数据的统计分析方法

scipy.stats模块

1.连续型随机变量及分布

```
from scipy import stats

uniform.pdf(x, a, b)  # [a, b]区间上的均匀分布
expon.pdf(x, scale=theta)  # 期望为theta的指数分布
chi2.pdf(x, n)  # 自由度为n的x^2分布
t.pdf(x, n)  #自由度为n的t分布
f.pdf(x, m, n)  # 自由度为m, n的f分布
gamma.pdf(x. a=A, scale=B)  # 形状参数为A，尺度参数为B的gamma分布

norm.pdf(x, mu, sigma)  # 均值为mu， 标准差为sigma的正态分布
norm.cdf(x, mu, sigma)  # 正态分布的分布函数
norm.ppf(x, mu, sigma)  # 正态分布的alpha分位数
norm.rvs(mu, sigma, size=N)  # 产生均值为mu， 标准差为sigma的N个正态分布的随机数
```

2.离散型随机变量及分布

```
from scipy import stats

binom.pmf(x, n, p)  # 计算x处的二项分布概率
geom.pmf(x, p)  # 计算第x次首次成功的几何分布概率
possiom.pmf(x, lambda)  # 计算x处的泊松分布概率
```

统计

1.常用统计量

均值:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

标准差:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

极差:

$$R = \underset{1\le i\le n}{max}\ x_i - \underset{1\le i\le n}{min}\ x_i$$

偏度（反映分布的对称性）：

$$v_1 = \frac{1}{s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

峰度（反映分布偏离正态分布的尺度）：

$$v_2 = \frac{1}{s^4}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

协方差:

$x = [x_1,\ x_2,\ \ldots,\ x_n]$ 和 $y = [y_1,\ y_2,\ \ldots\ y_n]$ 的协方差为

$$cov(x,\ y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

相关系数:

$x = [x_1,\ x_2,\ \ldots,\ x_n]$ 和 $y = [y_1,\ y_2,\ \ldots\ y_n]$ 的相关系数为

$$\rho_{xy} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}}$$

k阶原点距:

$$a_k = \frac{1}{n}\sum_{i=1}^{n}x_i^k$$

k阶中心距:

$$b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

样本p分位数:

$$x_p = \begin{cases} x_{([np]+1)}, & np\text{不是整数}, \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}), & np\text{是整数}. \end{cases}$$

样本均值标准误差：

$$SEM = \frac{s}{\sqrt{n}}$$

2.用Python计算统计量

```
import numpy as np
import pandas as pd
from scipy.stats import sem

a = pd.raed_csv('data.csv')
b = a.values

mu = np.mean(b, axis=1)  # 平均值
zw = np.median(b, axis=1)  # 中位数
jc = np.ptp(b, axis=1)  # 极差
fc = np.var(b, axis=1, ddof=1)  # 方差
bz = np.std(b ,axis=1, ddof=1)  # 标准差
xf = np.cov(b)  # 协方差矩阵
xs = np.corrcoef(b)  # 相关系数矩阵
sm = sem(b)  # 样本均值标准误差
```

3.参数估计和假设检验

正态总体标准差 σ 已知的 t 检测法

设总体 $X \sim N(\mu, \sigma^2)$，其中 μ 未知，σ 已知，
提出原假设 $H : \mu = \mu_0$，
检测统计量为

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

检验显著性水平为 α ，标准正态分布上的 $\alpha/2$ 分位数记为 $t_{\alpha/2}$ ，当 t 的观测值满足 $|t| < t_{\alpha/2}$ 时，接受原假设.

```
from scipy.stats import ttest_1samp

tstat, pvalue = ttest.1samp(a, popmean, alternative='two-sided')
```

a为检测样本数据，popmean表示假设的总体均值，tstats为 t 值.

正态总体标准差 σ 已知的 Z 检测法
设总体 $X \sim N(\mu,\, \sigma^2)$，其中 μ 未知，σ 已知，
提出原假设 $H : \mu = \mu_0$，
检测统计量为

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

检验显著性水平为 α ，标准正态分布上的 $\alpha/2$ 分位数记为 $z_{\alpha/2}$ ，当 Z 的观测值 z 满足 $|z| < z_{\alpha/2}$ 时，接受原假设.

Z 统计量与 t 统计量的关系为：

$$Z = \frac{s}{\sigma} \cdot t$$

χ^2检验
假设H_0 : 总体 X 分布函数为 $F(x)$；

将数轴分为 k 个区间，令 p_i 为分布函数 $F(x)$ 的总体 X 在第 i 个区间内取值的概率，设 f_i 为 n 个样本观察值中落入第 i 个区间上的个数；

选取统计量：

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{f_i^2}{np_i} - n$$

若 H_0 为真，则 $\chi^2 \sim \chi^2(k-1-r)$ ，其中 r 为分布函数 $F(x)$ 中未知参数的个数；

对于给定的显著性水平 α ，确定 χ^2_α ，使其满足 $P\{\chi^2(k-1-r) > \chi^2_\alpha\} = \alpha$ ，并依据样本统计量计算 χ^2 的观察值；

若 $\chi^2 < \chi^2_\alpha$ ，则接受 H_0 .

Kolmogorov-Smirnov检验
经验分布函数 $F_n(x)$ 观察值为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ \dfrac{k}{n}, & x_{(k)} \leq x \leq x_{(k+1)}, k = 1, 2, \ldots, n-1 \\ 1, & x \geq x_{(n)} \end{cases}$$

假设 $H_0 : F(x) = F_0(x)$；

选取检验统计量

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|$$

当 H_0 为真时，D_n 有偏小趋势；

给定显著性水平 α， 根据 D_n 极限分布表，求出 t_α 满足

$$P\{\sqrt{n}D_n \geq t_\alpha\} = \alpha,$$

作为临界值；

若 $\sqrt{n}D_n < t_\alpha$，则接受 H_0 .