相关性分析

Pearson皮尔逊相关系数

对于两组总体数据$X:\{X_1,X_2,\cdots,X_n\}$和$Y:\{Y_1,Y_2,\cdots Y_n\}$，其总体均值为：

$$\begin{aligned} E(X)&=\frac{\sum_{i=1}^{n}X_i}{n},\\ E(Y)&=\frac{\sum_{i=0}^{n}Y_i}{n}. \end{aligned} \tag{1}$$

因此总体Covariance协方差为：

$$\mathrm{Cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-E(X))(Y_i-E(Y))}{n}, \tag{2}$$

协方差可以通俗的理解为：两个变量在变化过程中是同方向变化，还是反方向变化，同向或反向程度如何[1]。

总体皮尔逊相关系数定义为：

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}\frac{X_i-E(X)}{\sigma_X}\frac{Y_i-E(Y)}{\sigma_Y}}{n}, \tag{3}$$

上式中，σ_X和σ_Y分别表示总体数据X和Y的标准差，其计算公式为：

$$\begin{aligned} \sigma_X&=\sqrt{\frac{\sum_{i=1}^{n}(X-E(X))^2}{n}},\\ \sigma_Y&=\sqrt{\frac{\sum_{i=1}^{n}(Y-E(Y))^2}{n}}. \end{aligned} \tag{4}$$

皮尔逊相关系数也可以看成是剔除了两个变量量纲影响，即将X和Y标准化后的协方差。当X和Y不是总体数据而是样本数据时，其计算同理。对于皮尔逊相关系数取值的含义如图1所示[2]，图中的r即位皮尔逊相关系数值。

r value	Interpretation
$r=1$	Perfect positive linear correlation
$1>r\geq 0.8$	Strong positive linear correlation
$0.8>r\geq 0.4$	Moderate positive linear correlation
$0.4>r>0$	Weak positive linear correlation
$r=0$	No correlation
$0>r\geq -0.4$	Weak negative linear correlation
$-0.4>r\geq -0.8$	Moderate negative linear correlation
$-0.8>r>-1$	Strong negative linear correlation
$r=-1$	Perfect negative linear correlation

图1 维基百科Strength of Correlation

当然，对于计算出来的相关系数是否有效，要看每个相关系数对应的p值是否显著才行。需要注意的是，皮尔逊相关系数只能用来衡量两个变量的线性相关程度，其适用场景为**连续数据**，**正态分布**，**线性关系**，这三个条件均满足才能使用皮尔逊相关系数，否则就用斯皮尔曼相关系数[3]。

Spearman斯皮尔曼相关系数

斯皮尔曼相关系数也被视为是皮尔逊相关系数的一种变形，即它先把原始数据变成对应的秩（排序），然后再计算这些秩之间的皮尔逊相关系数[4]。斯皮尔曼相关系数也被称为斯皮尔曼值秩相关系数，其适用条件比皮尔逊相关系数要广，只要数据满足**单调关系**（例如线性函数、指数函数、对数函数等）就能够使用，当数据不满足皮尔逊相关系数的要求时，如离散变量之间，或离散与连续变量之间，人们常常使用斯皮尔曼相关性系数进行计算。

对于两个随机变量的取值$\{X_i\}_{i=1}^n$和$\{Y_i\}_{i=1}^n$，对其从小到大进行排序，并将最小的值对应秩为1，第二小的值对应秩为2，依此类推，直到最大的值对应秩n。若有相同值，可以将这些相同值的秩取平均后再赋给它们（这是处理**并列秩**的方法）[4]。

将样本X_i的秩$R(X_i)$和样本Y_i的秩$R(Y_i)$代入公式(3)可得斯皮尔曼相关系数的计算公式：

$$\rho = \frac{\sum_{i=1}^n (R(X_i) - \overline{R(X)})(R(Y_i) - \overline{R(Y)})}{\sqrt{\sum_{i=1}^n (R(X_i) - \overline{R(X)})^2}\sqrt{\sum_{i=1}^n (R(Y_i) - \overline{R(Y)})^2}}. \tag{5}$$

对于离散数据还可以具体分为两种，一种是无序的，比如学院中的物联专业、信计专业、数据专业等；另一种则是定序的，比如成绩分段中的优良中差。这两种类型的数据，斯皮尔曼相关系数都可以应对。更进一步地，对于无序数据，斯皮尔曼相关系数有一个更加简洁的计算公式：

已知：

$$\overline{R(X)} = \overline{R(Y)} = \frac{n+1}{2}. \tag{6}$$

则公式(5)的分子部分可化简为：

$$\begin{aligned} &\sum_{i=1}^n (R(X_i) - \frac{n+1}{2})(R(Y_i) - \frac{n+1}{2}) = \\ &\sum_{i=1}^n R(X_i)R(Y_i) - \frac{n+1}{2}\sum_{i=1}^n R(X_i) - \frac{n+1}{2}\sum_{i=1}^n R(Y_i) + n\frac{(n+1)^2}{4}. \end{aligned} \tag{7}$$

由于：

$$\sum_{i=1}^n R(X_i) = \sum_{i=1}^n R(Y_i) = \frac{n(n+1)}{2}, \tag{8}$$

代入公式(7)可得：

$$\sum_{i=1}^n (R(X_i) - \frac{n+1}{2})(R(Y_i) - \frac{n+1}{2}) = \sum_{i=1}^n R(X_i)R(Y_i) - \frac{n(n+1)^2}{4}. \tag{9}$$

定义等级差$d_i = R(X_i) - R(Y_i)$，则等级差的平方和为：

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (R(X_i) - R(Y_i))^2 = \sum_{i=1}^n (R(X_i)^2 + R(Y_i)^2 - 2R(X_i)R(Y_i)). \tag{10}$$

即：

$$\sum_{i=1}^{n} R(X_i)R(Y_i) = \frac{1}{2}\left[\sum_{i=1}^{n} R(X_i)^2 + \sum_{i=1}^{n} R(Y_i)^2 - \sum_{i=1}^{n} d_i^2\right]. \tag{11}$$

对于秩$R(X_i)$和$R(Y_i)$，有：

$$\sum_{i=1}^{n} R(X_i)^2 = \sum_{i=1}^{n} R(Y_i)^2 = \frac{n(n+1)(2n+1)}{6}. \tag{12}$$

结合公式(9)和(11)可将公式(5)的分子部分化简为：

$$\sum_{i=1}^{n}(R(X_i) - \frac{n+1}{2})(R(Y_i) - \frac{n+1}{2}) = -\frac{1}{2}\sum_{i=1}^{n} d_i^2. \tag{13}$$

对于公式(5)的分母部分第一项中的根号，可以化简为：

$$\begin{aligned} &\sum_{i=1}^{n}(R(X_i) - \frac{n+1}{2})^2 \\ &= \sum_{i=1}^{n} R(X_i)^2 - n(\frac{n+1}{2})^2 \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \\ &= \frac{n(n^2-1)}{12}. \end{aligned} \tag{14}$$

同理，第二项中的根号，可以化简为：

$$\sum_{i=1}^{n}(R(Y_i) - \frac{n+1}{2})^2 = \frac{n(n^2-1)}{12}. \tag{15}$$

所以分母可以化简为：

$$\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{n(n^2-1)}{12}. \tag{16}$$

综上所述，对于无序情况下的斯皮尔曼相关系数，其计算公式可以化简为：

$$\begin{aligned} \rho &= \frac{-\frac{1}{2}\sum_{i=1}^{n} d_i^2}{\frac{n(n^2-1)}{12}} \\ &= \frac{-6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}. \end{aligned} \tag{17}$$

但是公式(17)的形式中，当$\sum_{i=1}^{n} d_i^2$为0时，计算结果为0，不符合直觉。因此通过1来调整，最终的计算公式如下：

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}. \tag{18}$$

Kendall肯德尔相关系数

肯德尔相关系数适用于定序或连续变量情况下的计算，相比斯皮尔曼相关系数，更适用于小样本的场景。斯皮尔曼相关是基于秩差来进行相关关系的评估，而肯德尔相关则是基于样本数据对之间的关系来进行相关系数的强弱的分析，数据对可以分为一致对（Concordant）和分歧对（Discordant）[5]。

一致对和分歧对的概念如图2所示[6]。

1. 数据对的比较： 对于每对观测值 (X_i, Y_i) 和 (X_j, Y_j)：

- 如果 $X_i > X_j$ 且 $Y_i > Y_j$，则该对被称为 一致对
- 如果 $X_i < X_j$ 且 $Y_i < Y_j$，则该对被称为 一致对
- 如果 $X_i > X_j$ 且 $Y_i < Y_j$，则该对被称为 不一致对
- 如果 $X_i < X_j$ 且 $Y_i > Y_j$，则该对被称为 不一致对
- 如果 $X_i = X_j$ 且 $Y_i = Y_j$，则该对被称为 完全平局对
- 如果 $X_i = X_j$ 且 $Y_i \neq Y_j$，则该对被称为 仅在x中的平局对
- 如果 $X_i \neq X_j$ 且 $Y_i = Y_j$，则该对被称为 仅在y中的平局对

图2 一致对和分歧对的概念

对于肯德尔相关系数τ的计算，首先考虑所有可能的观测值对的数量：

$$C_n^2 = \frac{n(n-1)}{2}. \tag{19}$$

计算一致对的数量C，和分歧对的数量D，其计算公式为：

$$\tau = 2\frac{C-D}{n(n-1)}. \tag{20}$$

这里以斯皮尔曼相关系数的计算和可视化为例，绘制的热力图如图3所示。

```python
from matplotlib import rcParams

rcParams['font.family'] = 'Microsoft YaHei'
rcParams['axes.unicode_minus'] = False

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import spearmanr

df = pd.read_excel(r'八年级男生体测数据.xls')

correlation_matrix, p_value_matrix = spearmanr(df)

correlation_df = pd.DataFrame(correlation_matrix, columns=df.columns, index=df.columns)
p_value_df = pd.DataFrame(p_value_matrix, columns=df.columns, index=df.columns)

plt.figure(figsize=(8, 8))
sns.heatmap(correlation_df, cmap='Blues', annot=True, fmt='.3f', vmin=-1, vmax=1)
plt.title('Spearman Correlation Heatmap of All Columns')
plt.tight_layout()
```

```
plt.savefig(r'Spearman_Correlation_Heatmap_of_All_Columns.png', dpi=600)
plt.show()

print("各列之间的 p 值矩阵：")
print(p_value_df)

p_value_df.to_excel(r'Spearman_Correlation_p_values.xlsx', sheet_name='P-Values', index=True)

```

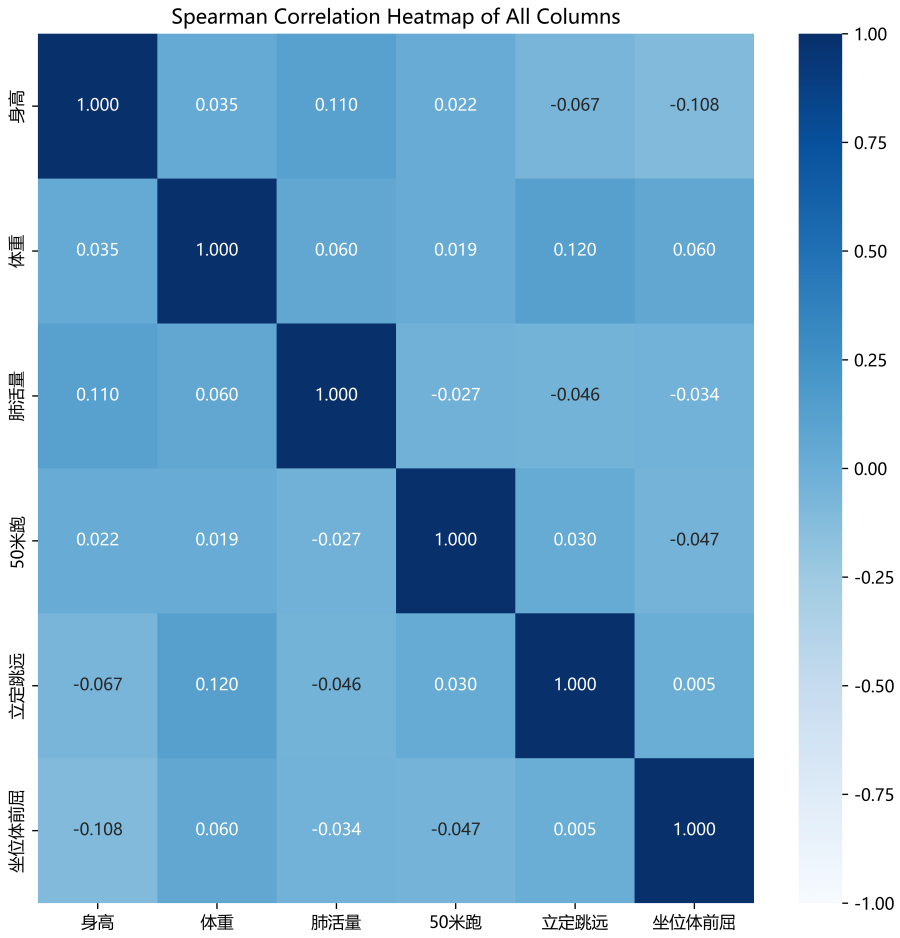

图3 斯皮尔曼相关系数热力图

计算的p值如图4所示：

	身高	体重	肺活量	50米跑	立定跳远	坐位体前屈
身高	0.000	0.346	0.003	0.562	0.072	0.004
体重	0.346	0.000	0.103	0.599	0.001	0.103
肺活量	0.003	0.103	0.000	0.462	0.216	0.355
50米跑	0.562	0.599	0.462	0.000	0.416	0.202
立定跳远	0.072	0.001	0.216	0.416	0.000	0.882
坐位体前屈	0.004	0.103	0.355	0.202	0.882	0.000

图4 各列的p值

正态性检验

Jarque-Bera test 雅克-贝拉检验

雅克-贝拉检验适用于大样本，即样本量$n \geq 30$的情况[7]，其基本思想为：如果样本来自正态分布，那么样本的偏度和峰度应该接近正态分布的偏度和峰度。正态分布中不同偏度和峰度的可视化如图5所示，其中，通过均值来调整偏度，通过标准差来调整峰度。

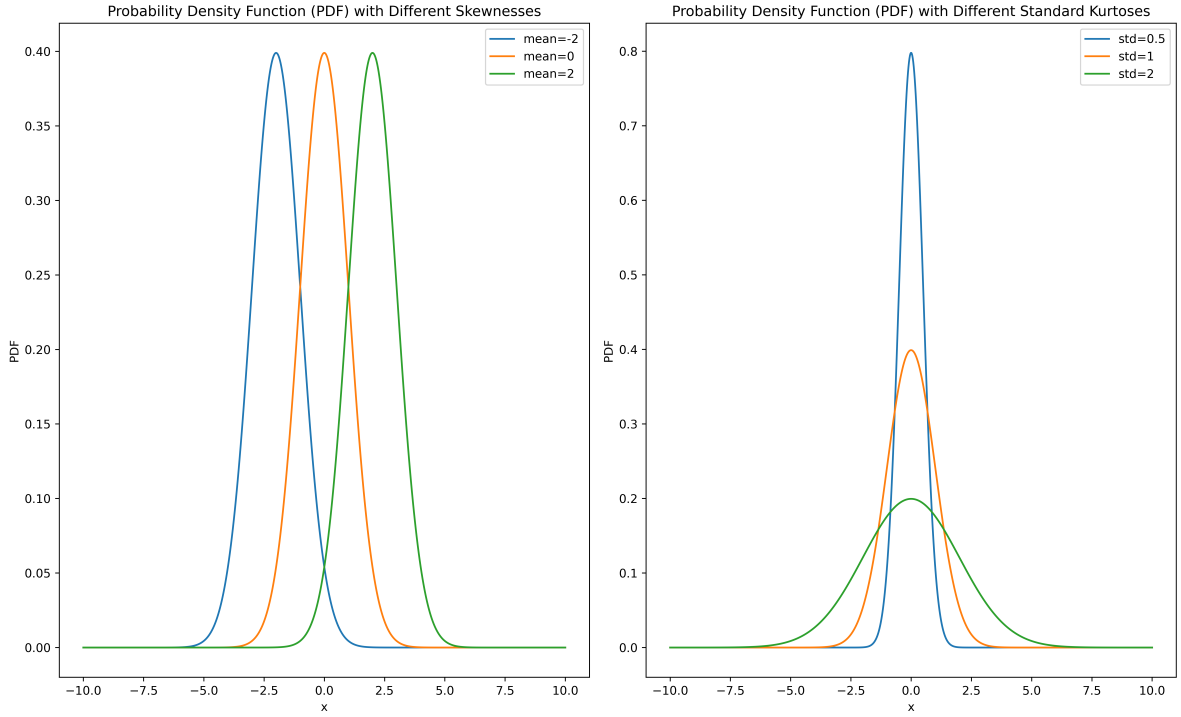

图5 正态分布中不同偏度和峰度的可视化

对于一个随机变量X，其总体偏度的定义为：

$$\gamma_1 = E[(\frac{X-\mu}{\sigma})^3], \tag{21}$$

上式中，μ为均值，σ为标准差。

在样本中，用样本均值$\bar{x}$代替总体均值μ，样本标准差s代替总体标准差σ。因此，样本偏度S可用如下公式计算：

$$S = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3}, \tag{22}$$

上式中，$\frac{n}{(n-1)(n-2)}$一个修正系数，当样本量n较大时，这个系数趋近于1，它的存在是为了使样本偏度成为总体偏度的无偏估计量。

对于一个随机变量X，其总体峰度的定义为：

$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]. \tag{23}$$

在样本中，样本峰度公式K可用如下公式计算：

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}, \tag{24}$$

系数$n(n+1)/[(n-1)(n-2)(n-3)]$和$-3(n-1)^2/[(n-2)(n-3)]$是修正系数，它们的存在是为了使样本峰度成为总体峰度的无偏估计量。当样本量n较大时，这些系数的作用使得样本峰度能较好地估计总体峰度。

雅克贝拉检验的原假设是样本数据服从正态分布，使用Python进行雅克贝拉检验和可视化绘制如图6所示。

```python
import numpy as np
from scipy.stats import jarque_bera, gaussian_kde, skew, kurtosis
import matplotlib.pyplot as plt

# 原始数据
data = np.array(
    [1.26, 0.34, 0.70, 1.75, 50.57, 1.55, 0.08, 0.42, 0.50, 3.20, 0.15, 0.49, 0.95, 0.24, 1.37, 0.17, 6.98, 0.10, 0.94,
     0.38])

# 计算原始数据的Jarque-Bera检验、偏度和峰度
statistic_data, p_value_data = jarque_bera(data)
skewness_data = skew(data)
kurt_data = kurtosis(data, fisher=True)  # Fisher's definition of kurtosis (normal ==> 0.0)

print(f"Original Data - Skewness: {skewness_data:.4f}, Kurtosis: {kurt_data:.4f}")

# 确保所有数据都是正值，然后对数据取自然对数
log_data = np.log(data)  # 过滤掉任何可能存在的非正值

# 计算对数变换后数据的Jarque-Bera检验、偏度和峰度
statistic_log_data, p_value_log_data = jarque_bera(log_data)
skewness_log_data = skew(log_data)
kurt_log_data = kurtosis(log_data, fisher=True)  # Fisher's definition of kurtosis (normal ==> 0.0)

print(f"Log Transformed Data - Skewness: {skewness_log_data:.4f}, Kurtosis: {kurt_log_data:.4f}")

# 创建一个2x1的子图布局
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(14, 8.6))

# 绘制原始数据的概率密度函数
kde_data = gaussian_kde(data)
x_data = np.linspace(min(data), max(data), 1000)
axes[0].plot(x_data, kde_data(x_data), label='PDF')
axes[0].hist(data, density=True, alpha=0.5, bins=10)  # 添加直方图以对比PDF
axes[0].set_title(
    f'Original Data PDF - Jarque-Bera Stat: {statistic_data:.4f}, P-value: {p_value_data:.4f}\nSkewness: {skewness_data:.4f}, Kurtosis: {kurt_data:.4f}')
axes[0].set_xlabel('Value')
axes[0].set_ylabel('Density')
```

```
39
40  # 绘制对数变换后数据的概率密度函数
41  kde_log_data = gaussian_kde(log_data)
42  x_log_data = np.linspace(min(log_data), max(log_data), 1000)
43  axes[1].plot(x_log_data, kde_log_data(x_log_data), label='PDF')
44  axes[1].hist(log_data, density=True, alpha=0.5, bins=10)  # 添加直方图以对比PDF
45  axes[1].set_title(
46      f'Log Transformed Data PDF - Jarque-Bera Stat: {statistic_log_data:.4f},
    P-value: {p_value_log_data:.4f}\nSkewness: {skewness_log_data:.4f},
    Kurtosis: {kurt_log_data:.4f}')
47  axes[1].set_xlabel('Log Value')
48  axes[1].set_ylabel('Density')
49
50  # 调整子图之间的间距
51  plt.tight_layout()
52
53  # 显示图表
54  plt.savefig(r'J-B test.png', dpi=600)
55  plt.show()
```

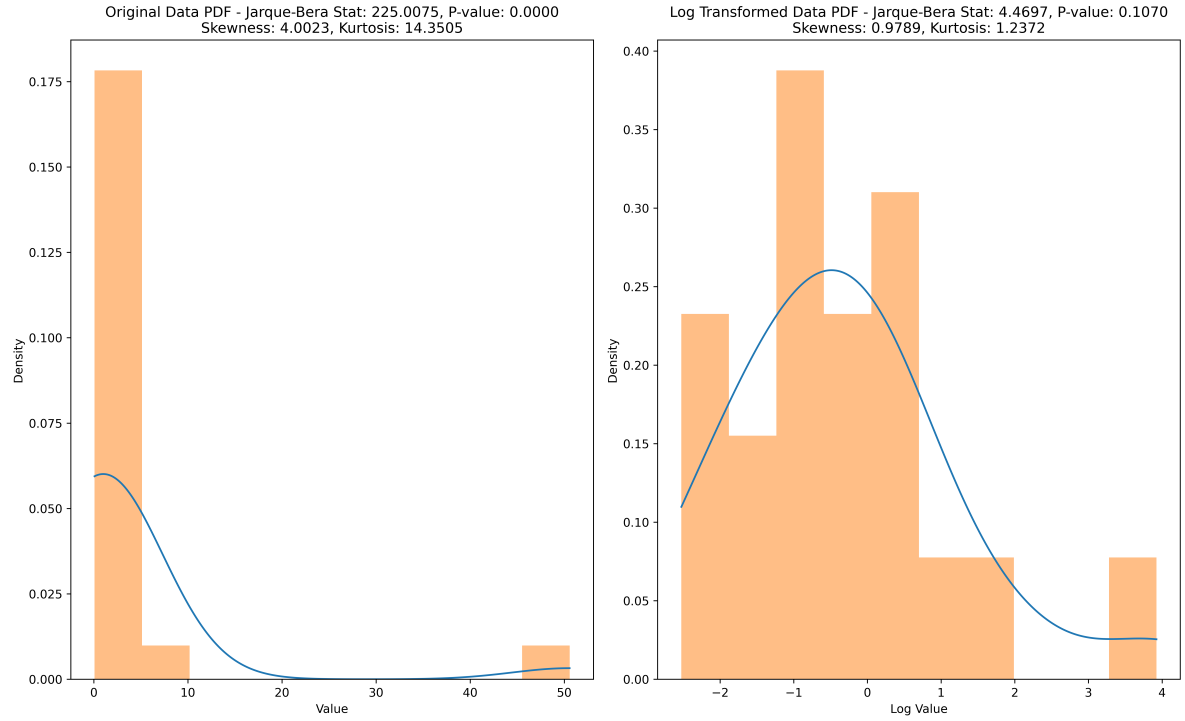

图6 雅克贝拉检验及可视化

由图4可知，取对数前的偏度为4.0023，峰度为14.3505，p值小于0.0001，显著拒绝原假设，不服从正态分布；而在取对数之后，偏度降低至0.9789，峰度降低至1.2372，p值为0.1070，相对于取对数前接近正态分布。

Shapiro-wilk 夏皮洛-威尔克检验

夏皮洛-威尔克检验适用于小样本，即样本量$3 \leq n \leq 50$的情况[8]，其统计量为：

$$W = \frac{(\sum_{i=1}^{n} \alpha_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{25}$$

上式中，$x_{(i)}$表示抽样样本$x_1, x_2, \cdots x_n$从小到大排序得到顺序统计量，$\bar{x}$表示样本均值，系数α_i由如下公式计算：

$$(\alpha_1, \alpha_2, \cdots, \alpha_n) = \frac{m^T V^{-1}}{C}. \tag{26}$$

假设从标准正态分布$N(0,1)$中抽取n个样本$x_1, x_2, \cdots x_n$，$m = (m_1, m_2, \cdots, m_n)$由从标准正态分布中独立同分布抽样得到的顺序统计量（order statistics）的期望值组成，即$m_i = E(x_{(i)})$；V表示正态顺序统计量（normal order statistics）的协方差矩阵，即$V_{ij} = \mathrm{Cov}(x_{(i)}, x_{(j)})$；$C$是向量范数，由如下公式计算：

$$C = \|V^{-1}m\| = (m^T V^{-1} V^{-1} m)^{1/2}. \tag{27}$$

Shapiro-wilk检验的原假设是样本数据符合正态分布，使用Python进行Shapiro-Wilk检验：

```
import numpy as np
from scipy.stats import shapiro

data = np.array(
    [1.26, 0.34, 0.70, 1.75, 50.57, 1.55, 0.08, 0.42, 0.50, 3.20, 0.15,
0.49, 0.95, 0.24, 1.37, 0.17, 6.98, 0.10, 0.94,
     0.38])

# 进行 Shapiro-Wilk 检验
statistic, p_value = shapiro(data)

print(f"Shapiro-Wilk 统计量: {statistic}")
print(f"P 值: {p_value}")
```

代码输出结果为：

```
Shapiro-Wilk 统计量: 0.3246904615503522
P 值: 1.163381439796133e-08
```

因此可以在99%的显著水平下拒绝原假设，该样本数据不服从正态分布。

Kolmogorov-Smirnov 柯尔莫哥洛夫-斯米尔诺夫检验

K-S检验是一种非参数检验方法，主要用于比较一个样本的经验分布函数和一个理论分布函数（比如正态分布、均匀分布等），或者用于比较两个独立样本的经验分布函数是否来自同一分布。下面以博客园Arkenstone[9]的博客为例介绍经验分布函数与理论分布函数的比较。

现有两列样本：

```
controlB = {1.26, 0.34, 0.70, 1.75, 50.57, 1.55, 0.08, 0.42, 0.50, 3.20,
0.15, 0.49, 0.95, 0.24, 1.37, 0.17, 6.98, 0.10, 0.94, 0.38}
treatmentB= {2.37, 2.16, 14.82, 1.73, 41.04, 0.23, 1.32, 2.91, 39.41, 0.11,
27.44, 4.51, 0.51, 4.50, 0.18, 14.68, 4.66, 1.30, 2.06, 1.19}
```

首先分别对controlB和treatmentB进行升序，经验分布函数是对样本数据中小于等于给定值的样本点所占比例的一种累积函数，它通过对样本数据进行排序和计算累积频率来近似表示总体的分布情况，两列样本取对数前后的经验分布函数可视化如图7所示。

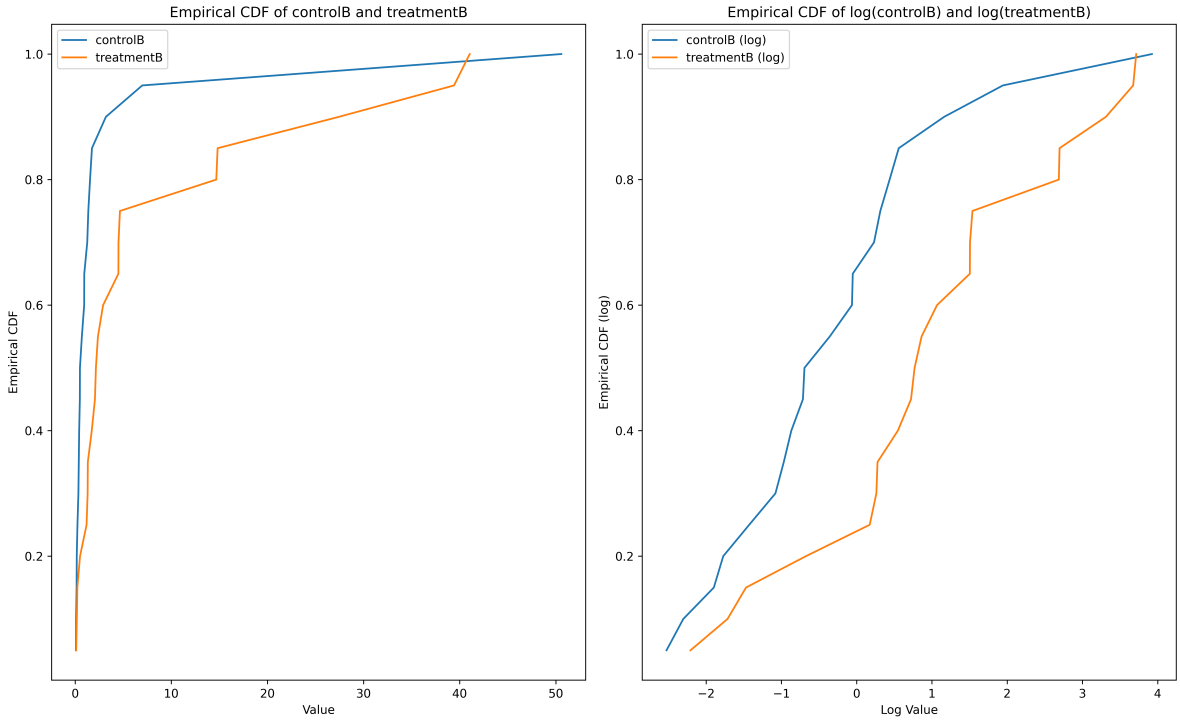

图7 经验分布函数可视化

K-S检验使用的是两条累计分布曲线之间的最大垂直差作为统计量D_n，其计算公式如下：

$$D_n = \sup_x |F_n(x) - F(x)|. \tag{28}$$

图4的K-S统计量出现在$x = 1$附近，取值为0.45（0.65-0.25）。K-S检验的原假设是两个样本数据的分布之间没有显著差异，使用Python计算的统计量和P值：

```python
import numpy as np
from scipy.stats import ks_2samp

# 定义两个样本
controlB = np.array(
    [1.26, 0.34, 0.70, 1.75, 50.57, 1.55, 0.08, 0.42, 0.50, 3.20, 0.15, 0.49, 0.95, 0.24, 1.37, 0.17, 6.98, 0.10, 0.94,
     0.38])
treatmentB = np.array(
    [2.37, 2.16, 14.82, 1.73, 41.04, 0.23, 1.32, 2.91, 39.41, 0.11, 27.44, 4.51, 0.51, 4.50, 0.18, 14.68, 4.66, 1.30,
     2.06, 1.19])

# 进行 K-S 检验
statistic, p_value = ks_2samp(controlB, treatmentB)

print(f"K-S 统计量: {statistic}")
print(f"P 值: {p_value}")
```

代码输出结果为：

```
K-S 统计量: 0.45
P 值: 0.0335416594061465
```

因此在95%的置信水平下显著，可以拒绝原假设，这两列样本数据的分布有显著差异。

Lilliefors 利利福斯检验

Lilliefors检验在小样本情况下相对更稳健，它是K-S检验的改进版本，其原假设为样本数据服从一般正态分布，应用。K-S检验的理论分布是确定的，比如为标准正态分布。而Lilliefors检验的理论分布并不确定，比如服从某个均值和标准差下的正态分布。

Anderson-Darling 安德森-达林检验

A-D检验基于累积分布函数与样本数据的经验累积分布函数之间的差异检验数据是否符合某种特定分布，其原假设为服从正态分布。连续问题上的A-D检验统计量如下：

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x). \tag{29}$$

K-S统计量关注的是经验分布函数和理论分布函数之间的最大垂直距离，而A-D统计量考虑的是在整个分布范围内经验分布函数和理论分布函数的差异，通过积分的方式综合评估这种差异，对分布的形状差异更为敏感。故与K-S检验相比，A-D检验对尾部数据，即数据分布的极端部分更加敏感，在许多情况下比K-S检验表现得更好。（这个检验我没学会）

下面以实例展示上述五种不同的正态性检验的结果，其直方图如图8所示。

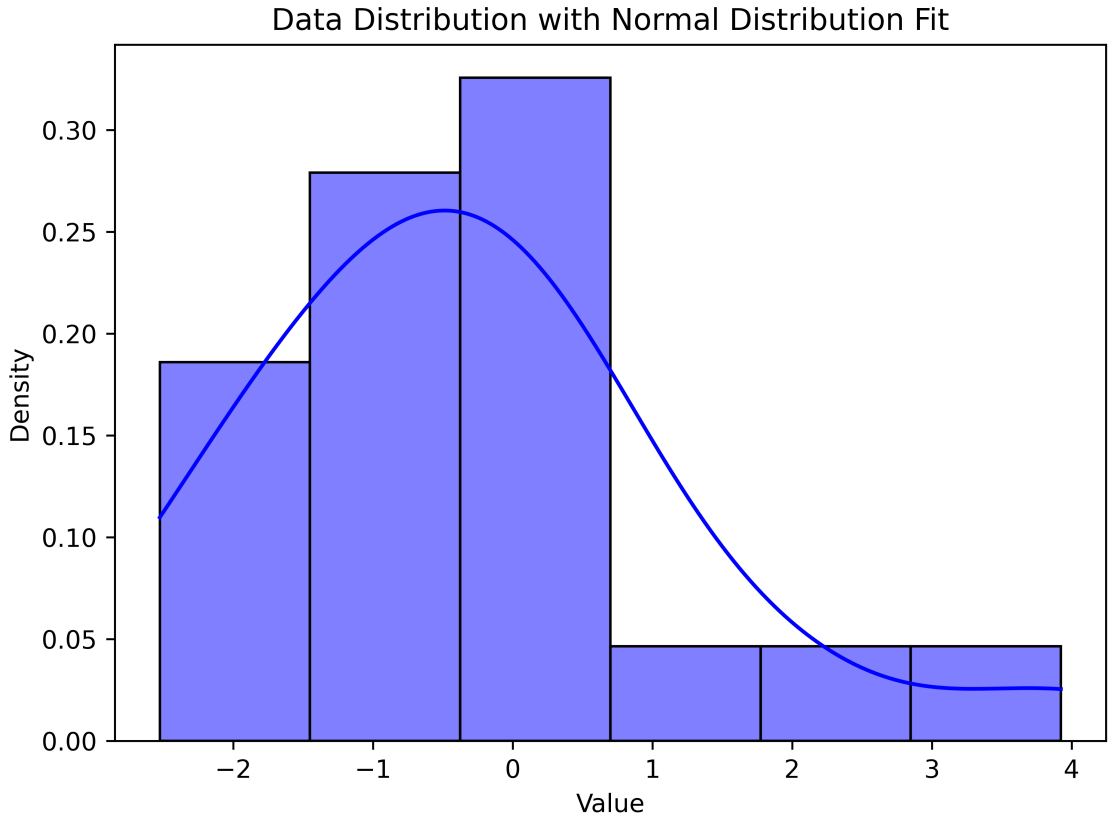

图8 样本数据直方图

```
import numpy as np
from scipy import stats
from statsmodels.stats.diagnostic import lilliefors
```

```
import matplotlib.pyplot as plt
import seaborn as sns

# 数据
data = [0.23111172096338664, -1.0788096613719298, -0.35667494393873245,
0.5596157879354227, 3.9233585150918917,
        0.4382549309311553, -2.5257286443082556, -0.8675005677047231,
-0.6931471805599453, 1.1631508098056809,
        -1.8971199848858813, -0.7133498878774648, -0.05129329438755058,
-1.4271163556401458, 0.3148107398400336,
        -1.7719568419318752, 1.9430489167742813, -2.3025850929940455,
-0.06187540371808753, -0.9675840262617056]

# 1. **Jarque-Bera (J-B) 检验**
jb_statistic, jb_p_value = stats.jarque_bera(data)
print(f"Jarque-Bera 检验统计量: {jb_statistic:.3f}, p 值: {jb_p_value:.3f}")

# 2. **Shapiro-Wilk (S-W) 检验**
sw_statistic, sw_p_value = stats.shapiro(data)
print(f"Shapiro-Wilk 检验统计量: {sw_statistic:.3f}, p 值: {sw_p_value:.3f}")

# 3. **Kolmogorov-Smirnov (K-S) 检验**
# 假设数据符合正态分布，使用均值和标准差作为参数来进行检验
ks_statistic, ks_p_value = stats.kstest(data, 'norm', args=(np.mean(data),
np.std(data)))
print(f"Kolmogorov-Smirnov 检验统计量: {ks_statistic:.3f}, p 值:
{ks_p_value:.3f}")

# 4. **Lilliefors 检验**
lilliefors_statistic, lilliefors_p_value = lilliefors(data)
print(f"Lilliefors 检验统计量: {lilliefors_statistic:.3f}, p 值:
{lilliefors_p_value:.3f}")

# 5. **Anderson-Darling (A-D) 检验**
ad_result = stats.anderson(data, dist='norm')
print(f"Anderson-Darling 检验统计量: {ad_result.statistic:.3f}")

# A-D 检验的临界值和对应的显著性水平
print("Anderson-Darling 检验临界值：")
for i in range(len(ad_result.critical_values)):
    print(f"{ad_result.significance_level[i]}% 级别的临界值:
{ad_result.critical_values[i]:.3f}")

# 可视化数据的直方图和正态分布拟合
sns.histplot(data, kde=True, stat='density', color='blue', label='Data
Histogram')
plt.title('Data Distribution with Normal Distribution Fit')
plt.xlabel('Value')
plt.ylabel('Density')
plt.show()

```

代码输出结果为：

```
Jarque-Bera 检验统计量: 4.470, p 值: 0.107
Shapiro-Wilk 检验统计量: 0.937, p 值: 0.207
Kolmogorov-Smirnov 检验统计量: 0.129, p 值: 0.853
Lilliefors 检验统计量: 0.134, p 值: 0.450
Anderson-Darling 检验统计量: 0.359
Anderson-Darling 检验临界值:
15.0% 级别的临界值: 0.506
10.0% 级别的临界值: 0.577
5.0% 级别的临界值: 0.692
2.5% 级别的临界值: 0.807
1.0% 级别的临界值: 0.960
```

这意味着上述Jarque-Bera、Shapiro-Wilk、Kolmogorov-Smirnov和Lilliefors检验的原假设均为服从正态分布，所有检验的p值都大于0.05，这意味着在显著性水平0.05下无法拒绝原假设，结果表明数据没有显著偏离正态分布。对于Anderson-Darling检验而言，统计量为0.359，均远小于各个临界值下的值，这说明小于所有显著性水平下的临界值，数据没有显著偏离正态分布。

差异性分析

Chi-squared test 卡方检验

卡方分布概率密度函数的推导

卡方分布的概率密度函数的推导，知乎用户SlipLe的帖子[10]讲解得非常清楚，甚至连推导过程需要的伽马函数和卷积操作都有公式介绍。因此，我在这里基于帖子省略的步骤进行补充。

卡方分布是由一组独立的标准正态分布变量的平方和所定义的。若$Z_1, Z_2, \ldots, Z_k$是k个独立的标准正态分布随机变量，即每个$Z_i \sim N(0,1)$，则卡方分布的随机变量X定义为：

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2. \tag{31}$$

标准正态分布$Z \sim N(0,1)$的概率密度函数是：

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad -\infty < z < \infty. \tag{32}$$

首先推导自由度为1的卡方分布，即当$X \sim N(0,1)$时，证明$X^2 \sim \chi^2(1)$：

$$\begin{aligned} P\{X^2 < y\} &= P\{-\sqrt{y} < X < \sqrt{y}\} \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dX \\ &= \frac{1}{2^{\frac{1}{2}}\sqrt{\pi}} y^{(\frac{1}{2}-1)} e^{-\frac{y}{2}}. \end{aligned} \tag{33}$$

对于伽马函数有：

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt. \tag{34}$$

为证明$\Gamma(1/2) = \sqrt{\pi}$，令$t = u^2$：

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-t} t^{-\frac{1}{2}} dt = 2\int_0^\infty e^{-u^2} du, \tag{35}$$

由此可知，积分项是半个高斯积分，求解高斯积分的方法是先计算$\Gamma^2(1/2)$：

$$[\Gamma\left(\frac{1}{2}\right)]^2 = 4\int_0^\infty \int_0^\infty e^{-(u^2+v^2)} du dv, \tag{36}$$

在上式中，因为u和v都是不相关的，所以可以把两个积分相乘写成二重积分的形式。

接下来使用极坐标变换，令$u = rcos\theta$，$v = rsin\theta$，由于$r = \sqrt{u^2+v^2}$，所以r的取值范围是0到∞，由于u和v都是大于等于0的，因此θ在第一象限，故取值范围为0到$\pi/2$，可将公式(36)化简为如下形式：

$$[\Gamma\left(\frac{1}{2}\right)]^2 = 4\int_0^{\frac{\pi}{2}} \int_0^\infty e^{-r^2} r dr d\theta = \pi. \tag{37}$$

因此，公式(33)可以用伽玛函数表示：

$$P\{X^2 < y\} = \frac{(\frac{1}{2})^{\frac{1}{2}} y^{(\frac{1}{2}-1)} e^{-\frac{y}{2}}}{\Gamma(\frac{1}{2})}. \tag{38}$$

现定义伽玛分布的概率密度函数：

$$f(y,\lambda,\alpha)=\frac{\lambda^{\alpha}y^{(\alpha-1)}e^{-\lambda y}}{\Gamma(\alpha)}. \tag{39}$$

因此，自由度为1的卡方分布的概率密度函数也可以写成如下形式：

$$\frac{(\frac{1}{2})^{\frac{1}{2}}y^{(\frac{1}{2}-1)}e^{-\frac{y}{2}}}{\Gamma(\frac{1}{2})}=Ga(\frac{1}{2},\frac{1}{2}). \tag{40}$$

下面推导自由度为n时的卡方分布的概率密度函数，因为涉及到多项相加，因此可以先由卷积公式证明如下性质：

若有$Y_1\sim Ga(\lambda,\alpha_1)$和$Y_2\sim Ga(\lambda,\alpha_2)$，那么$Y_1+Y_2\sim Ga(\lambda,\alpha_1+\alpha_2)$。根据卷积的定义，$Y=Y_1+Y_2$的概率密度函数$f(y)$为：

$$f(y)=f_{Y_1}(y)*f_{Y_2}(y)=\int_0^y\frac{\lambda^{\alpha_1}t^{\alpha_1-1}e^{-\lambda t}}{\Gamma(\alpha_1)}\cdot\frac{\lambda^{\alpha_2}(y-t)^{\alpha_2-1}e^{-\lambda(y-t)}}{\Gamma(\alpha_2)}dt. \tag{41}$$

两个独立随机变量的和的概率密度函数是这两个随机变量的概率密度函数的卷积。对于连续随机变量Y_1和Y_2，Y的概率密度函数$f_Y(y)$由$f_{Y_1}(y)$和$f_{Y_2}(y)$的卷积给出：

$$f_Y(y)=\int_{-\infty}^{\infty}f_{Y_1}(t)f_{Y_2}(y-t)dt. \tag{42}$$

对于公式(41)，首先，将指数项合并，再将系数和幂次项合并，积分变为：

$$f(y)=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda y}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^y t^{\alpha_1-1}(y-t)^{\alpha_2-1}dt. \tag{41}$$

接着对积分项进行化简，令$t=yu$，当$t=0$时有$u=0$，当$t=y$时有$u=1$，因此变量替换后的积分区域是0到1，化简后合并指数项可以得到：

$$\int_0^y t^{\alpha_1-1}(y-t)^{\alpha_2-1}dt=\int_0^1 y^{\alpha_1+\alpha_2-1}u^{\alpha_1-1}(1-u)^{\alpha_2-1}du. \tag{42}$$

结合公式(42)可将公式(41)化简为如下形式：

$$f(y)=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda y}y^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_2-1}du.. \tag{42}$$

对于Beta函数有如下定义：

$$B(p,q)=\int_0^1 t^{p-1}(1-t)^{q-1}dt=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{43}$$

观察公式(42)可知，其积分项即为$B(\alpha_1,\alpha_2)$，现证明公式(43)。

对于公式(34)，令$t=x^2$，因此$\Gamma(p)$为：

$$\Gamma(p)=\int_0^{\infty}e^{-x^2}(x^2)^{p-1}2xdx=2\int_0^{\infty}e^{-x^2}x^{2p-1}dx. \tag{44}$$

类似的，令$t=y^2$，因此$\Gamma(q)$为：

$$\Gamma(q)=\int_0^{\infty}e^{-y^2}(y^2)^{q-1}2ydy=2\int_0^{\infty}e^{-y^2}y^{2q-1}dy. \tag{45}$$

公式(44)和公式(45)相乘可得：

$$\Gamma(p)\Gamma(q)=4\int_0^{\infty}\int_0^{\infty}e^{-x^2-y^2}x^{2p-1}y^{2q-1}dxdy. \tag{46}$$

利用极坐标变换，令$x=r\cos\theta$，$y=r\sin\theta$，公式(46)可化简为：

$$\begin{aligned}\Gamma(p)\Gamma(q)&=4\int_0^{\infty}\int_0^{\frac{\pi}{2}}e^{-r^2}(rcos\theta)^{2p-1}(rsin\theta)^{2q-1}rdrd\theta\\&=4\int_0^{\infty}e^{-r^2}r^{2(p+q)-1}dr\int_0^{\frac{\pi}{2}}(cos\theta)^{2p-1}(sin\theta)^{2q-1}d\theta\\&=4\int_0^{\infty}e^{-r^2}r^{2(p+q)-1}dr\int_0^{\frac{\pi}{2}}(cos^2\theta)^{p-1}\cdot cos\theta\cdot(sin^2\theta)^{q-1}\cdot sin\theta d\theta.\end{aligned} \tag{47}$$

对于r部分的积分，即为$\Gamma(p+q)$；对于θ部分的积分，令$cos^2\theta=t$，可化简为：

$$\int_1^0 t^{p-1}\cdot\sqrt{t}(1-t)^{q-1}\cdot\sqrt{1-t}\cdot\frac{-1}{2\sqrt{t}\sqrt{1-t}}dt. \tag{48}$$

把公式(48)中可约分的项消掉，以及负号与负号抵消后，可发现公式(48)与公式(43)形式一致，即为$B(p,q)$。

因此可以证明$\Gamma(p)\Gamma(q)=\Gamma(p+q)\cdot B(p,q)$。

至此，已经证明了若有$Y_1\sim Ga(\lambda,\alpha_1)$和$Y_2\sim Ga(\lambda,\alpha_2)$，那么$Y_1+Y_2\sim Ga(\lambda,\alpha_1+\alpha_2)$。接下来只需要让所有的

$$\lambda=\frac{1}{2},\alpha_1=\alpha_2=\ldots=\alpha_n=\frac{1}{2}. \tag{49}$$

于是：

$$\sum_{k=1}^{n}X_k^2+X_2^2+\ldots+X_n^2=\sum_{k=1}^{n}Y_1+Y_2+\ldots+Y_n\sim Ga\left(\frac{1}{2},\frac{n}{2}\right). \tag{50}$$

综上所述，得到卡方分布的概率密度函数：

$$f(y)=\begin{cases}\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}y^{\left(\frac{n}{2}-1\right)}e^{-\frac{y}{2}} & 0<y\\0 & y\le 0\end{cases}. \tag{51}$$

卡方检验

上面的推导过程是对知乎帖子“概率论-卡方分布推导[10]”的补充，属于是基础知识的巩固了，下面介绍卡方检验的应用[11]。

卡方检验（Pearson Chi-Square）的原假设是两个分类变量相互独立，是一种非参数检验，适用于总样本量大于等于40，期望频数大于等于5时的情况。假如总样本量大于等于40，有期望频数大于等于1且小于5时，选择Yates校正卡方检验（Continuity correction）。

卡方检验在差异性分析上的应用是列联表检验，假设有两个分类变量A和B，A有r个类别，B有c个类别。现将样本数据整理成一个$r\times c$的列联表，其中n_{ij}表示变量A属于第i类且变量B属于第j类的观察频数。

计算期望频数：

$$E_{ij}=\frac{n_{i.}n_{.j}}{n}, \tag{52}$$

其中，$n_{i.}=\sum_{j=1}^{c}n_{ij}$，$n_{.j}=\sum_{i=1}^{r}n_{ij}$，$n=\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij}$。

构造卡方检验统计量：

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}, \tag{53}$$

这个统计量在两个变量相互独立的情况下近似服从自由度为$df=(r-1)(c-1)$的卡方分布。

在进行列联表的卡方检验时，当样本量较小时，卡方分布是一个连续分布，而列联表中的频数是离散的。这种离散-连续的差异可能会导致卡方检验统计量的分布与理论卡方分布有偏差。Yates校正卡方检验主要适用于2×2列联表，特别是当期望频数不是很大时（有期望频数大于等于1且小于5时），可以减少这种偏差，使检验结果更加准确[12]。

对于一个下面这个列联表：

	类别1	**类别2**
组1	a	b
组2	c	d

未校正的卡方统计量计算公式为：

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}. \tag{54}$$

Yates校正后的卡方统计量计算公式为：

$$\chi^2_{Yates} = \frac{(|ad-bc|-n/2)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}, \tag{55}$$

其中，$n=a+b+c+d$是总样本量。

假设以学生性别和成绩等级为例：

	优秀	**良好**	**合格**	**不合格**
男	10	15	20	5
女	8	12	18	2

```python
import numpy as np
from scipy.stats import chi2_contingency

# 构建列联表
contingency_table = np.array([[10, 15, 20, 5], [8, 12, 18, 2]])

# 未校正的卡方检验
chi2_uncorrected, p_value_uncorrected, dof_uncorrected, expected_uncorrected = chi2_contingency(contingency_table)
print("卡方统计量:", chi2_uncorrected)
print("p值:", p_value_uncorrected)
print("自由度:", dof_uncorrected)
print("期望频数:", expected_uncorrected)
```

其输出结果为：

```
卡方统计量: 0.8458646616541353
p值: 0.8384682368484039
自由度: 3
期望频数: [[10 15 21.11111111 3.88888889]
 [ 8 12 16.88888889 3.11111111]]
```

假设显著性水平为0.05，p值远大于该显著性水平，则不拒绝原假设，认为性别和成绩等级是独立的。

假设上面的列联表只有2行2列：

	优秀	**良好**
男	10	15
女	8	12

```python
import numpy as np
from scipy.stats import chi2_contingency

# 构建一个2x2列联表
contingency_table = np.array([[10, 15], [8, 12]])

# 未校正的卡方检验
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print("未校正卡方统计量:", chi2)
print("未校正p值:", p_value)

# Yates校正卡方检验手动计算
a, b, c, d = contingency_table[0][0], contingency_table[0][1], contingency_table[1][0], contingency_table[1][1]
n = a + b + c + d
chi2_yates = ((np.abs(a * d - b * c) - n / 2) ** 2 * (a + b + c + d)) / ((a + b) * (c + d) * (a + c) * (b + d))

# 自由度对于2x2列联表为1
p_value_yates = 1 - np.sum(
    [np.exp(- chi2_yates / 2) * (chi2_yates / 2) ** k / np.math.factorial(k) for k in range(int(dof + 1))])
print("Yates校正卡方统计量:", chi2_yates)
print("Yates校正p值:", p_value_yates)
```

其输出结果为：

```
未校正卡方统计量: 0.0
未校正p值: 1.0
Yates校正卡方统计量: 0.09375
Yates校正p值: 0.001064896563505946
```

由此可见，未校正的卡方检验失效。

独立样本t检验

配对样本t检验

方差分析

单因素方差分析

多因素方差分析

多元方差分析

非参数检验

Wilcoxon 威尔科克森秩和检验

Kruskal-Wallis 克鲁斯卡尔-沃尔利斯检验

方差齐性检验

Levene 列文检验

Bartlett 巴特利特球形检验

Fligner-Killeen 弗莱纳 - 基林检验

引用

[1] 如何通俗易懂地解释「协方差」与「相关系数」的概念？ - GRAYLAMB的回答 - 知乎

[2] Strength of Correlation

[3] 数学建模：相关性分析学习——皮尔逊（pearson）相关系数与斯皮尔曼（spearman）相关系数

[4] Spearman等级相关系数 - 越来越好的文章 - 知乎

[5] 肯德尔（Kendall）相关系数概述及Python计算例

[6] 【时间序列分析】肯德尔（Kendall）相关系数基础理论及python代码实现

[7] 【125】正态性检验 - AnkiStudy的文章 - 知乎

[8] Shapiro–Wilk test

[9] KS-检验（Kolmogorov-Smirnov test） -- 检验数据是否符合某种分布

[10] 概率论-卡方分布推导

[11] 统计之美：如何优雅理解卡方分布与卡方检验之精髓所在？(重磅)

[12] Yates校正卡方检验