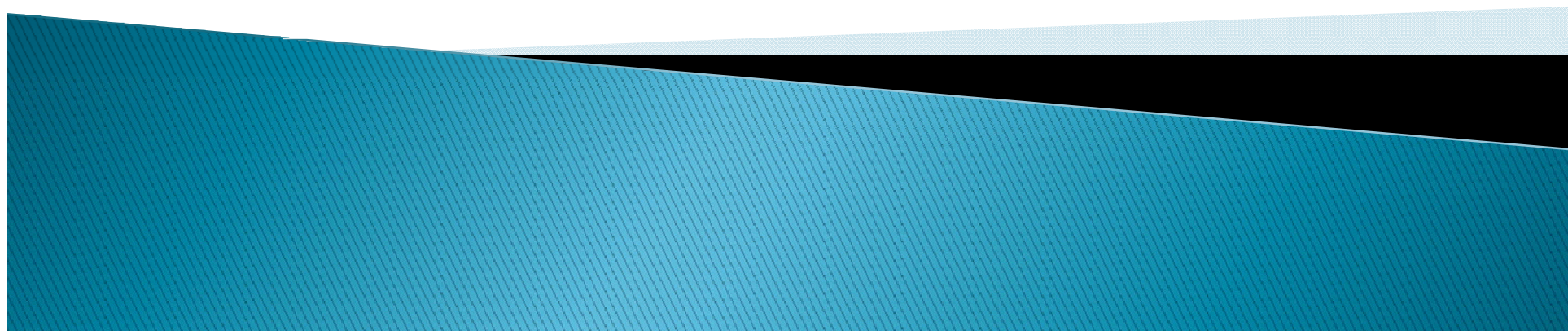


R语言与生物统计学

基础生物统计学与统计计算分析



统计学的定义

从试验设计、数据收集、总结、推断到决策

▶ 统计学的基本分类

- 描述性统计 (descriptive statistics)
- 概率 (probability) : 分布 (distribution)
- 推断统计 (inferential statistics) : 参数估计 (parameter estimation)、假设检验 (hypothesis testing)、统计决策 (decision making)

R语言

- ▶ CRAN

<http://cran.r-project.org>

- ▶ R语言的进化史

- S语言
- S-PLUS

- ▶ 其他统计学语言

- SPSS
- SAS

数据分类

- ▶ 比例（ratio或者proportion）
- ▶ 连续型数据（continuous）
 - 取值范围（range）
- ▶ 离散型数据（discrete）
 - 有序的（ordinal）
 - 命名的（nominal）
 - Binary（dichotomous）
- ▶ 生存数据（survival）
 - 有截断（truncation）
 - 有删失（censoring）

“Missing data”

描述性统计

- ▶ 位置信息 (location)
 - 均值 (mean)
 - 中位值 (median)
 - Mode
- ▶ 分布信息 (dispersion)
 - 范围 (range) : (min, max)
 - 标准偏差 (standard deviation)
 - IQR(inter-quartile range): 3rd quartile - 1st quartile
 - 百分位数 (percentile)
 - 变异系数 (coefficient of variation) : $sd/mean$
- ▶ 分布形状信息 (shape)
 - 方差 (variance) : 反映数据的变化程度大小
 - 偏度 (skewness) : 反映数据是否对称
 - 峭度 (Kurtosis) : 反映数据的集中程度
 - Moments
 - L-moments

R语言

- ▶ 下载R的安装包
- ▶ 安装R
- ▶ 启动R
- ▶ 退出R
 - `q()`

基本数值计算

- ▶ $1+2$, $3-1$, $4*5$, $3/4$, $5\%/\%2$
- ▶ 取余: $5\%\%3$
- ▶ 幂运算: $2**2$, 2^3
- ▶ 绝对值: $\text{abs}(3.1)$, $\text{abs}(-4.5)$
- ▶ 精度: $\text{signif}(1.345, 2)$
- ▶ 四舍五入: $\text{round}(0.345, 2)$
- ▶ 取整数部分: $\text{trunc}(3.4)$, $\text{trunc}(-3.4)$
- ▶ 向下取整: $\text{floor}(-2.3)$, $\text{floor}(2.3)$
- ▶ 向上取整: $\text{ceiling}(-2.3)$, $\text{ceiling}(2.3)$
- ▶ 三角函数: $\text{sin}(2*\text{pi})$, $\text{cos}(\text{pi})$, $\text{tan}(\text{pi}/4)$, $\text{asin}(1)$,
 $\text{acos}(0)$, $\text{atan}(1)$
- ▶ 符号函数: $\text{sign}(1.5)$, $\text{sign}(0)$, $\text{sign}(-3.4)$

基本数值计算（续）

- ▶ 比较运算： $3 > 4$, $5 \leq 8$, $3 > 1$, $2.5 < 0.8$,
 $(3 == 4) == \text{FALSE}$, $(3 == 4) != \text{TRUE}$
- ▶ 逻辑运算： $3 > 1 \ \& \ 2 < 5$, $3 > 1 \ | \ 2 < 5$
- ▶ 指数运算： $\text{exp}(1)$, $\text{exp}(-10)$
- ▶ 对数运算： $\log(\text{exp}(1))$, $\log_{10}(1 \text{e}-10)$,
 $\log_2(2^31)$, $\log(2^3, 2)$

帮助信息

获取solve()函数的帮助信息

```
help('solve')
```

或者

```
?solve
```

查找与“solve”有关的帮助条目

```
help.search('solve')
```

列出MASS包中的帮助条目

```
help(package=MASS)
```

向量 (vector)

创建向量

```
x <- c(1,3,6,5,7,9,2)
```

向量长度与总和

```
length(x), sum(x)
```

均值、范围、最大值、最小值

▶ `mean(x)`, `range(x)`, `max(x)`, `min(x)`

中位数、中位绝对偏差、四分位差

```
median(x), mad(x), IQR(x), cummin(x), cummax(x)
```

计算百分位数

```
quantile(x, .10), quantile(x, .25)
```

```
quantile(x, c(.10,.25,.50,.75,.90))
```

计算方差与标准偏差

```
var(x), sd(x)
```

一元描述统计

见分发下来的材料

向量的总结函数

```
## 安装e1071包
```

```
install.packages("e1071")
```

```
## 导入其他package
```

```
library(e1071)
```

```
## 计算偏度，偏度为0表示没有严重拖尾，左拖尾
```

```
skewness(x)
```

```
## 计算峭度
```

```
kurtosis(x)
```

其他产生向量的函数

```
## 用a:b构建向量(a, a+-1, ..., b)
```

```
1:10
```

```
10:1
```

```
### sequences
```

```
seq(0, 10, by=1)
```

```
seq(0, 10, length=11)
```

```
x <- 1:4
```

```
seq(1, 41, along=x)
```

```
### repeat
```

```
rep(10, 5)
```

```
rep(c(10,3,5), 5)
```

```
rep(c(10,3,5), each=5)
```

```
rep(1:3, 3:1)
```

画图

```
#### 产生长度为1000的随机向量x
x <- rnorm(1000)
#### 密度分布曲线
plot(density(x), col="blue", lty=2)
#### 直方图
ht <- hist(x, freq=F, add=T, col="gray",
  border="lightgray")
ht
density(x)
```

离散化处理

```
### 用table()函数  
table(x >= 3)
```

```
### all()函数和any()函数  
all(x > 3)  
any(x > 3)
```