title: "<9e><8c><94> <94><9f><89><81><80><8d><95> <9f><88><86><9e><90>" author: "<83><91><8f><98> <8e>" date: "2019/11/9" output: pdf_document: default html_document: default

实验五 生物信息简单统计分析

实验说明

- 实验类型：综合性
- 实验目的：
    1. 熟悉R 语言进行简单统计分析。
    2. 熟悉R 语言格式化输出统计分析结果。
- 实验内容：

1. 使用t.test 函数进行t 检验。
2. 使用aov 函数进行方差分析。
3. 使用lm 函数进行回归分析。
4. 使用summary 函数格式化输出分析结果。
5. 使用dplyr包进行数据整理与分析
6. rep和seq命令

实验项目

1. rep和seq

```
# seq: 之前看到的1：10实际上就是一个序列，用seq函数可以写成seq(from=1,to=10,by=1)。
# 补全以下代码，输出1到20的偶数
seq(from=__,to=__,by=__)

# rep: rep(重复的元素,重复次数)
# 进行统计分析前，一般要对数据进行整理。比如下面的一个dataframe
set.seed(12)
df = data.frame(g1 = round(rnorm(20,3,1),2),g2=round(rnorm(20,5,2),2))
# 这个dataframe有g1和g2两个变量，现在这种形式是没有办法做t检验的，首先你需要将这个数据框整理成group变量和value变量：
# 完成这种变形的办法有很多，这里使用比较原始的rep+rbind方法：
# 使用df的g1列生成dataframe:df1，另外加上一个group列，内容全部等于"group1"
df1 = data.frame(value = df$__,group=rep(__))
# 使用df的g2列生成dataframe：df2，另外加上一个group列，内容全部等于group2
df2 = __
# 使用rbind将df1和df2合并
df_reshaped = rbind(__,__)
df_reshaped
```

2. 使用dplyr和tidyr进行数据整理

到现在我们已经可以:

- 读取外部数据(如read.csv), 写入数据(如write.csv)
- 筛选记录(基于筛选向量), 选择记录(基于名称, 序号), 排序(如order)
- 对变量进行修改(sapply等)
- 将两个表合并(如rbind和cbind)
- 基于分组进行统计(如aggregate, aggregate除了设置formula以外,还可以使用by参数来设置分组依据)

还有很多操作都可以使用r的基本函数实现. 但是目前更加流行的是tidyverse系列的各种数据处理包. 尤其我们接下来还要联系ggplot, 所以适当的学习一下dplyr以及tidyr还是很有帮助的.

参考学习资料

- dplyr 简要说明
- 使用tidyr和dplyr整理数据

3. pcr表格

pcr的一种比较常用的相对量化是用2^-ddCt法。首先每一个样本的目标基因（比如数据中的cyclinD和cyclinE）的Ct减去内参（如数据中的actin）的Ct获得dCt。之后分别求出实验组dct平均值，标准差，对照组dct平均值，标准差。实验组减去对照组的dct就可以得到ddct。之后使用2的-ddCt法。首先每一个样本的目标基因（比如数据中的cyclinD和cyclinE）的Ct减去内参（如数据中的actin）的Ct获得dCt。之后分别求出实验组dct平均值，标准差，对照组dct平均值，标准差。实验组减去对照组的dct就可以得到ddct。之后使用2的-ddct次方就可以计算出相对表达差异量（relative fold change）

3.1 前次实验我们使用for loop和筛选工具做成了join的一个naive implementation。但是对于这种表和表的合并，使用现有的工具是最好的选择。使用left_join将actin表的Ct列加入cyclin表中

```
df_cyclin = read_excel('exp04/pcr.xlsx',sheet = 'cyclin')
df_actin = read_excel('exp04/pcr.xlsx',sheet ='actin')
# actin表只需要Ct列，另外也需要sample_id列作为匹配依据
df_actin_Ct = df_actin %>% select(__,__)
# 使用left_join将两个表合并。请注意cyclin表中样本id的表名称是sampleid，而actin表中则是sample_id
# 请不要修改源数据表格，而是在程序中通过命名向量的方式（named vector）解决这个问题。详细请参考?left_join说明
# 另外，默认的x和y后缀不是很好，最好手动设置同名列的后缀（Ct列在两个表里肯定是同名的）
df_cyclin_joined = left_join(df_cyclin,df_actin_Ct,by=(__),suffix=c(__,__))
```

3.2 目前Sample Name列实际上包含了三的内容：treatment(F还是control），over_inhibit（over还是inhibit）和days（3d，7d）。这不符合tidy data原则，也会给分析造成麻烦。请使用separate函数将该列拆分成两列。具体方法参见?separate。sep参数涉及正则表达式，这里就不作考察了。

```
df_cyclin_joined = df_cyclin_joined %>%
separate(col=__,into=c(__,__),sep='[\\-,]')
```

3.3 接下来需要将各行的Ct-cyclin减去Ct-actin获得dCt

```
df_cyclin_joined = df_cyclin_joined %>% mutate(dCt = __)
```

3.4 然后，按照刚才生成的treatment，over_inhibit, days和本来就有的Target Name进行分组，统计各组的dCt的平均值与标准差

请注意，变量名中间的空格会引起歧义，可以在变量名两边加上``解决这个问题。比如Target Name应该写成`Target Name`

```
df_cyclin_stat = df_cyclin_joined %>% group_by(__,__,__,__)  %>% summarise(mean =
__(__),sd = __(__))
```

3.5 接下来计算同个over_inhibit，同一个days，同一个Target Name下，F组的mean值减去control组的mean值的差。这个过程可以先用filter将`df_cyclin_stat`转成两个子表，然后通过`left_join`来合并

```
# 通过filter获取treatment为control的行
dfctrl = df_cyclin_stat %>% filter(__)
# 通过filter获取treatment为F的行
df_f = df_cyclin_stat %>% filter(__)
# 依据over_inhibit，days和Target Name这三列对两个表进行合并
df_cyclin_stat_F_ctrl_joined = left_join(__,__,by=__)
# 最后，计算同样的over_inhibit，days和Target Name组内，F组的mean减去control组的mean的
值
df_cyclin_stat_F_ctrl_joined = df_cyclin_stat_F_ctrl_joined %>% mutate(ddCt = __)
# 另外，还可以计算2^-ddCt的值（选做）
```

4. 方差分析

补全以下代码，解读结果

```
c1 = rnorm(10,2,1)
c2 = rnorm(11,3,2)
c3 = rnorm(16,4,1)
df = data.frame(grp = c(rep('a',10),rep('b',11),rep('c',16)),val=c(c1,c2,c3))
fit = aov(__~__,data=__)
summary(fit)
```

5. lm，线性拟合

课堂演示

6. 在纸质实验报告中总结陈述如何使用dplyr筛选记录，对变量进行修改，将两个表合并，分组和基于分组进行统计