

<p>GIB ARE THE NOS11TO17 1 WINS 1 FULINUMBERSTAKEALLSTAKES</p> <p>逢金色统吃</p> <p>大 一中一</p>	<div> </div> <p>INTHESE6SQUARES1WINS150 以上六門一中一百五十</p> <div> </div> <p>INTHESE3SQUARES1WINS8 以上三門一中八</p> <div> <p>3 2 1 3 3 2 2 1 1 6 5 4 6 6 5 5 4 4</p> <p>THIS SQUARE 1 WINS 24 一中二十四</p> </div> <div> </div> <p>INTHESE3SQUARES1WINS8 以上三門一中八</p>	<p>GIB ARE THE NOS11TO17 1 WINS 1 FULINUMBERSTAKEALLSTAKES</p> <p>逢金色统吃</p> <p>大 一中一</p>
<p>SMALL ARE THE NOS4TO10 1 WINS 1 FULINUMBERSTAKEALLSTAKES</p> <p>逢金色统吃</p> <p>小 一中一</p>	<p>17 16 15 14 13 12 11 10 9 8 7 6 5 4</p> <p>1 中 50 1 中 18 1 中 14 1 中 12 1 中 8 1 中 6 1 中 6 1 中 6 1 中 6 1 中 8 1 中 12 1 中 14 1 中 18 1 中 50</p> <p>INTHESE10 SQUARES 1 WINS 9 以上十門一中九</p> <div> </div>	<p>SMALL ARE THE NOS4TO10 1 WINS 1 FULINUMBERSTAKEALLSTAKES</p> <p>逢金色统吃</p> <p>小 一中一</p>

昵图网 www.nipic.com BY: pnhgg NO:20101213211843079656

数据分析与R语言 第3周

2012.5.19

随机事件与概率

- 随机试验与样本空间
- 随机事件
- 对立事件与互斥事件
- 随机事件的运算律
- 概率

■ 三个条件

- 1 可以重复进行
- 2 不能预知结果
- 3 知道所有可能的情况

■ 例子

- 1 投硬币，掷骰子
- 2 射击命中
- 3 身高、体重



样本空间

- 样本空间就是特定随机试验所有可能结果所组成的集合

- 例子

投硬币

掷骰子

身高体重

成绩

随机事件与必然事件

- 样本空间的子集称为随机事件
- 必然事件的例子
- 对立事件与互斥事件

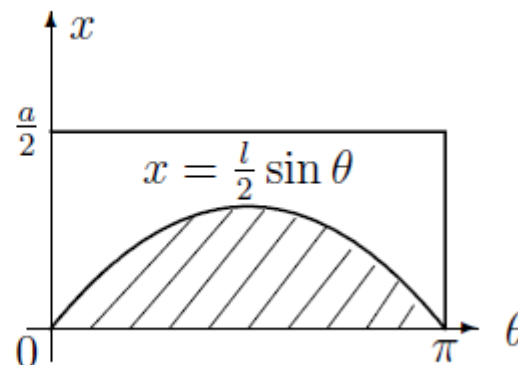
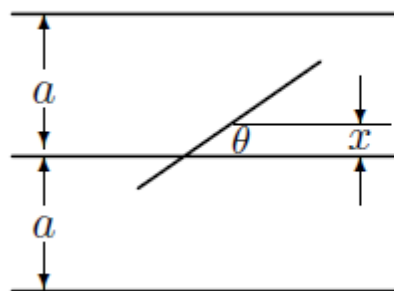
■ 直观的概率计算——古典概型

设随机事件 E 的样本空间中只有有限个样本点, 即 $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, 其中 n 为样本点总数. 每个样本点 $\omega_i (i = 1, 2, \dots, n)$ 出现是等可能的, 并且每次试验有且仅有一个样本点发生, 则称这类现象为古典概型 (classical probability model). 若事件 A 包含 m 个样本点, 则事件 A 的概率定义为

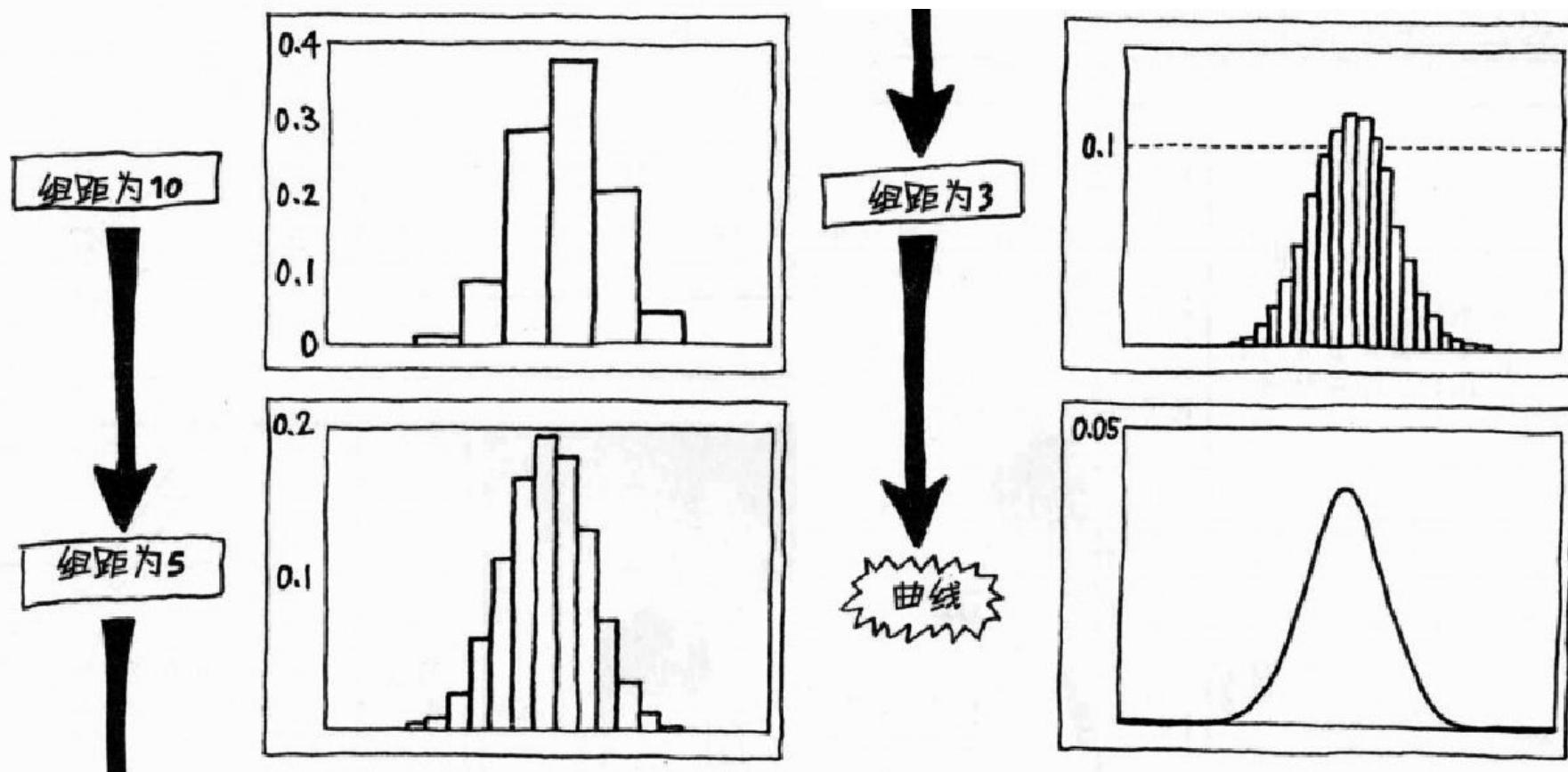
$$P(A) = \frac{m}{n} = \frac{\text{事件 } A \text{ 包含的基本事件数}}{\text{基本事件总数}}. \quad (1.12)$$

连续样本空间情形下的概率

■ 几何概型——布冯投针实验



连续样本空间情形下的概率：概率密度



2012.5.19

- 离散型分布：两点分布，二项分布，泊松分布
- 连续型分布：均匀分布，指数分布，正态分布
- 对于某一特定场景，其所符合的分布规律一般先验给出

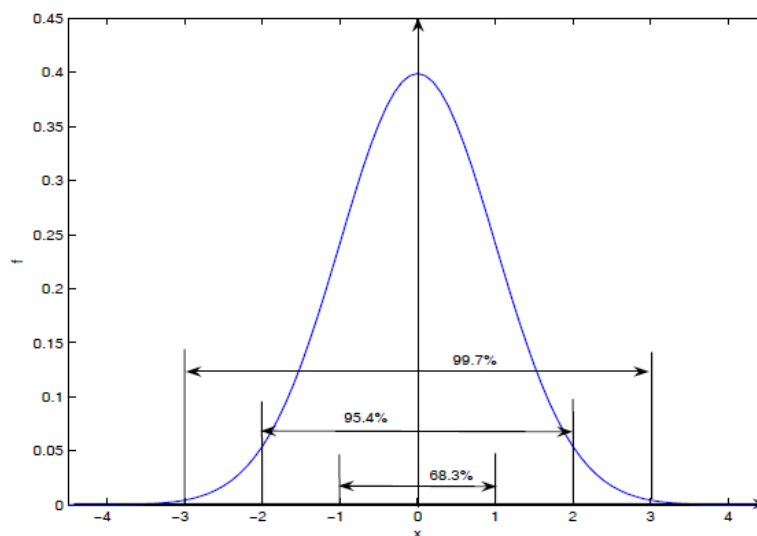
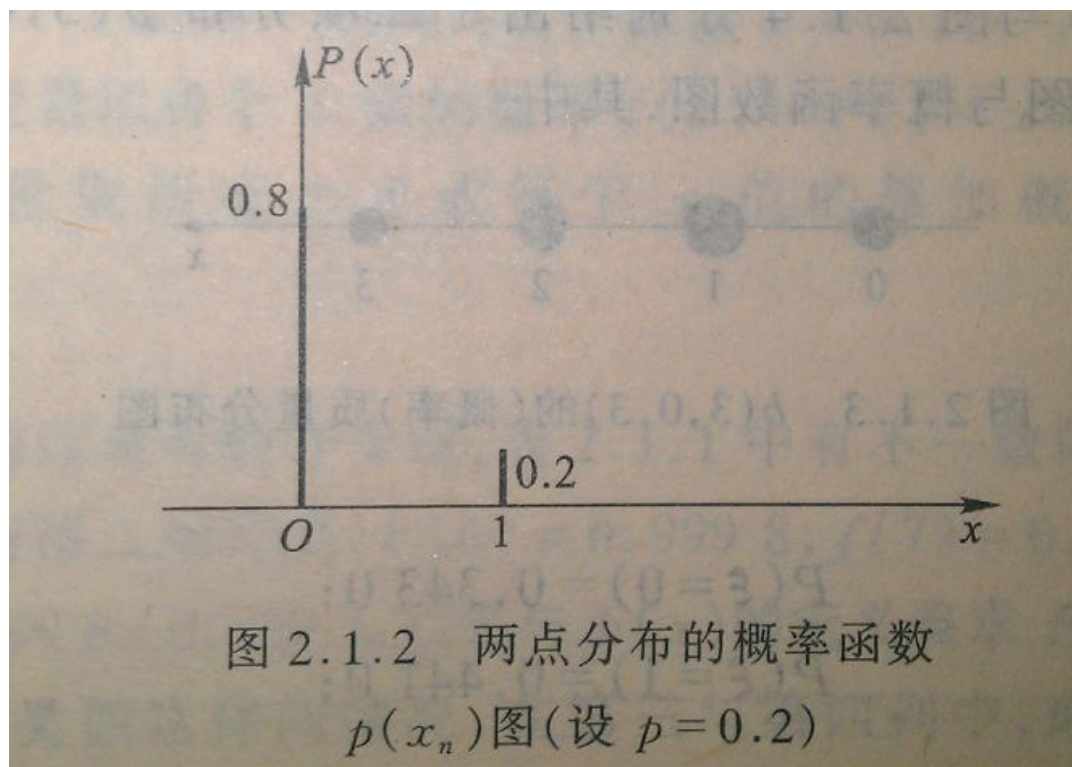


图 1.3: 标准正态分布和对应区间上积分 (面积) 的百分比



若随机变量 X 的分布律为

$$P\{X = k\} = C_n^k p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n, \quad (1.27)$$

则称 X 服从参数为 n, p 的二项分布 (binomial distribution), 记为 $X \sim B(n, p)$,

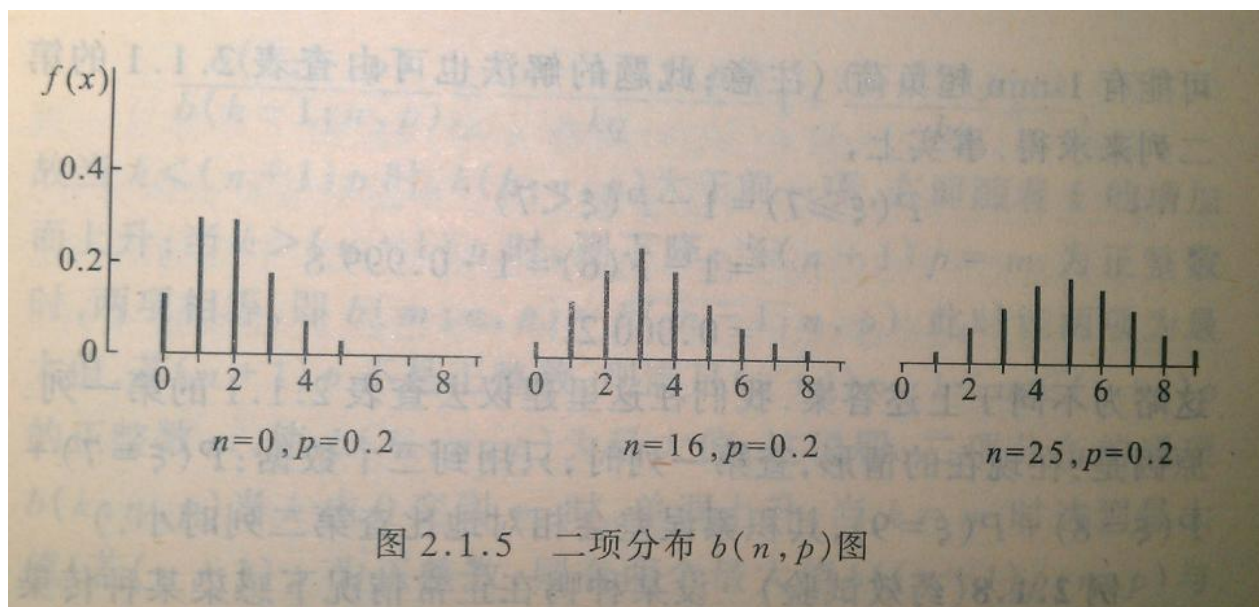


图 2.1.5 二项分布 $b(n, p)$ 图

若随机变量 X 的分布律为

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots,$$

则称 X 服从参数为 λ 的 Poisson (泊松) 分布 (Poisson distribution).

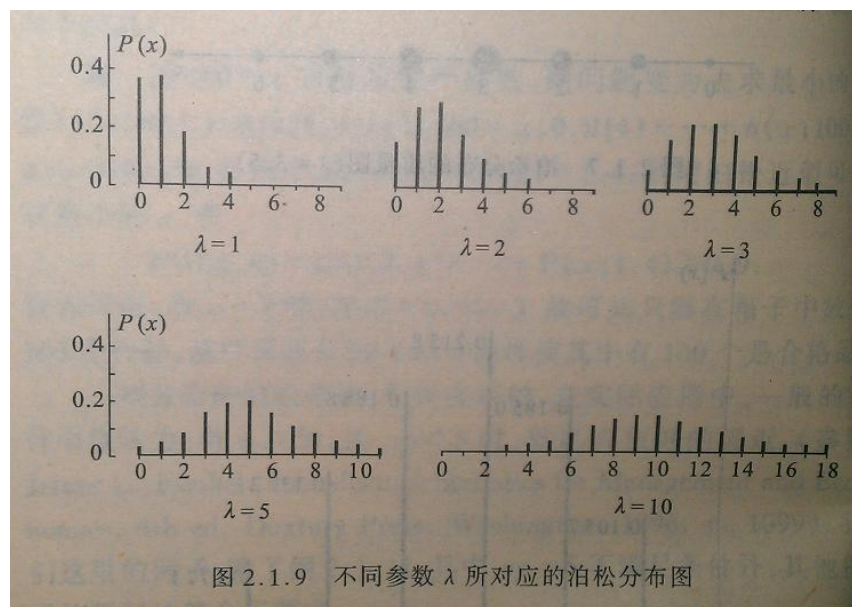
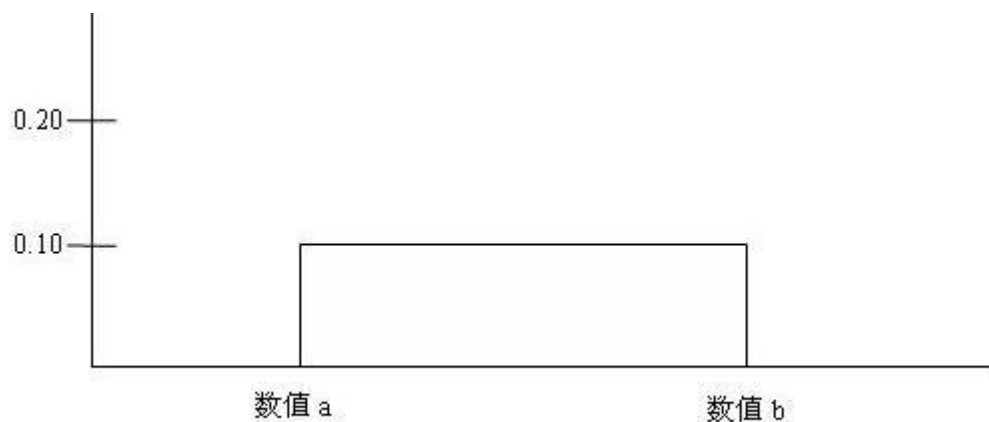


图 2.1.9 不同参数 λ 所对应的泊松分布图

若随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{其它}, \end{cases}$$

则称 X 服从区间 $[a, b]$ 上的均匀分布 (uniform distribution),



若随机变量 X 的概率密度函数为

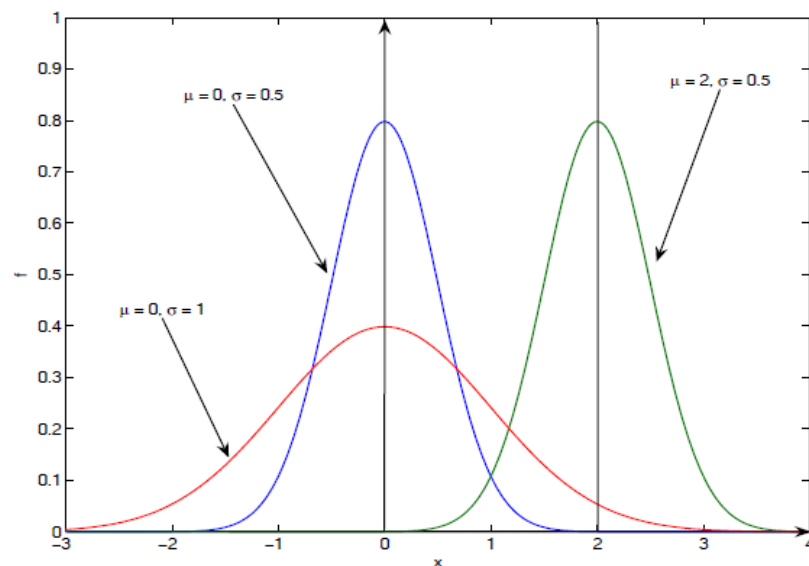
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

其中 $\lambda > 0$ 为常数, 则称 X 服从参数为 λ 的指数分布

若随机变量 X 的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}, \quad -\infty < x < +\infty, \quad (1.38)$$

其中 $\mu, \sigma (\sigma > 0)$ 是两个常数, 则称 X 服从参数为 μ, σ 的正态分布 (normal distribution), 也称为 Gauss 分布, 记作 $X \sim N(\mu, \sigma^2)$.



2012.5.19

R语言的各种分布函数

```

rnorm(n, mean=0, sd=1) 高斯(正态)
rexp(n, rate=1) 指数
rgamma(n, shape, scale=1)  $\gamma$  分布
rpois(n, lambda) Poisson 分布
rweibull(n, shape, scale=1) Weibull 分布
rcauchy(n, location=0, scale=1) Cauchy 分布
rbeta(n, shape1, shape2)  $\beta$  分布
rt(n, df)  $t$  分布
rf(n, df1, df2) F 分布
rchisq(n, df)  $\chi^2$  分布
rbinom(n, size, prob) 二项
rgeom(n, prob) 几何
rhyper(nn, m, n, k) 超几何
rlogis(n, location=0, scale=1) logistic 分布
rlnorm(n, meanlog=0, sdlog=1) 对数正态
rnbinom(n, size, prob) 负二项分布
runif(n, min=0, max=1) 均匀分布
rwilcox(nn, m, n), rsignrank(nn, n) Wilcoxon 分布

```


■ 期望（平均值）

定义 1.15 设离散型随机变量 X 的分布律为 $P\{X = x_i\} = p_i, i = 1, 2, \dots$, 若级数 $\sum_i |x_i|p_i$ 收敛, 则称级数 $\sum_i x_i p_i$ 的和为随机变量 X 的数学期望 (*mathematical expectation*), 记为 $E(X)$, 即

$$E(X) = \sum_i x_i p_i. \quad (1.61)$$

设连续型随机变量 X 的概率密度函数为 $f(x)$, 若积分 $\int_{-\infty}^{+\infty} |x|f(x)dx$ 收敛, 则称积分 $\int_{-\infty}^{+\infty} xf(x)dx$ 的值为随机变量 X 的数学期望, 记为 $E(X)$, 即

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx. \quad (1.62)$$

■ 方差

定义 1.16 设 X 为随机变量, 如果 $E\{[X - E(X)]^2\}$ 存在, 则称 $E\{[X - E(X)]^2\}$ 为 X 的方差 (*variance*), 记为 $\text{Var}(X)$, 即

$$\text{Var}(X) = E\{[X - E(X)]^2\}, \quad (1.64)$$

并称 $\sqrt{\text{Var}(X)}$ 为 X 的标准差 (*standard deviation*) 或均方差 (*root mean square*).

可以证明:

$$\text{Var}(X) = E(X^2) - [E(X)]^2.$$

- 大数定理与中心极限定理的意义
- 常用统计量：样本均值，样本方差，标准差，众数，最小值，最大值，分位数，中位数，上下四分位数

常见的数据描述性分析

- 中位数 median()
- 百分位数 quantile()

```
> quantile(x$x1)
      0%      25%      50%      75%     100%
61.00  74.00  80.50  84.25 101.00
> quantile(x$x1, probs = seq(0, 1, 0.2))
      0%     20%     40%     60%     80%    100%
61.0  73.0  78.0  82.4  87.0 101.0
> |
```

常见的数据描述性分析

■ 五数总括：

中位数 m_e , 下四分位数 Q_1 , 上四分位数 Q_3 , 最小值 min 和最大值 max.

```
> fivenum(x$x1, na.rm = TRUE)
[1] 61.0 74.0 80.5 84.5 101.0
> |
```

常见的数据描述性分析

- 正态性检验：函数shapiro.test()
- $P > 0.05$ ，正态性分布

```
> shapiro.test(x$x1)
```

```
Shapiro-Wilk normality test
```

```
data: x$x1
```

```
W = 0.9937, p-value = 0.9259
```

```
> shapiro.test(x$x3)
```

```
Shapiro-Wilk normality test
```

```
data: x$x3
```

```
W = 0.9444, p-value = 0.0003618
```

■ 方差与协方差、相关系数

$$s_{xx} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

$$s_{yy} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2,$$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

则称 s_{xx} 为变量 X 的观测样本的方差, 称 s_{yy} 为变量 Y 的观测样本的方差, 称 s_{xy} 为变量 X, Y 的观测样本的协方差. 称

$$S = \begin{bmatrix} s_{xx} & s_{xy} \\ s_{xy} & s_{yy} \end{bmatrix}$$

为观测样本的协方差矩阵. 称

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

为观测样本的相关系数.

协方差与相关系数计算

■ 函数cov()和cor()

```
> cov(x$x1,x$x2)
[1] 4.928283
> cor(x$x1,x$x2)
[1] 0.03982364

> cov(x[2:4])
      x1      x2      x3
x1 57.626263  4.928283 16.15152
x2  4.928283 265.759495 10.61010
x3 16.151515 10.610101 125.03030
> cor(x[2:4])
      x1      x2      x3
x1 1.00000000 0.03982364 0.19028099
x2 0.03982364 1.00000000 0.05820596
x3 0.19028099 0.05820596 1.00000000
> |
```



```
> cor.test(x$x1,x$x2)
```

```
Pearson's product-moment correlation
```

```
data: x$x1 and x$x2
```

```
t = 0.3945, df = 98, p-value = 0.694
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.1578290  0.2344082
```

```
sample estimates:
```

```
cor
```

```
0.03982364
```

相关分析与回归分析

■ 变量之间的关系

函数关系：有精确的数学表达式

相关关系：非确定性关系

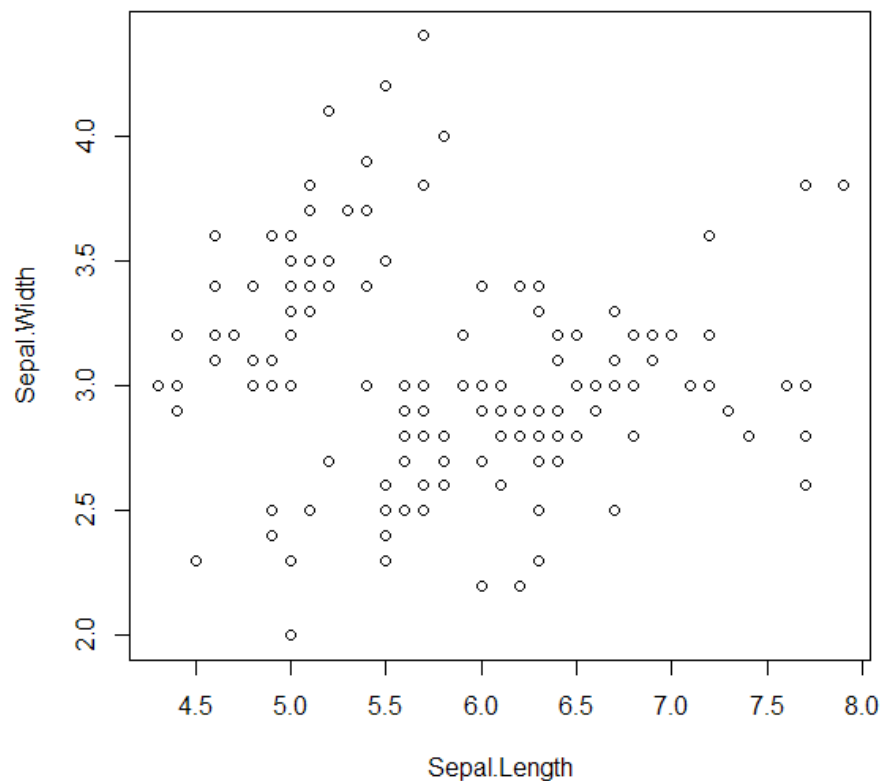
平行关系：相关分析（一元，多元）

依存关系：回归分析（一元，多元）

相关分析的例子

- Iris数据集
- 目测相关性

```
plot(iris[1,2])
```



2012.5.19

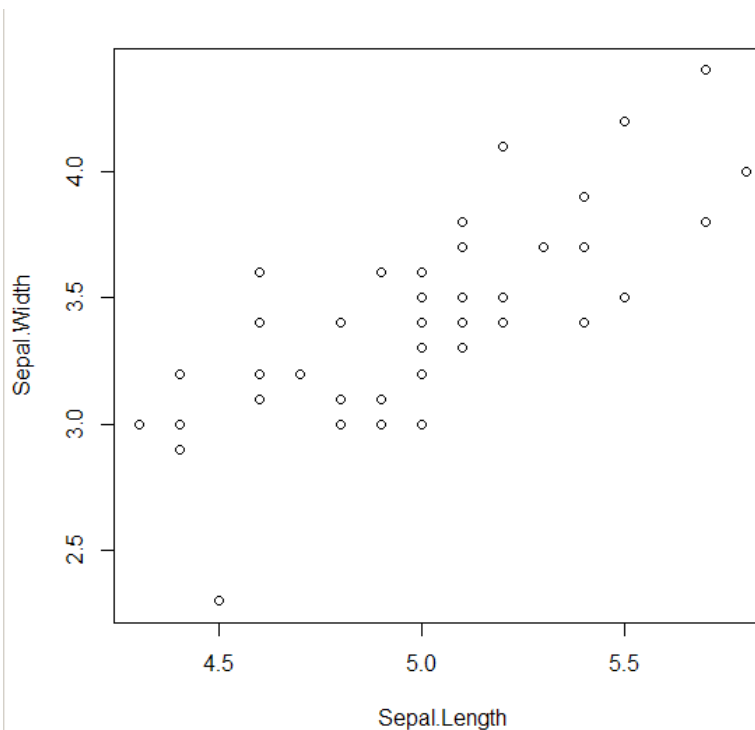
相关分析的例子

■ 分离种属

```
i1=iris[which(iris$Species=="setosa"),1:2]
```

```
plot(i1)
```

i1是提取出来的一个种属



2012.5.19

相关分析的例子

- 求相关系数
- 相关系数是否显著，不能只根据值的大小还需要进行假设检验

```
> cor(i1[1],i1[2])
                Sepal.Width
Sepal.Length    0.7425467
```

i1是之前用which函数提取出来的种属

相关分析的例子

- 相关系数显著性的假设检验
- 假设 r_0 为总体相关系数， $r_0=0$ 则说明没有相关关系，建立假设 $H_0:r_0=0$ ， $H_1:r_0 \neq 0$ ($\alpha=0.05$)
- 计算相关系数 r 的 t 值和 P -值

```
> cor.test(il$Sepal.Length, il$Sepal.Width)
```

```
Pearson's product-moment correlation
```

```
data:  il$Sepal.Length and il$Sepal.Width
t = 7.6807, df = 48, p-value = 6.71e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5851391 0.8460314
sample estimates:
      cor
0.7425467
```

一元线性回归分析

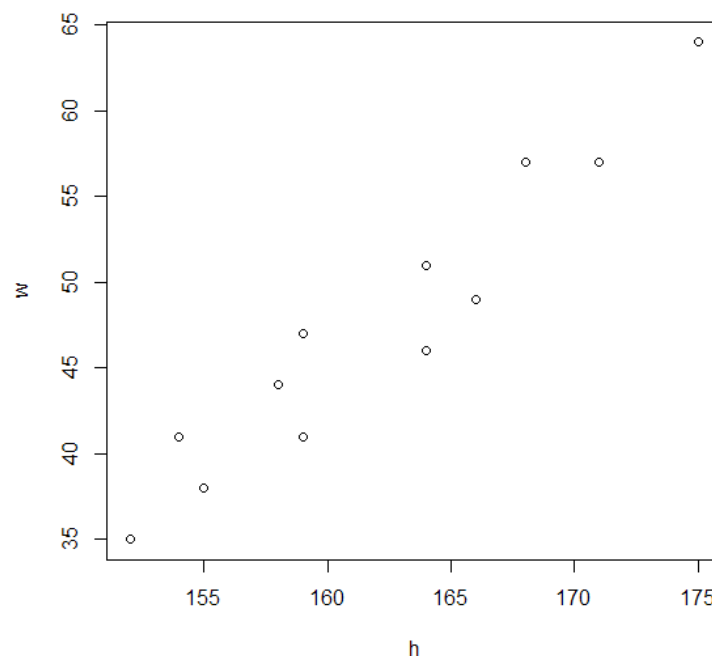
- 原理，最小二乘法
- 步骤：建立回归模型，求解回归模型中的参数，对回归模型进行检验
- 例子

数据：身高-体重

$h = c(171, 175, 159, 155, 152, 158, 154, 164, 168, 166, 159, 164)$

$w = c(57, 64, 41, 38, 35, 44, 41, 51, 57, 49, 47, 46)$

$\text{plot}(w \sim h + 1)$



一元线性回归分析

自定义函数 lxy <-

```
function(x,y){n=length(x);sum(x*  
y)-sum(x)*sum(y)/n}
```

假设 $w = a + bh$

则有

```
> b=lxy(h,w)/lxy(h,h)
```

```
> a=mean(w)-b*mean(h)
```

```
> a
```

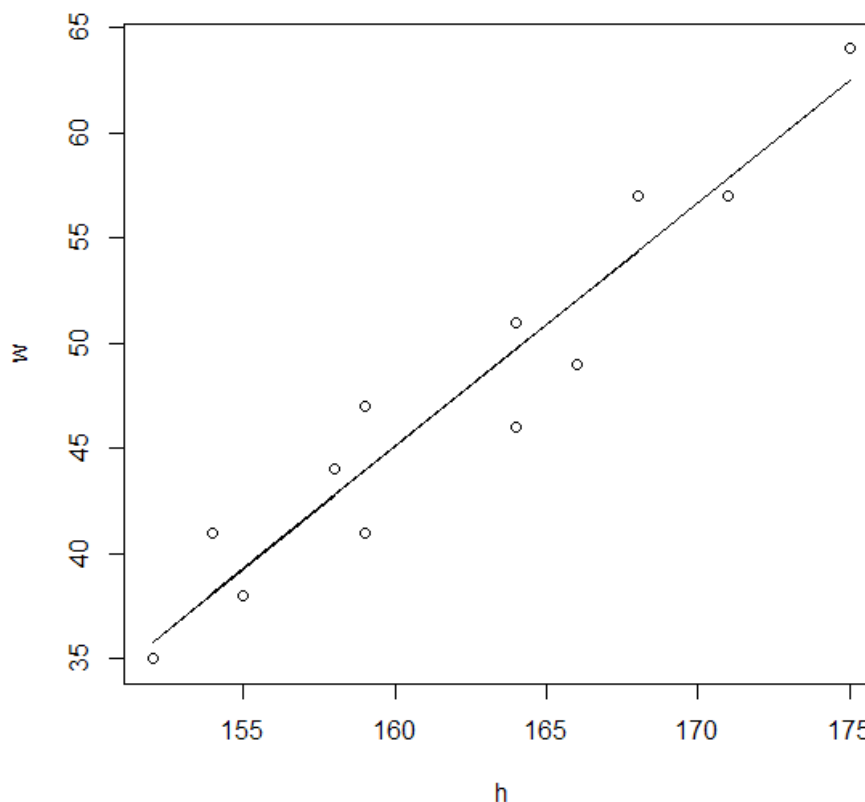
```
[1] -140.3644
```

```
> b
```

```
[1] 1.15906
```

作回归直线

```
lines(h,a+b*h)
```



一元线性回归分析

- 回归系数的假设检验
- 建立线性模型

lm是指 linear model, 即线性模型

这里的a指代了这个模型 `> a=lm(w~1+h)` `w~1+h`的意思是 $w = k+bh$ 的意思
`> a`

```
Call:
lm(formula = w ~ 1 + h)
```

```
Coefficients:
(Intercept)                h
   -140.364             1.159
      k
```

这个结果和前面手工算是一样的

■ 线性模型的汇总数据，t检验，summary()函数

```
> summary(a)  a已经是回归好的一个模型
```

```
Call:
lm(formula = w ~ 1 + h)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-3.721	-1.699	0.210	1.807	3.074

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-140.3644	17.5026	-8.02	1.15e-05	***
h	1.1591	0.1079	10.74	8.21e-07	***

显著性标记,***极度显著,**高度显著,*显著,圆点不太显著,没有记号不显著

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.546 on 10 degrees of freedom
Multiple R-squared: 0.9203,    Adjusted R-squared: 0.9123
F-statistic: 115.4 on 1 and 10 DF,  p-value: 8.21e-07
```

p值越小,模型越合理

2012.5.19

一元线性回归分析

- 汇总数据的解释
- Residuals : 参差分析数据
- Coefficients : 回归方程的系数 , 以及推算的系数的标准差 , t值 , P-值
- F-statistic : F检验值
- Signif : 显著性标记 , ***极度显著 , **高度显著 , *显著 , 圆点不太显著 , 没有记号不显著

■ 方差分析，函数anova()

```
> anova(a)
Analysis of Variance Table

Response: w
          Df Sum Sq Mean Sq F value    Pr(>F)
h           1  748.17   748.17   115.41 8.21e-07 ***
Residuals 10   64.83     6.48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

■ 预测：一个身高185的人，体重大约是多少？

> a+b*185

[1] 74.0618

>

lm()线性模型函数

适应于多元线性模型的基本函数是 `lm()`, 其调用形式是

```
fitted.model <- lm(formula, data = data.frame)
```

其中 `formula` 为模型公式. `data.frame` 为数据框. 返回值为线性模型结果的对象存放在 `fitted.model` 中. 例如

```
fm2 <- lm(y ~ x1 + x2, data = production)
```

适应于 y 关于 x_1 和 x_2 的多元回归模型 (隐含着截距项). 没专门提0的话, 结局项都是包括的

这两个是一个意思, 用法是一样的

- $y \sim 1 + x$ 或 $y \sim x$ 均表示 $y = a + bx$ 有截距形式的线性模型
- 通过原点的线性模型可以表达为 : $y \sim x - 1$ 或 $y \sim x + 0$ 或 $y \sim 0 + x$

参见 `help(formula)`

与线性模型有关的函数

建立数据：身高-体重

```
x=c(171,175,159,155,152,158,154,164,168,166,159,164)
```

```
y=c(57,64,41,38,35,44,41,51,57,49,47,46)
```

建立线性模型

```
a=lm(y~x)
```

求模型系数

```
> coef(a)
```

```
(Intercept)      x
-140.36436    1.15906
```

提取模型公式

```
> formula(a)
```

```
y ~ x
```

与线性模型有关的函数

计算残差平方和 (什么是残差平方和)

```
> deviance(a)
```

```
[1] 64.82657
```

绘画模型诊断图 (很强大 , 显示残差、拟合值和一些诊断情况)

```
> plot(a)
```

计算残差

```
> residuals(a)
```

1	2	3	4	5	6	7
-0.8349544	1.5288044	-2.9262307	-1.2899895	-0.8128086	1.2328296	2.8690708
8	9	10	11	12		
1.2784678	2.6422265	-3.0396529	3.0737693	-3.7215322		

与线性模型有关的函数

打印模型信息

```
> print(a)
```

Call:

```
lm(formula = y ~ x)
```

Coefficients:

(Intercept)	x
-140.364	1.159

计算方差分析表

```
> anova(a)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
x           1  748.17   748.17  115.41 8.21e-07 ***
Residuals  10   64.83     6.48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

提取模型汇总资料

```
> summary(a)
```

```
Call:
```

```
lm(formula = y ~ x)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-3.721	-1.699	0.210	1.807	3.074

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-140.3644	17.5026	-8.02	1.15e-05 ***
x	1.1591	0.1079	10.74	8.21e-07 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.546 on 10 degrees of freedom
```

```
Multiple R-squared: 0.9203, Adjusted R-squared: 0.9123
```

```
F-statistic: 115.4 on 1 and 10 DF, p-value: 8.21e-07
```

2012.5.19

与线性模型有关的函数

作出预测

```
> z=data.frame(x=185)
> predict(a,z)
1
74.0618
> predict(a,z,interval="prediction", level=0.95)
fit lwr upr
1 74.0618 65.9862 82.13739
```

课后阅读：薛毅书，p308，计算实例



Thanks

FAQ时间