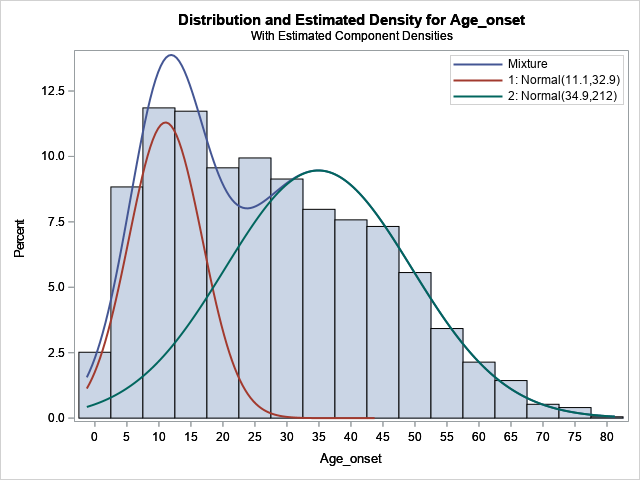
一.发病年龄双峰：

1、能否把整体的曲线和双峰放在一个图，如下图？

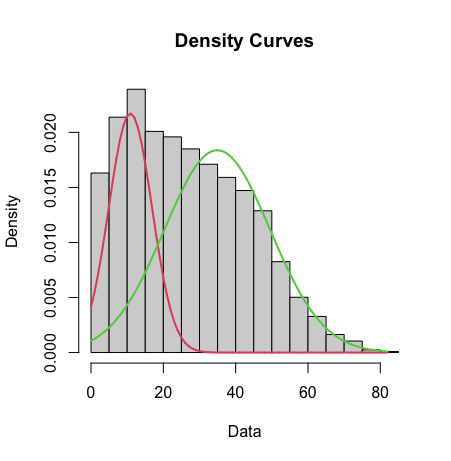
不好意思，我现在只能做出两个分曲线图。

整体的曲线图我一直查不到叫什么名字（英文）所以暂时没有头绪。

如果您可以给我一点类似的信息／线索，我可以再次尝试。



2、两个曲线的Mixing proportion ？

$lambda

[1] 0.3256286 0.6743714

$mu

[1] 10.89470 34.91403

$sigma

[1] 5.98363 14.63282

$loglik

[1] -16775.29

mixing proportion = lambda

Green Distribution = 0.6743

Red Distribution = 0.3256

3、文献里方法描述“we performed goodness of fit analyses using the finite mixture model (FMM) procedure in SAS version 9.4 [https://www.sas.com]. We set component distributions as normal, the maximum number of mixture components as seven, fit the FMMs using the maximum likelihood method, and chose the best fit model using the Bayesian information criterion (BIC)”

您用的什么方法？

* **Mixture Model Clustering** is used to estimate the entire distribution
* normalmixEM function was used under the package named mixtools in Rstudio(<https://rstudio.com/>)

Component distributions设置？

* normalmixEM function builds a **2 component Guassian Mixture Model(Normal Mixture Model)**
* 

k = 2 means 2 component distributions

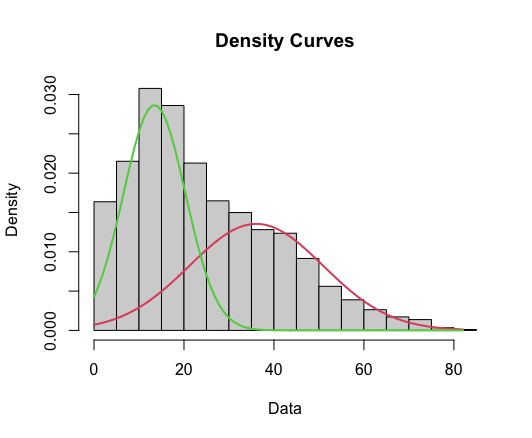
Maximum number of mixture components是多少？

* Maximum number of mixture component is not mentioned in this method

4. 按男性和女性两个cohort分别做发病年龄双峰，看看这两个cohort的early-onset和late-onset的mean+SD， mixing proportion?

Male:

> summary(malemix)

summary of normalmixEM object:

comp 1 comp 2

lambda 0.504876 0.495124

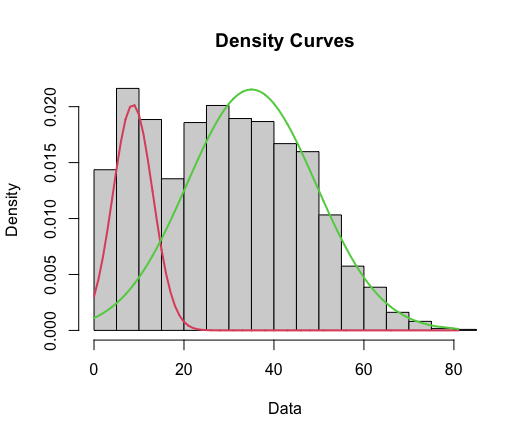
mu 36.082131 13.441243

sigma 14.866477 6.890332

loglik at estimate: -7175.529

|  |
| --- |
| lambda = mixing proportion |
| mu = mean |
| sigma = standard deviation/SD |
| Green Distribution = comp 2  Red Distribution = comp 1 |

Female:



> summary(femalemix)

summary of normalmixEM object:

comp 1 comp 2

lambda 0.224606 0.775394

mu 8.640215 34.997807

sigma 4.437238 14.361338

loglik at estimate: -9302.438

>

|  |
| --- |
| lambda = mixing proportion |
| mu = mean |
| sigma = standard deviation/SD |
| Green Distribution = comp 2  Red Distribution = comp 1 |

二、发病年龄随年度变化

1. 现有数据是否能分析出2个分界点？如果不行的话，是不是因为80年以前的数据不够？

1）没有数据的年份：1953-1961， 1963-1969，1974，1976-1978

2）拥有一个数据的年份：1952，1962, 1973，1975，1979，1983，1984，1991

3）拥有两个数据的年份: 1970, 1982 , 1984

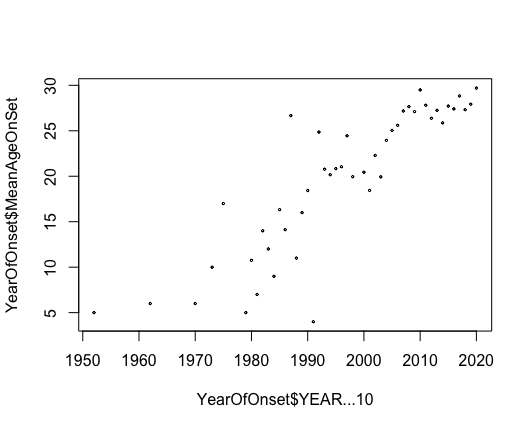
4) Central Limit Theorem(中心极限定理): distribution of sample means approximates a normal distribution as the sample size gets larger. Sample size equal to greater than 30 are considered sufficient for CLT to hold. (在统计学中一般默认抽取样本的数量要大于等于30, 数据会呈现正态分布)

综上，1990年以前的数据都非常缺乏，平均发病年龄在数据样本量（当年的发病年龄数据样本量)大于30才有普遍代表意义（simple, random sample). (也就是说至少要有当年30个病人的发病年龄然后再计算当年平均发病年龄才有代表意义)

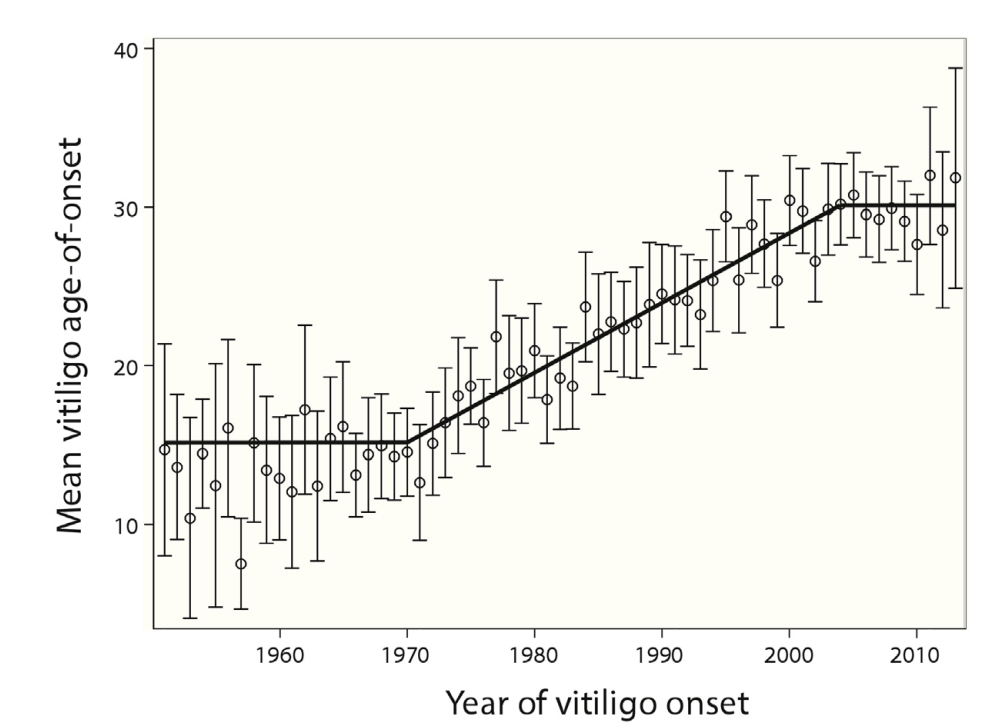
在所给的数据中，基本只有2000年以后的平均发病年龄才能满足这个要求。某些年份只有一份数据，只能按照唯一的发病年龄算作当年的平均发病年龄（参见2,），这其实**是不能满足在统计学中做回归曲线的前提的**，因为我们不能保证：

* 样本是随机抽取的（did we use simple random sample?）
* 样本之间是独立且没有关联性的（are the subject independent?）

参见scatter plot, 我们可以明显看出和2000年以后的数据相比，1990年以前的平均发病年龄是不规律分布的，因为很多这些平均发病年龄是没有代表意义的。尤其某些特殊的年份。（举例：比如只有一份发病年龄数据的年份， 1991）



2、能否按下图表示出每年的SD和变化折线？



因为无法满足做回归的前提，而且我也不知道这个图的英文名，所以这个图也没无法画出来。

每条竖线代表的不是SD而是 [ 95% confidence interval ], 如果您需要指定年段的95% confidence interval，我可以帮您用软件计算出来（比如有实际意义的2000年以后的数据）。

3、1980-2002年的斜率？SE？P值？

2002年以后的斜率？SE？P值？

> summary(lfit)

Call:

lm(formula = meanageonset ~ yearonset)

Residuals:

Min 1Q Median 3Q Max

-5.5522 -2.1474 -0.2263 1.1750 10.8349

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -464.92062 130.97417 -3.550 0.00291 \*\*

yearonset 0.24014 0.06594 3.642 0.00241 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.352 on 15 degrees of freedom

(3990 observations deleted due to missingness)

Multiple R-squared: 0.4692, Adjusted R-squared: 0.4339

F-statistic: 13.26 on 1 and 15 DF, p-value: 0.002411

Meaningful coefficients of the linear terms:

(Intercept) yearonset U1.yearonset

-551.3521 0.2839 -0.3123

Estimated Break-Point(s):

psi1.yearonset

2002

> sfit <- segmented(obj = lfit, seg.Z = ~yearonset, psi = 2)

Error in segmented.lm(obj = lfit, seg.Z = ~yearonset, psi = 2) :

starting psi out of the admissible range

> plot(sfit)

> summary(sfit)

\*\*\*Regression Model with Segmented Relationship(s)\*\*\*

Call:

segmented.lm(obj = lfit, seg.Z = ~yearonset)

Estimated Break-Point(s):

Est. St.Err

psi1.yearonset 2002.183 19.839

Meaningful coefficients of the linear terms:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -551.35208 192.27560 -2.868 0.0132 \*

yearonset 0.28387 0.09706 2.925 0.0118 \*

U1.yearonset -0.31225 0.65512 -0.477 NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.581 on 13 degrees of freedom

Multiple R-Squared: 0.4903, Adjusted R-squared: 0.3727

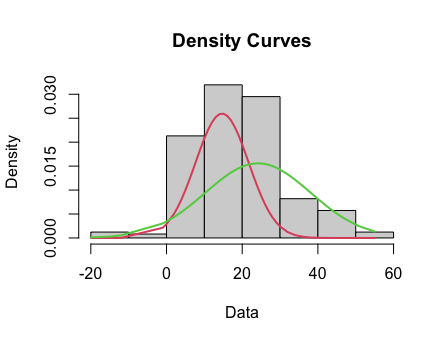
Convergence \*not\* attained in 1 iter. (rel. change -1.5078e-05)

>

4. 按1980-2002 和2002以后两个cohort分别做发病年龄双峰，看看这两个cohort的early-onset和late-onset的mean+SD， mixing proportion?

1980-2002的男性和女性分别做发病年龄双峰；2002以后的男性和女性分别做发病年龄双峰，看看以上参数的变化？

1980 – 2002: （应该是错误的，从图表看1980-2002的平均年龄分布不存在双峰分布，男女发病双峰也省去了，因为没有意义）



summary of normalmixEM object:

comp 1 comp 2

lambda 0.45663 0.54337

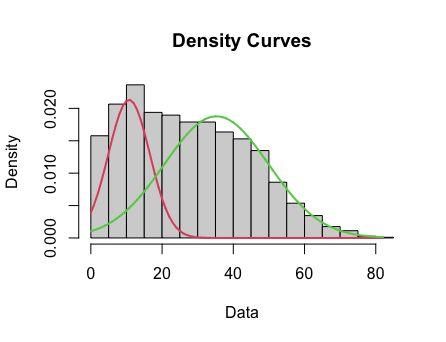
mu 14.69150 24.18205

sigma 7.01490 13.91288

loglik at estimate: -949.26

|  |
| --- |
| lambda = mixing proportion |
| mu = mean |
| sigma = standard deviation/SD |
| Green Distribution = comp 2  Red Distribution = comp 1 |

2002 – 至今：



summary of normalmixEM object:

comp 1 comp 2

lambda 0.311684 0.688316

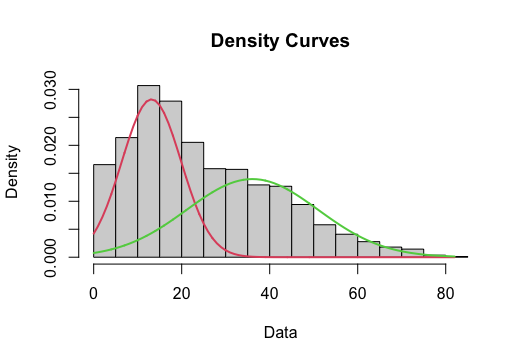
mu 10.750437 35.274607

sigma 5.831863 14.626367

loglik at estimate: -15735.6

|  |
| --- |
| lambda = mixing proportion |
| mu = mean |
| sigma = standard deviation/SD |
| Green Distribution = comp 2  Red Distribution = comp 1 |

2002-至今男性：



> summary(mymixageonsetmale2002)

summary of normalmixEM object:

comp 1 comp 2

lambda 0.476809 0.523191

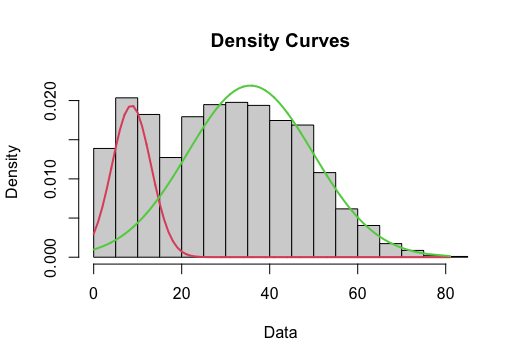
mu 13.135285 36.071868

sigma 6.741183 14.943068

loglik at estimate: -6817.467

|  |
| --- |
| lambda = mixing proportion |
| mu = mean |
| sigma = standard deviation/SD |
| Green Distribution = comp 2  Red Distribution = comp 1 |

2002-至今女性：



> summary(mymixfemale)

summary of normalmixEM object:

comp 1 comp 2

lambda 0.215775 0.784225

mu 8.589966 35.598922

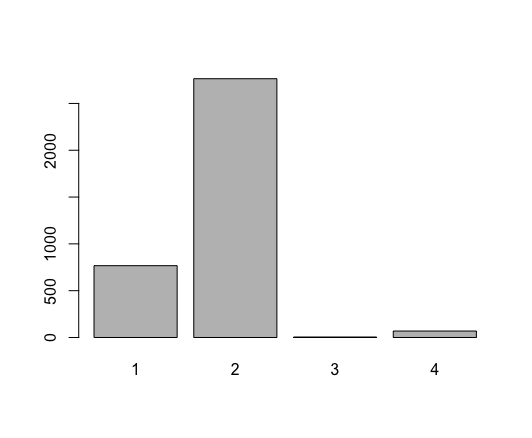
sigma 4.441533 14.276612

loglik at estimate: -8671.847

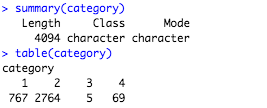
|  |
| --- |
| lambda = mixing proportion |
| mu = mean |
| sigma = standard deviation/SD |
| Green Distribution = comp 2  Red Distribution = comp 1 |

三、其他分析

1、发病类型比例随年度变化趋势

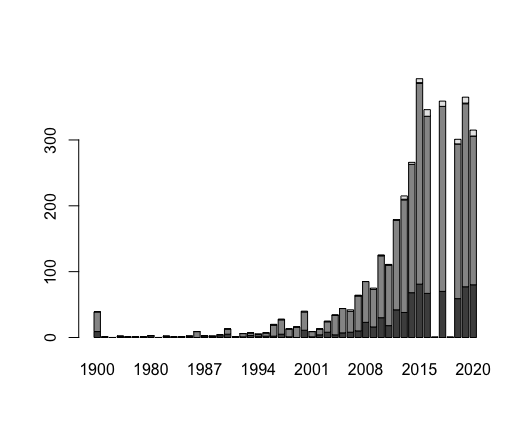
发病类型即表格中的西医分型，主要想看看各分型的比例（主要是1型和2型）是否随年度有变化规律？

**各分型的比例（主要是1型和2型）:**

****

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |  |
| Count | 767 | 2764 | 5 | 69 | Total: 4094 |
| Percentage | 18.73% | 67.51% | 0.12% | 1.68% | Total: 100% |

**是否随年度有变化规律？**

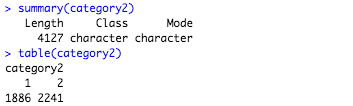


2、初发部位（暴露/非暴露部位）比例随年度变化趋势

表格中“基本信息\_初发部位”这部分来区分暴露部位和非暴露部位，其中“头、面、颈、手”属于暴露部位，其余属于非暴露部位。需要处理一下原始数据。看暴露部位/非暴露部位的比例有没有随年度变化的规律？

* Value contains “头、面、颈、手” = 1
* Value does not contain “头、面、颈、手”= 2





|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 |  |
| Count | 1886 | 2241 | Total:4127 |
| Percentage | 45.69% | 54.30% | Total:99.99% |

看暴露部位/非暴露部位的比例有没有随年度变化的规律？

