2025 March

数学月刊

Mathematics Monthly

Author :
Guanzhong Yang(student)
Yan Wang(teacher)

Mathematics changes our lives

PREFACE

This month, we are going to talk about three brand-new distributions, namely Chi-square distribution, t distribution and F distribution. And they all have a big impact on mathematics.

If you have other brilliant questions or knowledge willing to learn, email to anmiciuangray@163.com **for surprising rewards!**

1. 卡方分布

Proof

卡方分布就是由 n 个(维)标准正态随机变量的平方和的分布.

设 $X \sim N(0,1)$, 因此$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. 让 $Y = X^2$, for $y > 0$, 我们要求$f_Y(y)$:

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

$$f_Y(y) = \frac{d}{dy}[F_X(\sqrt{y}) - F_X(-\sqrt{y})] = \frac{1}{2\sqrt{y}}[f_X(\sqrt{y}) + f_X(-\sqrt{y})]$$

$$= \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}} = \Gamma(\frac{1}{2},2).$$

因为如果 $M \sim \Gamma(m,\theta)$, $N \sim \Gamma(n,\theta)$, $M+N \sim \Gamma(m+n,\theta)$:

如果 $Y = \sum_{i=1}^{n} X_i^2$, 其中X_i是来自总体 N(0,1)的样本, 则

$$f_Y(y) = \Gamma(\frac{n}{2},2) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}.$$

由此便得到了卡方分布的公式.

Afterwords

我们来回顾一下卡方分布的两个用法:

(1)

为了检验一组数据的分布概率是否符合我们的预期的分布概率, 我们先统计出来每一种情况的观察频数(O_i), 计算出来每一种情况的期望频数(E_i). 随后通过计算

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

并根据 significance level 确定我们的预期是否正确.

为什么可以这样做呢?

$(O_i - E_i)^2$表示了观察频数与期望频数之间的差异, 并且消除正负差异的抵消的可能性.

$\frac{1}{E_i}$是为了标准化差异, 这样每个类别的贡献被标准化为相对差异, 而不是绝对差异, 此时近似服从卡方分布.

(2)

我们想检验一组数据的方差是不是符合我们预期的方差σ_0^2. 我们先统计出来S^2, 随后通过计算

$$\chi^2 = \frac{n-1}{\sigma_0^2} S^2$$

并根据 significance level 确定我们的预期是否正确.

2. t 分布

Auxiliary Result

(1)
首先我们要证明出

$$\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1).$$

对于 n 维正态随机变量$X_1, X_2, ..., X_n$, 设其对应的标准化正态变量为$Z_i = \frac{X_i-\mu}{\sigma}$, $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, 则:

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}((X_i-\mu)-(\bar{X}-\mu))^2 = \sigma^2\sum_{i=1}^{n}(Z_i-\bar{Z})^2.$$

$$\frac{n-1}{\sigma^2}S^2 = \sum_{i=1}^{n}(Z_i-\bar{Z})^2 = \sum_{i=1}^{n}Z_i^2 - 2\sum_{i=1}^{n}Z_i\bar{Z} + n\bar{Z}^2 = \sum_{i=1}^{n}Z_i^2 - n\bar{Z}^2.$$

这里我们考虑一组新的随机变量$Y_1, Y_2, ..., Y_n$, 其满足$\vec{Y} = A\vec{Z}$, 其中$\vec{Y} = (Y_1\ Y_2\ \dots\ Y_n)^T$, $\vec{Z} = (Z_1\ Z_2\ \dots\ Z_n)^T$, A 为一个正交矩阵 orthonormal matrix, 且 A 的第一行的元素均为$\frac{1}{\sqrt{n}}$.
则对于 i = 1, 2, ..., n, 有$Y_i = \sum_{j=1}^{n} a_{ij}Z_j$, 因此$Y_1, Y_2, ..., Y_n$仍为正态变量.
因为对于 i, j = 1, 2, ..., n, 有 $E(Z_i) = 0$, $Cov(Z_i,Z_j) = \delta_{ij}$($\delta_{ij} = 1$ 当 i=j, 否则$\delta_{ij} = 0$):

$$E(Y_i) = E(\sum_{j=1}^{n} a_{ij}Z_j) = \sum_{j=1}^{n} a_{ij}E(Z_j) = 0$$

$$Cov(Y_i, Y_k) = Cov(\sum_{j=1}^{n} a_{ij}Z_j, \sum_{l=1}^{n} a_{kl}Z_l) = \sum_{j=1}^{n}\sum_{l=1}^{n} a_{ij}a_{kl}Cov(Z_j,Z_l)$$

$$= \sum_{j=1}^{n} a_{ij}a_{kj} = (\text{矩阵 A 的第 i 行})\bullet(\text{矩阵 A 的第 k 行}) = \delta_{ik}.$$

因此, $Y_1, Y_2, ..., Y_n$两两不相关. 又由于 n 维随机变量$(Y_1, Y_2, ..., Y_n)$是 n 维正态随机变量$(X_1, X_2, ..., X_n)$经过线性变换得到的, 因此$(Y_1, Y_2, ..., Y_n)$也是 n 维正态随机变量. 于是由$(Y_1, Y_2, ..., Y_n)$两两不相关推论到$(Y_1, Y_2, ..., Y_n)$相互独立, 且有$Y_i \sim N(0,1)$, i = 1, 2, ..., n. 而

$$Y_1 = \sum_{j=1}^{n} a_{1j}Z_j = \sum_{j=1}^{n}\frac{1}{\sqrt{n}}Z_j = \sqrt{n}\bar{Z}.$$

又因为

$$\sum_{i=1}^{n} Y_i^2 = \vec{Y}^T\vec{Y} = (A\vec{Z})^T A\vec{Z} = \vec{Z}^T A^T A\vec{Z} = \vec{Z}^T\vec{Z} = \sum_{i=1}^{n} Z_i^2.$$

$$\frac{n-1}{\sigma^2}S^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2.$$

由于$Y_2, ..., Y_n$相互独立且有$Y_i \sim N(0,1)$, i = 2, ..., n, 所以:

$$\sum_{i=2}^{n} Y_i^2 \sim \chi^2(n-1).$$

$$\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1).$$

(2)
如果 X 是一维连续型随机变量, 概率密度函数为$f_X(x)$, 则 Y = aX+b 的概率函数密度为:

$$f_{aX+b}(y) = \frac{1}{|a|}f_X(\frac{y-b}{a}).$$

(3)
如果(X,Y)是二维连续型随机变量, 概率密度函数为 f(x,y), 则 $Z = \frac{Y}{X}$的概率函数密度为:

$$f_{\frac{Y}{X}}(z) = \int_{-\infty}^{\infty} |x| f(x, xz)\,dx.$$

(4)
设随机变量具有概率密度函数$f_X(x)$, -∞<x<∞, 又设函数 g(x)处处可导且恒有 g'(x)>0, 则 Y = g(X)是连续型随机变量, 且其概率密度函数为:

$$f_Y(y) = f_X(h(y))|h'(y)|,$$

其中 min{g(-∞),g(∞)}<y<max{g(-∞),g(∞)}, h(x)为 g(x)的反函数.

Proof

t 分布就是为了检测小样本而诞生的.
我们想要找到一个拥有 n 个元素小样本的一些特征. 首先, 我们要假设: 该小样本的总体符合正态分布. 这是因为有中心极限定义: 一句话总结, 正态分布为卷积不动点, 因而卷积的极限趋近于正态分布.
在以上假设的前提下, 我们有:

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

现在的问题是: 我们不知道σ^2, 于是我们用S^2代替:

$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}.$$

因为$\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$, 所以:

$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma}}{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}^{\frac{1}{2}}} = \frac{Z}{(\frac{\chi^2(n-1)}{n-1})^{\frac{1}{2}}}.$$

因此, 如果 $Z \sim N(0,1)$, $Y \sim \chi^2(n)$, 且 Z 与 Y 相互独立, 那么随机变量

$$T = \frac{Z}{(\frac{Y}{n})^{\frac{1}{2}}}$$

服从自由度为 k 的 t 分布.

下面我们要证明 t 分布的概率密度函数.

已知

$$f_Y(y) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}.$$

则我们先求 $W = \frac{Y}{n}$的分布:

$$f_{\frac{Y}{n}}(w) = nf_Y(nw) = \frac{n}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} (nw)^{\frac{n}{2}-1} e^{-\frac{nw}{2}} = \frac{(\frac{n}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} w^{\frac{n}{2}-1} e^{-\frac{nw}{2}}.$$

下面来求 $V = \sqrt{W}$的分布:

$$f_{\sqrt{W}}(v) = 2zf_W(v^2) = \frac{2(\frac{n}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} v^{n-1} e^{-\frac{nv^2}{2}}.$$

最后我们来求 $T = \frac{Z}{V}$的分布:

$$f_{\frac{Z}{V}}(t) = \int_{-\infty}^{\infty} |v| f(v, vt) dv = \int_{-\infty}^{\infty} |v| f_V(v) f_Z(vt) dv.$$

因为 V 在定义的时候就是非负的, 因此:

$$f_{\frac{Z}{V}}(t) = \int_0^{\infty} v \frac{2(\frac{n}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} v^{n-1} e^{-\frac{nv^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2t^2}{2}} dv = \frac{2(\frac{n}{2})^{\frac{n}{2}}}{\sqrt{2\pi}\Gamma(\frac{n}{2})} \int_0^{\infty} v^n e^{-\frac{v^2}{2}(n+t^2)} dv.$$

令 $u = \frac{v^2}{2}(n+t^2)$, 则 $v = \frac{\sqrt{2u}}{\sqrt{n+t^2}}$, $\frac{du}{dv} = v(n+t^2) = \sqrt{2u(n+t^2)}$, $dv = \frac{du}{\sqrt{2u(n+t^2)}}$:

$$f_{\frac{Z}{V}}(t) = \frac{2(\frac{n}{2})^{\frac{n}{2}}}{\sqrt{2\pi}\Gamma(\frac{n}{2})} \int_0^{\infty} \frac{(2u)^{\frac{n}{2}}}{(n+t^2)^{\frac{n}{2}}} e^{-u} \frac{(2u)^{-\frac{1}{2}}}{\sqrt{(n+t^2)}} du = \frac{2(\frac{n}{2})^{\frac{n}{2}}}{\sqrt{2\pi}\Gamma(\frac{n}{2})} \frac{2^{\frac{n-1}{2}}}{(n+t^2)^{\frac{n+1}{2}}} \int_0^{\infty} u^{\frac{n-1}{2}} e^{-u} du$$

$$= \frac{1}{\sqrt{n\pi}\Gamma(\frac{n}{2})} (1+\frac{t^2}{n})^{-\frac{n+1}{2}} \int_0^{\infty} u^{\frac{n-1}{2}} e^{-u} du = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} (1+\frac{t^2}{n})^{-\frac{n+1}{2}}.$$

由此, 我们便得到了 t 分布的概率密度函数.

3. F 分布

Proof

F 分布用来比较两个样本的方差. 如果两个方差相等, 则这个比值在 1 附近; 这个值越大, 他们越可能不相等.

设 $U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且 U, V 相互独立. 那么我们便可以得到 $Y = \frac{U}{n_1}$, $X = \frac{V}{n_2}$的概率密度函数:

$$f_{\frac{U}{n_1}}(y) = \frac{(\frac{n_1}{2})^{\frac{n_1}{2}}}{\Gamma(\frac{n_1}{2})} y^{\frac{n_1}{2}-1} e^{-\frac{n_1 y}{2}},$$

$$f_{\frac{V}{n_2}}(x) = \frac{(\frac{n_2}{2})^{\frac{n_2}{2}}}{\Gamma(\frac{n_2}{2})} x^{\frac{n_2}{2}-1} e^{-\frac{n_2 x}{2}}.$$

最后我们便可以得到 $F = \frac{Y}{X}$的概率密度函数, 因为 X, Y 非负:

$$f_{\frac{Y}{X}}(z) = \int_0^\infty x f(x, xz) dx = \int_0^\infty x f_X(x) f_Y(xz) dx$$

$$= \int_0^\infty x \frac{(\frac{n_2}{2})^{\frac{n_2}{2}}}{\Gamma(\frac{n_2}{2})} x^{\frac{n_2}{2}-1} e^{-\frac{n_2 x}{2}} \frac{(\frac{n_1}{2})^{\frac{n_1}{2}}}{\Gamma(\frac{n_1}{2})} (xz)^{\frac{n_1}{2}-1} e^{-\frac{n_1 xz}{2}} dx$$

$$= \frac{(\frac{n_2}{2})^{\frac{n_2}{2}} (\frac{n_1}{2})^{\frac{n_1}{2}}}{\Gamma(\frac{n_2}{2})\Gamma(\frac{n_1}{2})} z^{\frac{n_1}{2}-1} \int_0^\infty x^{\frac{n_1+n_2}{2}-1} e^{-\frac{x(n_1 z+n_2)}{2}} dx$$

让 $u = \frac{x(n_1 z+n_2)}{2}$, 所以 $x = \frac{2u}{n_1 z+n_2}$, $dx = \frac{2}{n_1 z+n_2} du$:

$$f_F(z) = \frac{(\frac{n_2}{2})^{\frac{n_2}{2}} (\frac{n_1}{2})^{\frac{n_1}{2}}}{\Gamma(\frac{n_2}{2})\Gamma(\frac{n_1}{2})} z^{\frac{n_1}{2}-1} \int_0^\infty (\frac{2u}{n_1 z+n_2})^{\frac{n_1+n_2}{2}-1} e^{-u} \frac{2}{n_1 z+n_2} du$$

$$= \frac{(\frac{n_2}{2})^{\frac{n_2}{2}} (\frac{n_1}{2})^{\frac{n_1}{2}}}{\Gamma(\frac{n_2}{2})\Gamma(\frac{n_1}{2})} z^{\frac{n_1}{2}-1} 2^{\frac{n_1+n_2}{2}} (\frac{1}{n_1 z+n_2})^{\frac{n_1+n_2}{2}} \int_0^\infty u^{\frac{n_1+n_2}{2}-1} e^{-u} du$$

$$= \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_2}{2})\Gamma(\frac{n_1}{2})} z^{\frac{n_1}{2}-1} 2^{\frac{n_1+n_2}{2}} \frac{(\frac{n_2}{2})^{\frac{n_2}{2}} (\frac{n_1}{2})^{\frac{n_1}{2}}}{(n_1 z+n_2)^{\frac{n_1+n_2}{2}}} = \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_2}{2})\Gamma(\frac{n_1}{2})} z^{\frac{n_1}{2}-1} \frac{(\frac{n_2}{2})^{\frac{n_2}{2}} (\frac{n_1}{2})^{\frac{n_1}{2}} (\frac{n_2}{2})^{-\frac{n_1+n_2}{2}}}{(n_1 z+n_2)^{\frac{n_1+n_2}{2}} n_2^{-\frac{n_1+n_2}{2}}}$$

$$= \frac{\Gamma(\frac{n_1+n_2}{2}) (\frac{n_1}{n_2})^{\frac{n_1}{2}}}{\Gamma(\frac{n_2}{2})\Gamma(\frac{n_1}{2}) (\frac{n_1}{n_2} z+1)^{\frac{n_1+n_2}{2}}} z^{\frac{n_1}{2}-1}.$$

由此, 我们便得到了 F 分布的概率密度函数.