# 统计中的计算方法 第一次作业 修订

傅长青　信息与计算科学　13300180003

May 9, 2016

**1. 为估计一件物体的重量$\mu$,将其称了10 次,得到的重量(单位:kg)为10.1, 10, 9.8, 10.5, 9.7, 10.1, 9.9, 10.2, 10.3, 9.9 假设所称的物体重量服从$N(\mu, \sigma^2)$,求该物体重量$\mu$的置信系数为0.95 的置信区间。**

**解：**

样本方差S = sqrt((∑(x_i - μ)^2)/(n-1))

置信区间为

$$(\bar{X} - t_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{0.025}\frac{\sigma}{\sqrt{n}})$$

其中$\bar{X} = 10.05, \sigma = 0.58, n = 10, t_{0.025} = 1.96$.故置信区间为

S = 0.241　~~(10.01, 10.09)~~ (9.90, 10.20)

**2. 据以往经验,新生儿染色体异常率一般为1%,某医院观察了当地400 名新生儿,只有一例染色体异常,问该地区新生儿染色体异常是否低于一般水平?**

**解：**

设随机变量

$$X = \begin{cases} 0 & \text{染色体正常} \\ 1 & \text{染色体异常} \end{cases}$$

H0: =1%

H1: <1%

根据统计结果，用样本均值估计总体均值，

P(data | H0) = 400*.99^399*.01= 7.2%>5% 接受假设

$$\mu \approx \bar{X} = \frac{1}{400} < 1\%$$

即低于一般水平

故高于一般水平

**3. 假设$X = (X_1, \ldots, X_n)$是n个来自于三个二元正态分布的混合分布的独立样本,试推导出用EM方法估计三个二元正态分布参数的迭代步骤。**

**解：**

设$X$服从混合高斯分布

$$X \sim p_1 N(\mu_1, \sigma_1) + p_2 N(\mu_2, \sigma_2) + p_3 N(\mu_3, \sigma_3)$$

$$or,\ X \sim f(x|\theta) = \sum_{i=1}^{3} p_i N(x; \mu_1, \sigma_i) \qquad X_i = (X_{i1}, X_{i2})$$

**The E step:**

$$
\begin{aligned}
\log L(\theta; X, Z) &= P(X, Z|\theta) = \prod_{i,j} (\tau f(x_i, \mu_j, \Sigma_j))^{z_{ij}} \\
Q(\theta|\theta_t) &:= \mathbb{E} \log L(\theta; X, Z) \\
&= \mathbb{E} \sum_{i,j} z_{ij}((\log \tau_j - 0.5 \log |\Sigma_j| - 0.5(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j) + const) \\
&= \sum_{i,j} T_{ij}^{(t)}((\log \tau_j - 0.5 \log |\Sigma_j| - 0.5(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j) + const) \\
\text{where } T_{ij}^{(t)} &= \frac{\tau^{(t)} f(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_j \tau^{(t)} f(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}
\end{aligned}
$$

**the M step:** Maximize $Q(\theta|\theta_t)$above, we get

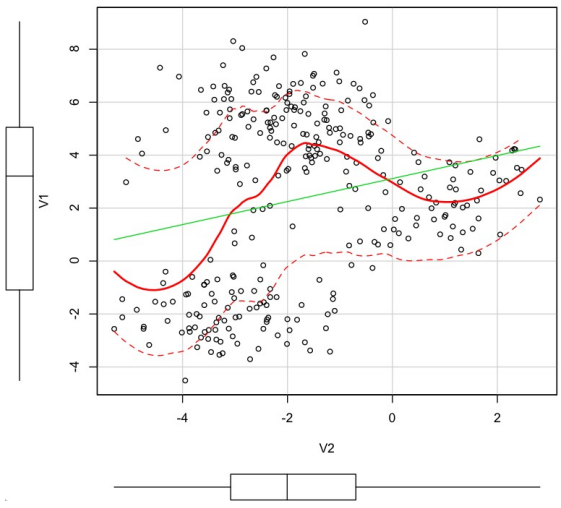
$$
\begin{aligned}
\tau_j^{(t+1)} &:= \frac{1}{n} \sum_i T_{ij}^{(t)} \\
\mu_j^{(t+1)} &= \frac{\sum_i T_{ij}^{(t)} x_i}{\sum_i T_{ij}^{(t)}} \\
\Sigma_j^{(t+1)} &= \frac{\sum_i T_{ij}^{(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T}{\sum_i T_{ij}^{(t)}}
\end{aligned}
$$

**4. 对数据Data1.csv用3中方法估计参数。**

**解：**

1)画出草图估计数据

$$\widehat{V1} = \widehat{\beta} + \alpha V_2 = 0.4348 V_2 + 3.1162$$




相关的代码为：

```
> data4 = read.table("~/Desktop/STAT/hw/1/Data1'.csv",sep=",",header
    =TRUE)#delete the first column of data1.csv, save as data1'.csv.
> plot(data4$V1,data4$V2)
> linear = lm(V1~V2,data=data4)
```

```
> library(car)
> scatterplot(V1 ~ V2, data=data4)
```

2)KMeans

```
kmeans(dat, 3)
K-means clustering with 3 clusters of sizes 135, 73, 92

Cluster means:
          V1          V2
1  5.333055 -2.1202332
2  2.241573  0.6286581
3 -1.931235 -3.1101372
```

3)EM:

```
##Multi-dimentional Estimation-Maximization method for Mixed
    Gaussian Distribution of n terms(clusters):
library("mvtnorm")
EM = function(data, cluster_number) {
  ##setting primitive values
  n = nrow(data)
  dim = ncol(data)
  mu = matrix(runif(cluster_number * dim),nrow = dim,ncol = cluster_
      number)##mean
  sigma = rep(c(diag(dim)), length = cluster_number * dim * dim);dim
      (sigma) = c(dim, dim, cluster_number)##covariance matrix
  tau = runif(cluster_number)##coefficients of mixed gaussian
      distribution

  step = 150
  for (s in 1:step) {
    ##the E Step:
    T = matrix(nrow = cluster_number, ncol = n)##membership prob
    for (i in 1:n) {
      P = 0
      for (j in 1:cluster_number) {
        P = P + tau[j] * dmvnorm(data[i, ], mu[, j], sigma[, , j])
      }
      for (j in 1:cluster_number) {
        T[j, i] = tau[j] * dmvnorm(data[i, ], mu[, j], sigma[, , j])
            / P
      }
    }
    ##the M Step:
    tau = apply(T, 1, mean)
    sum = apply(T, 1, sum)
    for (j in 1:cluster_number) {
      P = rep(0, dim)
      Q = rep(0, dim * dim);dim(Q) = c(dim, dim)
      for (i in 1:n) {
        P = P + T[j, i] * data[i, ]
        Q = Q + T[j, i] * t(as.matrix(data[i, ] - mu[,j])) %*% as.
            matrix(data[i, ] - mu[,j])
      }
      mu[, j] = t(as.matrix(P / sum[j]))
      sigma[,,j] = as.matrix(Q / sum[j])
    }
```

```
  }
  return(list(mu=mu,sigma=sigma,tau=tau))
}
data = read.table("~/Desktop/STAT/hw/1/Data1'.csv",
                  sep = ",",
                  header = TRUE)
EM(data,3)
```

结果如下：

```
$mu
          [,1]       [,2]     [,3]
[1,] -1.964060  5.068487 2.087015
[2,] -3.113237 -2.013263 0.986269

$sigma
, , 1

            [,1]        [,2]
[1,] 0.99377298 0.01606908
[2,] 0.01606908 0.96803571

, , 2

           [,1]       [,2]
[1,]  2.049798 -0.111201
[2,] -0.111201  1.172887

, , 3

           [,1]       [,2]
[1,] 1.4512558 0.6590272
[2,] 0.6590272 0.8731801


$tau
[1] 0.3025644 0.5003484 0.1970871
```

**5. 假设我们有许多成人的身高数据：171 174 159 176 164 169 170 173 159 172 166 175 161 186 160 168 166 174 159 178 165 189 164 168 165 185 160 175 172 168 167 171 160 174 168 174 167 175 162 177 已知人的身高服从正态分布,且男性比女性的平均身高,试估计男性和女性的平均身高。**

**解：**

$$X \sim f(x|\theta) = \sum_{i=1}^{2} p_i N(x; \mu_1, \sigma_i)$$

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, 2, ..., n f(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_i - \mu)^2}{2\sigma^2})$$

```
data5 = matrix(c(171, 174, 159, 176, 164, 169, 170, 173, 159,
                 172, 166, 175, 161, 186, 160, 168, 166, 174,
                 159, 178, 165, 189, 164, 168, 165, 185, 160,
                 175, 172, 168, 167, 171, 160, 174, 168, 174,
                 167, 175, 162, 177))
plot(data5)
EM(dat, 2)
```

EM为上题中的函数。其中dmvnorm替换为mvnorm.
结果为$\mu = (0.92855130.07144865), \sigma = (168.3369186.7152), \tau = (33.289932.908326)$

# 1 习题课

$$\frac{\partial tr(BA)}{\partial A} = B^T \frac{\partial det(A)}{\partial A} = A^{*T} trace(ABC)可交换$$