

## 数据分析、展现与R语言 第3周

## 法律声明



【声明】本视频和幻灯片为炼数成金网络课程的教学资料,所有资料只能在课程内使用,不得在课程以外范围散播,违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

## 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

## 随机试验



- 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

## 概率——刻画随机事件出现可能性的指标



#### ■ 直观的概率计算——古典概型

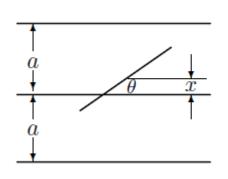
设随机事件 E 的样本空间中只有有限个样本点,即  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ ,其中 n 为样本点总数. 每个样本点  $\omega_i (i=1,2,\dots,n)$  出现是等可能的,并且每次试验有且仅有一个样本点发生,则称这类现象为古典概型 (classical probability model). 若事件 A 包含 m 个样本点,则事件 A 的概率定义为

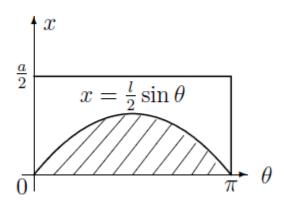
$$P(A) = \frac{m}{n} = \frac{\text{$\mathbb{P}$}(A) + A + 2 + 2 + 2}{\text{$\mathbb{P}$}(A)}.$$
 (1.12)

## 连续样本空间情形下的概率



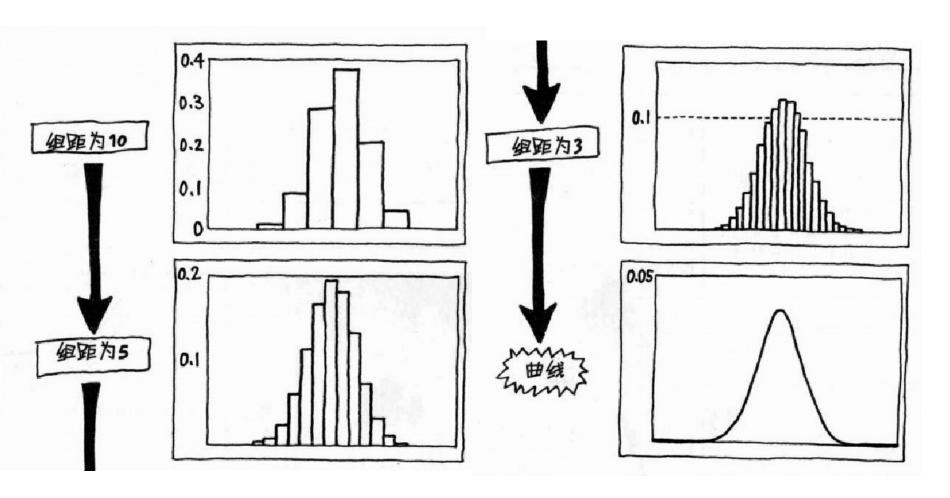
■ 几何概型——布冯投针实验





## 连续样本空间情形下的概率: 概率密度





## 分布



■ 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

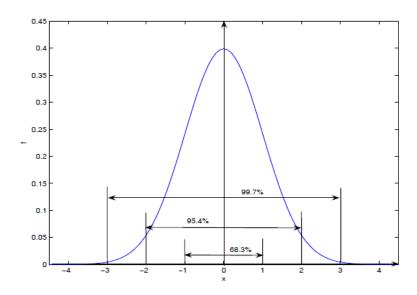
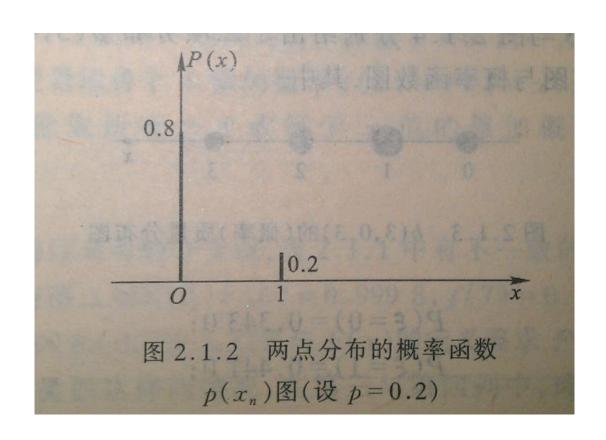


图 1.3: 标准正态分布和对应区间上积分 (面积) 的百分比

## 两点分布





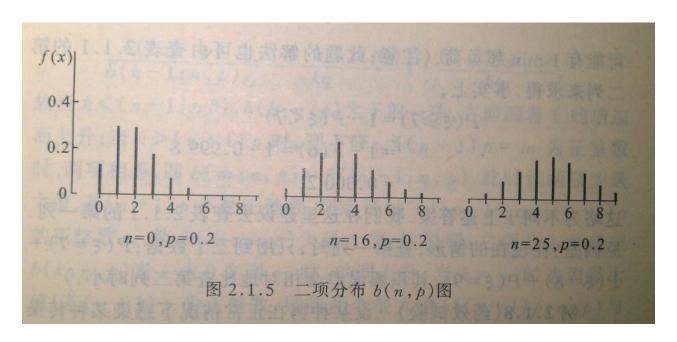
## 二项分布



若随机变量 X 的分布律为

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

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



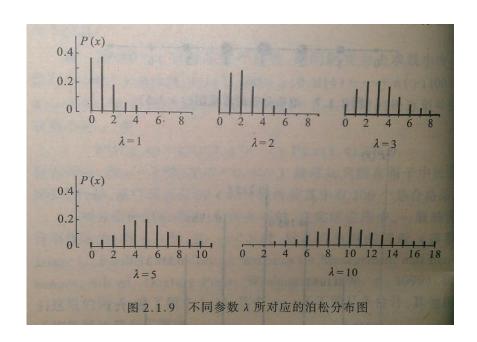
## 泊松分布



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  $\lambda$  的 Poisson (泊松) 分布 (Poisson distribution)



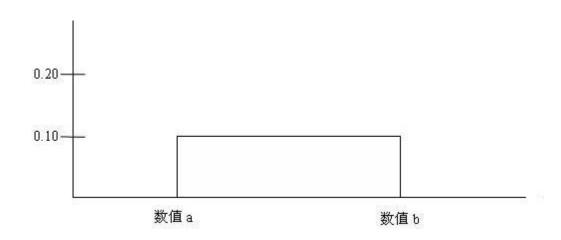
## 均匀分布



若随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \sharp \Xi, \end{cases}$$

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



## 指数分布



若随机变量 X 的概率密度函数为

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

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

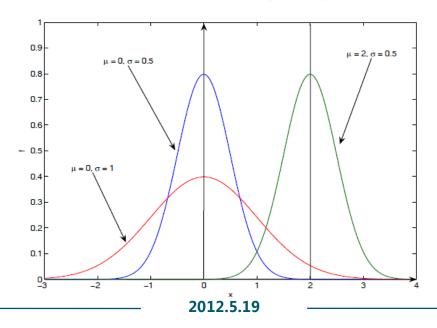
## 正态分布



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  $\mu, \sigma$  的正态分布 (normal distribution), 也称为 Gauss 分布,记作  $X \sim N(\mu, \sigma^2)$ .



## 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)γ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) β 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
```

## 随机变量的数字特征



#### ■ 期望(平均值)

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

$$E(X) = \sum_{i} x_i p_i. \tag{1.61}$$

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  $\int_{-\infty}^{+\infty} x f(x) dx$  的值为随机变量 X 的数学期望,记为 E(X), 即

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

## 随机变量的数字特征



#### ■ 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 Var(X), 即

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

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

可以证明:

$$Var(X) = E(X^2) - [E(X)]^2$$
.

## 总体与抽样



- 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
```

## 多元数据的数据特征



#### ■ 方差与协方差、相关系数

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

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

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

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])
          \times 1
                      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
>
```

## 相关性检验



#### > cor.test(x\$x1,x\$x2)

Pearson's product-moment correlation

## 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

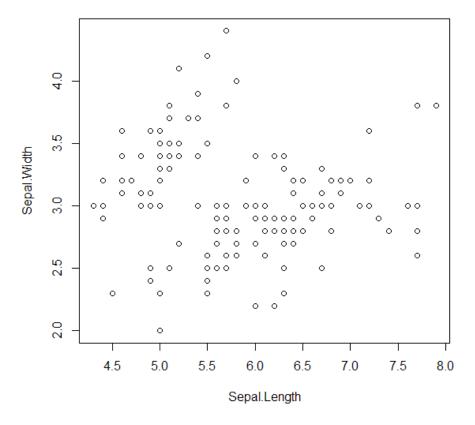
平行关系:相关分析(一元,多元)

依存关系:回归分析(一元,多元)



- Iris数据集
- 目测相关性

plot(iris[1,2])

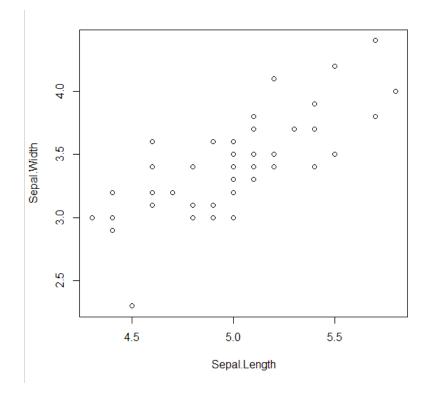




#### ■ 分离种属

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

plot(i1)





- 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



- 相关系数显著性的假设检验
- 假设r0为总体相关系数,r0=0则说明没有相关关系,建立假设H0:r0=0, H1:r0<>0(alpha=0.05)

> cor.test(i1\$Sepal.Length,i1\$Sepal.Width)

■ 计算相关系数r的t值和P-值

cor

0.7425467

Pearson's product-moment correlation

data: i1\$Sepal.Length and i1\$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:



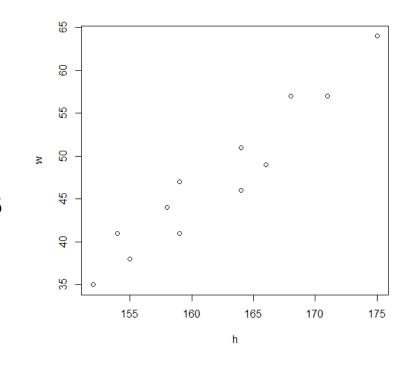
- 原理,最小二乘法
- 步骤:建立回归模型,求解回归模型中的参数,对回归模型进行检验
- 例子

数据:身高-体重

h=c(171,175,159,155,152,158,154,16 4,168,166,159,164)

w=c(57,64,41,38,35,44,41,51,57,49,4 7,46)

 $plot(w \sim h + 1)$ 



2012.5.19



```
自定义函数 lxy<-
```

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

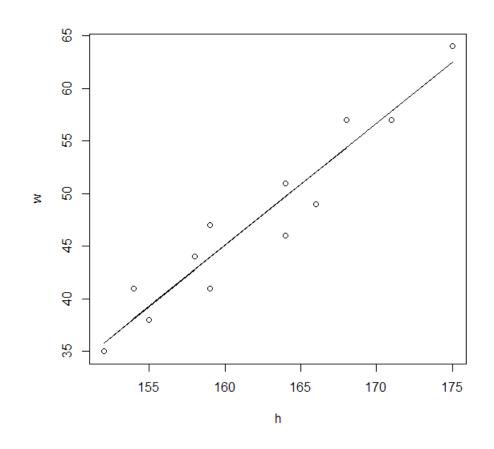
假设w=a+bh

#### 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)



2012.5.19



- 回归系数的假设检验
- 建立线性模型



■ 线性模型的汇总数据,t检验,summary()函数

```
> summary(a)
Call:
lm(formula = w \sim 1 + h)
Residuals:
  Min 10 Median 30 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 ***
              1.1591 0.1079 10.74 8.21e-07 ***
h
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
```



■ 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。 显著



■ 方差分析,函数anova()



■ 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)</pre>

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

 $fm2 \leftarrow lm(y \sim x1 + x2, data = production)$ 

适应于 y 关于 x1 和 x2 的多元回归模型(隐含着截距项)。

- y~1+x或y~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~x-1或y~x+0或y~0+x

参见help(formula)



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

 $a=Im(y\sim x)$ 

#### 求模型系数

> coef(a)

(Intercept) x

-140.36436 1.15906

#### 提取模型公式

> formula(a)

y ~ x



#### 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

-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



#### 打印模型信息

> print(a)

Call:

 $Im(formula = y \sim x)$ 

Coefficients:

(Intercept) x

-140.364 1.159



#### 计算方差分析表

1



#### 提取模型汇总资料

```
> summary(a)
Call:
lm(formula = v \sim x)
Residuals:
  Min 10 Median 30 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 ***
              1.1591 0.1079 10.74 8.21e-07 ***
X
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
```



#### 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

## 炼数成金逆向收费式网络课程



- Dataguru (炼数成金)是专业数据分析网站,提供教育,媒体,内容,社区,出版,数据分析业务等服务。我们的课程采用新兴的互联网教育形式,独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围,重竞争压力的特点,同时又发挥互联网的威力打破时空限制,把天南地北志同道合的朋友组织在一起交流学习,使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成于上万的学习成本,直线下降至百元范围,造福大众。我们的目标是:低成本传播高价值知识,构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情,请看我们的培训网站 http://edu.dataguru.cn





# Thanks

## FAQ时间