DOI:10.13471/j.cnki.j.sun.yat-sen.univ(med.sci).1999.0102

·论著写作·

近期医学论文中常见统计错误及其纠正

方积乾[1] 凌 莉[1] 张敏瑞[2]

(1 中山医科大学卫生统计学教研室 2 中山医科大学学报;广州,510089)

摘 要 **目的**:剖析近期医学论文统计学方面的常见错误。**方法**:围绕7个实例,从反面展示不恰当做法,解释理由,并从正面提议规范的方法。**结果**:用本文提出的办法可避免所犯统计错误。**结论**:向研究人员普及中、高级统计学知识和加强统计咨询服务是医学和统计学双方发展的需要。

主题词 统计学;论文,学位;随机对照实验;方差分析;剂量效应关系,药物;重复测量资料①;多维列联表①;随访研究

中图号 O 213

Frequently Appearing Statistical Mistakes in Recent Medical Articles and Relevant Suggestions

Fang Jiqian Ling Li Zhang Minrui

(1 Department of Medical Statistics 2 Department of Acad J SUMS, Sun Yat-sen University of Medical Science, Guangzhou, 510089)

Abstract **Objective**: Analyze the frequently appearing mistakes on statistics in recent medical articles. **Method**: Show the inappropriate approaches, explain the reason and suggest the proper methods for seven typical examples. **Result**: The concerned mistakes can be avoided with the proposed methods. **Conclusion**: To popularize medium and high level of statistical knowledge and strengthen the service of statistical consultation are the needs for further development of medical sciences as well as statistical science.

Subject headings statistics; dissertations, academic; randomized controlled trials; analysis of variance; dose-response relationship, drug; repeated measuresdata①; multiple dimensional contingency table①; follow-up; studies

随着统计学知识的普及,公开发表的医学论文中有关初等统计学的错误已不多见。但由于现代医学研究的深入,医学论文中又时常出现一些新的统计学错误,影响着高质量医学研究资料的有效与充分利用。本文针对作者近几年在医学论文评审中看到的具有共性的统计学问题,借助7个实例展示不恰当做法,阐明理由,并提议正确方法,希望对医学论文的撰写和编审有所帮助。

1 随机区组设计:不可用成组分析方法

1.1 例1

设有5件标本Ⅰ,Ⅱ,Ⅲ,Ⅳ和Ⅴ,每一件均匀分成4份,随机分配给4种处理*A*,*B*,*C*和*D*,试比较4种处理间的差别(表1)。

表1 随机区组设计下的实验结果

Table 1 The results of randomized block designed experiment

Block	Treatment			
	A	*B*	*C*	*D*
Ⅰ	0.80	0.36	0.17	0.28
Ⅱ	0.74	0.38	0.42	0.36
Ⅲ	0.31	0.20	0.38	0.25
Ⅳ	0.48	0.18	0.44	0.22
Ⅴ	0.76	0.26	0.28	0.13

① 自由词

1.2　不恰当做法

视纵向 4 列为 4 组数据，进行 4 组间比较。①多次 t 检验：分别对 A 与 B，A 与 C，A 与 D，B 与 C，B 与 D，C 与 D 作两组比较的 t 检验，得 P 值依次为 0.02，0.03，0.01，0.37，0.63 和 0.19，据此认为 B，C，D 与 A 之间的差异具有统计学意义，但 B 与 C，B 与 D，C 与 D 的差异均无统计学意义。②单因素方差分析：同时比较 A，B，C，D 4 个处理组均数间的差异，得 $F=0.55$，$P=0.59$；据此认为 4 个处理组均数的差异无统计学意义。

1.3　理　由

①辜负设计的苦心：此资料是随机区组设计，同一件标本为一个区组，同一区组内 4 种处理有较好的可比性；无视区组的存在，组内个体间变异较大，识别差异的能力降低。②不能翻来覆去比较：如果规定每次 t 检验犯第一类错误的概率为 0.05，则粗略估计，6 次 t 检验犯第一类错误的概率$=1-(1-0.05)^6=0.27$。

1.4　正确作法

随机区组资料的方差分析[1,2]：将处理组间的差别与扣除区组间变异后的随机误差进行比较，识别差异的能力提高。本例，处理组间，$F=8.23$，$P=0.003$，差别有统计学意义；区组间，$F=1.21$，$P=0.36$，差别无统计学意义。

2　析因设计：不能用单因素方差分析

2.1　例 2

收集甲状腺术中正常甲状腺组织做成 16 份标本，随机分为 4 个处理组。将处理后的标本分别行免疫组织化学染色，作单位面积 HLA-Ⅱ抗原阳性细胞计数。得数据如下：第 1 组（经体积分数为 95%空气和 5% CO_2 培养 7 d）：2.5，2.4，2.6，2.5。第 2 组（经体积分数为 95%空气和 5%CO_2 培养 14 d）：1.8，1.9，1.7，1.8。第 3 组（经体积分数为 95% O_2 和 5%CO_2 培养 7 d）：2.2，2.3，2.1，2.2。第 4 组（经体积分数为 95%O_2 和 5%CO_2 培养 14 d）：1.9，1.8，1.7，1.9。

2.2　不恰当做法

单因素方差分析方法：直接比较 4 组资料间的差别，得 $F=60.00$，$P=0.00$；据此认为 4 组间差别具有统计学意义。

2.3　理　由

典型的两因素 2 水平的析因设计，可将数据整理，其中纵向两列和横向两列均有可比性，不分析各因素的效应，有负设计初衷；如用单因素方差分析，不能扣除各因素的效应而得到单纯由误差引起的变异，4 组间比较不甚敏感（表 2）。

表 2　析因设计下的实验结果

Table 2　The results of factorial designed experiment　($\bar{x}$)

Factor B	Factor A	
	φ (95% air and 5% CO_2)	φ (95% O_2 and 5% CO_2)
7 d	2.5(Group 1)	2.2(Group 3)
14 d	1.8(Group 2)	1.9(Group 4)

2.4　正确作法

两因素 2 水平析因设计的方差分析[1,2]：除讨论 4 组资料间的差别外，还可进一步分析处理的主效应和交互效应。①主效应：指某一因素各水平间的平均差别。本例，A 因素的主效应$=(2.2+1.9)/2-(2.5+1.8)/2=-0.1$；$B$ 因素的主效应$=(1.8+1.9)/2-(2.5+2.2)/2=-0.5$。②交互效应：若某因素的单独效应随另一因素的变化而有较大变化，则称这两个因素存在交互效应。本例，AB 的交互效应$=[(1.9-1.8)-(2.2-2.5)]/2=0.2$。③检验：A 的主效应，$F=1.37$，$P=0.26$；B 的主效应，$F=47.13$，$P=0.00$；$A$ 与 B 交互效应：$F=0.03$，$P=0.87$。可见，只有 B 因素的效应具有统计学意义。

3　剂量反应关系：不能用单因素方差分析

3.1　例 3

有人欲分析蛇毒因子（CVF）剂量对血液白细胞噬菌率的影响，得表 3 的数据。

表 3　蛇毒因子（CVF）剂量对血液白细胞噬菌率的影响

Table 3　The blood neutrophils bacteriophagic rate in different CVF dose

Group	$\frac{D(\text{CVF})}{(\mu g/kg)}$	n	neutrophils bacteriophagic rate(%) ($\bar{x}\pm s$)
1	0	6	60.0±17.0
2	10	6	57.0±15.2
3	20	6	54.0±16.6
4	40	6	51.0±17.2
5	80	6	48.0±16.0
6	160	6	45.0±16.4

3.2 **不恰当做法**

①用单因素方差分析：$F=0.701$，$P>0.5$，认为均数间差别无统计学意义。②作各组均数关于lg(剂量)的回归分析。回归方程为 Y(噬菌率)$=61.784-6.885$ lg(剂量$+1$)，决定系数为 $R^2=0.914$。

3.3 **理　由**

①多个剂量组比较答非所问：由多个剂量组的比较只能得知均数间是否有差异，有差异不等于有剂量反应关系，也未必能概括出是否有趋势和有什么样的趋势。②以均数回归易出假象：回归方程是否有统计学意义与反应的变异状况有关，以诸个体反应值的均数作回归计算掩盖了反应的变异性。

3.4 **正确作法**

①利用个体资料作回归分析：回归方程为 Y(噬菌率)$=61.784-6.885$ lg(剂量$+1$)，决定系数为 $R^2=0.095$。②适当变换：若反应和剂量间散点图呈曲线状，可对原数据作适当变换后作回归分析(图1,2)。

4 重复测量资料的分析：不能反复地作两两比较

4.1 **例** 4

在一系列规定时间点分别测量18名钩虫病所致贫血患者和18名对照的血糖浓度。

4.2 **不恰当做法**

①逐个时间点计算均数和标准差。②均数连线，并附“误差线”。③翻来覆去作两组比较。

4.3 **理　由**

①均数连线无代表性：各个体的曲线形状和位置互不相同，各时间点的均数连成折线并不能很好地代表诸个体。②前后观察值有相关性：对任一个体，前后观察值之间是相关的，不同时间点的数据并非互相独立的多组资料；各时间点的两组比较结果也非互相独立。③难下综合结论：各时间点比较的结果常参差不一，难以综合出两组间究竟有否差异的结论。

4.4 **正确作法**

①重复测量资料的特定方法：对任一个体按时间顺序反复进行一系列同样的测量，所得资料称为重复测量(repeated measures)资料；这类资料具有时间依赖性，统计方法不同一般[2,3]。②特征指标法：若反应曲线均呈现一个高峰，则可采用曲线下面积、峰高、达峰时等取代原始数据；若反应曲线均呈上升(或下降)趋势，则可采用升(或降)的速率、最终值、最终值与最初值之差、达到某特定值所需时间等取代原始数据。③其他：如重复测量资料的方差分析、多元方差分析和特殊模型等方法均可选用。

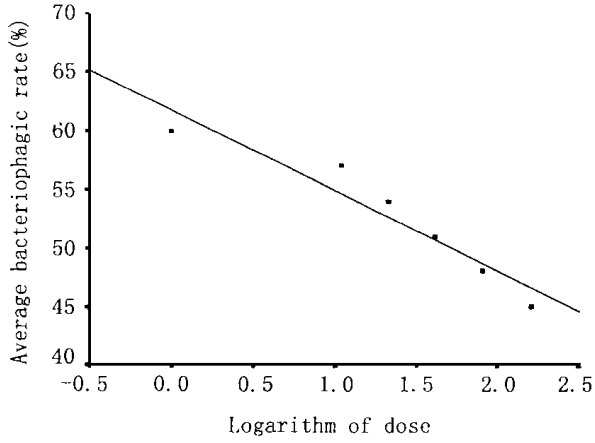

图1　均数关于剂量的回归

Fig·1　Regression of average bactcriophagic ratc on logarithm of dose

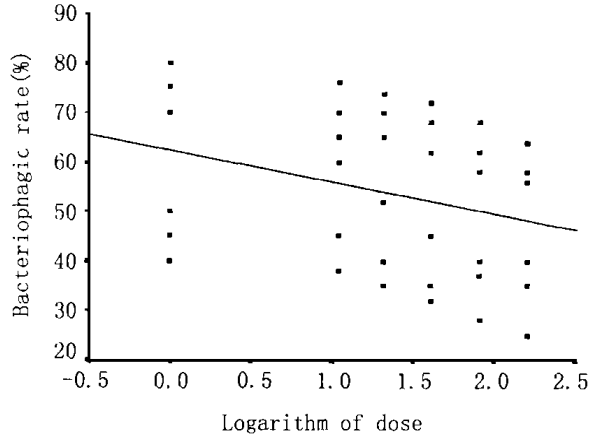

图2　原数据关于剂量的回归

Fig·2　Regression of bacteriophagic rate on logarithm of dose

5 多维列联表资料的分析：不能盲目地作合并分析

5.1 **例** 5

某地区呼吸系统疾病的患病率较高，有人怀疑与当地室内点香的习惯有关，经抽样调查，得表4中的数据。

5.2 **不恰当做法**

盲目合并后作 χ^2 检验：将多维列联表盲目合并为多个二维列联表，进行 χ^2 检验。本例，将各年

龄组合并,形成一个患病状况 (A) 与点香状况(B)的 2×2 列联表,经 χ^2 检验,认为患病与否和点香无关;将点香与不点香合并,形成一个患病状况(A)与年龄状况(C)的 2×4 列联表,经 χ^2 检验,认为患病机会和年龄有关。

表 4　呼吸道疾病与点香、年龄的观察资料

Table 4　The data of respiratory disease, incensation and age

Age	A_1 (respiiratory disease)		A_2 (no respiratory disease)		OR	P
	B_1 (incensation)	B_2 (noincensation)	B_1 (incensation)	B_2 (noincensation)		
C_1 (<20)	360	30	400	40	1.20	0.470
C_2 (20,40)	140	30	200	60	1.40	0.190
C_3 (40,60)	290	60	150	40	1.29	0.280
C_4 (≥60)	110	60	62	58	1.72	0.030
Total	900	180	812	198	1.22	0.081

5.3　理　由

①合并与分层结果相悖:若按年龄分层,在年龄组 C_1、C_2 和 C_3 中,患病状况与点香状况的 3 个 2×2 列联表分别有优越比(odds ratio, OR) OR_1 = 1.20, OR_2 =1.40, OR_3 =1.29,经统计检验, P 值分别为 0.47,0.19,0.28,对于这 3 个年龄组尚不能认为患病与点香有关。而在年龄组 C_4 组中, OR_4 =1.72, P 值为 0.03,说明这个年龄组中,患病与点香有关。然而,合并 4 个年龄组后,却得到 OR = 1.22,经统计检验,不能认为患病与点香有关。实际上,分层的结果是正确的。②年龄结构不同,患病率不可比:表 4 中,从 1 712 例点香者和 378 例不点香者的年龄结构来看,点香者年轻的多,不点香者年长的多,这两组的患病率不可比较。

5.4　正确作法

①logistic 回归[1]:为探讨患病机会和点香、年龄两个因素的关系,可将患病这个二值变量视为反应变量,点香和年龄视为解释变量:如欲讨论点香和年龄的交互效应,亦可将此二变量的乘积定义为另一个解释变量。本例分析的结果(表 5),年龄效应和点香效应皆有统计学意义。②对数线性模型:离散型变量的多维列联表资料可采用频数的对数线性模型来讨论变量间的关系[1,4]。本例中,利用该模型来平等地讨论 3 个因素的不同状况对列联表各个格子中频数的影响,得到极大似然方差分析表(表 6)。由最后一列可见,年龄和点香效应均有统计学意义,与 logistic 回归的结果一致。此外,年龄与患病、点香与患病的交互效应皆有统计学意义,表明不同年龄组及点香与否患病的机会均不同。一般说来,仅当变量 1 和变量 2 的交互效应为零时,方可将变量 2 的各水平合并后讨论变量 1 与其他变量的联系。本例中,由于年龄与点香存在交互效应,不可合并各年龄组而讨论点香与患病的关系。

表 5　用 logistic 回归分析例 5 资料的结果

Table 5　The results of logistic regression for e·g·5

Source of variation	υ [1)]	χ^2	P
Constant	1	0.22	0.638
Age	3	47.41	0.000
Incensation	1	7.27	0.007
Age×incensation	3	1.20	0.754

1) DF, degree of freedom

表 6　用对数线性模型分析例 5 资料的结果

Table 6　The results of log-linear model for e·g·5

Source of variation	υ [1)]	χ^2	P
Age (A)	3	37.40	0.000
Incensation (I·)	1	520.80	0.000
$A \times I$	3	138.65	0.000
Disease (D)	1	0.22	0.638
$A \times D$	3	47.41	0.000
$I \times D$	1	7.27	0.007
$A \times I \times D$	3	1.20	0.754

1) DF, degree of freedom

6　诊断效果的评价:应用要小心

6.1　例 6

有人用已确诊的患者和非患者各 93 名的资料来考核一种新诊断方法,结果如表 7。

表 7 诊断效果考核资料

Table 7 Evaluation of diagnostic test

Situation of disease	Diagnostic test		Total
	$T+$	$T-$	
$D+$	92	1	93
$D-$	64	29	93
Total	156	30	186

6.2 不恰当做法

①用灵敏度(sensitivity, *sen*)和特异度(specificity, *spe*)评价诊断效果:*sen* $= P(T+/D+) = 92/93 = 0.989$, *spe* $= P(T-/D-) = 29/93 = 0.312$,因灵敏度较高而认为该方法可用于发现病人。②计算阳性预测价值(positive predictive value)和阴性预测价值(negative predictive value):阳性预测价值$= P(D+/T+) = 92/156 = 0.590$,阴性预测价值$= P(D-/T-) = 29/30 = 0.967$,因阴性预测价值较高而认为该方法可用于排除非病人。

6.3 理 由

①灵敏度和特异度不反映诊断效果:灵敏度是已确诊为患者的前提下该方法呈现阳性的概率,特异度是已肯定为非患者的前提下该方法呈现阴性的概率。两者不能直接说明该方法可否用于诊断实践。②欲反映诊断效果需采用阳性预测价值和阴性预测价值:阳性预测价值是某方法呈现阳性的前提下确系患者的概率,阴性预测价值是某方法呈现阴性的前提下确系非患者的概率。两者与诊断效果挂钩,直接说明该方法可否用于诊断实践。③本例资料不能用来计算阳性预测价值和阴性预测价值:本例中,156 名诊断结论为阳性者和 30 名诊断结论为阴性者均不是相应总体中的随机样本。欲评价诊断效果,最好分别从诊断结论为阳性和阴性的总体中抽取随机样本,据此计算阳性预测价值和阴性预测价值。④需灵敏度(*sen*)、特异度(*spe*)和患病率三结合:若上述随机样本不易获得,而只能从病人和非病人两个总体中抽样,则必须了解目标人群中所论疾病的患病率(P_0),将三者结合起来方能算出阳性预测价值和阴性预测价值[1]。本例,假设由文献查得,患病率$=0.062$,则

阳性预测价值$=\dfrac{sen\ P_0}{sen\ P_0+(1-spe)(1-P_0)}=0.08$;

阴性预测价值$=\dfrac{spe(1-P_0)}{spe(1-P_0)+(1-sen)P_0}=0.998$。

照此看来,该诊断方法只能用作排除非病人,而不能用作发现病人。

7 随访资料的分析:要用专门的方法

7.1 例 7

有人随访分别用两种手术方案 *A* 和 *B* 治疗的两组患者,记录了术后生存时间(日)如下:治疗组 *A*(15 例):364, 141, 1 446$^+$, 950, 570, 836$^+$, 498$^+$, 173$^+$, 312, 1 540$^+$, 836$^+$, 570, 173, 401, 86;治疗组 *B*(14 例):505, 615$^+$, 570$^+$, 1 205$^+$, 296, 1 726$^+$, 1 190$^+$, 822$^+$, 1 375, 1 408$^+$, 1 493$^+$, 1 645$^+$, 1 570$^+$, 688(其中,右上方标有+号者并非实际生存时间,例如,1 446$^+$表示该患者活到 1 446,此后情况不详)。欲比较这两种手术方案的术后生存时间。

7.2 不恰当做法

按是否死亡整理,治疗组 *A* 死亡 9 例,生存 6 例;治疗组 *B* 死亡 4 例,生存 10 例;用四格表的确切概率法求得 $P=0.139$,认为两组术后生存率无统计学差异。

7.3 理 由

须统一所论的时间范围:可以认为,治疗组 *A* 中 6 个带有+号者在术后 173 日时都活着,但那 9 位死亡者却在术后 950 日前皆死去。所论的时间范围不一致;治疗组 *B* 中,死亡和生存所论的时间范围也不一致。因而,上述按四格表资料进行分析是错误的。

7.4 正确作法

log-rank 检验[1]。①在每一死亡时刻统计死亡与历险人数:为了在同一时段内比较两组的死亡率,将所有确切的术后死亡时刻排队;逐一列出治疗组 *A* 和 *B* 在相应时刻死亡数和该时刻前一瞬间历险的人数。②合理综合各时刻的可比资料:按两组历险人数的比例剖分期望死亡数,用 Pearson χ^2 统计量检验两组实际与期望死亡数的差别。本例,$\chi^2=4.675$,自由度$=1$,$\chi^2>\chi^2_{0.05(1)}$,故两组生存时间具有统计学差异。

8 讨 论

以上我们通过 7 个实例展示了近几年医学论

(下转第 322 页)

TG 高于正常值[2]。

3 讨 论

心脏 X 综合征由 Kemp[3]于 1973 年首次提出,至今为止,虽然有不少关于 X 综合征病因的解释,如微血管病变、冠状动脉痉挛、心肌病、代谢异常、交感神经张力增加、冠状动脉储留功能降低、血管内皮功能异常和冠状动脉粥样硬化等[4],但它的确切发病机制仍不清楚。由于技术的原因以前对冠状动脉粥样硬化可能导致 X 综合征的推测是根据临床随访观察提出的[4]。EBCT 可检出即便是少量的钙化,即可对冠状动脉钙化作出精确的定量,而冠状动脉钙化为冠状动脉粥样硬化的重要标志[1]。因此,EBCT 为临床无创观察冠状动脉粥样硬化情况成为可能。本文的 18 例 X 综合征患者中 17 例有不同程度冠状动脉钙化,提示冠状动脉粥样硬化可能为 X 综合征的发病原因之一。血清 Chol、TG 及 LDL—C 升高、HDL—C 降低可导致冠状动脉粥样硬化发生已得到公认[5]。我们观察显示,18 例临床诊断为 X 综合征的患者虽然均无血清 Chol 升高,但其中 6 例表现为 HDC-C 降低,2 例 LDL-C 升高,5 例 TG 高于正常,由此进一步提示冠状动脉粥样硬化可能为 X 综合征的发病原因之一。

参 考 文 献

1 张少雄,戴汝平,高润霖,等. 电子束 CT 对冠状动脉钙化的定量研究. 中华心血管病杂志,1998,26(4):289

2 方 圻,王钟林,宁田海,等. 血脂异常防治建议. 中华心血管病杂志,1997,25(3):169

3 Kemp H G. Left ventricular function in patients with the anginal syndrome and normal coronary arteriograms. Am J Cardiol, 1973,32(3):375

4 Chauhan A, Mullins P A, Thuraisingham S I, *et al*. Clinical presentation and functional prognosis in syndrome X. Br Heart J, 1993,70(4):346

5 Alexander R W, Schlant R C, Fuster V. The Heart, 9th ed. New York:McGraw-Hill, 1998,1148～1153

(1999-04-27 收稿 1999-09-20 修回)

(上接第 318 页)

文中具有共性的统计学错误,涉及分析方法与设计的匹配、重复测量资料的分析、多维列联表资料的分析、诊断效果的评价和随访资料的分析等 5 个方面。产生这些错误的原因可能是:初等统计学的普及不深不透;想当然地"举一反三";统计学的普及长期停留在初级阶段。当务之急有二。一是对高层医学研究人员普及中高级统计学知识;二是加强统计咨询服务。普及的目标是使广大医学研究人员具备"点菜"和"吃菜"的本领,而不必学会"炒菜"。"点菜"指善于向统计专业人员寻求咨询服务;"吃菜"指善于结合本专业知识解释统计计算的结果。受过中高级统计学知识熏陶的医学研究人员在有效的咨询服务帮助下,定能减少医学论文中的统计学错误。

参 考 文 献

1 方积乾主编. 医学统计学与电脑实验. 上海:上海科学技术出版社, 1997. 153～155,163～164,322～329,360～371,339～340

2 胡良平主编. 现代统计学与 SAS 应用. 北京:军事医学科学出版社, 1996. 88～92,129～145

3 Everitt B S. The analysis of repeated measures: a practical review with examples. Statistician, 1995,44(1):113

4 Everitt B S. The analysis of contigency tables. 2nd ed. London: Chapman and Hall, 1992. 98～146

(1999-07-08 收稿 1999-09-20 修回)