Creative Education Studies 创新教育研究, 2021, 9(2), 420-425
Published Online April 2021 in Hans. http://www.hanspub.org/journal/ces
https://doi.org/10.12677/ces.2021.92066

Hans汉斯

样本方差自由度的多角度阐释

孙廷哲*，穆　丹

安庆师范大学，生命科学学院，安徽　安庆
Email: *confucian007@126.com

收稿日期：2021年3月7日；录用日期：2021年4月18日；发布日期：2021年4月25日

摘　要

生物统计学是生命科学各专业本科生的必修课程。样本方差是生物统计学中的一个重要概念，但对样本方差自由度的介绍一直是课堂教学的难点所在。运用MATLAB模拟、结合实例以及数学证明对自由度概念给予多角度阐释，不仅有助于学生正确理解样本方差自由度的概念，激发学生的学习兴趣，同时也能为统计推断的教学奠定基础。

关键词

生物统计学，MATLAB，样本方差，自由度

Explaining the Degree of Freedom in Sample Variance from Multiple Perspectives

Tingzhe Sun*, Dan Mu

School of Life Sciences, Anqing Normal University, Anqing Anhui
Email: *confucian007@126.com

Received: Mar. 7th, 2021; accepted: Apr. 18th, 2021; published: Apr. 25th, 2021

Abstract

Biostatistics is a compulsory major course for students. A key concept in biostatistics is sample variance, but it is difficult to clearly demonstrate the degree of freedom in sample variance during classroom teaching. MATLAB based simulations together with real-world examples and rigorous proof can explain the concept of degree of freedom from multiple perspectives. These combinatorial strategies not only help to accurately understand the concept of degree of freedom in sample variance, inspire students' interest but also assist in the study of statistical inference.

*通讯作者。

文章引用: 孙廷哲, 穆丹. 样本方差自由度的多角度阐释[J]. 创新教育研究, 2021, 9(2): 420-425.
DOI: 10.12677/ces.2021.92066

Keywords

Biostatistics, MATLAB, Sample Variance, Degree of Freedom

1. 引言

生物统计学是一门基于数据的科学，是统计学的一个分支，旨在运用数理统计方法，探究生命科学研究中的现象和实验数据[1]。生物统计学对提高信息探索和数据分析的能力起到了重要作用。然而生物统计学公式复杂、理论性强；若数学基础较弱，甚至对于某些统计学基础概念都无法正确理解。

在样本方差的定义中，需要调整自由度(degree of freedom, df)，即由 n 调整为 $n-1$。此处自由度调整对初学者而言可能具有一定的理解困难。解释自由度的真正含义需要建立在对“数学期望”概念的理解基础上，而对于非数学专业学衡而言，“数学期望”并不是一个十分熟悉的概念。生物统计学教材中常不加证明而直接给出样本方差抽样分布的形式，通过此分布性质进而解释样本方差的自由度调整[2]。这种方式既非严格的数学证明推导，也并未从直观上给予阐释。所以，样本方差的定义很难被真正的理解。自由度虽是统计学中的重要概念，但在常用生物统计学教材中却并未给予过多的解释，而仅仅以一种公式化的方式呈现[3]。部分教材和文献中给出了自由度的具体含义，并强调了自由度在统计过程中的应用[4] [5]，但仍缺少一种科学直观的方式使学生掌握方差概念中的自由度 $n-1$。

MATLAB 具有更接近自然语义的语法和数据结构、强大的绘图功能，并兼具卓越的科学计算能力[6]。近年来，MATLAB 在生物统计学课堂中得到了广泛的应用，可以在一定程度上增强教学效果[7]。利用 MATLAB 随机数发生器对样本抽取过程进行模拟，配合应用实例和严格的数学证明，从多角度阐释样本方差的自由度，有助于从不同程度上理解方差的定义，为进一步的统计推断学习奠定基础。

2. 样本方差自由度的证明

设 $\boldsymbol{x}=\left(x_1,x_2,\cdots,x_n\right)^{\mathrm{T}}$ 为取自正态总体 $N\left(\mu,\sigma^2\right)$ 的简单随机样本，记样本均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，样本方差 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$，则：$Es^2=\sigma^2$

证明：取 $\boldsymbol{y}=\boldsymbol{\Gamma x}$，$\boldsymbol{y}=\left(y_1,y_2,\cdots,y_n\right)^{\mathrm{T}}$，其中：

$$\boldsymbol{\Gamma}=\begin{pmatrix}\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}}\\ \frac{1}{\sqrt{2\cdot1}} & -\frac{1}{\sqrt{2\cdot1}} & \cdots & \cdots & 0\\ \frac{1}{\sqrt{3\cdot2}} & \frac{1}{\sqrt{3\cdot2}} & -\frac{2}{\sqrt{3\cdot2}} & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \frac{1}{\sqrt{n\cdot(n-1)}} & \frac{1}{\sqrt{n\cdot(n-1)}} & \frac{1}{\sqrt{n\cdot(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n\cdot(n-1)}}\end{pmatrix} \tag{1}$$

容易验证矩阵 $\boldsymbol{\Gamma}$ 为正交矩阵。

则：

$$E\boldsymbol{y}=\boldsymbol{\Gamma}E\boldsymbol{x}=\begin{pmatrix}\sqrt{n}\mu\\0\\\vdots\\0\end{pmatrix} \tag{2}$$

$\boldsymbol{y}$ 的方差 $Var(\boldsymbol{y})=\boldsymbol{\Gamma}Var(\boldsymbol{x})\boldsymbol{\Gamma}^{\mathrm{T}}=\boldsymbol{\Gamma}\left(\sigma^2 I_n\right)\boldsymbol{\Gamma}^{\mathrm{T}}=\sigma^2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}=\sigma^2 I_n$，这里 I_n 为 n 阶单位矩阵。

由 $\boldsymbol{y}$ 的方差可知，$y_1,y_2,\cdots,y_n$ 相互独立，且 $y_1\sim N\left(\sqrt{n}\mu,\sigma^2\right)$，$y_i\sim N\left(0,\sigma^2\right),i=2,\cdots,n$。

另：

$$\sum_{i=1}^{n}y_i^2=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{x}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}=\sum_{i=1}^{n}x_i^2 \tag{3}$$

$$(n-1)s^2=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2=\sum_{i=1}^{n}y_i^2-n\bar{x}^2=\sum_{i=1}^{n}y_i^2-y_1^2=\sum_{i=2}^{n}y_i^2 \tag{4}$$

则根据 χ^2 分布定义，得

$$\frac{(n-1)s^2}{\sigma^2}=\sum_{i=2}^{n}\frac{y_i^2}{\sigma^2}\sim\chi^2(n-1) \tag{5}$$

根据 χ^2 分布的性质：

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right]=n-1 \tag{6}$$

即 $Es^2=\sigma^2$，证毕。

3. 样本方差自由度的解释

由表达式(5)可知，有关样本方差的抽样分布为 χ^2 分布，其自由度为 $n-1$。样本方差为总体方差 σ^2 的无偏估计量。但对于绝大多数生命科学背景的学生而言，理解上述证明具有一定的难度。所以，在实际的讲授中，一般会通过形象化的举例描述来解释方差公式中的除数$(n-1)$，如：“$n-1$”在统计学上可称之为自由度，是指独立可自由变化的观测数个数。在计算 n 个观测数的样本标准差时，每个 x 与 $\bar{x}$ 比较，虽有 n 个离均差，但只有 $n-1$ 个是自由变动的，最后一个离均差由于受到一个条件 $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=0$ 的限制不能自由变动。如一个样本具有 5 个观测数，已知 4 个离均差为 2，3，1，−2，则第 5 个离均差必然为−4，才能使 $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=0$ 成立。由于能自由变动的离均差是 4，故自由度为 4，即自由度为 $n-1$。通过列举如上实例，可帮助部分基础薄弱学生对样本方差自由度形成一定的认识。

4. MATLAB 模拟样本方差自由度

利用软件辅助教学是课程设计的重要环节。将 MATLAB 融入生物统计学课堂讲学，不仅可以提升学生知识运用的能力，还能辅助进行复杂繁琐的计算，激发学生热情[8]。下面通过设计一个 MATLAB 模拟实例来介绍样本方差自由度的概念。设总体服从正态分布，均值为 μ，方差为 σ^2。从总体中随机抽取容量为 n 的样本，样本均值为 $\bar{x}$。记样本的离均差平方和为 $SS=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$。若重复此过程 N 次，记

$SS_{mean}=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}\right)^2$。当 $N\to+\infty$ 时，$SS_{mean}\to(n-1)\sigma^2$ [8]。当样本离均差平方和乘以 $\frac{1}{n}$，则得到偏小估计 $\frac{1}{n}SS_{mean}$，$N\to+\infty$ 时，偏差为 $\frac{1}{n}SS_{mean}-\frac{1}{n-1}SS_{mean}=-\frac{\sigma^2}{n}$。

运用 MATLAB 进行数值模拟，样本容量 $n=5$。简化起见，取标准正态分布随机变量，即 $x_i\sim N(0,1)$，$\boldsymbol{x}=(x_1,x_2,x_3,x_4,x_5)^{\mathrm{T}}$。简单随机样本(每行为 1 样本)，示例参见图 1。随机样本使用 MATLAB 中 *randn* 函数生成。无偏和有偏方差的计算基于 MATLAB 中 *std* 函数，通过设定 *std* (*X*, *Dim*)函数第二输入参数 *Dim* 实现(*Dim* = 0 为无偏，*Dim* = 1 为有偏)。每抽取一个容量 $n=5$ 的样本，计算一次样本的离均差平方和。若此时重复次数为 N，则依 $SS_{mean}=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}\right)^2$ 公式计算 SS_{mean}。因随机变量来自于标准正态分布 $\sigma^2=1$，故 $N\to+\infty$ 时，$\frac{1}{n-1}SS_{mean}\to1$，$\frac{1}{n}SS_{mean}\to0.8$，偏差 $-\frac{\sigma^2}{n}\to-0.2$。所以随着重复取样次数 N 的增加，$\frac{1}{n-1}SS_{mean}$、$\frac{1}{n}SS_{mean}$ 和偏差 $-\frac{\sigma^2}{n}$ 取值将逼近理论值(图 2)。因此，样本方差的自由度应为 $n-1$，若样本离均差平方和除以 n，则与理论值存在偏差，不符合参数估计中“无偏性”的要求。

随机样本生成和均方计算 MATLAB 脚本文件如图 3 所示。其中“RepNum”为重复数 N 的取值范围，以适应对数横坐标。

用于绘制图 2 的 MATLAB 代码如图 4 所示。“scatter”为绘制散点图的 MATLAB 函数。在实际的教学过程，可通过改变 scatter 函数的参数取值，调整散点图的效果，不仅可以使学生熟悉统计图谱绘制，同时也帮助学生了解统计图对生物统计学结果的增强效应。

	1	2	3	4	5
1	-0.3696	0.4312	-1.0713	-0.0928	0.3073
2	-0.5731	-0.4781	0.3325	-0.0379	0.4796
3	-0.2104	-0.1197	-0.6285	-1.1961	1.2486
4	-0.1474	0.0604	1.3319	-0.9844	1.2537
5	-1.4203	-1.8582	0.1104	1.2729	0.9605
6	1.4665	-0.6493	-1.5915	1.0988	0.4692
7	1.4937	0.5169	0.9270	-0.8269	0.8477
8	0.7294	0.2665	-0.1238	-0.2469	-0.0408
9	0.1267	0.3452	-0.0244	-1.2868	0.5738
10	0.4163	-0.7277	2.9394	0.9928	-1.1902
11	2.3127	0.3709	-0.1755	2.7872	-0.0836
12	1.5427	1.1384	-1.3430	0.7140	1.8856
13	-0.1217	2.1658	-1.2255	-0.1326	0.4474
14	0.6501	-0.2891	-0.4558	-1.0818	0.9345
15	-0.1840	-0.6209	0.8397	-0.3261	-0.5020
16	0.3783	0.1427	1.0606	1.2254	-1.5576
17	0.4552	0.7872	-0.9133	2.1641	-1.1202
18	0.3925	-0.1414	-2.0082	-0.1363	-0.0842
19	-0.0533	1.4989	2.8786	-0.2230	1.7957
20	-1.8244	-0.3791	1.0407	-0.3049	0.1516
21	2.0068	-0.2369	2.4188	-0.6956	1.3293
22	1.7218	-0.6086	-1.5872	-1.8244	0.3151
23	-0.3558	2.1661	1.1637	0.2570	0.3587
24	-0.3656	-0.6892	-0.6077	0.3571	1.4931
25	0.5273	0.9512	-0.3795	-0.0721	0.1629
26	-0.8988	-1.7464	-0.7905	-0.0075	0.4537
27	-0.5802	-0.7936	0.7466	0.8987	-0.6181
28	-0.5772	0.9987	-0.7224	1.6423	0.1706
29	1.0468	0.5754	0.4171	-0.3142	-2.1368
30	-0.9800	0.5093	-2.1211	-0.7705	0.1595

	1	2	3	4	5
43	-0.1605	-1.5230	0.5758	0.6576	-0.3914
44	1.8229	0.3283	0.3848	-1.1899	-2.4971
45	-0.4430	1.6824	-0.4848	0.8386	-0.2551
46	-0.0312	-1.1289	0.2953	0.9847	-0.9522
47	0.3158	-0.8892	-0.3886	-0.3363	0.8375
48	0.2468	0.0804	0.4717	-1.0705	0.2457
49	-0.2943	-2.2917	-0.7946	-1.4101	-0.4246
50	1.3493	0.0971	1.6273	0.1806	0.2001
51	0.3592	0.7099	2.1370	0.0785	1.1013
52	1.0344	-0.2313	0.7610	0.1445	0.1606
53	-0.2859	0.9166	-0.3755	0.1750	-0.1398
54	0.4800	-2.2500	1.5297	-0.2531	0.0815
55	-0.9543	1.5839	0.6013	-1.2908	-0.6553
56	0.0918	-1.1469	1.5477	-0.4710	-0.4269
57	0.2307	-0.1205	1.1957	-0.0951	-0.1964
58	1.7725	0.8049	-0.3211	-1.4119	-0.3939
59	2.0090	-0.6060	-0.2156	-0.8141	1.7982
60	0.7499	-1.0409	-0.0578	-0.1936	0.9905
61	-0.2405	-1.6581	-0.5160	1.2167	0.8248
62	-0.8412	1.5058	0.9492	-0.6177	0.9702
63	0.0610	-0.4288	1.4400	-0.4122	0.0614
64	0.4315	-1.1724	0.7530	0.6124	1.1056
65	0.9426	0.1178	0.7675	-0.4965	1.6460
66	0.8073	1.6455	0.3021	0.4641	-0.0959
67	0.2677	0.4005	0.6965	0.6498	1.1727
68	1.0402	-0.5487	-2.2161	-0.1835	1.7486
69	0.9570	0.0554	-0.1699	-0.7962	-0.7379
70	0.2365	-2.0591	0.0669	0.1879	0.7311
71	-1.5350	0.0318	-0.8373	-0.0104	-1.0079
72	0.3820	0.9943	0.3719	-1.9007	1.5057

Figure 1. Random variables with normal distribution generated in MATLAB
图 1. MATLAB 生成服从标准正态分布随机变量

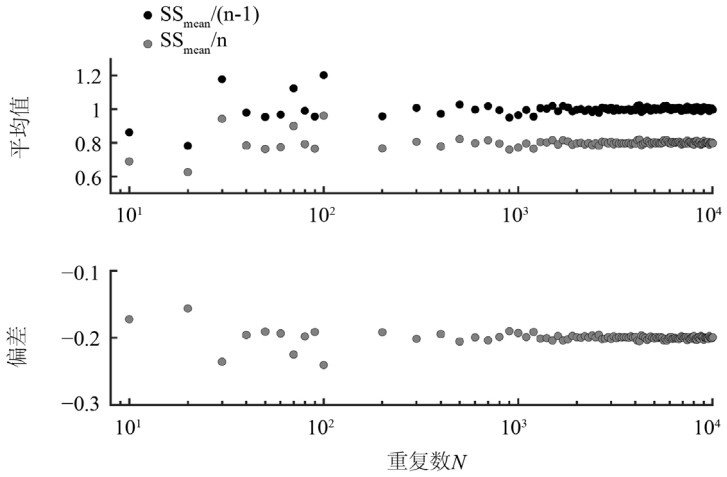

Figure 2. Simulations of unbiased, biased variances and deviations in MATLAB with increasing repeat number *N*
图 2. 无偏、有偏方差和偏差值随重复数 *N* 变化的 MATLAB 模拟

```
%% sample variance via simulations
RepNum=[10:10:100 200:100:1e4]; % Series of replicates
SSmean=zeros(length(RepNum),1);
SSmeanBias=zeros(length(RepNum),1);
Bias=zeros(length(RepNum),1);
n=5; %sample size
for i=1:length(RepNum)
N=RepNum(i);% No.of Replicates
x=randn(N,n);
SSmean(i)=mean((std(x,0,2)).^2);
SSmeanBias(i)=mean((std(x,1,2)).^2);
Bias(i)=SSmeanBias(i)-SSmean(i);
end
```

Figure 3. Codes for calculation using MATLAB
图 3. 用于计算的 MATLAB 源代码

```
subplot(2,1,1);
scatter(RepNum,SSmean,'filled','marker','o','markerfacecolor',...
    [0 0 0],'markeredgecolor','k','sizedata',120);
hold on;
scatter(RepNum,SSmeanBias,'filled','marker','o','markerfacecolor',...
    [0.5 0.5 0.5],'markeredgecolor','k','sizedata',120);
set(gcf,'color','w');
set(gca,'fontsize',45,'linewidth',3,'xtick',10.^[1 2 3 4],'ytick',...
    0:0.2:1.2,'xscale','log');
xlim([8 max(RepNum)+1]);
ylim([0.5 1.3]);
subplot(2,1,2);
scatter(RepNum,Bias,'filled','marker','o','markerfacecolor',...
    [0.5 0.5 0.5],'markeredgecolor','k','sizedata',120);
hold on;
set(gcf,'color','w');
set(gca,'fontsize',45,'linewidth',3,'xtick',10.^[1 2 3 4],'ytick',...
    -0.3:0.1:-0.1,'xscale','log');
xlim([8 max(RepNum)+1]);
ylim([-0.3 -0.1]);
```

Figure 4. Codes for plots using MATLAB
图 4. 用于绘图的 MATLAB 源代码

5. 结论

方差自由度是统计学中的重要基本概念，但一般的生物统计学教材中缺少对此自由度的证明和详尽

解释。因此，多数学生无法清楚理解基于 $n-1$ 自由度产生方差无偏估计的性质。通过修改脚本中 RepNum 向量的最大值，增加重复取样 N 的次数，可以实现对图 2 横轴的延伸，进而观察到更多的均方变化趋势，这不失为一种有益的课堂交互式体验。另外，由于 MATLAB 语法更接近于自然语义，所以上述均方计算代码相对简单易懂，有助于具有一定基础的学生理解和掌握。另外，MATLAB 代码的运行时间相对较短，生成图 2 所需循环时间仅为约 0.085 秒(Windows 10 操作系统，Intel CoreTM i5-8265 CPU，1.80 GHz，8.00 GB RAM)。即使增大 RepNum 向量的最大值 3 个数量级至 10^7，运行时间仍小于 0.1 秒。最近有文献以 Excel 2010 作为语言工具，运用 Excel 2010 内嵌的 VBA (Visual Basic for Application)编程功能进行了随机抽样过程并对自由度进行了模拟[9]。但 Excel 2010 的最大循环次数仅为 1048575 (~10^6)，与 MATLAB 最大允许循环数相距甚远；模拟时间来看，MATLAB 模拟相较于使用 Excel 2010 模拟可显著降低运行时间(降低约 3 个数量级) [9]。此外，MATLAB 作为高级编程语言，在数值计算上比 VBA 更具优势。结合严格的数学证明，辅以教学实例和 MATLAB 模拟，从多角度阐释方差自由度的含义，可适用于不同基础的学生。

值得注意的是，2018 年，高教司提出了“金课”的建设标准“两性一度”，即高阶性、创新性、挑战度[10]。高阶性要求授课内容需要“优于、高于大纲，培养学生解决复杂问题的综合能力和高级思维”[10]。鉴于生物统计学教材中常忽略方差自由度的数学证明，教师若课堂上讲解此证明过程可帮助具有一定基础的生命科学专业学生更深刻的认识方差自由度，符合“高阶性”的要求。

致　谢

感谢安庆师范大学生命科学学院朱亮亮老师对本研究的帮助。

基金项目

国家自然科学基金面上项目(31971185)；安徽省高等学校省级质量工程线下课程(原精品线下开放课程)示范项目(2020kfkc299)；安徽省高等学校省级质量工程大规模在线开放课程(MOOC)示范项目(2018mooc399)；安徽省高等学校省级质量工程教学研究重点项目(2017jyxm0307)。

参考文献

[1] 李春喜, 姜丽娜, 邵云, 等. 生物统计学[M]. 第 5 版. 北京: 高等教育出版社, 2018: 1-2.

[2] 杜荣骞. 生物统计学[M]. 第 4 版. 北京: 高等教育出版社, 2014: 69-70.

[3] Eisenhauer, J.G. (2008) Degrees of Freedom. *Teaching Statistics*, **30**, 75-78. https://doi.org/10.1111/j.1467-9639.2008.00324.x

[4] Cashing, D. (2017) An Informal Justification for Selected Degree-of-Freedom Formulae. *Teaching Statistics*, **40**, 12-15. https://doi.org/10.1111/test.12143

[5] 贾俊平, 何晓群, 金勇进. 统计学[M]. 第 4 版. 北京: 中国人民大学出版社, 2009: 99.

[6] Webb, C.R. and Domijan, M. (2019) Introduction to MATLAB for Biologists. Springer International Publishing, Cham, 7. https://doi.org/10.1007/978-3-030-21337-4

[7] 应智霞, 张欢, 葛刚, 邹志文. MATLAB 软件在生物统计理论教学中的应用——以抽样分布为例[J]. 生物学杂志, 2020, 37(4): 127-129.

[8] 解博丽, 雷英杰, 杨丽, 薛震. 概率论与数理统计引入 MATLAB 实验教学手段的必要性[J]. 教育教学论坛, 2020(23): 280-282.

[9] Liu, X.S. and Shin, H.H. (2020) Expectation and Degrees of Freedom for Sample Variance. *Teaching Statistics*, **42**, 54-57. https://doi.org/10.1111/test.12221

[10] 李孟军, 杨克巍, 赵青松, 葛冰峰. 本科教育课程质量建设的新视角——“金课”的开放性要求及闭环运行机制[J]. 高等教育研究学报, 2019, 42(3): 18-21.