

## R(2)

1.某医院分别用化学疗法和化疗结合放射治疗卵巢癌肿患者，结果如下，问两种疗法有无差别？

组别	有效	无效
化学疗法	18	25
化放结合疗法	35	10

输入如下代码：

```
all.data= matrix(c(18,35,25,10),nrow=2,ncol=2)
print(all.data)
chisq.test(all.data)
```

结果如下，可以看到p值较小，说明两种疗法有显著差别。

```
PS C:\Users\czh> Rscript "c:\Users\czh\Desktop\Code\try.r"
      [,1] [,2]
[1,]    18    25
[2,]    35    10

      Pearson's Chi-squared test with Yates' continuity correction

data:  all.data
X-squared = 10.39, df = 1, p-value = 0.001267
```

2.为检测某耐除草剂转基因玉米对除草剂的耐受性，在田间分别开展了未喷施除草剂（CK），喷施1倍除草剂（P1），喷施2倍除草剂（P2），喷施4倍除草剂（P4）四种处理，每种处理设计了3个重复，四周后对株高进行测定，每个重复测定了30个单株株高，试根据测定数据判断该转基因玉米对除草剂的耐受能力，并绘制箱线图进行展示。（见herbicide tolerance.txt文件）


注：文本文件内第三列列名为 plant height，但通过用R语言打印该表可以发现其列名为 plant.height !!!

首先读入该表并打印：

```
wheat_data=read.table("C:\\Users\\czh\\Desktop\\herbicide tolerance.txt",sep="\t",header=T)
print(wheat_data)
```

```
PS C:\Users\czh> Rscript "c:\Users\czh\Desktop\Code\try.r"
      Treat Block plant.height
1         CK     I          154
2         CK     I          180
3         CK     I          190
4         CK     I          180
5         CK     I          192
6         CK     I          176
7         CK     I          180
8         CK     I          171
9         CK     I          178
10        CK     I          182
```

注意到第三列名为 `plant.height` 而不是文本文件内的 `plant height` :



```
herbicide tolerance.txt - 记事本
文件(E) 编辑(E) 格式(O) 查看(V) 帮助(H)
Treat      Block      plant height
```

然后进行两因素方差分析:

```
rbd= aov(plant.height ~ Treat + Block, data=wheat_data)
summary(rbd)
```

```
      Df Sum Sq Mean Sq F value    Pr(>F)
Treat   3  20011    6670   21.032 1.47e-12 ***
Block   2    665     333    1.049   0.351
Residuals 354 112275     317
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

计算每种处理的均值:

```
tapply(wheat_data$plant.height, wheat_data$Treat, mean)
```

```
      CK      P1      P2      P4
178.7556 177.2556 168.4889 160.1889
```

进行多重比较:

```
pairwise.t.test(wheat_data$plant.height, wheat_data$Treat)
```

```
Pairwise comparisons using t tests with pooled SD

data:  wheat_data$plant.height and wheat_data$Treat

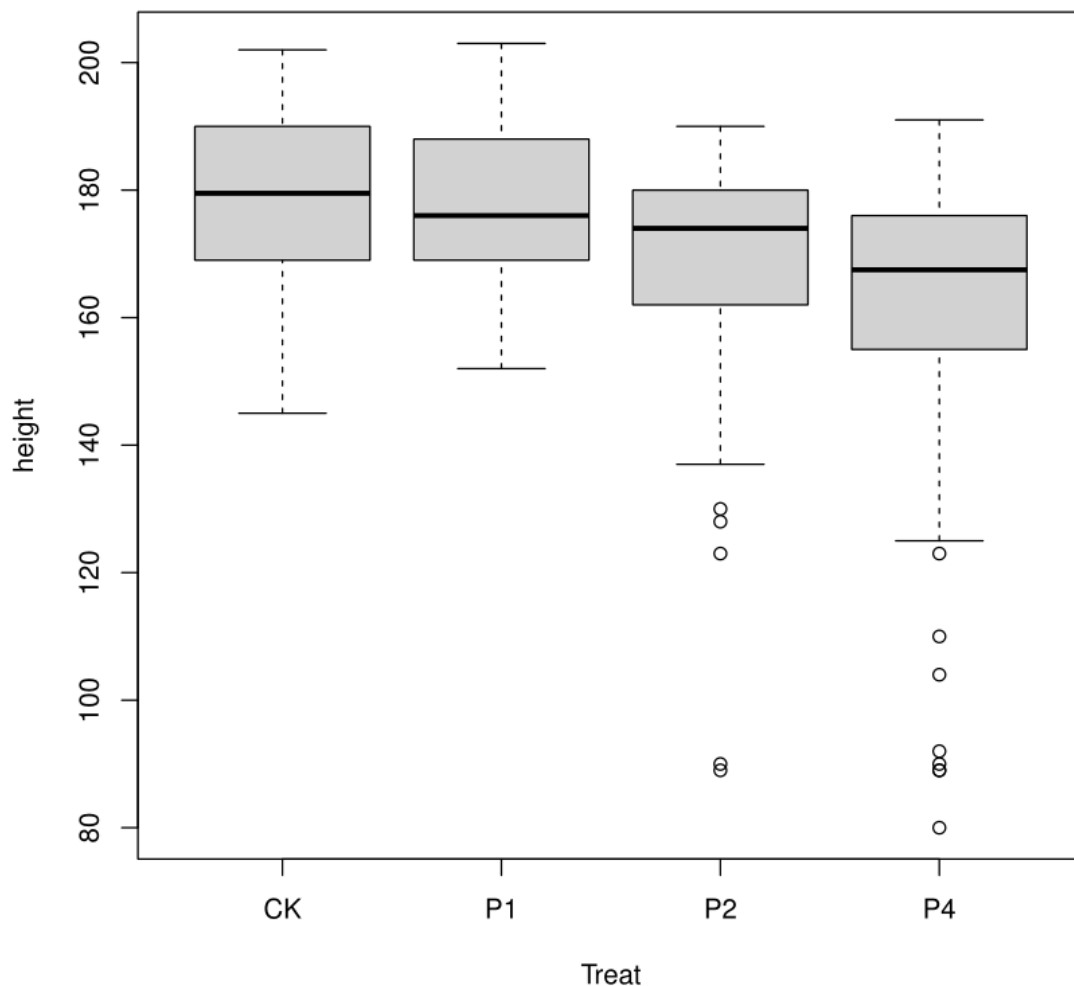
      CK      P1      P2
P1 0.57247 -        -
P2 0.00052 0.00317 -
P4 8.0e-11 2.1e-09 0.00384

P value adjustment method: holm
```

绘制箱线图, 由于原数据第一列为字符, 但绘制箱线图要求数据为数字, 故需要按照分组的方式来绘制:

```
wheat_data=read.table("C:\\Users\\czt\\Desktop\\herbicide tolerance.txt", sep="\t", header=T)
x<-c(wheat_data[,3])
f <- factor(rep(c("CK", "P1", "P2", "P4"), each=90))
data <- data.frame(x, f)
boxplot(x~f, data, xlab="Treat", ylab="height")
```

上述代码的意思是将360行数据分为4组, 每组90个 (按照 `Treat` 的不同情况进行分类)。只取出第三列 `plant.height` (因为第一列不是数字无法取出), 这样重新生成数据框后, 就会自动按照顺序画出4组箱线图:



3.文件plant\_ear\_height.txt为100株玉米的株高和穗位高的观测值。读入数据后，试做如下分析：

- (1) 试测验穗位高与株高是否相关，计算出相关系数和决定系数。
- (2) 试测验穗位高与株高是否存在直线回归关系，若存在，计算出直线回归方程，说明株高能解释穗位高多少的表型变异，并绘制两变量散点图。

代码如下，其中z代表穗位高，x代表株高：

```
y=read.table("C:\\Users\\czh\\Desktop\\plant_ear_height.txt",sep="\t",header=T)
x=y[,1]
z=y[,2]
print(x)
print(z)
cor.test(x,z)
test=lm(z~x)
summary(test)
plot(z~x)
abline(test)
```

可以看到输出的相关系数 $cor = 0.6303652$ ，决定系数为 $cor^2 = 0.3973603$

### Pearson's product-moment correlation

```
data: x and z
t = 8.0385, df = 98, p-value = 2.099e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4952689 0.7356933
sample estimates:
      cor
0.6303652
```

从上图可以看到p值很小，故两变量存在线性回归关系：

```
Call:
lm(formula = z ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-21.6430  -7.3771  -0.7735   6.8047  30.7417

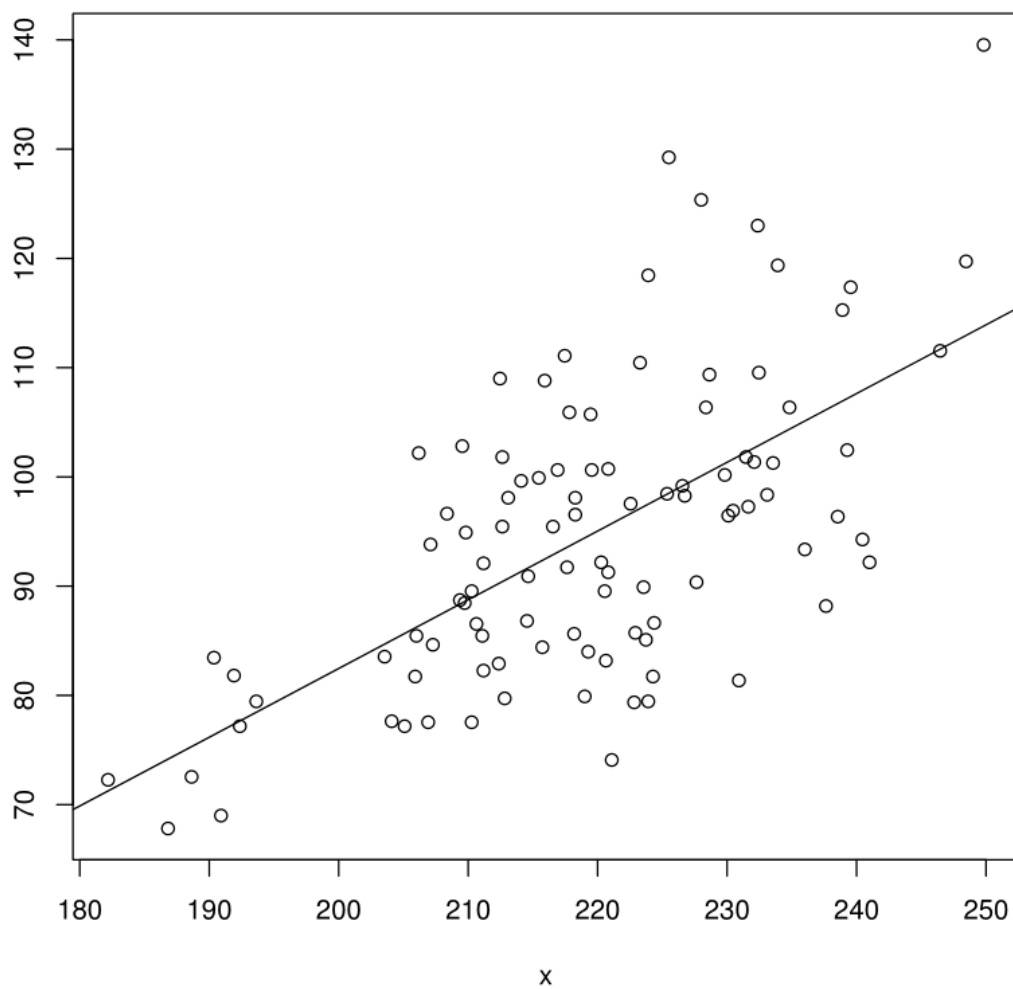
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.38507    17.18701  -2.524   0.0132 *
x             0.62924     0.07828   8.039 2.1e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.67 on 98 degrees of freedom
Multiple R-squared:  0.3974,    Adjusted R-squared:  0.3912
F-statistic: 64.62 on 1 and 98 DF,  p-value: 2.099e-12
```

由上图可以看到线性回归方程为 $z = 0.62924x - 43.38507$ ，且株高能解释穗位高39.74%的变异。

绘制的两个变量的散点图及回归图如下：

z



x