田纳西河流中鱼的污染问题

同济大学数学科学学院数学建模讲座　　王勇智

写在前面的话

* 本讲目的不是就这个问题——田纳西河流中鱼的污染问题得到一个漂亮的结果，而是展示在数据分析中我们可以采用的统计方法以及应该怎样分析问题。

内容

* 1、问题
* 2、数据集的认识
* 3、缺失值的处理
* 4、异常值的处理
* 5、数据集的描述
* 6、观察数据、发现问题
* 7、分析问题、解决问题

1、问题

化学工厂以及制造工厂向附近的河流和溪流中排放有毒的废物。一种称作DDT的污染物质对鱼有害，对人类也有间接性的影响。

若一条鱼中的含量的限量为百万分之五（ppm），若食用了超过这个限量的鱼，就认为对人体有潜在的危害。现在检测生活在田纳西河流（位于亚拉巴马）及其支流中的鱼类的DDT含量。得到数据集如下：

2、数据集的认识

- 田纳西河及三条支流分别用TRM，FCM，LCM，SCM表示，
- 鱼称重（单位：g）
- 鱼体长（单位： cm）
- DDT（单位：百万分之一）
- 鱼的种类CCATFISH，SMBUFFALO，LMBASS

2、数据集的认识

RIVER	MILE	SPECIES	LENGTH	WEIGHT	DDT
FCM	5	CCATFISH	42.5	732	10
FCM	5	CCATFISH	44	795	16
FCM	5	CCATFISH	41.5	547	23
FCM	5	CCATFISH	39	465	21
FCM	5	CCATFISH	50.5	1252	50
FCM	5	CCATFISH	52	1255	150
LCM	3	CCATFISH	40.5	741	28
LCM	3	CCATFISH	48	1151	7.7
LCM	3	CCATFISH	48	1186	2
LCM	3	CCATFISH	43.5	754	19
LCM	3	CCATFISH	40.5	679	16
LCM	3	CCATFISH	47.5	985	5.4
SCM	1	CCATFISH	44.5	1133	2.6
SCM	1	CCATFISH	46	1139	3.1
SCM	1	CCATFISH	48	1186	3.5
SCM	1	CCATFISH	45	984	9.1
SCM	1	CCATFISH	43	965	7.8
SCM	1	CCATFISH	45	1084	4.1
TRM	275	CCATFISH	48	986	8.4
TRM	275	CCATFISH	45	1023	15
TRM	275	CCATFISH	49	1266	25
TRM	275	CCATFISH	50	1086	5.6
TRM	275	CCATFISH	46	1044	4.6
TRM	275	CCATFISH	52	1770	8.2
TRM	280	CCATFISH	48	1048	6.1
TRM	280	CCATFISH	51	1641	13

```
> mean(DDT$LENGTH)
[1] 42.8125
```

数据集的变量

```
> names(DDT)
[1] "RIVER"   "MILE"    "SPECIES" "LENGTH"  "WEIGHT"  "DDT"
```

数据集的内部结构

```
> str(DDT)
'data.frame':   144 obs. of  6 variables:
 $ RIVER  : Factor w/ 4 levels "FCM","LCM","SCM",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ MILE   : int  5 5 5 5 5 5 3 3 3 3 ...
 $ SPECIES: Factor w/ 3 levels "CCATFISH","LMBASS",..: 1 1 1 1 1 1 1 1 1 1 .$
 $ LENGTH : num  42.5 44 41.5 39 50.5 52 40.5 48 48 43.5 ...
 $ WEIGHT : int  732 795 547 465 1252 1255 741 1151 1186 754 ...
 $ DDT    : num  10 16 23 21 50 150 28 7.7 2 19 ...
```

3、缺失值的处理

```
> ND<-is.na(DDT)
> colSums(ND,na.rm=TRUE)
  RIVER    MILE SPECIES  LENGTH  WEIGHT     DDT
      0       0       0       0       0       0
> nDDT<-DDT[complete.cases(DDT), ]
> nDDT
    RIVER MILE    SPECIES LENGTH WEIGHT     DDT
1     FCM    5  CCATFISH   42.5    732   10.00
2     FCM    5  CCATFISH   44.0    795   16.00
3     FCM    5  CCATFISH   41.5    547   23.00
```

4、异常值的处理

DDT含量是否存在异常值？怎样检验？

（1）计算z得分

（2）用盒子图

4、异常值的处理

（1）计算z得分

```
> DDTYC<-DDT[((DDT$DDT-mean(DDT$DDT))/sd(DDT$DDT))>=2, ]
> DDTYC
    RIVER MILE  SPECIES LENGTH WEIGHT  DDT
105   TRM  320 CCATFISH   49.5   1255  360
115   TRM  325 CCATFISH   46.0    863 1100
> DDTYC1<-DDT[((DDT$DDT-mean(DDT$DDT))/sd(DDT$DDT))>=1, ]
> DDTYC1
    RIVER MILE  SPECIES LENGTH WEIGHT  DDT
6     FCM    5 CCATFISH   52.0   1255  150
105   TRM  320 CCATFISH   49.5   1255  360
106   TRM  320 CCATFISH   47.0   1152  130
115   TRM  325 CCATFISH   46.0    863 1100
123   TRM  330 CCATFISH   51.5   1229  140
130   TRM  340 CCATFISH   45.0    911  180
```

4、异常值的处理

（2）用盒子图

```
> summary(DDT$DDT)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    0.110     3.375     7.150    24.355    13.000 1100.000
> sort(DDT$DDT,index.return = TRUE)
$x
  [1]    0.11    0.18    0.22    0.25    0.30    0.34    0.35    0.43    0.45
 [10]    0.48    0.58    0.74    0.80    0.99    1.20    1.30    1.40    1.50
 [19]    1.90    2.00    2.00    2.00    2.20    2.30    2.40    2.50    2.50
 [28]    2.60    2.80    2.80    2.80    3.00    3.00    3.10    3.30    3.30
 [37]    3.40    3.50    3.50    3.70    3.80    3.90    4.00    4.10    4.10
 [46]    4.20    4.20    4.30    4.50    4.60    4.70    4.70    4.80    4.80
 [55]    5.10    5.10    5.20    5.40    5.50    5.60    5.60    5.70    6.00
 [64]    6.00    6.00    6.00    6.10    6.10    6.60    6.80    6.80    7.10
 [73]    7.20    7.30    7.40    7.50    7.70    7.80    8.00    8.20    8.40
 [82]    8.70    8.80    9.10    9.10    9.30    9.40    9.70    9.90   10.00
 [91]   10.00   10.00   10.00   10.00   11.00   11.00   12.00   12.00   12.00
[100]   12.00   12.00   12.00   12.00   12.00   13.00   13.00   13.00   13.00
[109]   13.00   13.00   14.00   15.00   15.00   16.00   16.00   17.00   17.00
[118]   18.00   19.00   19.00   21.00   21.00   21.00   22.00   22.00   22.00
[127]   23.00   25.00   27.00   28.00   31.00   33.00   44.00   48.00   50.00
[136]   57.00   61.00   96.00  130.00  140.00  150.00  180.00  360.00 1100.00

$ix
  [1]  46  44  47 138 127  45 143 114 136  43 139 119  48 135 128  73  91 131
 [19] 144   9 126 140 141  34 132  35 137  13  42  93 118  33  69  14  96  97
 [37]  55  15  52  98  88 134  77  18 117  32 124 133  31  23  63  86  74  94
 [55]  75  76  65  12  29  22  57  95  27  39  62  87  25  92  28  36 100 129
 [73]  38  71 142  68   8  17  60  24  19  49 102  16 122  53 116  84  99   1
 [91]  40  78  81 112  30  82  41  58  61  79  85  90 109 125  26  51  56  70
[109] 101 107 120  20  72   2  11  83  89  67  10  37   4  54  59  50  80 121
[127]   3  21  66   7  64 110 113 111   5 103 108 104 102  17  76  45  55  12
```

4、异常值的处理

（2）用盒子图

```
> Q1<-quantile(DDT$DDT,0.25)
> Q1
  25%
3.375
> Q2<-median(DDT$DDT)
> Q2
[1] 7.15
> Q3<-quantile(DDT$DDT,0.75)
> Q3
75%
 13
> y<-median(DDT$DDT)+1.5*(Q3-Q1)
> y
    75%
21.5875
> DDTY<-DDT[DDT$DDT>y, ]
> DDTY
    RIVER MILE   SPECIES LENGTH WEIGHT  DDT
3     FCM    5  CCATFISH   41.5    547   23
5     FCM    5  CCATFISH   50.5   1252   50
6     FCM    5  CCATFISH   52.0   1255  150
7     LCM    3  CCATFISH   40.5    741   28
```

```
> DDTY
    RIVER MILE   SPECIES LENGTH WEIGHT  DDT
5     FCM    5  CCATFISH   50.5   1252   50
6     FCM    5  CCATFISH   52.0   1255  150
7     LCM    3  CCATFISH   40.5    741   28
64    TRM  295  CCATFISH   49.5   1084   31
103   TRM  320  CCATFISH   47.5    983   57
104   TRM  320  CCATFISH   51.5   1251   96
105   TRM  320  CCATFISH   49.5   1255  360
106   TRM  320  CCATFISH   47.0   1152  130
108   TRM  320  CCATFISH   47.0   1118   61
110   TRM  320 SMBUFFALO   34.5   1178   33
111   TRM  320 SMBUFFALO   44.5   1492   48
113   TRM  320 SMBUFFALO   46.0   1473   44
115   TRM  325  CCATFISH   46.0    863 1100
123   TRM  330  CCATFISH   51.5   1229  140
130   TRM  340  CCATFISH   45.0    911  180
```

4、异常值的处理

（2）用盒子图

Z<-boxplot.stats(DDT$DDT,coef=1.5,do.conf=TRUE,do.out=TRUE)
sub=which(DDT$DDT%in%Z$out)
DDTY<-DDT[sub,]； DDTZ<-DDT[-sub,]

```
> Z<-boxplot.stats(DDT$DDT,coef=1.5,do.conf=TRUE,do.out=TRUE)
> Z
$stats
[1]  0.11  3.35  7.15 13.00 27.00

$n
[1] 144

$conf
[1] 5.879417 8.420583

$out
 [1]   50  150   28   31   57   96  360  130   61   33   48   44 1100  140
[15]  180
```

4、异常值的处理

```
> DDTYY<-which(DDT$DDT%in%Z$out)
> DDTY
    RIVER MILE   SPECIES LENGTH WEIGHT  DDT
3     FCM    5  CCATFISH   41.5    547   23
5     FCM    5  CCATFISH   50.5   1252   50
6     FCM    5  CCATFISH   52.0   1255  150
7     LCM    3  CCATFISH   40.5    741   28
21    TRM  275  CCATFISH   49.0   1266   25
50    TRM  290  CCATFISH   44.0    886   22
64    TRM  295  CCATFISH   49.5   1084   31
66    TRM  295  CCATFISH   46.5    724   27
80    TRM  305  CCATFISH   51.0    353   22
103   TRM  320  CCATFISH   47.5    983   57
104   TRM  320  CCATFISH   51.5   1251   96
105   TRM  320  CCATFISH   49.5   1255  360
106   TRM  320  CCATFISH   47.0   1152  130
108   TRM  320  CCATFISH   47.0   1118   61
110   TRM  320 SMBUFFALO   34.5   1178   33
111   TRM  320 SMBUFFALO   44.5   1492   48
113   TRM  320 SMBUFFALO   46.0   1473   44
115   TRM  325  CCATFISH   46.0    863 1100
121   TRM  330  CCATFISH   32.0    556   22
123   TRM  330  CCATFISH   51.5   1229  140
> length(DDTY$MILE)
[1] 15
> sub
 [1]   5   6   7  64 103 104 105 106 108 110 111 113 115 123 130
```

```
> DDTY
    RIVER MILE   SPECIES LENGTH WEIGHT  DDT
5     FCM    5  CCATFISH   50.5   1252   50
6     FCM    5  CCATFISH   52.0   1255  150
7     LCM    3  CCATFISH   40.5    741   28
64    TRM  295  CCATFISH   49.5   1084   31
103   TRM  320  CCATFISH   47.5    983   57
104   TRM  320  CCATFISH   51.5   1251   96
105   TRM  320  CCATFISH   49.5   1255  360
106   TRM  320  CCATFISH   47.0   1152  130
108   TRM  320  CCATFISH   47.0   1118   61
110   TRM  320 SMBUFFALO   34.5   1178   33
111   TRM  320 SMBUFFALO   44.5   1492   48
113   TRM  320 SMBUFFALO   46.0   1473   44
115   TRM  325  CCATFISH   46.0    863 1100
123   TRM  330  CCATFISH   51.5   1229  140
130   TRM  340  CCATFISH   45.0    911  180
```

5、数据集的描述
——鱼的特征的描述

（1）数字特征的方法

（2）画图的方法

5、数据集的描述
——鱼的特征的描述

（1）数字特征的方法

中心趋势的度量：平均值、中位数、众数

波动的度量：极差、方差、标准差

相对位置的度量：百分位得分、z得分

检查异常值：盒子图和z得分

最大值、最小值

5、数据集的描述
——鱼的特征的描述

（2）画图的方法

定性数据的图形法：饼图、条形图等

定量数据的图形法：点图、茎叶图、直方图

由这张图分析得到什么？

* > plot(DDT)

多角度的观察数据集

```
> mean(DDT$LENGTH)
[1] 42.8125
```

所有鱼的平均身长

```
> mean(DDT$LENGTH[DDT$SPECIES=="CCATFISH"])
[1] 44.72917
```

品种为CCATFISH鱼的平均身长

```
> lapply(DDT[ ,4:6],FUN=mean)
$LENGTH
[1] 42.8125

$WEIGHT
[1] 1049.715

$DDT
[1] 24.355
```

```
> Z<-cbind(DDT$LENGTH,DDT$WEIGHT,DDT$DDT)
> colnames(Z)<-c("LENGTH","WEIGTH","DDT")
> summary(Z)
     LENGTH          WEIGTH            DDT
 Min.   :17.50   Min.   : 173.0   Min.   :   0.110
 1st Qu.:40.50   1st Qu.: 805.5   1st Qu.:   3.375
 Median :45.00   Median :1000.0   Median :   7.150
 Mean   :42.81   Mean   :1049.7   Mean   :  24.355
 3rd Qu.:47.50   3rd Qu.:1257.8   3rd Qu.:  13.000
 Max.   :52.00   Max.   :2302.0   Max.   :1100.000
```

多角度的观察数据集

```
> summary(DDT[ ,c(4,5,6)])
     LENGTH          WEIGHT            DDT
 Min.   :17.50   Min.   : 173.0   Min.   :   0.110
 1st Qu.:40.50   1st Qu.: 805.5   1st Qu.:   3.375
 Median :45.00   Median :1000.0   Median :   7.150
 Mean   :42.81   Mean   :1049.7   Mean   :  24.355
 3rd Qu.:47.50   3rd Qu.:1257.8   3rd Qu.:  13.080
 Max.   :52.00   Max.   :2302.0   Max.   :1100.000
```

DDT存在异常值

```
> table(DDT$MILE)

  1   3   5 275 280 285 290 295 300 305 310 315 320 325 330 340 345
  6   6   6   6  12  12  12   6  12   6  12   6  12   6   8  10   6
```

数据集中在
275-345km之间

```
> table(DDT$SPECIES)

 CCATFISH    LMBASS SMBUFFALO
       96        12        36
```

品种为
CCATFISH的鱼
最多

多角度的观察数据集

```
> table(DDT$SPECIES,DDT$RIVER)

            FCM LCM SCM TRM
  CCATFISH    6   6   6  78
  LMBASS      0   0   0  12
  SMBUFFALO   0   0   0  36
```

品种为CCATFISH且在田纳西河流的鱼最多

```
> DDT1<-DDT[DDT$SPECIES=="CCATFISH" & DDT$RIVER=="TRM", ]
> DDT1
    RIVER MILE  SPECIES LENGTH WEIGHT     DDT
19    TRM  275 CCATFISH   48.0    986    8.40
20    TRM  275 CCATFISH   45.0   1023   15.00
21    TRM  275 CCATFISH   49.0   1266   25.00
```

```
> mean(DDT1$WEIGHT)
[1] 996.7564
```

5、数据集的描述
——鱼的特征的描述

```
> DDTH1<-DDT[DDT$DDT<57, ]
> stem(DDTH1$DDT)

  The decimal point is at the |

   0 | 1223334455678023459
   2 | 00023455688800133455789
   4 | 01122356778811245667
   6 | 0000116881234578
   8 | 02478113479
  10 | 0000000
  12 | 00000000000000
  14 | 000
  16 | 0000
  18 | 000
  20 | 000
  22 | 0000
  24 | 0
  26 | 0
  28 | 0
  30 | 0
  32 | 0
  34 |
  36 |
  38 |
  40 |
  42 |
  44 | 0
  46 |
  48 | 0
  50 | 0
```

茎叶图说明？

6、观察数据、发现问题

```
> summary(DDT)
 RIVER          MILE             SPECIES       LENGTH          WEIGHT
 FCM:  6   Min.   :  1.0   CCATFISH :96   Min.   :17.50   Min.   : 173.0
 LCM:  6   1st Qu.:283.8   LMBASS   :12   1st Qu.:40.50   1st Qu.: 805.5
 SCM:  6   Median :300.0   SMBUFFALO:36   Median :45.00   Median :1000.0
 TRM:126   Mean   :268.6                  Mean   :42.81   Mean   :1049.7
           3rd Qu.:320.0                  3rd Qu.:47.50   3rd Qu.:1257.8
           Max.   :345.0                  Max.   :52.00   Max.   :2302.0
      DDT
 Min.   :   0.110
 1st Qu.:   3.375
 Median :   7.150
 Mean   :  24.355
 3rd Qu.:  13.000
 Max.   :1100.000
```

中位数大于5，受到了污染？

可能存在异常值怎样检验？

```
> cor(DDT[,4:6])
           LENGTH      WEIGHT         DDT
LENGTH 1.0000000  0.65461133  0.12610570
WEIGHT 0.6546113  1.00000000 -0.01190568
DDT    0.1261057 -0.01190568  1.00000000
```

身长和重量是否相关，怎样检验？

7、分析问题、解决问题

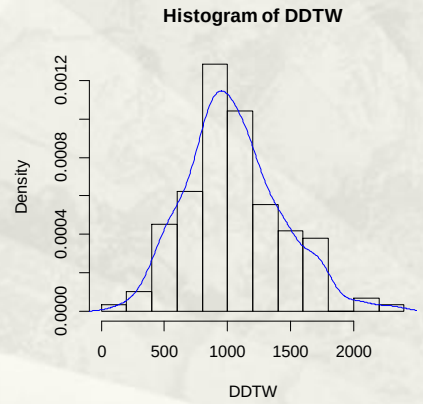
（1）问题：鱼的重量、体长服从什么分布？
画直方图
（2）问题：鱼的重量、体长是否服从正态分布？怎样检验？
画QQ图、KS-检验、正态性W检验、卡方拟合优度检
（3）问题：体重和身长之间有关系吗？
画散点图、求相关系数
（4）问题：怎样检验体重和身长之间的相关关系?
相关性检验
（5）不同种类的比例估计？
大样本下的估计
（6）不同种类的鱼的重量、身长的估计？
正态总体情形、非正态总体情形
（7）不同种类的鱼的重量、身长是否存在显著性差异？怎样检验？
方差分析，经典情形，非经典情形
（8）重量和身长能否建立回归方程？
回归系数显著性检验、方程显著性检验

7、分析问题、解决问题

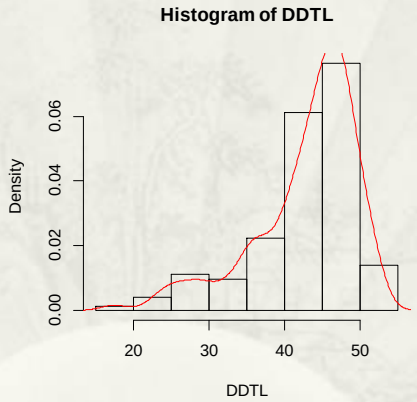
(1)问题：鱼的重量、体长服从什么分布？

* 画直方图

```
>  DDTW<-DDT$WEIGHT
> hist(DDTW, freq=FALSE)
> lines(density(DDTW),col="blue")
```




```
> hist(DDTL, freq=FALSE)
> lines(density(DDTL),col="red")
```

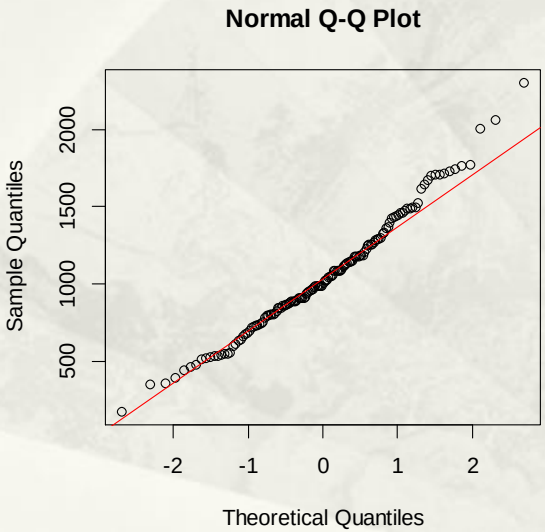


7、分析问题、解决问题

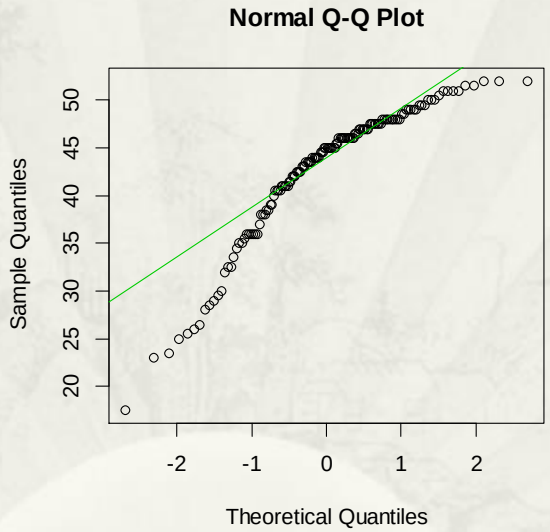
（2）问题：鱼的重量、体长是否服从正态分布？怎样检验？

* 画QQ图

```
> qqnorm(DDT$WEIGHT)
> qqline(DDT$WEIGHT,col=2)
```




```
> qqnorm(DDT$LENGTH)
> qqline(DDT$LENGTH,col=3)
```



重量可能服从正态分布，体长显然不服从正态分布

7、分析问题、解决问题

（2）问题：鱼的重量、体长是否服从正态分布？怎样检验？

* 正态性W检验(更为严谨的检验方法)
* KS-检验(更为严谨的检验方法)

```
> shapiro.test(DDTW)

        Shapiro-Wilk normality test

data:  DDTW
W = 0.9825, p-value = 0.06299
```

```
> shapiro.test(DDTL)

        Shapiro-Wilk normality test

data:  DDTL
W = 0.88524, p-value = 3.649e-09
```

```
> ks.test(DDTW,"pnorm",mean(DDTW),sd(DDTW))

        One-sample Kolmogorov-Smirnov test

data:  DDTW
D = 0.067034, p-value = 0.537
alternative hypothesis: two-sided
```

```
> ks.test(DDTL,"pnorm",mean(DDTL),sd(DDTL))

        One-sample Kolmogorov-Smirnov test

data:  DDTL
D = 0.1509, p-value = 0.002838
alternative hypothesis: two-sided
```

7、分析问题、解决问题

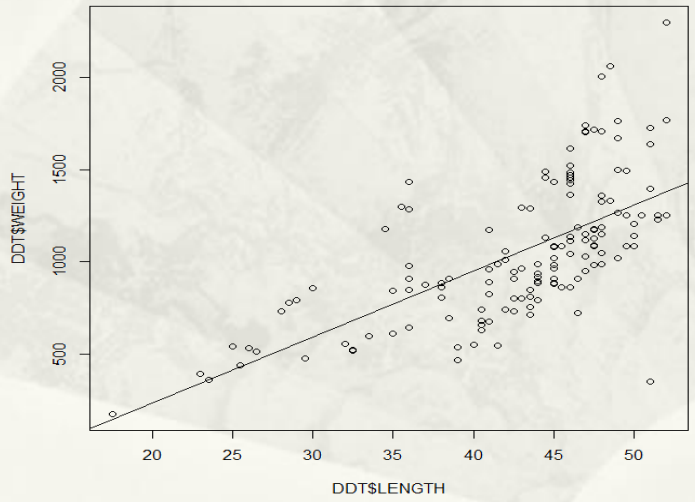
（2）问题：鱼的重量、体长是否服从正态分布？怎样检验？

* 还可以采用卡方拟合优度检验.....

7、分析问题、解决问题

（3）问题体重和身长之间有关系吗？

* 画散点图
* 求相关系数

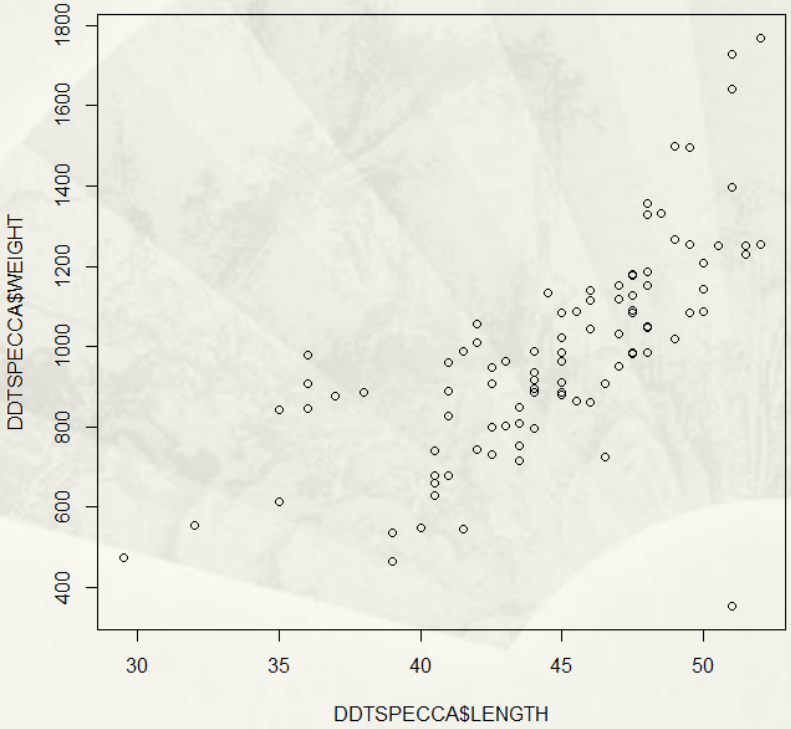



```
> cor(DDT[,c(4,5,6)])
                LENGTH     WEIGHT         DDT
LENGTH  1.0000000  0.65461133  0.12610570
WEIGHT  0.6546113  1.00000000  -0.01190568
DDT         0.1261057  -0.01190568   1.00000000
 abline(lm.wl)
```

7、分析问题、解决问题

种类为CCATFISH的重量和身长之间的关系

```
plot(DDTSPECCA$WEIGHT~DDTSPECCA$LENGTH)
```



7、分析问题、解决问题

（4）问题：怎样检验体重和身长之间的相关关系?
相关性检验

```
> cor.test(DDTL,DDTW)

        Pearson's product-moment correlation

data:  DDTL and DDTW
t = 10.319, df = 142, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5499201 0.7390507
sample estimates:
      cor
0.6546113
```

相关系数的点估计为：0.6546113
区间估计为：[0.5499201,0.7390507]
检验结果：相关系数显著不为零，即重量和体长相关。

7、分析问题、解决问题

（5）不同种类的鱼的比例估计？

略

7、分析问题、解决问题

（6）不同种类的鱼的重量、身长的估计？

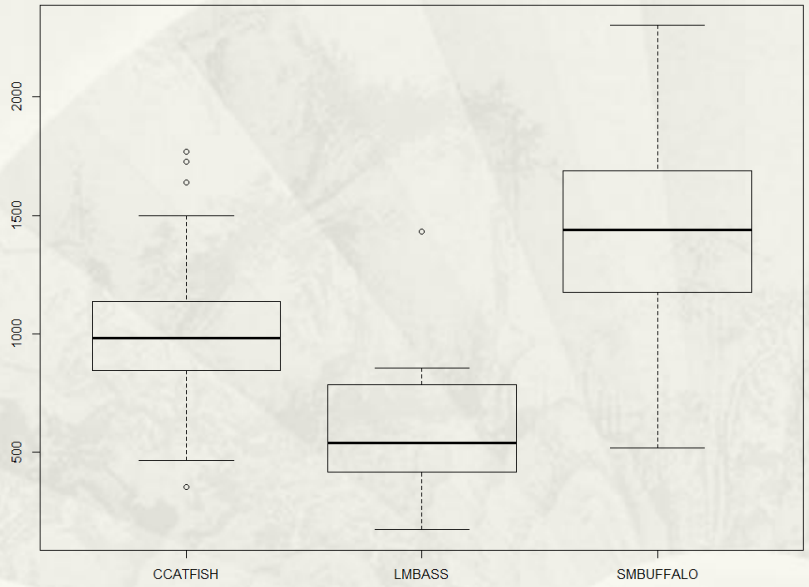
正态总体情形、非正态总体情形

7、分析问题、解决问题

（7）不同种类的鱼的重量、身长是否存在显著性差异？怎样检验？

方差分析

7、分析问题、解决问题



这张图告诉我们什么？

我们接着可以用什么方法检验分析？

```
boxplot(DDT$WEIGHT~DDT$SPECIES)
```

方差分析

进行方差分析的条件

1、误差服从正态分布，且相互独立

2、可加性，效应与随机误差可以叠加

3、方差齐性

正态性检验

```
> shapiro.test(DDT$WEIGHT[DDT$SPECIES=="CCATFISH"])

        Shapiro-Wilk normality test

data:  DDT$WEIGHT[DDT$SPECIES == "CCATFISH"]
W = 0.98162, p-value = 0.1982

> shapiro.test(DDT$WEIGHT[DDT$SPECIES=="LMBASS"])

        Shapiro-Wilk normality test

data:  DDT$WEIGHT[DDT$SPECIES == "LMBASS"]
W = 0.90268, p-value = 0.1718

> shapiro.test(DDT$WEIGHT[DDT$SPECIES=="SMBUFFALO"])

        Shapiro-Wilk normality test

data:  DDT$WEIGHT[DDT$SPECIES == "SMBUFFALO"]
W = 0.95588, p-value = 0.16
```

方差齐性检验

```
> attach(DDT)
The following object is masked _by_ .GlobalEnv:

    DDT

The following objects are masked from DDT (pos = 3):

    DDT, LENGTH, MILE, RIVER, SPECIES, WEIGHT

> bartlett.test(WEIGHT~SPECIES,data=DDT)

        Bartlett test of homogeneity of variances

data:  WEIGHT by SPECIES
Bartlett's K-squared = 14.644, df = 2, p-value = 0.0006608
```

拒绝原假设，怎么办？

采用Kruskal-Wallis检验

```
> kruskal.test(WEIGHT~SPECIES,data=DDT)

        Kruskal-Wallis rank sum test

data:  WEIGHT by SPECIES
Kruskal-Wallis chi-squared = 36.429, df = 2, p-value = 1.229e-08
```

拒绝原假设，可以认为不同种类鱼的重量存在显著性差异

7、分析问题、解决问题

（8）重量和身长能否建立回归方程？

回归系数显著性检验、方程显著性检验、回归诊断等

回归方程

```
> lm.wl<-lm(DDT$WEIGHT~1+DDT$LENGTH)
> summary(lm.wl)

Call:
lm(formula = DDT$WEIGHT ~ 1 + DDT$LENGTH)

Residuals:
    Min      1Q  Median      3Q     Max
-989.96 -189.45  -49.51  193.68  923.22

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -483.672    150.497  -3.214  0.00162 **
DDT$LENGTH    35.816      3.471  10.319  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 285.7 on 142 degrees of freedom
Multiple R-squared:  0.4285,    Adjusted R-squared:  0.4245
F-statistic: 106.5 on 1 and 142 DF,  p-value: < 2.2e-16
```

系数通过显著性检验

方程通过显著性检验

方程解释效果不是太好

回归方程

```
> lm2.sol<-lm(DDT$WEIGHT~DDT$LENGTH+I(DDT$LENGTH^2))
> summary(lm2.sol)

Call:
lm(formula = DDT$WEIGHT ~ DDT$LENGTH + I(DDT$LENGTH^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1065.02  -198.34   -69.67   182.84   826.91

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     654.8899   556.8450   1.176   0.2415
DDT$LENGTH      -26.3401    29.4958  -0.893   0.3734
I(DDT$LENGTH^2)   0.8099     0.3817   2.122   0.0356 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 282.2 on 141 degrees of freedom
Multiple R-squared:  0.4462,    Adjusted R-squared:  0.4383
F-statistic:  56.8 on 2 and 141 DF,  p-value: < 2.2e-16
```

常数项、一次项系数
没有通过显著性检验

方程解释效
果不是太好

回归方程

```
> lm3.sol<-lm(DDT$WEIGHT~I(DDT$LENGTH^2))
> summary(lm3.sol)

Call:
lm(formula = DDT$WEIGHT ~ I(DDT$LENGTH^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1036.56  -191.54   -63.58   200.54   863.90

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     163.68023   86.61239    1.89   0.0608 .
I(DDT$LENGTH^2)   0.47131    0.04434   10.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 282 on 142 degrees of freedom
Multiple R-squared:  0.4431,    Adjusted R-squared:  0.4391
F-statistic:   113 on 1 and 142 DF,  p-value: < 2.2e-16
```

常数项没有通过显著性检验

方程解释效果仍不是太好

回归方程

```
> lm4.sol<-lm(DDT$WEIGHT~I(DDT$LENGTH^2)+0)
> summary(lm4.sol)

Call:
lm(formula = DDT$WEIGHT ~ I(DDT$LENGTH^2) + 0)

Residuals:
     Min       1Q   Median       3Q      Max
-1082.66  -199.42   -54.27   197.53   809.48

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
I(DDT$LENGTH^2)  0.55197    0.01214   45.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 284.5 on 143 degrees of freedom
Multiple R-squared:  0.9353,    Adjusted R-squared:  0.9349
F-statistic:  2068 on 1 and 143 DF,  p-value: < 2.2e-16
```

回归系数通过显著性检验

方程通过显著性检验

方程解释效果很好

回归诊断

* 标准化残差有个别在[-2，2]之外，可以去掉异常值进一步找较优的回归方程。

7、分析问题、解决问题

* 还可以进一步分析寻找DDT和重量、身高、距离之间的回归方程。。。