实训 - 医学数据分析

计算机基础教研室
张国鹏 18092191446 (微信) zhanggp@fmmu.edu.cn

医学数据分析－现实需求

某医院检验科的部分血常规数据文件记录了多名病人的病人编号（zID）,性别（Gender），红细胞计数(RBC)和白细胞计数（WBC），血红蛋白(Hb)和淋巴细胞计数（LY）共6列数据。

其中zID为病人编号，用来唯一识别某一个病人

数据文件逻辑上 n 行6列数据表（含表头），

zID	Gender	RBC(10^12/L)	WBC(10^9/L)	Hb(g/L)	LY(10^9/L)
Z000023	男	3.8543	6.6968	156.1351	3.1976
……					

python

医学数据分析 – 现实需求

原始文件数据格式：

文本文件，后缀名.daz，编码方式'utf-8'

实际文件中表头占1行（制表符'\t'分隔的6列），第2行是所有的数据记录，记录内部和记录之间都用制表符'\t'分隔。患者ID长度为固定长度7位，以字符“Z”开头，数字部分为6位，不足位数补0

```
zID\tGender\tRBC(10^12/L)\tWBC(10^9/L)\tHb(g/L)\tLY(10^9/L)\nZ000010\t男\t4.0224\t9.7905\t136.5197\t3.4792\tZ000003\t ……
```

python

医学数据分析－现实需求

原始文件数据格式：

数据文件最少有1条数据记录，最多为999999条记录

zID	Gender	RBC(10^12/L)		WBC(10^9/L)		Hb(g/L)	LY(10^9/L)
Z000001	男	3.8543	6.6906	156.1351	3.1976		

zID	Gender	RBC(10^12/L)		WBC(10^9/L)		Hb(g/L)	LY(10^9/L)				
Z000010	男	4.0224	9.7905	136.5197	3.4792	Z000011	男	4.7786	6.7324	129.1999	0.1086
Z000043	女	6.0504	8.2686	127.2782	0.5001	Z000030	女	1.9106	4.6063	122.6804	1.9844
Z000015	男	1.5914	10.0966	111.7144	3.6703	Z000038	女	2.6294	7.3138	121.9451	2.2825
Z000036	男	5.4177	5.5367	145.1475	2.0372	Z000033	男	4.1337	4.3680	134.4502	2.2851
Z000020	女	2.9128	2.8126	156.8516	0.9403	Z000016	女	0.9001	8.8633	137.6296	3.2089
Z000018	男	4.1054	2.8274	126.0920	3.2561	Z000013	男	3.9172	10.1505	138.0053	3.7312
Z000024	男	3.3933	4.5287	124.9847	2.4259	Z000003	女	2.0027	5.0927	138.7247	3.7369
Z000042	女	4.7768	12.5906	138.4560	1.2300	Z000023	男	5.9997	10.5593	139.9406	2.9328
Z000035	女	2.7111	9.6683	137.0609	2.4561	Z000030	女	2.2823	4.1127	155.3371	4.8295
Z000024	男	4.1954	12.4295	131.6136	1.3285	Z000013	男	5.1196	5.5110	144.2287	5.3486
Z000031	女	5.8995	8.2830	135.3623	4.1414	Z000028	男	5.8272	5.2572	150.1703	1.1063
Z000015	男	5.3619	4.9540	153.8149	2.2653	Z000044	男	4.4305	5.7719	132.0353	1.6250
Z000032	女	3.9433	10.1627	139.1092	1.7493	Z000042	女	4.7907	9.0031	127.0224	2.5667
Z000023	男	8.1339	10.2874	155.4771	0.8816	Z000031	女	2.5008	4.0163	128.1123	2.9787
Z000001	女	2.3653	4.6850	142.5660	0.6067	Z000019	男	4.3241	7.6815	148.9888	2.3749
Z000007	男	3.7368	8.8386	116.6403	0.5945	Z000004	男	4.8518	5.6126	131.0019	2.7984
Z000002	女	4.8958	10.9668	132.6811	0.1667	Z000005	男	5.2751	9.3109	161.6172	0.3819

用记事本打开

（注意记事本会自动折行显示）

医学数据分析 – 现实需求

- 编写一个程序实现此类文件数据的统计分析

  任务1（必做）. 对用户指定的数据文件，分别统计并显示总人次，男/女人数，男/女的RBC、WBC平均值共 7 个统计指标。注意：若有多个zID相同，计入总人次，但人数只能算1人，统计RBC和WBC时只统计此人的第1条记录，其余记录不纳入统计。同时对各检查次数的人数分别进行统计，显示检查次数最多的所有人ID

  任务2（拓展选作，不计分，不上交）. 将数据重新整理成行列整齐的数据，存成 Excel 可以处理的文件，并用 Excel 进行分析

python

医学数据分析 – 编程要求

- **任务1**
  - 用户输入数据文件名
  - 程序读取文件中的数据并分别统计总人次、男/女人数、男/女RBC平均和WBC平均 7 个统计指标
  - 统计各有效检查次数的人数；按升序输出检查次数最多的人员编号

程序运行格式范例：

男/女红细胞平均计数

男/女白细胞平均计数

```
<请输入文件名>med1.daz
<总人次>1820<男女人数>645/593<RBC>4.672/4.343<WBC>7.338/7.323
<检查人次统计如下>
1次761人
2次380人
3次89人
4次8人
<检查4次的人按编号从小到大排列如下>
Z000664
Z000841
Z001033
Z001056
Z001112
Z001528
Z001582
Z001698
```

医学数据分析－格式要求

- 程序运行显示提示语<请输入文件名>，用户输入数据文件名（用户确保数据文件和程序文件在同一文件夹下）
- 程序打开数据文件并进行分析后，屏幕输出<总人次><男女人数><RBC><WBC>共7个统计指标。

  【注意】在一行输出，中间没有空格；<RBC>和<WBC>平均值保留3位小数；<男女人数> <RBC><WBC>的数据均是先是男性后是女性。例如<RBC>4.672/4.343表示男红细胞平均计数为4.672，女红为4.343
- 显示提示语<检查人次统计如下>并换行，逐行按次数升序输出有效（无人数不输出）检查次数对应人数
- 显示提示语<检查n次的人按编号从小到大排列如下>并换行，按编号升序逐行输出检查n次的人员编号zID。

  【注意】其中n为最多的检查次数，zID输出顺序是数值部分的升序

程序运行格式范例0：

男/女红细胞平均计数

男/女白细胞平均计数

```
<请输入文件名>med1.daz
<总人次>1820<男女人数>645/593<RBC>4.672/4.343<WBC>7.338/7.323
<检查人次统计如下>
1次761人
2次380人
3次89人
4次8人
<检查4次的人按编号从小到大排列如下>
Z000664
Z000841
Z001033
Z001056
Z001112
Z001528
Z001582
Z001698
```

医学数据分析－格式要求

- 程序运行显示提示语<请输入文件名>，用户输入数据文件名（用户确保数据文件和程序文件在同一文件夹下）
- 程序打开数据文件并进行分析后，屏幕输出<总人次><男女人数><RBC><WBC>共7个统计指标。

  【注意】在一行输出，中间没有空格；<RBC>和<WBC>平均值保留3位小数；<男女人数> <RBC><WBC>的数据均是先是男性后是女性。例如<RBC>4.672/4.343表示男红细胞平均计数为4.672，女红为4.343
- 显示提示语<检查人次统计如下>并换行，逐行按次数升序输出有效（无人数不输出）检查次数对应人数
- 显示提示语<检查n次的人按编号从小到大排列如下>并换行，按编号升序逐行输出检查n次的人员编号zID。

  【注意】其中n为最多的检查次数，zID输出顺序是数值部分的升序

程序运行格式范例1（无女性检查者的情况）：

<请输入文件名>med2.daz
<总人次>1<男女人数>1/0<RBC>3.854/NA<WBC>6.691/NA
<检查人次统计如下>
1次1人
<检查1次的人按编号从小到大排列如下>
Z000001

注意：若人数为0，则对应RBC和WBC统计指标应为NA

程序运行格式范例2

<请输入文件名>med3.daz
<总人次>680<男女人数>230/219<RBC>4.992/4.001<WBC>7.237/7.660
<检查人次统计如下>
1次269人
2次134人
3次42人
4次3人
5次1人
<检查5次的人按编号从小到大排列如下>
Z000533

实训(ID=9)：医学数据分析-上交

- 上交文件名：“py+实训ID+下划线+学号.py”

  例如学号为1234567的同学的上交文件名应为：

  py9_1234567.py

python

医学数据分析（ID=9）参考

- 可能用到的知识点：
  - 文件对象的方法：open，close，read，readline，write，seek
  - 内置函数：len、int
  - 字符串的方法： split
  - 循环结构、分支结构
  - 列表的遍历、索引和切片操作；字典的使用
  - 字符串的格式化输出

https://www.runoob.com/python/att-string-format.html

https://docs.python.org/3/library/string.html#formatstrings

python

医学数据分析－任务2

- 在任务1的基础之上进行拓展
- 编程创建新文件 stu+学号.csv（例如学号为12345678的同学新文件名应为stu12345678.csv），将原始数据中的前4列数据写入新文件。写入格式，第1行为表头，后面每1行写入1条记录，列之间用英文逗号分隔
- **在线参考资料：**

  https://docs.python.org/3/library/csv.html?highlight=csv

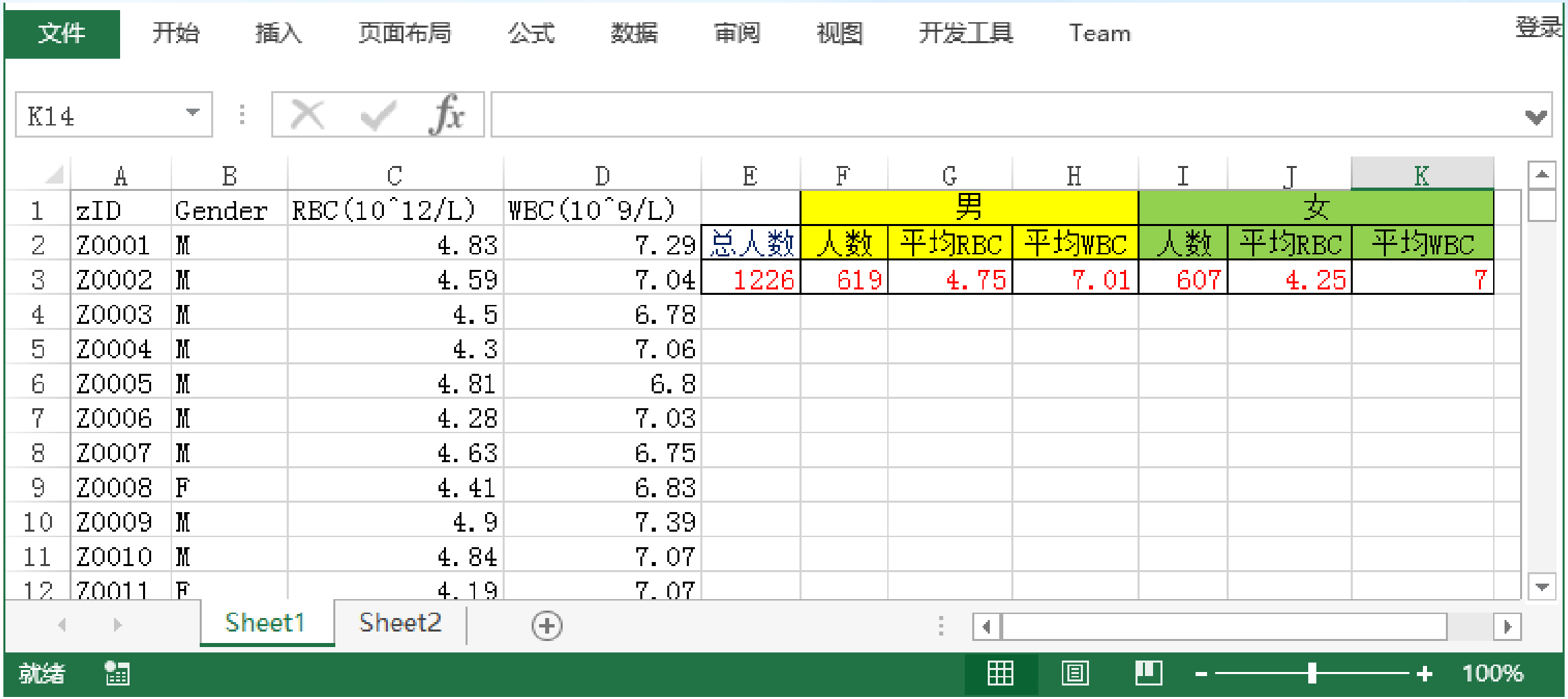
python

医学数据分析－任务2

1. 生成.csv文件后，关闭程序。
2. 用Excel打开生成文件stu12345678.csv中的数据，另存为stu12345678.xlsx
3. 在Excel中对stu12345678.xlsx按病人编号排序，用Excel的公式统计总人数和任务1统计的6个指标，对比Excel统计的结果和程序统计结果
4. 思考题：能不能用Excel去掉重复的ID？

python

完成效果示例图：
请注意：E3:K3区域均为公式计算所得，并非直接填入数字



	A	B	C	D	E	F	G	H	I	J	K
1	zID	Gender	RBC(10^12/L)	WBC(10^9/L)		男			女		
2	Z0001	M	4.83	7.29	总人数	人数	平均RBC	平均WBC	人数	平均RBC	平均WBC
3	Z0002	M	4.59	7.04	1226	619	4.75	7.01	607	4.25	7
4	Z0003	M	4.5	6.78							
5	Z0004	M	4.3	7.06							
6	Z0005	M	4.81	6.8							
7	Z0006	M	4.28	7.03							
8	Z0007	M	4.63	6.75							
9	Z0008	F	4.41	6.83							
10	Z0009	M	4.9	7.39							
11	Z0010	M	4.84	7.07							
12	Z0011	F	4.19	7.07							