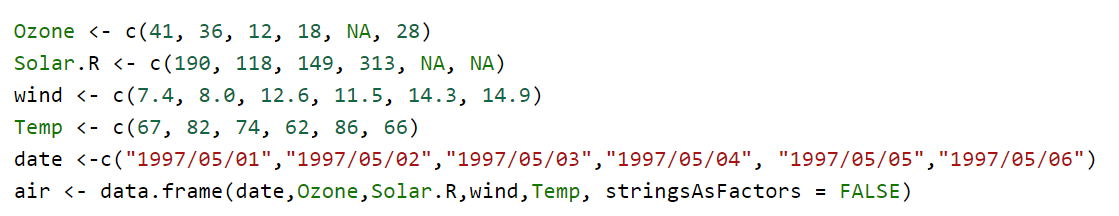
实验二 R语言与描述性统计

【实验目的】

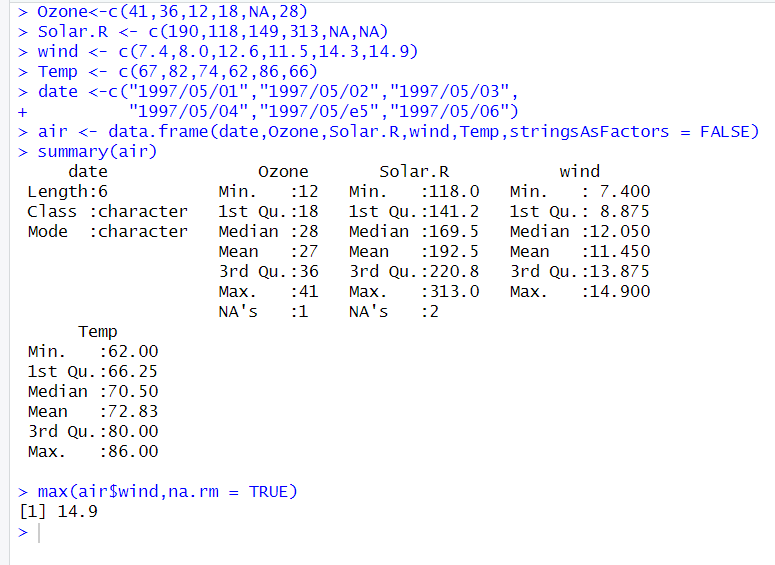
1. 掌握数据预处理的相关操作，包括查看数据集的缺失值，对缺失值进行处理，对异常变量重新赋值。
2. 掌握使用可视化图表方法进行探索性统计分析。
3. 掌握常用的概率分布，并学会用R语言编程求解概率分布问题。
4. 掌握常用的抽样方法，并学会用R语言编程实现抽样。

【实验内容与实现】

1. 根据下述代码，生成数据。

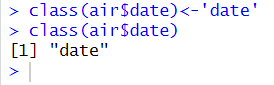


1. 使用summary()函数查看哪一列含有缺失值的个数，并查看‘wind’列的最值。



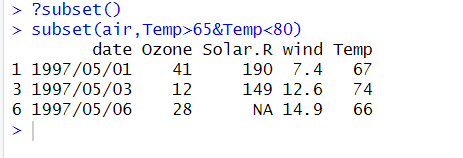
1. 使用class()函数查看date的类型，并将其转化为date型，格式为年/月/日。



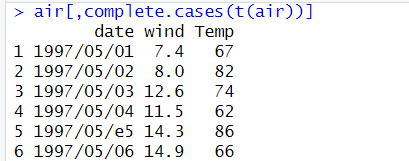


注意数据类型输入应该用单引号。

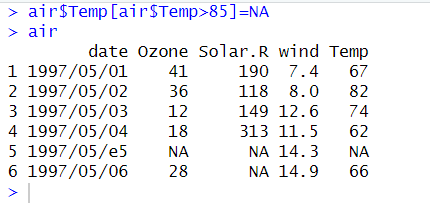
1. 使用subset()根据‘Temp’列获取air中‘Temp’大于65小于80的数据。



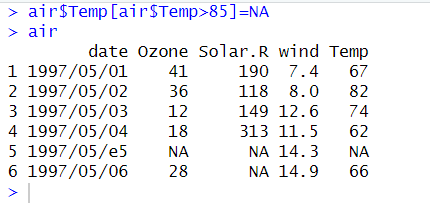
1. 去掉air中含有缺失值的列。



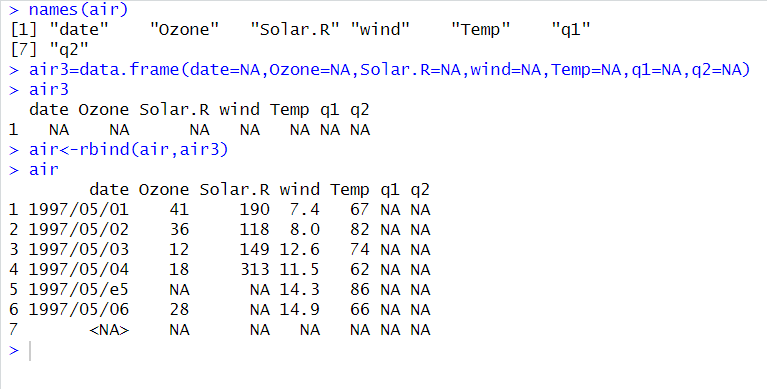
1. 找出 ‘Temp’ 中大于 85 的值将其设置为缺失值。



1. 创建一个数据框 air1 包含两个列 q1，q2。用 cbind 函数添加到 air 中。



1. 使用 names（） 函数查看 air 的列名，再自己创建一个新的数据框，新数据框的列名要与air的列名相同。使用 rbind（） 函数添加到 air 中。



1. 科学家正在开展一项模拟车窗着色对侧窗视觉影响的实验，测得的数据在data.txt，数据中含有以下几列：

case：观测序号；

id：研究车辆的识别号（1-26）；

age：受测者年龄；

sex：表示受测者性别的因子变量，具有f（女性）和m（男性）两个水平；

tint：表示着色程度的有序因子变量，no<lo<hi；

target：表示对比度的因子变量，水平为：locon（低对比度），hicon（高对比度）；

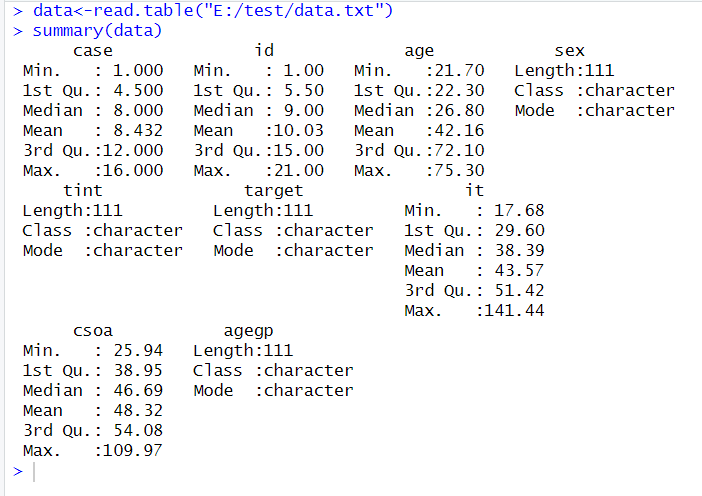
it：检查时间，执行简单识别任务所需的时间（单位为毫秒）；

csoa：临界刺激开始异步，识别字母数字目标的时间（单位为毫秒）；

agegp：年龄阶段因子变量，水平为younger（21-27岁）、older（70-78岁）；

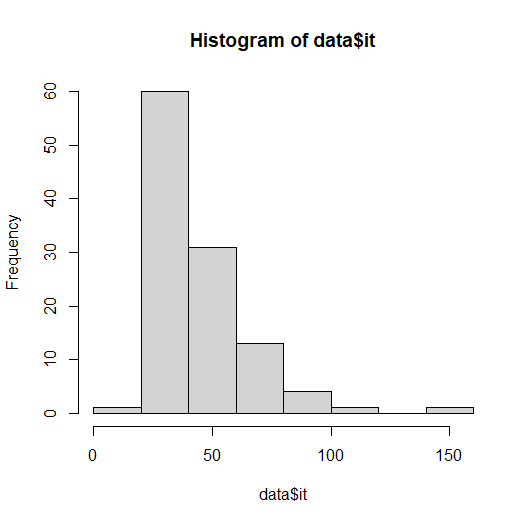
完成下述任务：

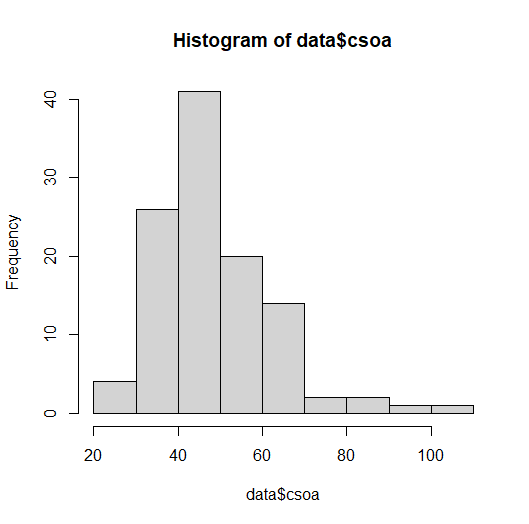
1. 读取数据，并逐列对数据进行描述性统计分析；



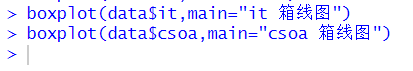
1. 分别探索it和csoa的数据分布情况；

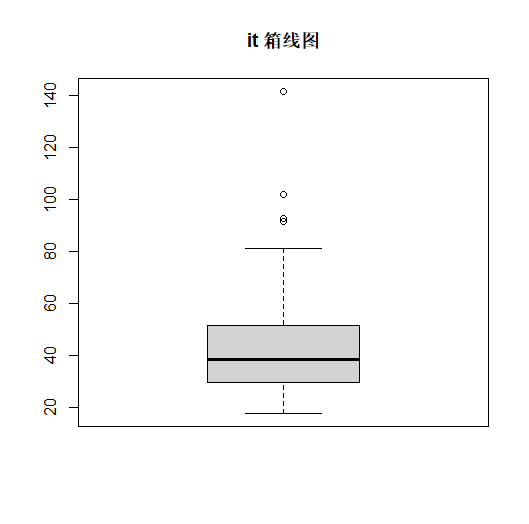
对it和csoa进行探索性数据分析，分别画出其直方图和箱线图。

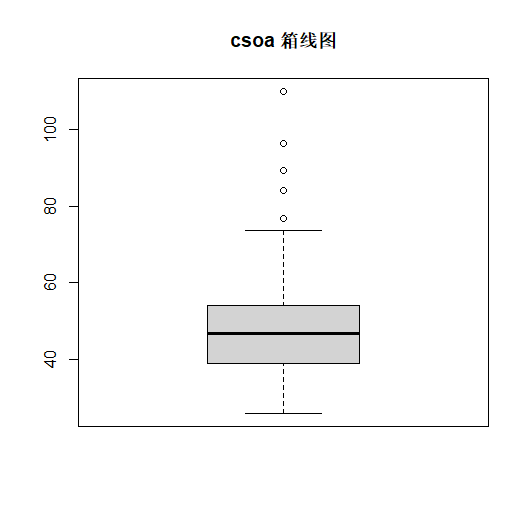




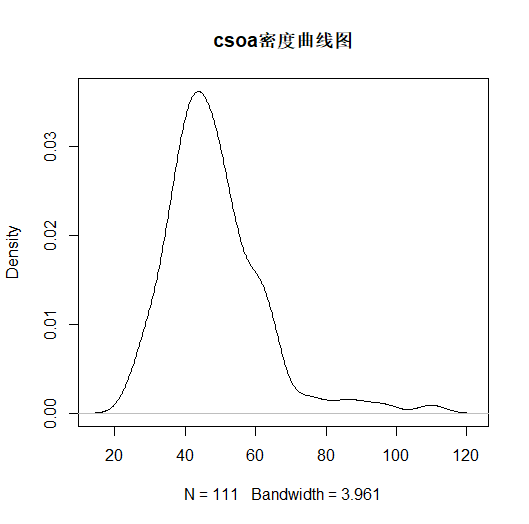
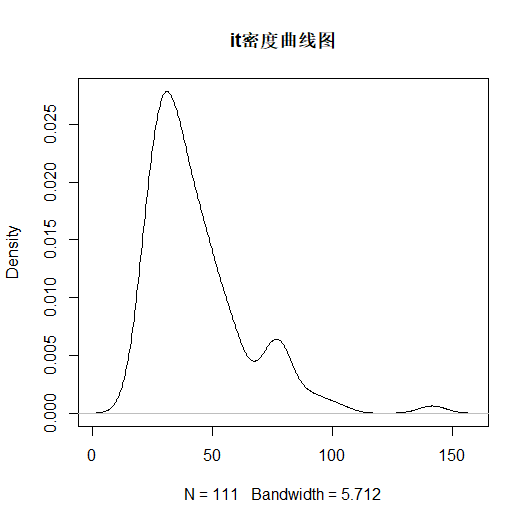
画出箱线图首先需要一个“ggplot2”的程序包，首先在Rstudio中安装这个包，然后调用这个包。





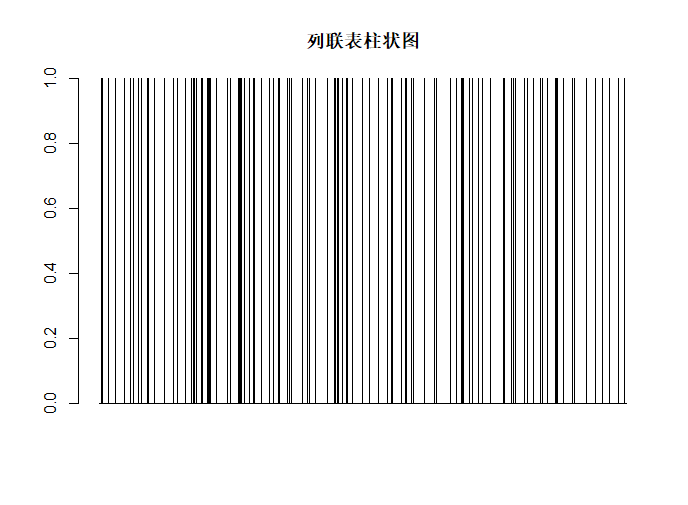


密度曲线图如下：



1. 利用列联表和联合柱状图，探索变量tint对it和csoa的影响；

Barplot（），beside=TRUE，每个类都分开画，xlab，ylab等等，要画图例。Lenged=TRUE



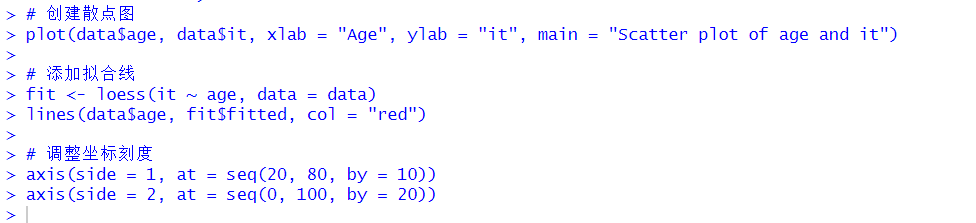


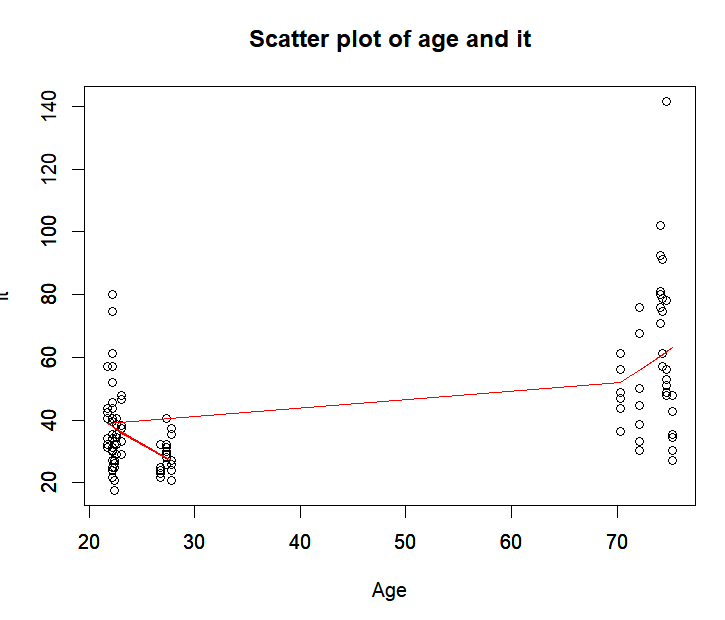
1. 作出age和it的散点图, 并进一步完成下面的操作:

Plot（），ggplot（）

i. 用loess()函数作出拟合线;

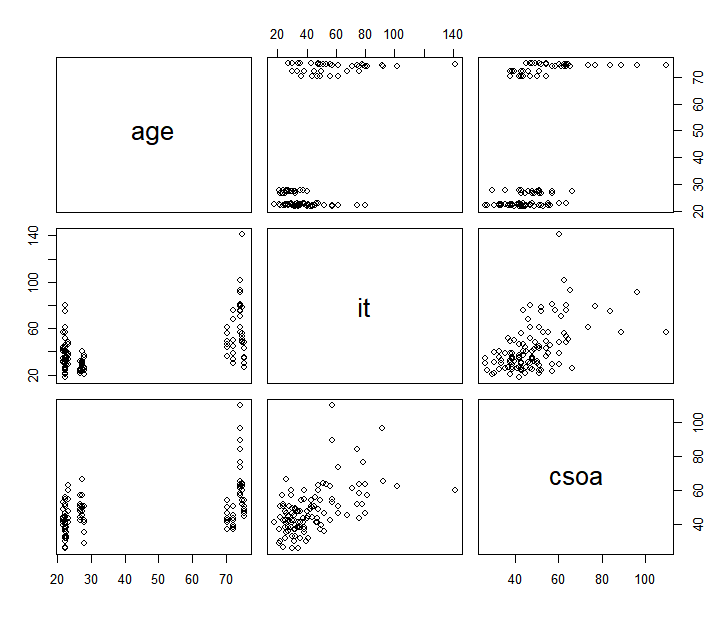
ii. 对图的坐标刻度进行调整;



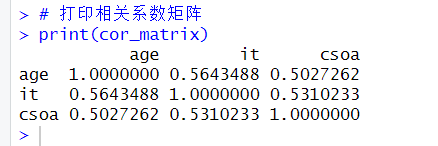


1. 探索描述变量age、it和csoa两两之间是否存在线性关系。

散点图（矩阵散点图）



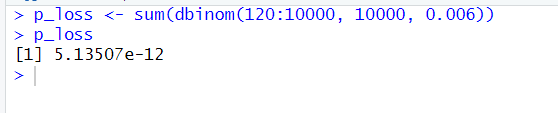




从以上相关性矩阵中可以看出，age和it之间不存在明显的线性相关关系，it和csoa之间存在着一定的线性相关关系。

1. 某保险公司有10000人参加保险，每人每年交12元保险费，在一年内一个人死亡的概率为0.006，死亡后其家属可向保险公司领到1000元。保险公司亏本的概率是多少？说明过程，并用R解答。

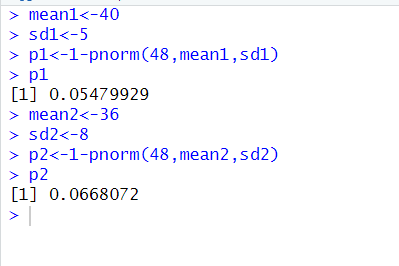
解答过程：首先考虑保险公司的收入，10000人参保，每人12元，那么总收入为120000元，死亡概率P=0.006，那么存活概率为1-P,那么保险公司亏本时应有120000-1000\*n<0，也就是有至少120死亡时，保险公司就会出现亏本情况。我用二项分布来计算这个概率。



即保险公司亏本的概率为5.13507\*10-12

1. 考虑两个相互独立的施工单位。在单位甲中，假定每个员工的工作量服从平均产量40、标准差为5的正态分布。单位乙中每个员工工作量服从平均产量为36、标准差为8的正态分布。一个高效的员工的工作量定义为至少48。请问从哪个施工单位中随机挑选员工可以获得更大的工作量？请提供用R语言分析本题的过程。

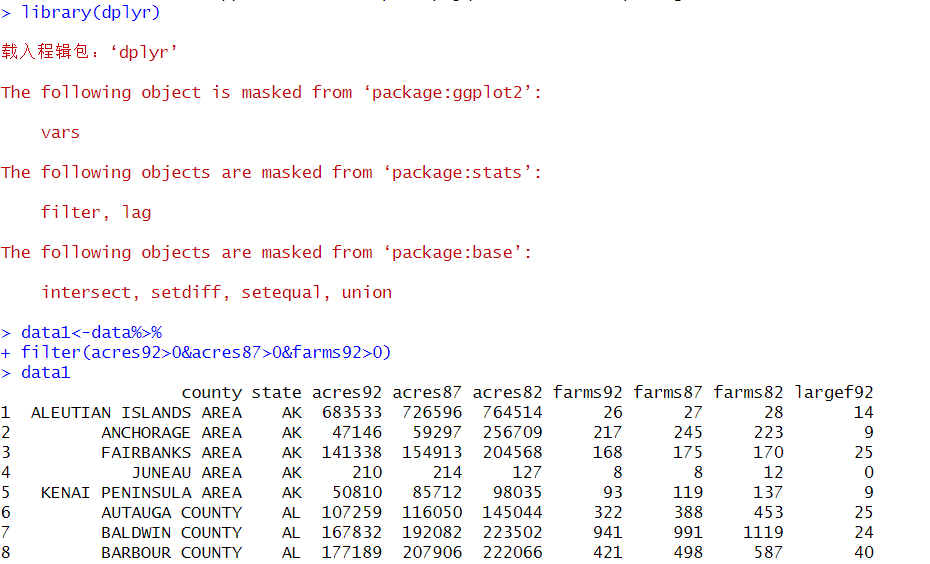
由题目可知，两个员工的工作量都服从正态分布，那么根据其对应的均值和方差使用pnorm（）函数就能计算出，工作量大于48的概率。R语言如下：



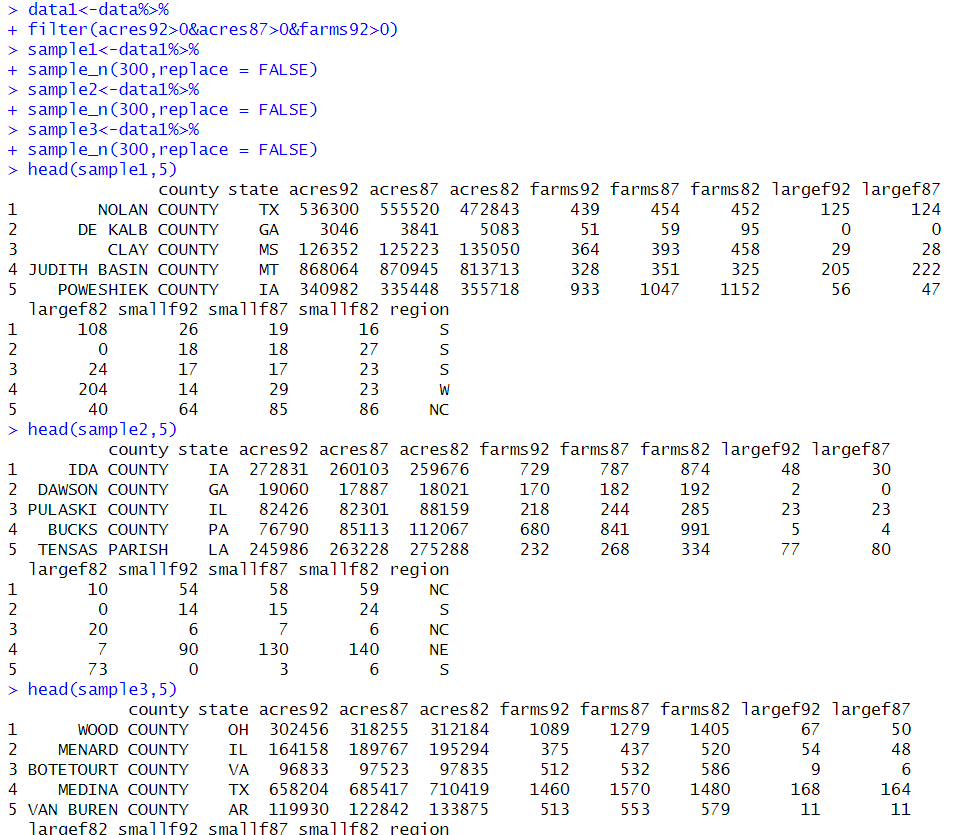
从中可以看出乙工作量大于48的概率更大，因此从乙施工单位中随机挑选员工比较合适。

1. agpop.csv文件是美国政府每五年做一次有关农业的普查，收集50个州的所有农场的有关数据，共3078个美国县级或与县级规模相当的农场数据。共有4个区域（region/rnum），50个州（state/snum）,以及3041个县（county/cnum）。我们用到的变量有:县（county/cnum）, 州（state/snum）, 区域（region/rnum）,1992年每个县的耕地面积（acres92）, 1987年每个县的耕地面积（acres87）,1992年每个县拥有的农场个数（farms92）。完成下述任务：
   * 1. 导入原始文件并对进行异常值做剔除处理；

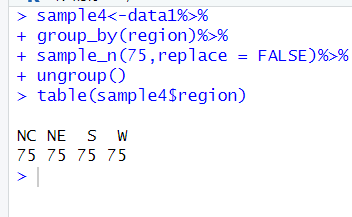
调用程序包dplyr进行异常值处理，首先把负值去除。



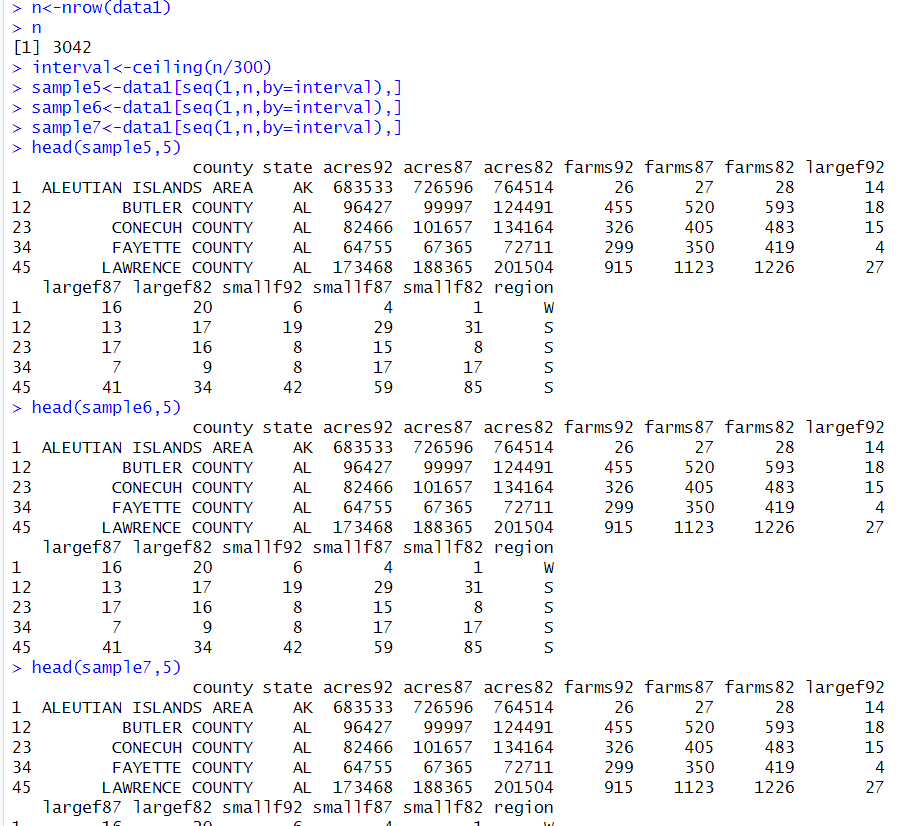
* + 1. 利用以下抽样方法，获取抽样样本：
       1. 利用不放回简单随机抽样，抽取三个样本容量为300的样本，并分别输出每个样本的前五行数据；



* + - 1. 使用分层抽样方法，以‘region’为分层变量，从每层简单随机抽取75个样本单元，构成一个样本，并查看样本中‘region’每层的样本数量；

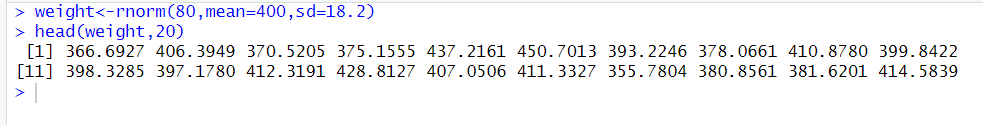


* + - 1. 采用等距抽样方法抽取三个样本容量为300的样本，并分别输出每个样本的前五行数据；

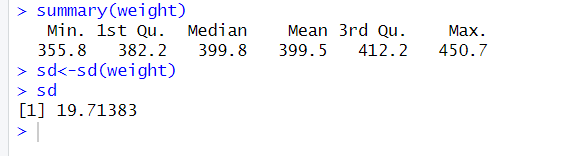


【作业】

1. 我们以奶牛为例，执行下列操作：
   1. 生成如下的数据：80个正态分布的随机数（即长度为80的向量），该正态分布满足均值为400.0以及标准差为18.2的条件。将该向量命名为‘weight’并输出他的前20个值。

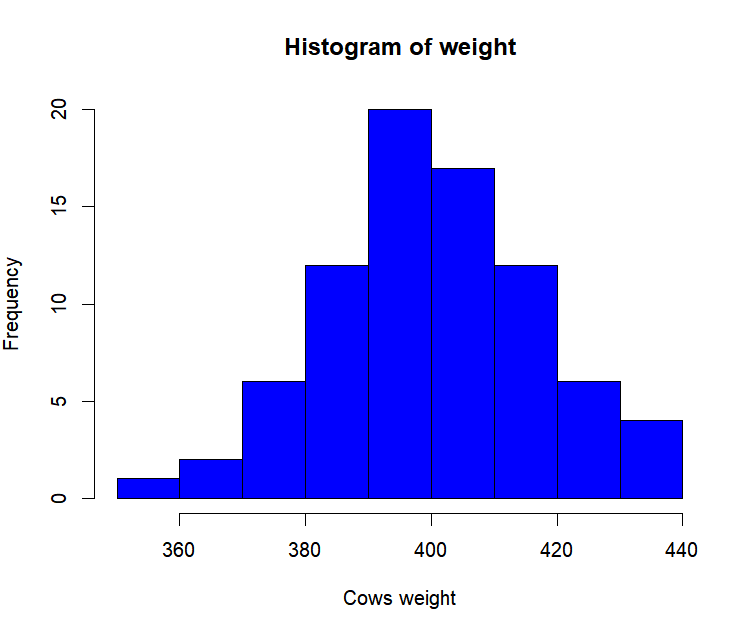


* 1. 计算并输出‘weight’向量的均值、中位数、值域、标准差以及该向量的分位数。



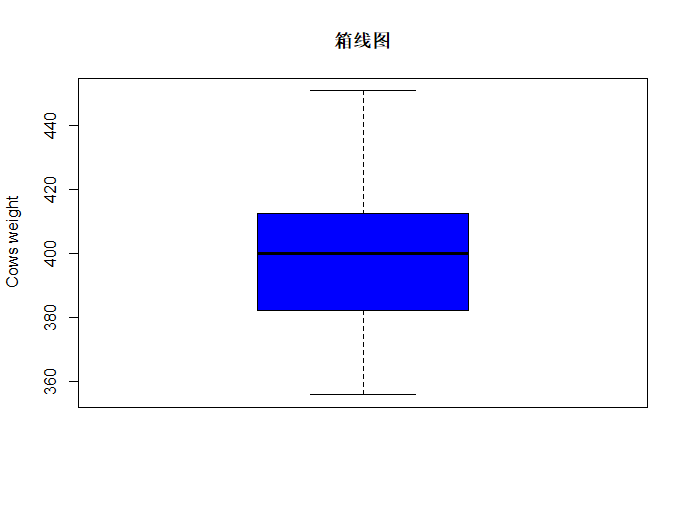
* 1. 通过‘weight’向量生成一个直方图。要求x轴标签为‘Cows weight’且直方图的柱状体为蓝色。





* 1. 生成‘weight’向量的箱线图。要求y轴对应标签为‘Cows weight’，设置颜色为蓝色。





* 1. 分析c,d绘制的两种图。通过你的分析，那种图生成了更多的信息？（70字以内）

直方图和箱线图都可对数据进行探索性分析，能够较好的反映出数据的分布情况，但是相比较来说，箱线图在处理大数据方面更有优势，异常点信息十分直观。

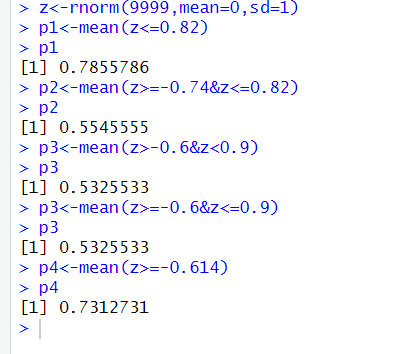
设Z〜N（0,1）。请使用R模拟(通过产生的随机数求得) 来计算以下概率。

a. P(Z ≤ 0.82)

b. P(-0.74 ≤ Z ≤ 0.82)

c. P(-0.6 ≤ Z ≤0.9)

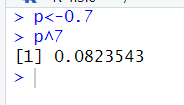
d. P(-0.614 ≤ Z)



1. Jack是一名猎人，他每天出门成功捕捉到猎物的概率是0.7。假设他每天捕猎成功的概率是相互独立的，我们统计他一周的捕猎情况，请问Jack一周七天全部捕猎成功的概率为0.0823543吗？请首先列出手动计算过程，并列出R语言的求解过程。

手动计算：0.7^7

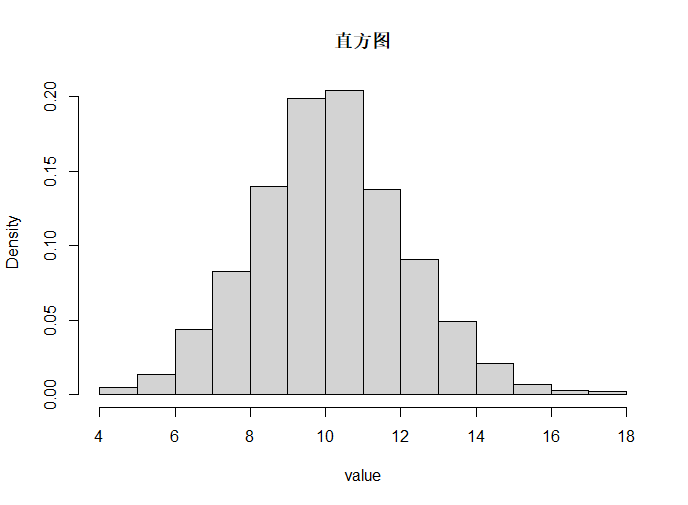
R语言求解：

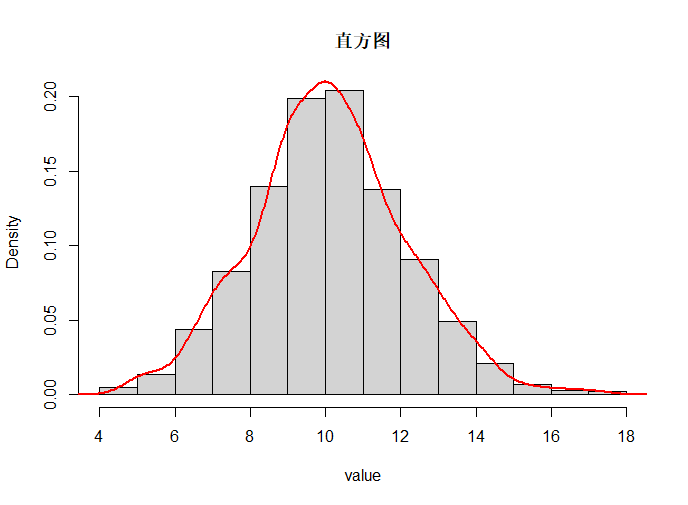


全部捕猎成功的概率是0.0823543

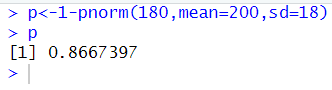
1. 用命令rnorm( )命令产生1000个均值为10, 方差为4的正态分布随机数,

用直方图呈现数据的分布并添加核密度曲线。





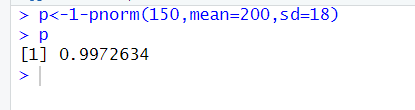
1. 已知从某一批材料中任取一件时，取得这件材料的强度服从.请用R语言计算下列题目。
   1. 计算取得的这件材料强度不低于180的概率。



* 1. 如多所用的这批材料要求以99%的概率保证强度不低于150，问这批材料是否符合这个要求。

假设：这批材料强度低于150

备择假设：这批材料的保证强度不低于150



由于强度不低于150的概率大于0.99故拒绝原假设。这批材料符合要求。