# HW1

1. 已知某工厂某批次水泥重量服从正态分布，总体方差为2.65公斤， 从该工厂随机抽取 18 袋水泥，其平均重量为24.9公斤， 试求该工厂水泥平均重量的95%和99%的置信区间。

$$
\begin{gathered}
SE_X = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2.65}}{\sqrt{18}} = 0.38 \\
95\%\text{置信区间的显著性水平}\alpha = 0.05, Z_{\alpha/2} = 1.96 \\
95\%\text{置信区间为：}24.9 - 1.96 \times 0.38 \le u \le 24.9 + 1.96 \times 0.38, \text{即}[24.16, 25.64] \\
\\
99\%\text{置信区间的显著性水平}\alpha = 0.01, Z_{\alpha/2} = 2.58 \\
99\%\text{置信区间为：}24.9 - 2.58 \times 0.38 \le u \le 24.9 + 2.58 \times 0.38, \text{即}[23.92, 25.88]
\end{gathered} \tag{1}
$$

2. 从某学校抽取20名高一学生，经测量， 这 20名学生的平均身高为155cm，标准差为10cm， 假设平均身高服从正态分布，试求该学校高一学生总平均身高的95%和99%的置信区间。

$$
\begin{gathered}
SE_X = \frac{S}{\sqrt{n-1}} = \frac{10}{\sqrt{19}} = 2.29 \\
t_{0.05/2(19)} = 2.093, t_{0.01/2(19)} = 2.861 \\
95\%\text{置信区间为：}155 - 2.093 \times 2.29 \le u \le 155 + 2.093 \times 2.29, \text{即}[150.21, 159.79] \\
99\%\text{置信区间为：}155 - 2.861 \times 2.29 \le u \le 155 + 2.861 \times 2.29, \text{即}[148.45, 161.55]
\end{gathered} \tag{2}
$$

3. 1.2 节P19，证明被抽样总体分布未知情况下的无偏估计。

**例 设$x_1, x_2, \dots, x_3$是均值为$x$、方差为$\sigma^2$的随机变量X的n个观测值的随机样本，证明：样本方差$s^2$是总体方差$\sigma^2$的一个无偏估计，其中：**

**a) 被抽样总体为正太分布**

**b) 被抽样总体的分布未知**

设样本均值为$\mu$

$$
\begin{aligned}
s^2 &= \frac{\Sigma(x_i-\mu)^2}{n-1} \\
&= \frac{\Sigma(x_i-x)^2+2(x-\mu)\times\Sigma(x_i-x)+n(\mu-x)^2}{n-1} \\
&= \frac{\Sigma(x_i-x)^2+2(x-\mu)\times(n(\mu-x))+n(\mu-x)^2}{n-1} \\
&= \frac{\Sigma(x_i-x)^2-n(x-\mu)^2}{n-1}
\end{aligned}
$$

$$
\begin{aligned}
E[(x-\mu)^2] &= Var[x-\mu]+([E(x-\mu)]^2) \\
&= Var[\frac{\Sigma(x-x_i)}{n}]+(x-\frac{Ex_i}{n})^2 \\
&= \frac{\Sigma Var[x_i-x]}{n^2}+(x-\frac{nx}{n})^2 \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n} \\
E[s^2] &= E[\frac{\Sigma(x_i-x)^2-n(x-\mu)^2}{n-1}] \\
&= \frac{n\times(E[\frac{\Sigma(x_i-x)^2}{n}]-E[\mu-x]^2)}{n-1} \\
&= \frac{n\times\sigma^2-n\times\frac{\sigma^2}{n}}{n-1} \\
&= \sigma^2
\end{aligned} \tag{3}
$$

4. 证明Lemma 1.3.1

**Lemma 1.3.1**. Let $P$ be the transition probability matrix for a connected Markov Chain. The $n\times(n+1)$ matrix $A=[P-I,1]$ obtained by augmenting the matrix $P-I$ with an additional column of ones has rank $n$.

反证法. 若 $rank(A)<n$

则 $A\vec{x}=0$ 存在两个线性无关解

注意到 $\sum_j P_{ij}=0$. 因此 $(\vec{1}, 0)^T$ 是方程的一个解

因此存在第二个解，设为 $(\vec{y}, \alpha)'$

$\Rightarrow A \cdot (\vec{y}, \alpha)^T = 0 \Rightarrow (P-I, \vec{1}) \cdot (\vec{y}, \alpha)^T = 0$

$\Rightarrow (P-I) \cdot \vec{y} + \alpha \cdot \vec{1} = 0$

对第i行：$\sum_j P_{ij} \cdot y_j - y_i + \alpha = 0$

$\Rightarrow \quad y_i = \sum_j P_{ij} \cdot y_j + \alpha$

由于 $(\vec{y}, \alpha)^T$ 与 $(\vec{1}, 0)$ 线性无关

因此 $y_i$ 互不相等

$\Rightarrow \exists i_m, i_M$ s.t. $\forall k \in N, k \le n \quad y_{i_m} \le y_k \le y_{i_M}$ 且 $y_{i_M} > y_{i_m}$

$\Rightarrow y_{i_M} > \sum_j P_{i_M j} \cdot y_j, \quad y_{i_m} < \sum_j P_{i_m j} \cdot y_j.$

$\Rightarrow 0 < \alpha < 0$，矛盾.

实验1: PageRank最高20份node:

```
● (py) jyjs@192 HW % python pagerank.py
[('7237', np.float64(0.007402252556857881)), ('7498', np.float64(0.006691202515011874)), ('7339', np.float64(0.006377762845352468)), ('716
2', np.float64(0.0031767569339574593)), ('7224', np.float64(0.0029302963203570952)), ('7435', np.float64(0.002750701330228721)), ('6519',
np.float64(0.0027171361605782102)), ('7100', np.float64(0.0025939892857066867)), ('7595', np.float64(0.002575726801676199)), ('7199', np.f
loat64(0.002575281129187989)), ('4811', np.float64(0.0024343041872750726)), ('6101', np.float64(0.0023085323564403337)), ('7536', np.float
64(0.002238758251594308)), ('4785', np.float64(0.0022117014168385888)), ('7489', np.float64(0.0019699795701771067)), ('7587', np.float64(0
.001849670865613301)), ('7488', np.float64(0.0018347441540418777)), ('6617', np.float64(0.0017652065393296442)), ('7226', np.float64(0.001
7385075407141284)), ('7416', np.float64(0.0017249734913823893))]
```

实验2:

结果统计:

```
recall: 0.6818181818181818
precision: 0.5284974093264249
f1 score: 0.5954465849387041
```

实验代码在压缩包中