·科研方法专题·

如何正确运用χ^2检验——χ^2分布及相关内容

胡纯严[1]，胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
*通信作者：胡良平，E-mail：lphu927@163.com）

【摘要】 本文目的是介绍χ^2分布及相关内容，包括χ^2分布和非中心χ^2分布。着重展示了两种χ^2分布的定义、概率密度函数的图形和主要性质，其中，两个最重要的性质分别是：①χ^2分布的极限分布为正态分布；②$\frac{(n-1)s^2}{\sigma^2}$服从自由度为$n-1$的$\chi^2$分布。除此之外，还阐释了$\chi^2$分布与正态分布、$\chi^2$检验统计量与$Z$检验统计量之间的关系。最后，基于SAS软件中的两个SAS函数呈现了χ^2分布的计算方法。

【关键词】 χ^2分布；正态分布；概率密度函数；自由度；χ^2检验；拟合优度检验
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20210115002

How to use χ^2 test correctly——χ^2 distribution and the related contents

Hu Chunyan[1], *Hu Liangping*[1,2*]

（1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*

*Corresponding author: *Hu Liangping*, *E-mail*: *lphu927@163.com*）

【Abstract】 The purpose of this article was to introduce the χ^2 distribution and related contents, including χ^2 distribution and non-central χ^2 distribution. It focused on showing the definition of two χ^2 distributions, the graph and the main properties of the probability density function. Among them, the two most important properties were: first, the limiting distribution of the χ^2 distribution was the normal distribution; second, $\frac{(n-1)s^2}{\sigma^2}$followed the χ^2 distribution with $n-1$ degrees of freedom. In addition, it also explained the relationship between the χ^2 distribution and the normal distribution, the relationship between χ^2 test statistic and Z test statistic. Finally, it illustrated the computational approaches of the χ^2 distribution based on the two SAS functions in SAS software.

【Keywords】 χ^2 distribution; Normal distribution; Probability density function; Degrees of freedom; χ^2 test; Goodness of fit test

在临床资料中，定性资料（特指结果变量及其取值）[1-6]出现的频率高。分析定性资料的统计分析方法主要有“差异性分析（通常适用于原因变量的个数≤2）[1-3]”和“logistic回归分析（通常被用于原因变量的个数>2）[4-6]”两大类。在前述提及的两类统计分析方法中，“χ^2检验”都是不可或缺的。

在经典统计学和贝叶斯统计学中，“概率分布”是统计分析的重要基础[7-11]，若离开了它，假设检验、区间估计、回归分析、判别分析和多元分析几乎无法进行。由数理统计知识可知，一旦掌握了某个随机变量的概率分布，就等于掌握了其变化规律。概率分布的种类很多，通常可以分为两大类，即“离散型随机变量的概率分布”和“连续型随机变量的概率分布”。本文介绍的χ^2分布是在统计分析中被广泛使用的一个连续型随机变量的概率分布，它具有两种表现形式，即“χ^2分布”和“非中心χ^2分布”[10-11]。本文对χ^2检验的基础知识，即“χ^2分布及相关内容”进行介绍。

1 χ^2分布

1.1 χ^2分布的历史

χ^2分布是从正态分布派生出来的一个连续型概率分布。尽管如此，由于许多分布可以用χ^2分布来近似，甚至在多元统计分析中也常用到它，故χ^2分布在数理统计中一直占有重要地位[11]。

χ^2分布分别由I. J. Bienayme（1858）、F. R. Helmert（1876）和K. Pearson（1900）发现，开始主要用于列联表资料的“独立性”分析和“评价回归模型对资料拟合效果好坏”的拟合优度检验[12]；在定性资料回归模型的构建过程中，χ^2分布常用于筛选自变量[13]；在广义线性回归模型和混合效应回归模型的构建中，χ^2分布常用于两个回归模型对同一个资

料拟合效果的比较[13]。

1.1.1 χ^2分布的定义

设随机变量$Y_1,Y_2,\cdots,Y_n$独立同分布，且$Y_i\sim N(0,1)$，则随机变量的分布称为具有n个自由度的χ^2分布，并记为$\chi^2\sim\chi^2_n$。见式(1)：

$$\chi^2=\sum_{i=1}^{n}Y_i^2 \tag{1}$$

1.1.2 χ^2分布的概率密度函数及其图形

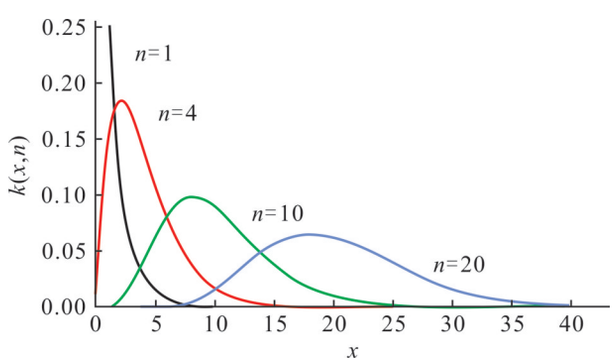
χ^2分布的概率密度函数$k(x,n)$如下：

$$\chi^2_{(x,n)}=\begin{cases}0 & ,x\leqslant 0\\ \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x^2}{2}} & ,x>0\end{cases} \tag{2}$$

式中$\Gamma(1/2)=\sqrt{\pi}$，$\Gamma(1)=\Gamma(2)=1$，当$n\geqslant 3$时，见下式：

$$\Gamma(n)=(n-1)!\ ,\Gamma(n/2)=1\times 3\times 5\times\cdots\times(n-2)2^{-(n-1)/2}\sqrt{\pi} \tag{3}$$

χ^2分布概率密度函数的图形见下图：



注：横坐标轴上的变量为服从χ^2分布的随机变量x；纵坐标轴上的变量为χ^2分布的概率密度函数$k(x,n)$(即任何一条曲线上与随机变量x对应的高度)

图1 具有几种不同自由度的χ^2分布概率密度函数的图形

1.1.3 χ^2分布的性质

1.1.3.1 χ^2分布的极限分布为正态分布

由图1可看出如下特点：①自由度n越大，曲线越趋于对称；②当自由度$n\to\infty$时，χ^2分布趋向于正态分布。

1.1.3.2 χ^2分布的期望和方差

若$X\sim\chi^2_n$，则

$$E(X)=n,\mathrm{Var}(X)=2n \tag{4}$$

式(4)的“E”代表“期望(通俗的表述为‘均值’)”；“Var”代表“方差”。

1.1.3.3 χ^2分布具有可加性

若$X_1\sim\chi^2_n$、$X_2\sim\chi^2_m$，且X_1与X_2相互独立，则

$$X_1+X_2\sim\chi^2_{n+m} \tag{5}$$

1.2 非中心χ^2分布

1.2.1 非中心χ^2分布的定义

设随机变量$Y_1,Y_2,\cdots,Y_n$相互独立，且$Y_i\sim N(\mu_i,1)$，则随机变量$\chi^2_{n,\delta}$的分布称为具有n个自由度且非中心参数为$\delta\equiv\mu_1^2+\mu_2^2+\cdots+\mu_n^2$的$\chi^2$分布，并记为$\chi^2_{n,\delta}$。见式(6)：

$$\chi^2_{n,\delta}=\sum_{i=1}^{n}Y_i^2 \tag{6}$$

在上式中，当$\delta=0$时，非中心χ^2分布$\chi^2_{n,\delta}$就退化成为前面定义的χ^2分布χ^2_n。

1.2.2 非中心$\chi^2_{n,\delta}$分布的概率密度函数

非中心$\chi^2_{n,\delta}$分布的概率密度函数如下：

$$\chi^2_{n,\delta}=\begin{cases}0 & ,x\leqslant 0\\ e^{-\frac{\delta+x}{2}}\sum_{j=0}^{\infty}\dfrac{1}{j!}\left(\dfrac{\delta}{2}\right)^j\dfrac{x^{\frac{n}{2}+j-1}}{2^{\frac{n}{2}+j}\Gamma\left(\frac{n}{2}+j\right)} & ,x>0\end{cases} \tag{7}$$

$$\frac{X_n-n}{\sqrt{2n}}\xrightarrow{L}N(0,1) \tag{8}$$

【说明】“$\xrightarrow{L}$”表示依分布收敛，且$\sqrt{2X_n}-\sqrt{2n}\xrightarrow{L}N(0,1)$。

另一方面，由来自正态分布的均值与离均差平方和或方差可构造出服从χ^2分布的随机变量。事实上，若$X_1,X_2,\cdots,X_n(n\geqslant 2)$是从总体$N(\mu,\sigma^2)$中抽出的样本，令：

【说明】因篇幅所限，该分布的性质从略。

2 χ^2分布与正态分布之间的关系

χ^2分布是从正态分布派生出来的一个分布；χ^2分布的极限分布为标准正态分布[7,11]。用数学语言表述如下：

若$X_n\sim\chi^2_n,n=1,2,\cdots$，则当$n\to\infty$时，有下面的关系式成立：

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\ sum=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \tag{9}$$

则有下式成立：

$$\chi^2=\frac{sum}{\sigma^2}\sim\chi^2_{n-1},\ df=n-1 \tag{10}$$

由式(10)定义的随机变量χ^2服从自由度为$n-1$的χ^2分布。该式的另一种表达形式如下：

$$\chi^2=\frac{(n-1)s^2}{\sigma^2}\sim\chi^2_{n-1},\ df=n-1 \tag{11}$$

由式(11)定义的随机变量χ^2服从自由度为$n-1$的χ^2分布。在此式中，s^2为样本方差，即$s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$。

【说明】因篇幅所限，式(8)、式(10)和式(11)的数学证明从略。

基于式(11)定义的随机变量“χ^2”可以构造出总体方差σ^2的100(1-α)%置信区间，见下式：

$$\left[\frac{(n-1)s^2}{\chi^2_{\left(1-\frac{\alpha}{2}\right)(n-1)}},\frac{(n-1)s^2}{\chi^2_{\left(\frac{\alpha}{2}\right)(n-1)}}\right] \tag{12}$$

在上式中，置信区间下限的分母是自由度为“$n-1$”的χ^2分布曲线下左侧尾端概率为“$1-\frac{\alpha}{2}$”时横坐标轴上的分位数；而置信区间上限的分母是自由度为“$n-1$”的χ^2分布曲线下左侧尾端概率为“$\frac{\alpha}{2}$”时横坐标轴上的分位数；“s^2”为样本方差；当α分别取0.05与0.01时，基于式(12)求出的分别是总体σ^2的95%与99%置信区间。

3 χ^2检验统计量与Z检验统计量之间的关系

由本文式(1)定义的χ^2分布可知，它是由n个互相独立且都服从标准正态分布的随机变量的平方之和构成的，故当其自由度为1时，χ^2检验统计量的平方根就是Z检验统计量(说明：在SAS软件和部分统计学教科书中，通常用Z表示服从标准正态分布的随机变量或检验统计量)。

4 χ^2分布的计算

4.1 χ^2分布曲线下累计概率的计算

在SAS软件中，χ^2分布的分布函数为：

probchi(x,df,nc)

该函数计算服从自由度为df，非中心参数为nc的χ^2分布的随机变量小于给定x的事件的概率。如果nc没有规定或取为0，那么被计算的就是中心χ^2分布曲线下累计概率。

【例1】试计算自由度为5，中心χ^2分布曲线下χ^2值小于20的概率值。

【分析与解答】所需要的SAS程序如下：

```
/*以下程序计算卡方分布累计概率*/
data a;
x=20;
df=5;
nc=0;
p=probchi(x,df,0);
run;
proc  print data=a noobs;
var x df nc p;
run;
```

【SAS输出结果及解释】

x	df	nc	p
20	5	0	0.99875

以上结果表明，当χ^2=20、自由度df=5、非中心参数nc=0的条件下，χ^2分布曲线下且位于横坐标轴上“0~20”区间内的累计概率为0.99875。

4.2 χ^2分布曲线下横坐标轴上P分位数的计算

在SAS软件中，χ^2分布的分位数函数为：

cinv(P,df,nc)(0≤P≤1，df>0，nc≥0)

该函数计算自由度为df，非中心参数为nc的χ^2分布的P分位数。取nc=0或不规定此项参数时，表明是中心χ^2分布。

【例2】试计算自由度为3，非中心参数为4.5的χ^2分布的P=0.95的分位数。

【分析与解答】所需要的SAS程序如下：

```
/*以下程序计算卡方分布的p分位数*/
data b;
p=0.95;
df=3;
nc=4.5;
x=cinv(0.95,3,4.5);
run;
proc print data=b noobs;
var p df nc x;
run;
```

【SAS输出结果及解释】

p	df	nc	x
0.95	3	4.5	16.8463

以上结果表明，当累计概率P=0.95、自由度df=3、非中心参数nc=4.5的条件下，χ^2分布曲线下横坐标轴上的分位数x=16.8463（注意：这里的“x”是一个服从自由度df=3、非中心参数nc=4.5的χ^2分布的随机变量）。

【例3】试基于SAS函数“cinv(P,df,nc)”产生χ^2分布临界值表。

【分析与解答】在很多统计学教科书的附录中，一般都会给出常用统计用表，其中，χ^2分布临界值表可以利用SAS函数“cinv(P,df,nc)”且令nc=0计算出来。现给出所需要的SAS程序如下：

```
OPTIONS LS=120 PS=50;
DATA abc;
ARRAY X(46,13);
g=1;
DO d=1 TO 40,50,60,70,80,90,100;
w=1;
DO alpha=0.995, 0.990, 0.975, 0.950, 0.900, 0.750, 0.500, 0.250, 0.100, 0.050, 0.025, 0.010, 0.005;
p=1-alpha;
b=CINV(p,d);
b=ROUND(b,0.01);
X(g,w)=b;
w=w+1;
OUTPUT;
END;
g=g+1;
END;
FILE PRINT;
DO L=1 TO 46;
c=L;
PUT #c @5 X(L,1) 5.2 @11 X(L,2) 5.2 @17 X(L,3) 5.2
@23 X(L,4) 5.2 @29 X(L,5) 5.2 @35 X(L,6) 5.2
@41 X(L,7) 5.2 @47 X(L,8) 6.2 @54 X(L,9) 6.2
@61 X(L,10) 6.2 @68 X(L,11) 6.2 @75 X(L,12) 6.2
@82 X(L,13) 6.2;
END;
RUN;
```

【程序说明】“DO d=1 TO 40,50,60,70,80,90,100”语句表明，自由度df的取值为1、2、…、39、40、50、60、70、80、90、100，共46种取值，即产生的χ^2分布临界值表有46行；“alpha=0.995,0.990,0.975,0.950,0.900,0.750,0.500,0.250,0.100,0.050,0.025,0.010,0.005;p=1-alpha;”两个语句表明，χ^2分布曲线下右侧尾端概率分别为0.995、0.990、…、0.005，共13种取值。也就是说，以上SAS程序共计算出46×13=498个χ^2分布临界值（本质上就是χ^2分布曲线下横坐标上的“分位数”的数值）。

【说明】因输出的数据较多，此处从略。

5 讨论与小结

5.1 讨论

χ^2分布是一种连续型随机变量的概率分布，然而，它不同于其他连续型随机变量的概率分布（如“正态分布”等）。因研究者应用统计学的过程中，诸如“正态分布”“t分布”和“F分布”的连续型概率分布常作为“Z检验”“t检验”和“方差分析（或称F检验）”的理论依据，直接应用于定量资料的差异性分析；而χ^2分布作为“χ^2检验”的理论依据，一般只应用于定性资料的差异性分析（如各种列联表资料的差异性分析）、不同统计模型对同一个统计资料拟合优度的比较等场合。但也有例外，即χ^2检验可应用于单因素多水平设计一元定量资料多个方差的齐性检验之中。χ^2分布和χ^2检验之所以可以应用于前述提及的各种场合，因为在那些场合下所构造出的“检验统计量”服从χ^2分布。

5.2 小结

本文针对处理定性资料所需要的χ^2检验，介绍了与其有关的理论基础，即χ^2分布和非中心χ^2分布。重点展示了χ^2分布的定义、概率密度函数的图形和主要性质；基于SAS软件中的两个SAS函数呈现了χ^2分布的计算方法（包括累计概率的计算和分位数的计算）和结果解释。

参考文献

[1] 张炳智，郑在江，田国娇，等.雅安市社区在管严重精神障碍患者现状研究[J].四川精神卫生，2020，33(1)：53-56，60.

[2] 易海，杨琼花，杜育如，等.重性精神疾病社区管理服务背景下湛江地区患者及家属对疾病知晓情况的分析[J].四川精神卫生，2020，33(1)：57-60.

[3] 邓兰芳，余金鸣，黄彩英，等.中山市流动人口与户籍人口自杀率和自杀方式特征分析[J].四川精神卫生，2020，33(1)：67-70，75.

[4] 徐海婷，刘嫣然，吕婧，等. 未治疗抑郁障碍患者自杀风险与认知情绪调节策略的关系[J]. 四川精神卫生，2020，33(1)：44-48.
[5] 缪楹，徐晓津，周勇杰，等. 社会支持对重性抑郁障碍患者自杀意念的影响[J]. 四川精神卫生，2020，33(2)：142-145.
[6] 宋晓灵，赵东梅，胡一君，等. 卒中类型、卒中部位与卒中后癫痫的多因素关系[J]. 四川精神卫生，2020，33(2)：164-167.
[7] 陈希孺. 数理统计引论[M]. 北京：科学出版社，1981：238-357.
[8] 詹姆斯·O·伯杰. 统计决策论及贝叶斯分析[M]. 贾乃光，译. 北京：中国统计出版社，1998：623-627.
[9] 伊冯·M·毕晓普，斯蒂芬·E·芬伯格，保罗·W·霍兰德. 离散多元分析理论与实践[M]. 张尧庭，译. 北京：中国统计出版社，1998：530-555.
[10] 茆诗松. 统计手册[M]. 北京：科学出版社，2006：1-33.
[11] 方开泰，许建伦. 统计分布[M]. 北京：科学出版社，1987：136-211.
[12] 胡良平. 医学统计学运用三型理论分析定量与定性资料[M]. 北京：人民军医出版社，2009：22-41.
[13] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary，NC：SAS Institute Inc，2018：3405-3608，5749-6006，6533-6728.

（收稿日期：2021-01-15）
（本文编辑：戴浩然）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955 年 8 月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事、中国生物医学统计学会副会长、北京大学口腔医学院客座教授和《中华医学杂志》等 10 余种杂志编委；现任世界中医药学会联合会临床科研统计学专业委员会会长、国家食品药品监督管理局评审专家和 3 种医学杂志编委；主编统计学专著 48 部、参编统计学专著 10 部；发表第一作者和通信作者学术论文 300 余篇、发表合作论文 130 余篇；获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作、参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告 100 余场，举办数十期全国统计学培训班，培养 20 多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析和 SAS 与 R 软件实现、各种层次的统计学教学培训和咨询工作。