# Distribution-hard

zwq

2022.10.22

本文介绍比较复杂但是在应用中非常重要的5个分布: $\Gamma$分布, Beta分布, $\chi^2$分布, $t$分布, $F$分布。另外介绍了一下Dirichlet分布. 本文收集了其概率密度函数(probability density function, pdf)的证明(在大多数的工科教科书中并没有).

此外，没有依赖现有的概率密度函数计算库，仅根据每个分布pdf的数学定义,分别绘制了这些分布的概率密度曲线.

## 预备知识

**$\Gamma$函数和Beta函数**

由Euler总结的2个著名的反常积分：

$$\begin{aligned}\Gamma(\alpha) &= \int_0^{+\infty} x^{\alpha-1}e^{-x}dx \\ \mathrm{Beta}(\alpha,\beta) &= \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx\end{aligned} \tag{1}$$

这两个反常积分有很多变体(换元积分法). 例如Beta函数可以方便地计算三角函数的积分(类似于Wallis公式)

$$B(\frac{m+1}{2},\frac{n+1}{2}) = 2\int_0^{\frac{\pi}{2}} \cos^m\theta\cdot\sin^n\theta d\theta$$

此外，还有一些基本结论:

$$\begin{gathered}\Gamma(\alpha+1)=\alpha\Gamma(\alpha),\ \alpha\in\mathbb{R}^+ \\ \Gamma(n+1)=n!,\ \Gamma(n+\tfrac{1}{2})=\tfrac{\sqrt{\pi}}{2^{2n}}\tfrac{\Gamma(2n+1)}{\Gamma(n+1)},\ n\in\mathbb{N}^+ \\ \Gamma(1)=1,\ \Gamma(\tfrac{1}{2})=\sqrt{\pi} \\ \mathrm{Beta}(\alpha,\beta)=\mathrm{Beta}(\beta,\alpha)=\tfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \\ \mathrm{Beta}(\tfrac{1}{2},\tfrac{1}{2})=\pi,\ \mathrm{Beta}(1,1)=\tfrac{1}{2}\end{gathered}$$

## 1 $\Gamma$分布

$$\begin{aligned}\frac{\Gamma(\alpha)}{\Gamma(\alpha)} &= \frac{1}{\Gamma(\alpha)}\int_0^{+\infty} t^{\alpha-1}e^{-t}dt \\ &\overset{t=\beta x}{=} \int_0^{+\infty}\frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}e^{-\beta x}dx \\ &= 1\end{aligned} \tag{2}$$

本pdf和实验代码可在`https://github.com/wqzh/math/dist`下载

version 1.0

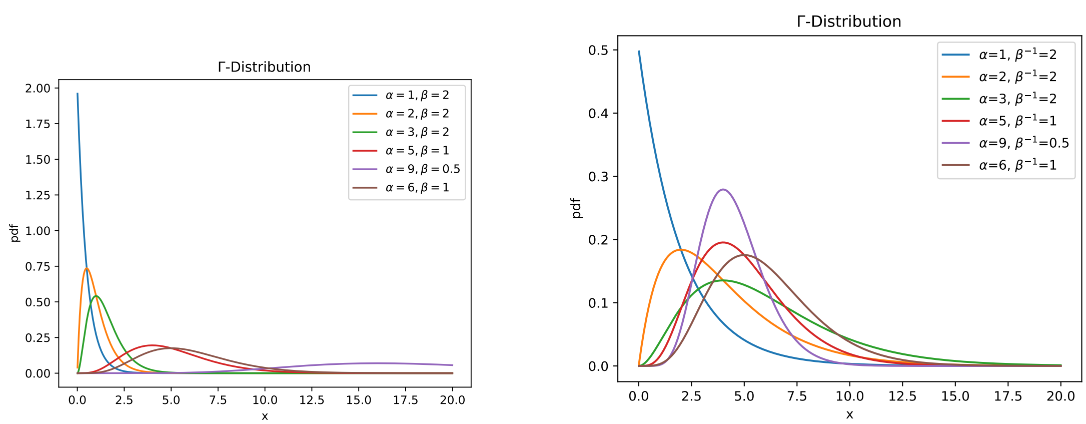
那么就得到了$\Gamma$分布的概率密度函数:

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{1}{\Gamma(\alpha)}\beta^{\alpha}x^{\alpha-1}e^{-\beta x}, & x>0\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

其中两个参数$\alpha,\beta>0$, $\alpha$为形状参数,$\beta$为尺度(逆尺度)参数.
就是说，也可以写成如下的形式(这种形式也更常用, 下图(b)就是基于这个pdf绘制的):

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-\frac{1}{\beta}x}, & x>0\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

当$\alpha<1$, $f(x;\alpha,\beta)$为递减函数;
当$\alpha=1$, $f(x;\alpha,\beta)$为递减函数;
当$\alpha>1$, $f(x;\alpha,\beta)$为单峰函数;




(a) (b)

Gamma 分布

在绘制Gamma分布的pdf曲线时，有一个非常有意思现象: 如果按照$f(x;\alpha,\beta)$绘图，得到的结果不好看Fig. 1a, 但是如果用$\lambda=\beta^{-1}$代替Gamma分布的$\beta$时，可得到$f(x;\alpha,\beta^{-1})$的曲线更美观,也更常用. Fig. 1b. 这可能是$\beta$被称为为尺度(逆尺度)参数的原因.

# 2 Beta分布

n-Bernouli试验，每一次事件发生的概率为p且互相独立, 那么这个试验X服从n重伯努利分布: $X\sim B(n,p)$

$$Pr(x=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

又因为在实验之前，没有任何先验知识,我们不知道p的值是什么。因此假设p服从均匀分布$p\sim U(0,1)$, 那么可得p的概率密度函数$p(x)\equiv 1$. 那么X的概率累计函数为: 当$x\le 0 (x\ge 1)$时， $F_X(x)=0(1)$. 当$0<x<1$时,有

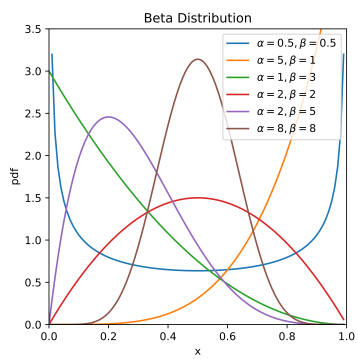
$$F_X(x)=Pr(p\le x|k,n)=\frac{Pr(k,n,p\le x)}{Pr(k,n)}$$

$$
\begin{aligned}
Pr(k,n) &= \int_0^1 Pr(k,n,p=x)p(x)dx \\
&= \int_0^1 \binom{n}{k} x^k(1-x)^{n-k}dx \\
&= C, \text{ is a constant.}
\end{aligned} \tag{5}
$$

因此

$$
F_X(x) = \frac{\int_0^x \binom{n}{k} t^k(1-t)^{n-k}dt}{Pr(k,n)}
$$

$$
\begin{aligned}
f_X(x) &= F'_X(x) \\
&= \left(\frac{\int_0^x \binom{n}{k} t^k(1-t)^{n-k}dt}{Pr(k,n)}\right)' \\
&= \frac{\binom{n}{k} x^k(1-x)^{n-k}}{Pr(k,n)} \\
&= \frac{x^k(1-x)^{n-k}}{\int_0^1 u^k(1-u)^{n-k}du}
\end{aligned} \tag{6}
$$




Beta 分布

令$k=\alpha-1, n-k=\beta-1$,那么就得到Beta分布的概率密度函数

$$
f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\mathrm{Beta}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0<x<1 \\ 0, & \text{otherwise} \end{cases} \tag{7}
$$

因为$\mathrm{Beta}(\alpha,\beta)=\mathrm{Beta}(\beta,\alpha)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, 因此可以替换为

$$
f(x;\alpha,\beta) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0<x<1 \\ 0, & \text{otherwise} \end{cases}
$$

**Dirichlet分布**

Beta分布是基于二项分布Binomial (n重伯努利分布)，即：n次实验，每次的结果只有两种(0,1; 发生或不发生; 掷硬币).

推而广之，多项分布Multinomial, 即多次实验，每次的结果只有n种情况(掷骰子6种情况).

类比于Beta分布，可得到Dirichlet分布的pdf大致如下:

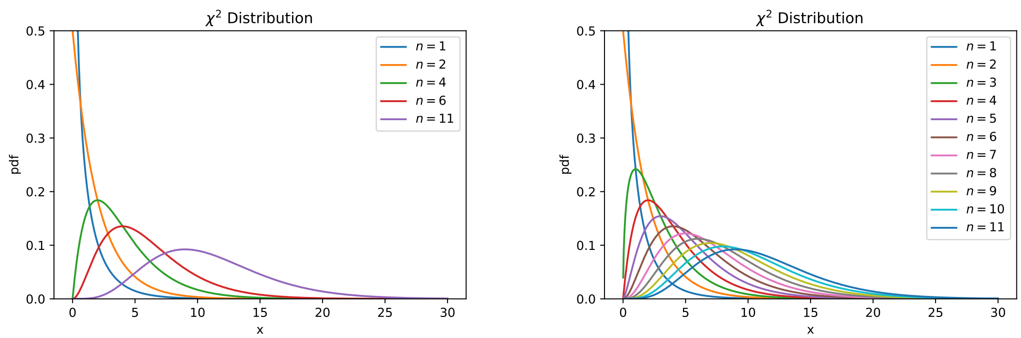
$$f(x;\alpha)=\begin{cases}\dfrac{\Gamma(\sum_{i=1}^{n}\alpha_i)}{\prod_{i=1}^{i=n}\Gamma(\alpha_i)}\prod_{i=1}^{i=n}x_i^{\alpha_i-1},\ 0<x_i<1,\sum_{i=1}^{n}x_i=1\\ 0,\quad \text{otherwise}\end{cases}\tag{8}$$

多维空间，不好绘制图像

# 3 $\chi^2$分布

假设随机变量$X_1,X_2,\cdots,X_n$独立同分布于标准正态分布$N(0,1)$.
那么随机变量$X=\sum_{i=1}^{n}X_i^2$满足$X\sim\chi^2(n)$.其概率密度函数为:

$$f(x;n)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}},\ x>0\\ 0,\quad \text{otherwise}\end{cases}\tag{9}$$




(a) (b)

$\chi^2$分布

**证明:**

$$Y=X_1^2+X_2^2+\cdots+X_n^2$$

$$\begin{aligned}F_Y(y)&=Pr(Y\le y)=Pr(X_1^2+X_2^2+\cdots+X_n^2\le y)\\&=\int\cdots\int_{x_1^2+x_2^2+\cdots+x_n^2\le y}\frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{x_1^2+x_2^2+\cdots+x_n^2}{2}}dx_1dx_2\cdots dx_n\\&=\int_0^{\sqrt{y}}\int_{\theta_1=\psi_1}^{\varphi_1}\cdots\int_{\theta_{n-1}=\psi_{n-1}}^{\varphi_{n-1}}\frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{r^2}{2}}r^{n-1}drd\theta_1\cdots d\theta_{n-1}\\&=C_n\int_0^{\sqrt{y}}e^{-\frac{r^2}{2}}r^{n-1}dr\end{aligned}\tag{10}$$

求导得概率密度函数:

$$\begin{aligned}f_Y(y)&=(F_Y(y))'\\&=\frac{\partial(C_n\int_0^{\sqrt{y}}e^{-\frac{r^2}{2}}r^{n-1}dr)}{\partial y}\\&=C_ne^{-\frac{y}{2}}y^{\frac{n-1}{2}}(\sqrt{y})'\\&=\frac{C_n}{2}e^{-\frac{y}{2}}y^{\frac{n}{2}-1}\end{aligned}\tag{11}$$

又因为概率密度积分为1,可得

$$\begin{aligned}
&\int_0^{+\infty} f_Y(y)dy \\
&= \frac{C_n}{2}\int_0^{+\infty} e^{-\frac{y}{2}} y^{\frac{n}{2}-1}dy \\
&\overset{x=\frac{y}{2}}{=} \frac{C_n}{2}\int_0^{+\infty} e^{-x}(2x)^{\frac{n}{2}-1}2dx \\
&= C_n 2^{\frac{n}{2}-1}\int_0^{+\infty} e^{-x}x^{\frac{n}{2}-1}dx \\
&= \frac{C_n}{2}2^{\frac{n}{2}}\Gamma(\frac{n}{2}) = 1
\end{aligned} \tag{12}$$

即$\frac{C_n}{2} = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}$
代入上式可得$\chi^2$分布的pdf:

$$f_Y(y) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}e^{-\frac{y}{2}}y^{\frac{n}{2}-1},\ y > 0$$

**引理 1 二维连续型随机变量商的分布** 二维随机变量$(X,Y)$的联合概率密度为$p(x,y)$. 那么$Z = \frac{X}{Y}$的概率密度$f_Z(z) = \int_{-\infty}^{+\infty} p(zy,y)|y|dy$.

**证明:**

$$\begin{aligned}
F_Z(z) &= Pr(Z \le z) = Pr(\frac{X}{Y} \le z) \\
&= \int_{-\infty}^{0} dy \int_{yz}^{+\infty} p(x,y)dx + \int_0^{+\infty} dy \int_{-\infty}^{yz} p(x,y)dx \\
&\overset{x=uy}{=} \int_{-\infty}^{0} dy \int_{z}^{+\infty} p(uy,y)ydu + \int_0^{+\infty} dy \int_{-\infty}^{z} p(uy,y)ydu \\
&= \int_{z}^{+\infty} du \int_{-\infty}^{0} p(uy,y)ydy + \int_{-\infty}^{z} du \int_0^{+\infty} p(uy,y)ydy
\end{aligned} \tag{13}$$

求导得:

$$\begin{aligned}
f_Z(z) &= (F_Z(z))' \\
&= -\int_{-\infty}^{0} p(zy,y)ydy + \int_0^{+\infty} p(zy,y)ydy \\
&= \int_{-\infty}^{+\infty} p(zy,y)|y|dy
\end{aligned} \tag{14}$$

**证毕.**

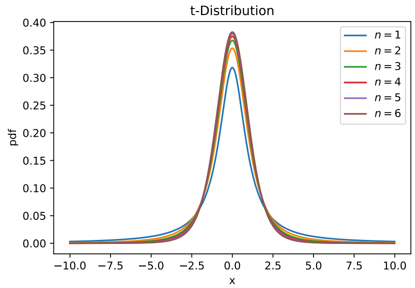
# 4 $t$分布

$X \sim N(0,1),\ Y \sim \chi^2(n)$, 那么随机变量$Z = \frac{X}{\sqrt{Y/n}}$服从自由度为n的t分布, 即:$Z \sim t(n)$

$$f(t;n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{t^2}{n})^{-\frac{n+1}{2}},\ -\infty < t < \infty \tag{15}$$

因为$X \sim N(0,1), Y \sim \chi^2(n)$. 记随机变量$W = \sqrt{\frac{Y}{n}}$, 显然W恒大于零. 当$w < 0$时, $f_W(w) = 0$. 当$w > 0$时,因为Y服从卡方分布,由卡方分布的pdf有:

$$\begin{aligned}
F_W(w) &= Pr(W \le w) = Pr(Y \le nw^2) \\
&= \int_0^{nw^2} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx
\end{aligned} \tag{16}$$




t分布

求导得$W=\sqrt{\frac{Y}{n}}$的pdf:(在第5节-F分布中,还会求得$W=\frac{Y}{n}$的pdf)

$$
\begin{aligned}
f_W(w;n) &= (F_W(w))' \\
&= \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}(nw^2)^{\frac{n}{2}-1}e^{-\frac{nw^2}{2}}2nw \\
&= \frac{1}{\Gamma(\frac{n}{2})}2^{1-\frac{n}{2}}n^{\frac{n}{2}}w^{n-1}e^{-\frac{nw^2}{2}}
\end{aligned}
\tag{17}
$$

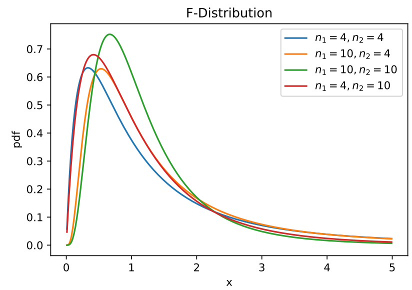
$Z=\frac{X}{\sqrt{Y/n}}$,可记为 $T=\frac{X}{W}$. 已知X的pdf(标准正态分布)和W的pdf, 又因为X和W相互独立，那么$(X,W)\sim p(x,w)=p(x)\cdot f_W(w)$. 那么由引理 1可求出两个随机变量商的分布:

$$
\begin{aligned}
f_T(t) &= \int_{-\infty}^{+\infty} p(tw,w)|w|dw = \int_{-\infty}^{+\infty} p_X(tw)p_W(w)|w|dw \\
&= \int_0^{+\infty} p_X(tw)p_W(w)wdw \\
&= \int_0^{+\infty}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2w^2}{2}}\right)\left(\frac{1}{\Gamma(\frac{n}{2})}2^{1-\frac{n}{2}}n^{\frac{n}{2}}w^{n-1}e^{-\frac{nw^2}{2}}\right)wdw \\
&= \left(\frac{1}{\sqrt{\pi}\Gamma(\frac{n}{2})}2^{\frac{1-n}{2}}n^{\frac{n}{2}}\right)\int_0^{+\infty}e^{-\frac{(n+t^2)}{2}w^2}w^n dw \\
&\overset{w=\sqrt{\frac{2z}{n+t^2}}}{=} \left(\frac{1}{\sqrt{\pi}\Gamma(\frac{n}{2})}2^{\frac{1-n}{2}}n^{\frac{n}{2}}\right)\int_0^{+\infty}e^{-z}\frac{1}{n+t^2}\left(\frac{2}{n+t^2}\right)^{\frac{n-1}{2}}z^{\frac{n-1}{2}}dz \\
&= \left(\frac{1}{\sqrt{\pi}\Gamma(\frac{n}{2})}2^{\frac{1-n}{2}}n^{\frac{n}{2}}\right)\frac{1}{2}\left(\frac{2}{n+t^2}\right)^{\frac{n+1}{2}}\int_0^{+\infty}e^{-z}z^{\frac{n-1}{2}}dz \\
&= \left(\frac{1}{\sqrt{\pi}\Gamma(\frac{n}{2})}2^{\frac{1-n}{2}}n^{\frac{n}{2}}\right)\frac{1}{2}\left(\frac{2}{n+t^2}\right)^{\frac{n+1}{2}}\Gamma(\frac{n+1}{2}) \\
&= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})}n^{\frac{n}{2}}(\frac{n^{\frac{1}{2}}}{\sqrt{n}})\left(\frac{1}{n+t^2}\right)^{\frac{n+1}{2}} \\
&= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}
\end{aligned}
\tag{18}
$$

# 5 $F$分布

$X\sim\chi^2(n_1),\ Y\sim\chi^2(n_2)$, 那么随机变量$Z=\frac{X/n_1}{Y/n_2}$服从自由度为$n_1,n_2$的F分布, 即:$Z\sim F(n_1,n_2)$

$$f(t;n_1,n_2)=\begin{cases}\dfrac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}t^{\frac{n_1}{2}-1}(n_1t+n_2)^{-\frac{n_1+n_2}{2}},\ t>0\\ 0,\quad \text{otherwise}\end{cases} \tag{19}$$




F分布

因为$X\sim\chi^2(n_1)$, $Y\sim\chi^2(n_2)$, 记新的随机变量$X'=X/n_1$, $Y'=Y/n_2$, 那么有$T=\frac{X/n_1}{Y/n_2}=\frac{X'}{Y'}$,由卡方分布的pdf,可得X,Y的pdf为:

$$f_X(x)=\frac{1}{2^{\frac{n_1}{2}}\Gamma(\frac{n_1}{2})}e^{-\frac{x}{2}}x^{\frac{n_1}{2}-1},\ x>0$$

$$f_Y(y)=\frac{1}{2^{\frac{n_2}{2}}\Gamma(\frac{n_2}{2})}e^{-\frac{y}{2}}y^{\frac{n_2}{2}-1},\ y>0$$

可得$X',Y'$的pdf为:

$$f_{X'}(x)=\frac{n_1^{\frac{n_1}{2}}}{2^{\frac{n_1}{2}}\Gamma(\frac{n_1}{2})}e^{-\frac{n_1x}{2}}x^{\frac{n_1}{2}-1},\ x>0$$

$$f_{Y'}(y)=\frac{n_2^{\frac{n_2}{2}}}{2^{\frac{n_2}{2}}\Gamma(\frac{n_2}{2})}e^{-\frac{n_2y}{2}}y^{\frac{n_2}{2}-1},\ y>0$$

**证明**

$X\sim\chi^2(n)$, 则$W=\frac{X}{n}$的pdf为(证明如下):

$$\begin{aligned}F_W(w)&=Pr(W\le w)=Pr(X\le nw)\\&=\int_0^{nw}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx\end{aligned} \tag{20}$$

求导得$W=\frac{X}{n}$的pdf:

$$\begin{aligned}f_W(w;n)&=(F_W(w))'\\&=\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}(nw)^{\frac{n}{2}-1}e^{-\frac{nw}{2}}n\\&=\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}n^{\frac{n}{2}}w^{\frac{n}{2}-1}e^{-\frac{nw}{2}}\end{aligned} \tag{21}$$

**证毕**

记随机变量$T=\frac{X'}{Y'}$, 显然T恒大于零. 当$t\le 0$时, $f_T(t)=0$. 当$t>0$时,有

$$F_T(t)=Pr(T\le t)=Pr(\frac{X'}{Y'}\le t)=Pr(X'\le Y't)$$

由于X,Y相互独立, 因此$X'$, $Y'$也相互独立,根据引理 1有:

$$
\begin{aligned}
f_T(t) &= \int_{-\infty}^{+\infty} p(ty,y)|y|dy = \int_0^{+\infty} p_{X'}(ty)p_{Y'}(y)ydy \\
&= \int_0^{+\infty} \left( \frac{{n_1}^{\frac{n_1}{2}}}{2^{\frac{n_1}{2}}\Gamma(\frac{n_1}{2})} e^{-\frac{n_1 ty}{2}} (ty)^{\frac{n_1}{2}-1} \right) \left( \frac{{n_2}^{\frac{n_2}{2}}}{2^{\frac{n_2}{2}}\Gamma(\frac{n_2}{2})} e^{-\frac{n_2 y}{2}} y^{\frac{n_2}{2}-1} \right) ydy \\
&= \left( \frac{{n_1}^{\frac{n_1}{2}}{n_2}^{\frac{n_2}{2}}}{2^{\frac{n_1+n_2}{2}}\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \right) t^{\frac{n_1}{2}-1} \int_0^{+\infty} e^{-\frac{(n_1 t+n_2)y}{2}} y^{\frac{n_1+n_2}{2}-1} ydy \\
&\overset{y=\frac{2z}{n_1 t+n_2}}{=} \left( \frac{{n_1}^{\frac{n_1}{2}}{n_2}^{\frac{n_2}{2}}}{2^{\frac{n_1+n_2}{2}}\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \right) \int_0^{+\infty} e^{-z} \left( \frac{2}{n_1 t+n_2} \right)^{\frac{n_1+n_2}{2}} z^{\frac{n_1+n_2}{2}-1} t^{\frac{n_1}{2}-1} dz \\
&= \left( \frac{{n_1}^{\frac{n_1}{2}}{n_2}^{\frac{n_2}{2}}}{2^{\frac{n_1+n_2}{2}}\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \right) \left( \frac{2}{n_1 t+n_2} \right)^{\frac{n_1+n_2}{2}} t^{\frac{n_1}{2}-1} \int_0^{+\infty} e^{-z} z^{\frac{n_1+n_2}{2}-1} dz \\
&= \left( \frac{{n_1}^{\frac{n_1}{2}}{n_2}^{\frac{n_2}{2}}}{2^{\frac{n_1+n_2}{2}}\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \right) \left( \frac{2}{n_1 t+n_2} \right)^{\frac{n_1+n_2}{2}} t^{\frac{n_1}{2}-1} \Gamma(\frac{n_1+n_2}{2}) \\
&= \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} {n_1}^{\frac{n_1}{2}} {n_2}^{\frac{n_2}{2}} (\frac{1}{n_1 t+n_2})^{\frac{n_1+n_2}{2}} t^{\frac{n_1}{2}-1} \\
&= \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} {n_1}^{\frac{n_1}{2}} {n_2}^{\frac{n_2}{2}} t^{\frac{n_1}{2}-1} (n_1 t+n_2)^{-\frac{n_1+n_2}{2}}
\end{aligned}
\tag{22}
$$