

## 산 · 학 · 연 논문

## ‘식품산업과 영양’ 학회지에서 사용된 통계방법에 대한 고찰

명성민<sup>1</sup> · 안정좌<sup>2\*</sup><sup>1</sup>중원대학교 보건행정학과<sup>2</sup>중원대학교 식품공학과

## Review of Statistical Methods in the Food Industry and Nutrition

Sungmin Myoung<sup>1</sup> and Joungjwa Ahn<sup>2\*</sup><sup>1</sup>Department of Health Administration and <sup>2</sup>Department of Food Science and Technology,  
Jungwon University, Chungbuk 28024, Korea

## 서론

통계학은 본질적으로 자료 분석에 적용되는 수학의 한 분야이다. 일반적으로 식품과학 분야에서는 실험 작업의 계획, 분석 및 해석에 통계적 절차가 필요하다고 알려져 있다(1). 이러한 통계적 절차는 식품 및 음료의 물리화학적, 관능적(sensory) 및 미생물학적 데이터를 기반으로 하는 결과 도출 및 최적화 결과로서의 특성 변화 고찰, 자료 분석이 필요한 변수들 간의 연관성 조사 등이 해당된다. 또한, 적절한 통계적 방법을 사용하여 반응에 영향을 주는 요인을 추론하고 평가할 수도 있다. 예를 들어 과일 추출물의 농도를 증가시켜 제품의 산성도를 평가하거나, 다른 용량의 식품 추출물로 처리된 동물연구에서 염증지표 및 산화 스트레스 등과 같은 생화학적 지표들에 대한 영향 평가 등이 있다.

위의 내용들을 일반화하면, 통계학은 기술통계(descriptive statistics)와 추측통계(inferential statistics)로 구분할 수 있다. 기술통계의 역할은 실험결과를 정확하게 요약하는 것이며, 추측통계는 실험결과(데이터)를 대상으로 일반적인 사실로 의사결정을 하는 것으로서, 이러한 통계기법을 이용한 실험 및 평가는 연구보고서의 객관적인 신뢰성을 높이기 위한 필수적인 방법이라고 알려져 있다(2,3).

식품과학 분야에서 통계기법 적용 및 수학적/통계적 모형의 적용은 급격하게 증가하는 추세이다. Nunes 등(4)에 의하면 1997년부터 2014년까지 ScienceDirect에 등재된 논문들을 대상으로 ‘통계분석(statistical analysis)’을 키워드로 하여 검색한 결과 1997년에는 약 40,000건에 비해 2014년에는 190,000건으로 약 4.5배 증가하는 추세로 나타났다(4). 우리나라의 경우 한국학술정보(Koreanstudies Information Service System; KISS)에

서 ‘통계분석’ 키워드로 학술지만을 고려하여 검색한 결과 1996년에는 147건이었으나, 2016년에는 1,563건으로 약 10배 이상 증가하는 추세로 나타났으며, 세부적으로는 2013년까지는 급격하게 증가하고 2014년부터는 비슷하게 유지되는 추세로 나타났다(Fig. 1). 이러한 경향을 갖는 가장 큰 이유는 컴퓨터의 가격이 낮아짐과 동시에 복잡하고 많은 양의 실험결과를 통계 분석하기 위한 성능이 향상되었기 때문이라고 할 수 있다. 이러한 의미에서 통계학자 및 전산학자들은 의학, 화학, 농업, 생물학, 식품과학 등과 같은 여러 학문의 문제를 해결하기 위한 새로운 방법을 최적화하고 개발하여, 이를 SAS, SPSS, R 등과 같은 통계 package로 구현하였다. 그러나 통계분석에 대한 관심 부족, 결과의 잘못된 해석 또는 부적절한 통계분석방법의 선택 등으로 인하여 연구목적에 부합하는 결과를 도출하지 못하는 경우가 많다. 또한, Passari 등(5)은 많은 연구자가 통계계산 프로그램에서의 중요한 통계적 개념과 결과를 이해하고 해석하는 데 어려움을 겪고 있다고 주장하기도 하였다(5). 이는 일반적으로 통

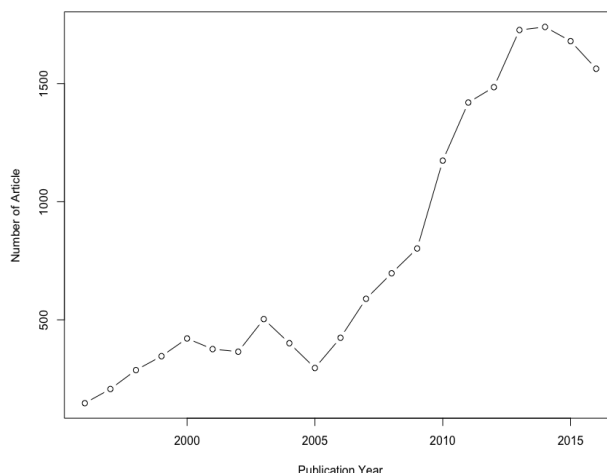


Fig. 1. Number of published articles in KISS from 1996 to 2016 using the term ‘statistical analysis’.

\*Corresponding author

E-mail: jjahn@jwu.ac.kr, Phone: 043-830-8611

계 package의 사용과 통계 및 모델링 개념의 이해를 기반으로 하는 다학제적 접근방식을 요구하기 때문에 이에 대한 필요성이 최근 많은 연구에서 제안되어지고 있다.

이에 본 논문에서는 ‘식품산업과 영양’ 학회지에 게재된 논문들의 주요 통계학적 사용방법들을 정리하고 통계적 오류의 종류 및 빈도를 제시하여, 학회 내부로는 올바른 통계사용을 장려하여 좋은 연구 성과를 얻는 계기를 마련하고, 외부로는 본 학회지의 학술적 발전에 기여하고자 본 연구를 수행하였다.

## 연구대상 및 방법

본 연구는 2007년부터 2017년도까지 11년간 발간된 총 22권의 ‘식품산업과 영양’ 학회지를 대상으로 조사하였다. 통계적 방법의 유형과 활용현황은 학회지에 게재된 155편의 논문을 대상으로 하였는데, 155편에 해당되는 세부 제시는 ‘특집’과 ‘산학연 논문’에 해당되는 논문들이었으며, ‘연구소 탐방’ 및 ‘인턴수기’, ‘기타’로 분류된 게시물은 평가 대상 논문에서 제외하였다.

게재된 논문에 사용된 통계적 활용현황은 기술통계와 추측 통계적 기법을 적용한 경우 그 통계적 기법을 조사하였으며, 설문 혹은 실험을 통하여 획득한 자료가 아닌 자료출처를 통하여 표로 제시한 논문은 비록 통계적 기법을 적용하였더라도 산정하지 않았다. 게재된 논문에 대한 사용정도를 빈도 분석하였으며, 이를 토대로 본 논문지에서 범한 몇 가지 사용오류를 정리하였다. 또한 식품과학 분야에서 주로 사용되는 통계적 방법들을 요약하여 그 사용방법을 고찰하였다.

## 학회지에 사용된 통계기법의 분류

2007년부터 2017년까지 11년간 게재된 155편 중 통계적 기법이 사용된 논문은 총 46편이었으며, 이들 논문에서 사용된 기술통계 기법은 28편, 추측통계기법이 사용된 논문은 18편이었다(Table 1). 본 학회지에 사용된 통계기법 중 t-검정(Student's t-test) 및 분산분석(Analysis of Variance; ANOVA)이 가장 많이 사용되었으며, 그다음으로 주성분 분석(Principal Component Analysis; PCA)으로 나타났다. 기타 항목으로는 카이제곱 검정(chi-square test), 선형계획법, 상관분석(correlation

analysis) 등이 있었다. 한 개의 논문에 A 통계방법이 n번 사용되었다면 통계방법 A는 해당 논문에 1회 사용되었다고 계산하였다.

위와 같이 식품과학 분야에서 주로 사용되는 통계분석 기법은 자료의 통계적 분포, 표본크기, 동질성에 대한 평균비교기법을 이용한다. 일반적으로 집단에 대한 평균검정방법을 모수적 검정(parametric test)이라고 하는데, 이는 자료가 정규분포를 따르며 동일한 분산을 가진다는 기본가정이 존재한다. 만약 이에 대한 가정이 만족하지 않는다면 비모수적 검정(nonparametric test)방법을 적용한다. Fig. 2는 Granato 등(1)에 의하여 제시된 식품과학 분야에서 집단에 대한 비교방법을 적용할 때의 통계적 절차를 제시한 것이다. 본 논문에서는 본 학회지에서 가장 높은 빈도로 나타난 t-검정과 분산분석을 중심으로 다음과 같이 요약하였다.

### t-검정(t-test)

t-검정법은 일반적으로 두 집단의 평균이 동일하다고 결론 내리는 것이 타당한지를 결정할 때 자주 쓰이는 방법이다. t-검정을 수행하기 위해서는 기본적으로 비교하고자 하는 두 집단의 값이 연속형 자료이면서 정규분포를 따른다는 가정이 필요하다.

두 집단에 대한 t-검정법은 자료의 유형에 따라 2가지 방법으로 분류될 수 있는데, 독립표본 t-검정(independent two sample t-test), 스튜던트 t-검정(Student's t-test) 또는 두 집단에 대한 t-검정(two-sample t-test)과 쌍을 이룬 자료(짝지어진 자료)에 대한 t-검정(paired t-test) 또는 대응표본 t-검정으로 나뉜다. 먼저 독립표본 t-검정의 경우 선행작업으로 두 집단에 대한 모분산이 같은지에 대한 검정을 실시하여야 한다. 등분산성 여부에 따라서 등분산일 때의 독립표본 t-검정에 대한 검정통계량 적용공식과 이분산일 때의 독립표본 t-검정에 대한 검정통계량 적용공식이 다르므로 주의하여야 한다. 예를 들어 Kim(6)의 일본간장의 맛 성분배에 대한 연구에서, *Aspergillus sojae*와 *Aspergillus oryzae* 두 종류의 간장 국균(麹菌)에 대한 글루탐산과 젖산수치를 비교하는 검정방법이 독립표본 t-검정을 적용한 경우이다.

쌍을 이룬 자료에 대한 t-검정의 경우는 동일한 개체를 대상으로 짝지어진 자료, 즉 실험단위를 동질적인 쌍으로 묶은 다음 각 쌍에서 모은 관측값의 차이를 추론하는 방법

Table 1. Classification of articles in the Food Industry and Nutrition (2007~2017)

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Total number of articles	17	9	10	14	17	12	13	14	14	18	17
Descriptive Statistics	9	3	0	1	2	4	4	1	3	0	1
Inferential Statistics	4	0	2	0	4	0	2	2	1	1	2
t-test	2	0	1	0	2	0	1	0	0	0	1
ANOVA	2	0	0	0	2	0	0	1	1	0	1
PCA	0	0	0	0	0	0	0	1	0	1	0
Etc	0	0	1	0	0	0	1	0	0	0	0

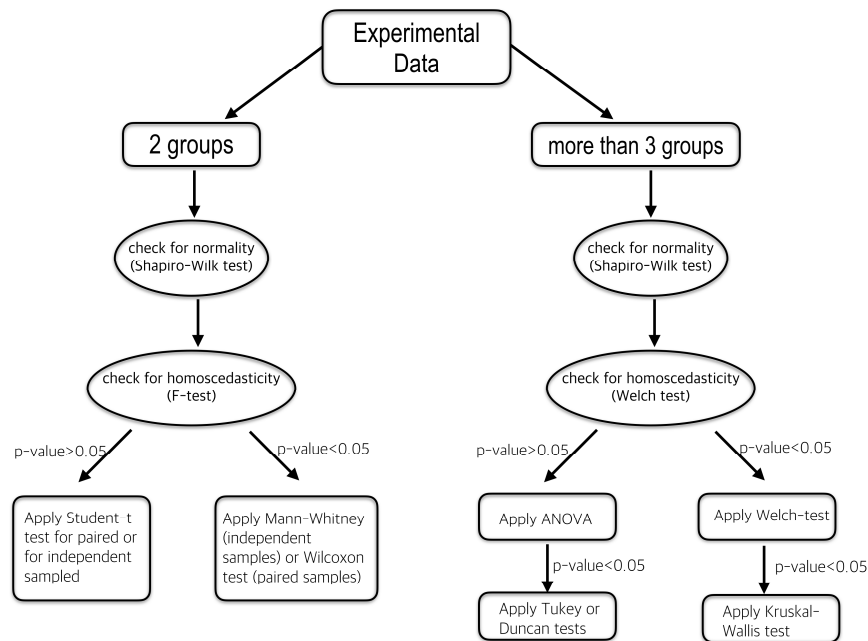


Fig. 2. Statistical procedures and test to compare two or more samples in food science.

이다. 예를 들어 마카(maca; *Lepidium meyenii*)의 효능을 확인하기 위하여 24~44세의 건강한 남성들을 대상으로 마카섭취 전후의 정액량, 정자수, 정자 움직임의 차이여부를 검정한 것이 대응표본 t-검정을 적용한 경우이다(7).

### 분산분석

두 집단의 경우 평균의 비교는 그 차이를 확인하여 모평균이 동일한지 확인할 수 있으나 집단이 3개가 있을 경우  ${}_3C_2=3$ 개의 쌍을 고려해야 하므로 전체 검정 신뢰도는 떨어지는 문제가 발생한다. 그러므로 다른 방법을 고려해야 하는데, 분산분석은 비교 대상 집단이 3개 이상일 때 그 평균들이 같은지 비교하는 방법이다. 특성값에 영향을 미칠 것이라고 여겨지는 효과를 인자(factor)라고 하며, 그 인자가  $k$ 개의 수준을 가진다고 할 때 이에 대한 분석방법을 일요인 분산분석(One-way ANOVA)이라 한다. 예를 들어 감각적 특성에 대한 시료 간의 유의적 차이를 알아보기 위해 수행된 연구에서 깨뜨려 먹는 유자푸딩 원형을 'Low Yogurt Low Yuzu(LYLY)', 'Low Yogurt Medium Yuzu(LYMY)', 'Medium Yogurt High Yuzu(MYHY)', 'High Yogurt Low Yuzu(HYLY)', 'High Yogurt Medium Yuzu(HYMY)', 'High Yogurt High Yuzu(HYHY)'로 나누어 소비자 전체 기호도에 대한 차이여부를 확인하였다면(8), 인자는 '유자푸딩 원형'이 되며 수준은 6개가 된다. 즉 이는 6개의 유자푸딩 원형에 대한 소비자 전체 기호도의 평균을 검정하는 문제로 귀결되며, 일요인 분산분석을 적용할 수 있다. 분산분석의 이론적 배경은 집단 간의 표본평균의 차이에 대한 변동(집단 간 변동)과 동일 집단 내에서의 변동(집단 내 변동)을 구하고, 이에 대한 차이를 비교하여 의사결정을 하는 것이다. 만약 분산분석

의 결과가 통계적으로 유의한 경우 특정집단 간의 비교를 수행할 수 있는데, 이를 사후검정(post-hoc test)이라 한다. 여러 가지 사후검정방법들 중에서 식품과학 분야에서는 Duncan의 방법 또는 Tukey의 방법을 많이 이용하는데, 일반적으로 반복수가 같은 경우에는 Tukey의 방법이 가장 좋은 것으로 알려져 있다(1). 한편 비모수적 방법인 Kruskal-Wallis 방법은 세 집단 이상에 대한 평균비교가 아닌 중위수에 대한 검정방법으로 알려져 있으며, 마지막으로 일요인 분산분석에서 등분산가정이 만족되지 않는 경우 Welch 방법을 고려할 수 있다(1,9).

또 다른 분산분석 방법으로 반복측정 자료에 대한 분산분석(repeated measures ANOVA)이 있는데, 이 방법은 어떤 실험을 실시할 때 동일한 실험대상에 대하여 실험조건을 3회 이상 다르게 하여 반복적으로 측정한 경우에 조건 간의 평균 차이여부를 검정할 때 사용한다.

### 범주형 자료분석

범주형 자료(categorical data)는 개체의 특성에 대한 관측이 유한개의 범주로 주어지는 변수로서 성별(남, 여), 혈액형(A형, B형, O형, AB형) 또는 환자의 치유상태(좋은, 나쁜) 등을 의미한다. 이에 대한 검정방법으로 카이제곱 검정이 대표적인데, 이는 개체의 두 특성이 범주형 자료일 때 이를 표로 요약한 분할표(contingency table)를 대상으로 두 변수 간의 동질성(homogeneity)을 확인할 때 이용하는 방법이다. 일례로 마카 추출물의 효능을 확인하기 위하여 21~56세의 건강한 성인 남자를 대상으로 4, 8, 12주간 대조군인 위약(placebo)과 마카 추출물을 섭취한 실험군에 대하여 성욕여부를 조사한 연구가 카이제곱 검정을 수행한 경우인데, 실험군/대조군은 범주형

변수의 형태이고 성육여부 또한 마찬가지로 범주형 변수의 형태이므로 이를 분할표로 요약할 수 있으며, 이에 대한 검정을 4, 8, 12주 별로 카이제곱 검정을 제시하였다(7).

### 상관분석

상관분석은 두 변수 사이의 관련 유무 또는 관련 정도를 알아보는 통계적 분석방법으로, 이는 선형적으로 얼마나 연관성을 가졌는지 알아보는 분석이며, 일반적으로 피어슨 상관계수(Pearson correlation coefficient)를 이용한다. 이변량 상관계수는 선형적 관계를 -1과 1 사이의 수치로 나타내는데, 양의 값을 가지는 경우 ‘양의 상관관계를 가진다’로 하며, 음의 값을 가지는 경우 ‘음의 상관관계를 가진다’고 한다. 예를 들어 Kim 등(10)의 국내 시판 발효액의 품질현황에 대한 연구에서 발효액의 대표특성인 총산, 당도, 알코올 간의 품목 간 상관성을 확인한 경우가 상관분석을 적용한 경우이다. 그 결과 총산과 당도 간의 상관계수 값이 0.175로 나타났는데, 이는 총산이 증가하면 당도도 증가하는 양의 상관관계가 존재한다고 할 수 있다. 한편 총산과 알코올의 상관계수 값은 -0.333으로서 총산이 증가하면 알코올은 감소하는 음의 상관관계가 존재한다고 할 수 있다. 또한 당도와 알코올 간에는 -0.421로서 음의 상관관계가 존재하였다. 주의할 점은 Pearson의 상관계수는 모집단의 분포가 정규분포일 때 적용하며, 그렇지 않은 경우 Spearman의 순위 상관계수를 사용한다.

### 주성분 분석

주성분 분석은 다변량 자료 분석(multivariate data analysis) 기법중 하나로, 여러 변수의 변량을 주성분(principal component)이라고 하는 서로 상관성이 높은 변수들의 선형결합(linear combination)으로 만들어 기존의 큰 자료군을 압축 요약하는 기법이다. 즉 여러 변수 간에 내재하는 상관관계, 연관성을 이용해 소수의 주성분으로 차원을 축소함으로써 데이터를 이해하기 쉽게 하는 것이다. 예를 들어 진통민속소주의 향기성분 분석을 위하여 15종의 증류주에 대하여 SPE(solid phase extraction)로 전처리한 시료를 GC-MS로 분석하여 ester 및 aldehyde 성분 13종의 정량값을 이용하여 주성분 분석을 통

하여 2개의 주성분으로 요약/분리한 결과 78.28%의 기여율로 나타났고, 민속소주는 제1 주성분 점수가 비슷하였으나 제2 주성분 점수가 편차를 보여 고유한 특성이 존재함을 확인하였다(11). 추가적으로 주성분 분석을 통하여 연관성이 높은 변수를 차원을 축소하여 군집분석 등을 통한 다른 다변량 분석 기법 등을 적용할 수 있다.

## ‘식품산업과 영양’ 학회지에 사용된 통계적 방법의 오류 사례

2007~2017년 동안 조사한 ‘식품산업과 영양’ 학회지에 게재된 논문의 통계적 오류의 사례를 정리하였다(Table 2). 빈도를 제시하지 않았으나 가장 많은 빈도를 차지한 오류 사례는 ANOVA로, 이 중 특히 많은 오류를 보인 것은 사후검정 결과만을 제시하고 분산분석을 수행한 검정 통계량 또는  $P$ -값을 제시하지 않은 것이다. 분산분석의 경우 우선적으로 고려해야 할 사항은 분산분석 수행 시의  $F$ -검정 결과이지 사후검정 결과는 아니라는 점이다.  $F$ -검정 결과가 유의했음 경우 사후검정을 실시하는 것이므로 사후검정에 더 큰 의미를 부여하면 안 된다. 그다음의 오류사항으로 분산분석 시 Student's  $t$ -검정을 3번 수행한 경우인데, 이는 위 절에서 제시하였듯이 전체적인 검정 신뢰도를 하락시키기 때문에  $F$ -검정을 실시하여야 한다는 것이다.

Student's  $t$ -검정을 적용한 경우 대응표본 자료인데 독립적인 두 집단으로 간주하여  $t$ -검정을 수행했던 사례가 있었으며, 독립적인 두 집단에 대한  $P$ -값을 제시하였으나 검정방법을 제시하지 않은 경우가 있었다. Table 2에 제시하지는 않았지만 추가적으로 잘못된 통계분석방법을 기술한 경우(반복측정자료에 대한 분산분석 기법인데 Student's  $t$ -검정으로 잘못 기재)가 있었으며, 또한 분석에 이용한 통계 소프트웨어를 기술하지 않은 경우, 가설검정 시 필수적으로 제시해야 하는 유의수준(significance level)을 제시하지 않은 경우도 존재하였다.

이 외로 본 학회지에는 없었지만, 많은 연구에서 나타날 수 있는 통계기법의 오류는 다음과 같다(12). 첫째로 통계기법의 생략 오류로서, 기초자료의 불충분한 기술 및 사용된 통계기법에 대한 타당성의 불충분한 검증 등이

Table 2. Type of errors for statistical methods in the Food Industry and Nutrition

Analysis method	Type of error
Student's $t$ -test	명칭오류 (Student's two-tailed $t$ -test) $P$ -value를 제시하였으나 분석기법을 제시하지 않음 paired 자료인데 independent 자료에 이용하는 방법 적용
ANOVA	유의성 차이는 있으나 분석기법 제시하지 않음 3 집단에 대한 평균비교인데 Student $t$ -test를 3번 시행 반복수가 같은 경우에 대한 사후검정 시 Duncan test 사용 Tukey 사후검정 명칭 오류 (Turkey test) 사후검정 결과만을 제시하고 분석기법에 대해서 제시하지 않음
Chi-square test	$P$ -value를 제시하였으나 분석기법을 제시하지 않음

있다. 둘째로, 통계기법의 시행 오류로서 대표값 및 산포의 정도에 대한 부적절한 기술, 불충분한 표본 수에서 카이제곱 검정을 사용하는 경우, 비모수적 자료를 모수적 기법으로 분석하는 경우, 표본에 대한 통계처리 없이 모집단에 대한 결론을 유추하는 경우 및 연구결과를 일반 집단에 적용하면서 그 한계를 언급하지 않은 경우 등이 있다.

## 결론

식품과학 분야의 실험 논문의 증가와 더불어 확보된 결과의 정확한 고찰과 활용을 위해서 다양한 통계기법은 매우 유용한 도구이다. 일반적으로 통계적 방법의 타당성을 판단하기 위하여 고려해야 할 주요 사항은 다음과 같다. 첫째, 연구대상이 표본일 경우 통계검정을 통하여 결론이 일반화되었는지, 둘째, 결론 도출 시 적절한 통계방법이 적용되었는지, 셋째, 변수의 분포에 맞는 통계량이 사용되었는지 등이다(12,13). 본고는 ‘식품산업과 영양’ 학회지에서 2007~2017년 동안 투고된 논문들을 대상으로 식품과학 분야에서 사용되는 통계적 분석기법에 대하여 개괄적으로 소개하고, 사용 현황 및 주로 나타난 오류에 대하여 정리하였다. 본 학회지는 주로 특집과 산학연 논문으로 구성되는 특징을 지니고 있다. 따라서 통계 기법을 활용한 적은 수의 논문을 고찰하여 그 결과를 적용하기에는 어려움이 있을 것으로 생각된다. 그러나 본고는 연구자들에게 일반적으로 많은 오류를 범하는 통계처리 사례를 소개하고, 실험데이터 분석 시 적합한 통계 방법을 제시하고자 하였다. 통계학을 전공하지 않은 연구자의 입장에서 볼 때 통계적 방법을 정확하게 적용하는 것이 쉬운 일은 아니지만, 기법 적용의 오류를 개선하기 위해서는 연구자 스스로의 노력과 연구계획단계에서부터 통계 전문가의 조언을 구하여 진행하는 방법도 적극적으로 활용해야 할 것이다.

## 참고문헌

1. Granato D, de Araújo Calado VM, Jarvis B. 2014. Observations on the use of statistical methods in Food Science and Technology. *Food Res Int* 55: 137-149.
2. Dawson B, Trapp RG. 1994. *Basic and clinical biostatistics*. McGraw-Hill, New York, NY, USA. p 20-248.
3. Park TY, Heo TY, Shin BC. 2010. Statistical errors of articles published in the Journal of Oriental Rehabilitation Medicine (I). *J Oriental Rehab Med* 20: 105-130.
4. Nunes CA, Alvarenga VO, de Souza Sant'Ana A, Santos JS, Granato D. 2015. The use of statistical software in food science and technology: Advantages, limitations and misuses. *Food Res Int* 75: 270-280.
5. Passari LMZG, Soares PK, Bruns RE, Scarminio IS. 2011. Statistics applied to chemistry: ten common doubts. *Quim Nova* 34: 888-892.
6. Kim ND. 2007. Trend of research papers on the soy sauce tastes in Japan. *Food Industry and Nutrition* 12(1): 40-50.
7. Kwon D, Lee SH, Lee S. 2009. Biological effects of Maca (*Lepidium meyenii*). *Food Industry and Nutrition* 14(2): 19-24.
8. Seo JH, Lee SY, Kang BA, Jeon EJ, Kim YS. 2014. Development of new types of Yuzu dessert and market potential assessment. *Food Industry and Nutrition* 19(2): 28-34.
9. Nam CM, Chung SY. 2012. Statistical methods for medical studies. *J Korean Med Assoc* 55: 573-581.
10. Kim YW, Im HE, Jeong ST, Yeo SH, Baek SY, Park HY. 2013. Quality status of commercial fermented liquid in Korea. *Food Industry and Nutrition* 18(2): 23-32.
11. Lee DS, Park HS, Kim K, Lee TS, Noh BS. 1994. Determination and multivariate analysis of flavour components in the Korean folk soju using GC-MS. *Korean J Food Sci Technol* 26: 750-758.
12. Kim YS, Oh HS, Lim EM. 2012. A review of statistical methods in the Journal of Oriental Obstetrics & Gynecology. *J Orient Obstet Gynecol* 25: 70-78.
13. Corry JEL, Jarvis B, Passmore S, Hedges A. 2007. A critical review of measurement uncertainty in the enumeration of food micro-organisms. *Food Microbiol* 24: 230-253.