연속형 자료에 대한 일치도를 평가할 목적으로 ICC를 계산하는 것은 적절한 방법이지만 측정치의 크기와 관련된 일치도의 변동성을 파악하지 못하기 때문에 다른 연구와 비교하기 어려운 단점이 있다. 측정치 간 불일치의 정도를 계량화하기 위해서는 Bland-Altman plot을 작성하는 것이 좋다. 여섯째, 측정치 간 절대차이를 평가하는 경우 측정치 간 계통적 오차(systematic bias)를 확인하고 일치도에 대한 신뢰구간을 계산하는 것이 바람직하다.

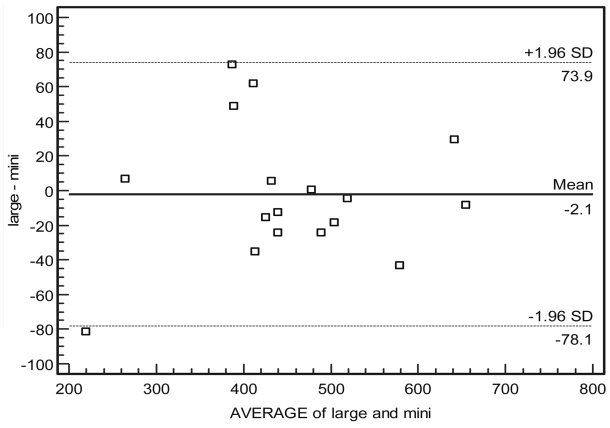

Fig 3. A plot of the difference between the peak expiratory flow rate (PEER) using a Wright peak flow meter (large) and a Mini Wright meter (mini).

감사의 글

본 연구는 강원대학교 동물의학종합연구소의 지원에 의해 이루어졌으며 이에 감사드립니다.

참 고 문 헌

1. Bass LM, Yu DY, Cullen LK. Comparison of femoral and auricular arterial blood pressure monitoring in pigs. Vet Anaesth Analg 2009; 36: 457-463.
2. Bland M, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; I: 307-310.
3. Deinzer M, Faissner R, Metzger T, Kaminski WE, Löhr M, Neumaier M, Brinkmann T. Comparison of two different methods for CA19-9 antigen determination. Clin Lab 2010; 56: 319-325.
4. Dey D, Schepis T, Marwan M, Slomka PJ, Berman DS, Achenbach S. Automated Three-dimensional Quantification of Noncalcified Coronary Plaque from Coronary CT Angiography: Comparison with Intravascular US. Radiology 2010; 257: 516-522.
5. German AJ, Holden SL, Morris PJ, Biourge V. Comparison of a bioimpedance monitor with dual-energy x-ray absorptiometry

for noninvasive estimation of percentage body fat in dogs. Am J Vet Res 2010; 71: 393-398.
6. Halvorsen PS, Sokolov A, Cvancarova M, Hol PK, Lundblad R, Tønnessen TI. Continuous cardiac output during off-pump coronary artery bypass surgery: pulse-contour analyses vs pulmonary artery thermodilution. Br J Anaesth 2007; 99: 484-492.
7. Hanneman SK, Jesurum-Urbaitis JT, Bickel DR. Comparison of methods of temperature measurement in swine. Lab Anim 2004; 38: 297-306.
8. Hausfater P, Brochet C, Freund Y, Charles V, Bernard M. Procalcitonin measurement in routine emergency medicine practice: comparison between two immunoassays. Clin Chem Lab Med 2010; 48: 501-504.
9. Klenner S, Bauer N, Moritz A. Evaluation of three automated human immunoturbidimetric assays for the detection of C-reactive protein in dogs. J Vet Diagn Invest 2010; 22: 544-552.
10. Lacara T, Domagtoy C, Lickliter D, Quattrocchi K, Snipes L, Kuszaj J, Prasnikar M. Comparison of point-of-care and laboratory glucose analysis in critically ill patients. Am J Crit Care 2007; 16: 336-346.
11. Liehr P, Dedo YL, Torres S, Meininger JC. Assessing agreement between clinical measurement methods. Heart Lung 1995; 24: 240-245.
12. Lopes PC, Sousa MG, Camacho AA, Carareto R, Nishimori CT, Santos PS, Nunes N. Comparison between two methods for cardiac output measurement in propofol-anesthetized dogs: thermodilution and Doppler. Vet Anaesth Analg 2010; 37: 401-408.
13. MacIsaac AI, McDonald IG, Kirsner KL, Graham SA, Gill RW. Quantification of mitral regurgitation by integrated Doppler backscatter power. J Am Coll Cardiol 1994; 24: 690-695.
14. Mosing M, Staub L, Moens Y. Comparison of two different methods for physiologic dead space measurements in ventilated dogs in a clinical setting. Vet Anaesth Analg 2010; 37: 393-400.
15. Petrie A, Watson P. Statistics for veterinary and animal science. Oxford: Blackwell Science. 1999: 170-173.
16. Proverbio D, Groppetti D, Spada E, Perego R. Comparison of the VIDAS and IMMULITE-2000 methods for cortisol measurement in canine serum. Vet Clin Pathol 2009; 38: 332-326.
17. Szaflarski NL, Slaughter RE. Technology assessment in critical care: understanding statistical analyses used to assess agreement between methods of clinical measurement. Am J Crit Care 1996; 5: 207-216.
18. Whittemore JC, Flatland B. Comparison of biochemical variables in plasma samples obtained from healthy dogs and cats by use of standard and microsample blood collection tubes. J Am Vet Med Assoc 2010a; 237: 288-292.
19. Whittemore JC, Flatland B. Comparison of complete blood counts in samples obtained from healthy dogs and cats by use of standard and microsample blood collection tubes. J Am Vet Med Assoc 2010b; 237: 281-287.
20. Zar JH. Biostatistical analysis. 4th ed. Upper Saddle River: Prentice Hall Inc. 1999: 129-131.