

统计学概论

续本达

清华大学 工程物理系

2023-11-08

统计学概论

续本达

复习

数据

描述统计学

直方图

箱线图

推断统计学

复习

使用 **随机数** 来解决 **计算问题** 方法的统称。

- ↓ 已知的信息越来越少，适用范围越来越广

	方法	已知的分布信息	维度
✓	逆函数法	完全已知	低
✓	舍选法	与标准分布的比值	低
✓	马尔可夫链法	与自身的比值	高
	对抗生成网络法*	无解析形式，无法计算比值	高

蒙特卡罗方法的普适性

- 蒙特卡罗方法在物理之外的几乎所有学科中都有应用。

统计学概论

续本达

复习

数据

描述统计学

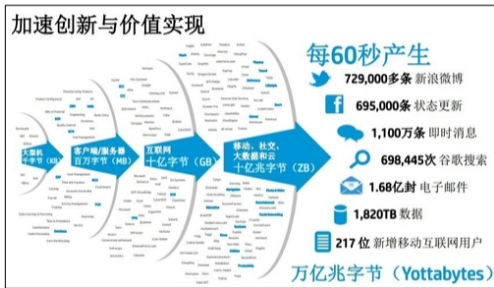
直方图

箱线图

推断统计学

数据

我们正进入大数据时代

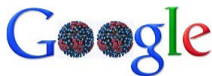


智能手机
 社交媒体
 摄像头
 基于位置服务
 GPS...

国际数据公司IDC报告：2011 年全球被创建和被复制的数据总量超过1.8 ZB，且增长趋势遵循新摩尔定律(全球数据量大约每两年翻一番)。

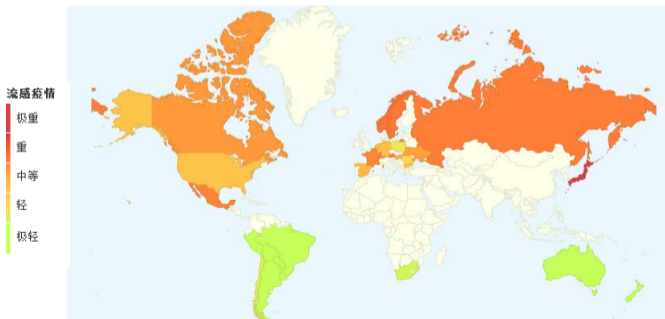
1 ZB : 10亿 GB

Google流感趋势预测



Google利用大数据来应对流感

■ Google是大数据时代的奠基者，也是行业大数据技术架构的标杆和示范。Google具有丰富的数据资源，所以Google Flu Trends（GFT）能够早于全球健康部门（如CDC）预测疫情趋势



统计学概论

续本达

复习

数据

描述统计学

直方图

箱线图

推断统计学

描述统计学

统计学

收集、分析、表述和解释数据的科学

- ① 数据搜集
- ② 数据分析
- ③ 数据表述（图表）
- ④ 数据解释

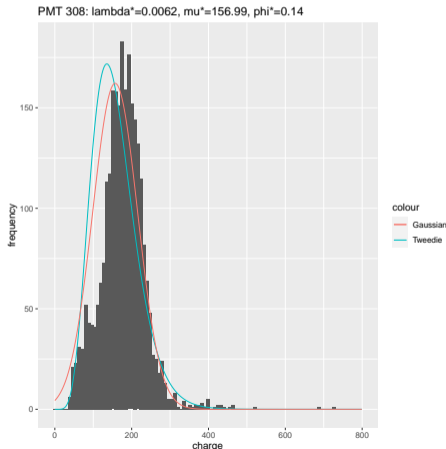
- 对随机现象进行观测、试验，以取得有代表性的观测值
- 收集、加工数据，并用图形、表格和数值方法来汇总数据的统计学

内容

- ① 搜集数据
- ② 整理数据
- ③ 展示数据
- ④ 描述性分析

目的

- ① 描述数据特征
- ② 找出数据的基本规律



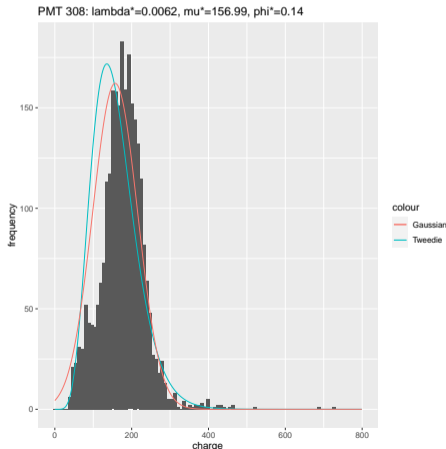
- 对随机现象进行观测、试验，以取得有代表性的观测值
- 收集、加工数据，并用图形、表格和数值方法来汇总数据的统计学

内容

- ① 搜集数据
- ② 整理数据
- ③ 展示数据
- ④ 描述性分析

目的

- ① 描述数据特征
- ② 找出数据的基本规律



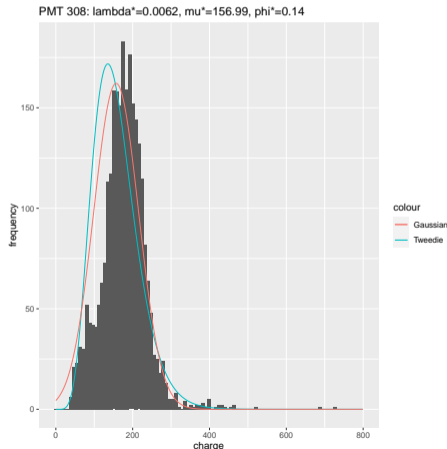
- 对随机现象进行观测、试验，以取得有代表性的观测值
- 收集、加工数据，并用图形、表格和数值方法来汇总数据的统计学

内容

- ① 搜集数据
- ② 整理数据
- ③ 展示数据
- ④ 描述性分析

目的

- ① 描述数据特征
- ② 找出数据的基本规律



统计学概论

续本达

复习

数据

描述统计学

直方图

箱线图

推断统计学

直方图

复习

数据

描述统计学

直方图

箱线图

推断统计学

- 实验采集的原始数据往往是一个个离散的样本点，很难从中直接得到关键信息。
- 只有经过整理后，用常用的表示方法展示后，数据中的信息才会变得直观。

常用的表示方法

列表法 频数分布表、频率分布表

图示法 频数直方图、频率直方图、箱线图

- 实验采集的原始数据往往是一个个离散的样本点，很难从中直接得到关键信息。
- 只有经过整理后，用常用的表示方法展示后，数据中的信息才会变得直观。

常用的表示方法

列表法 频数分布表、频率分布表

图示法 频数直方图、频率直方图、箱线图

复习

数据

描述统计学

直方图

箱线图

推断统计学

```
etruscan <- read.csv("https://hep.tsinghua.edu.cn/~orv/teaching/statistics/etruscan.csv")
a_etruscan <- etruscan[etruscan$group=='ancient', 'width']
a_etruscan
```

```
[1] 141 148 132 138 154 142 150 146 155 158 150 140 147 148 144 150 149 145 149
[20] 158 143 141 144 144 126 140 144 142 141 140 145 135 147 146 141 136 140 146
[39] 142 137 148 154 137 139 143 140 131 143 141 149 148 135 148 152 143 144 141
[58] 143 147 146 150 132 142 142 143 153 149 146 149 138 142 149 142 137 134 144
[77] 146 147 140 142 140 137 152 145
```

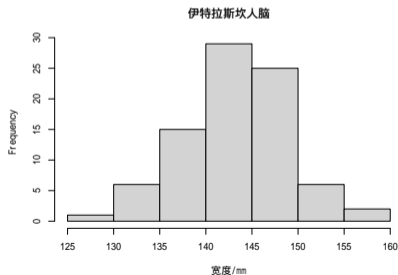
```
stem(a_etruscan) # 字符的总结
```

```
The decimal point is 1 digit(s) to the right of the |
```

```
12 | 6
13 | 1224
13 | 5567777889
14 | 00000001111112222222333333444444
14 | 555666666777788888999999
15 | 000022344
15 | 588
```

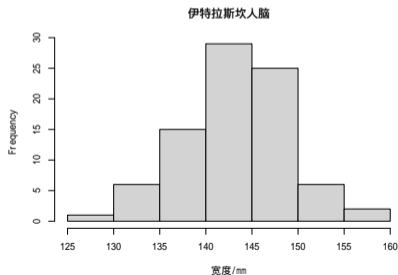

- ① 找出最小值 126 和最大值 158，取区间 $[124.5, 159.5]$;
- ② 将选定区间分为 $k = 7$ 个小区间，宽度为 $\Delta = \frac{159.5 - 124.5}{7} = 5$;
- ③ 画图每个区间内频数 f_i

```
hist(a_etruscan, main="伊特拉斯坎人脑", xlab="宽度/mm")
```



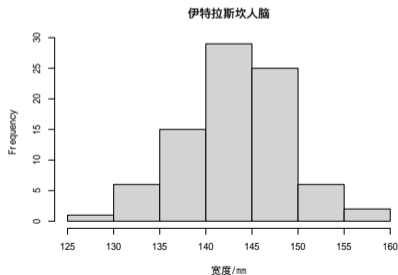
- ① 找出最小值 126 和最大值 158，取区间 $[124.5, 159.5]$;
- ② 将选定区间分为 $k = 7$ 个小区间，宽度为 $\Delta = \frac{159.5-124.5}{7} = 5$;
- ③ 画图每个区间内频数 f_i

```
hist(a_etruscan, main="伊特拉斯坎人脑", xlab="宽度/mm")
```



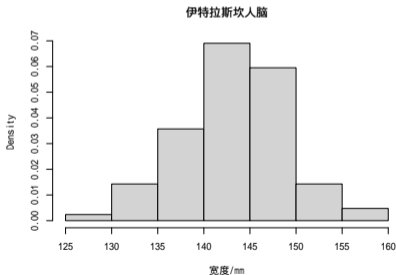
- ① 找出最小值 126 和最大值 158，取区间 $[124.5, 159.5]$;
- ② 将选定区间分为 $k = 7$ 个小区间，宽度为 $\Delta = \frac{159.5 - 124.5}{7} = 5$;
- ③ 画图每个区间内频数 f_i

```
hist(a_etruscan, main="伊特拉斯坎人脑", xlab="宽度/mm")
```



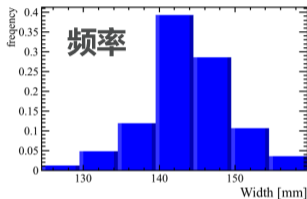
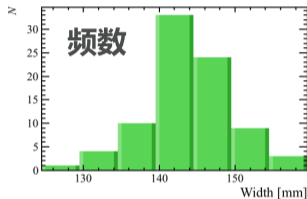
- 令直方图各区间加和为 1，则成为 **频率直方图**
- 在各个小区间上作以频率 $\frac{f_i}{n}$ 为高的小矩形。

```
hist(a_etruscan, freq=FALSE, main="伊特拉斯坎人脑", xlab="宽度/mm")
```



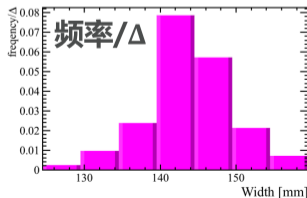
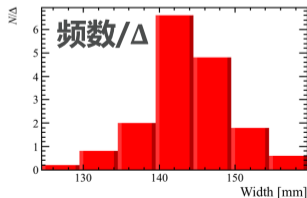
直方图中每个小矩形的高，可用频数或单位组距的频数，称**频数直方图**，也可用频率或单位组距的频率，称**频率直方图**。作图时要在直方图纵坐标注明。

频数累加
等于样本
总量 n



频率累加
等于1

面积累加
等于样本
总量 n



面积累加
等于1

统计学概论

续本达

复习

数据

描述统计学

直方图

箱线图

推断统计学

箱线图

复习

数据

描述统计学

直方图

箱线图

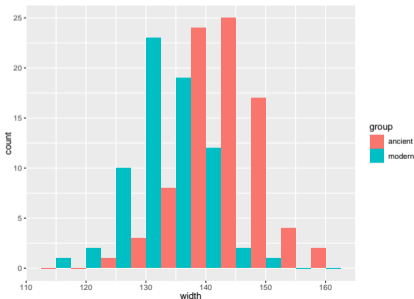
推断统计学

```
unique(etruscan$group)
```

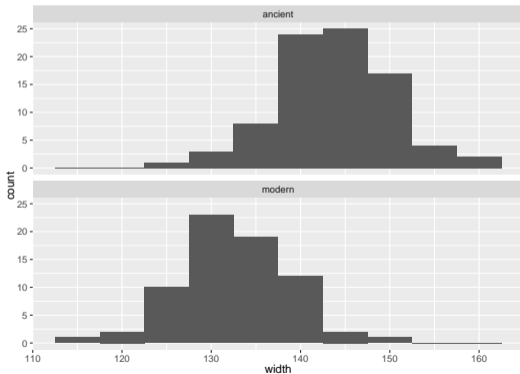
```
[1] "ancient" "modern"
```

叠叠画

```
library(ggplot2)
p <- ggplot(etruscan, aes(x=width, fill=group)) + geom_histogram(binwidth=5, position = "dodge")
print(p)
```



```
p <- ggplot(etruscan, aes(x=width)) + geom_histogram(binwidth=5) + facet_wrap(~group, nrow=2)
print(p)
```



复习

数据

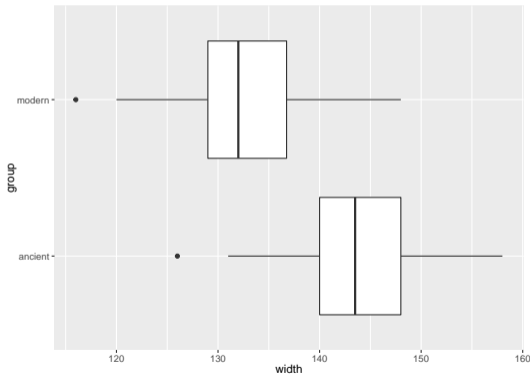
描述统计学

直方图

箱线图

推断统计学

```
p <- ggplot(etruscan, aes(y=group, x=width)) + geom_boxplot(binwidth=5)  
print(p)
```



设 $0 < \alpha < 1$, 若 x_α 满足

$$P(X \leq x_\alpha) = F(x_\alpha) = \alpha$$

则称 x_α 为 X 服从的分布的 α -分位数, 亦称下侧 α -分位数。若 x'_α 满足

$$P(X \geq x'_\alpha) = \alpha$$

则 x'_α 被称为上侧 α -分位数。

特点

- 因为 $X \leq x_\alpha$ 的可能性为 α , 所以 $X > x_\alpha$ 的可能性为 $1 - \alpha$ 。连续变量 $x'_\alpha = x_{1-\alpha}, x_\alpha = x'_{1-\alpha}$ 。
- 对离散型分布, 不一定存在 α -分位数。
- $\alpha = 0.5$ 时的 α -分位数 $x_{0.5}$ 称为 **中位数**。

数据集的箱线图是箱子和直线组成的图形，它的定义用到了数据的最小最大值和几个 **样本分位数**。

样本分位数：设有容量为 n 的样本观察值 (x_1, x_2, \dots, x_n) ，样本 p 分位数 $(0 < p < 1)$ 记为 x_p ，它具有以下性质：

- ① 至少有 np 个观察值小于或等于 x_p ；
- ② 至少有 $n(1-p)$ 个观察值大于或等于 x_p 。

复习

数据

描述统计学

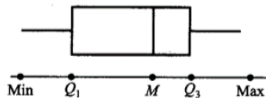
直方图

箱线图

推断统计学

箱线图是基于数据最小值 Min 、最大值 Max 、三个常用分位数 Q_1, M, Q_3 ，用箱子和线画出的图形。做法如下：

- ① 画一水平数轴，轴上标上 $\text{Min}, Q_1, M, Q_3, \text{Max}$ 。在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于 Q_1, Q_3 的上方。在 M 点上方箱子内部画一条垂直线段。
- ① 自箱子左侧引一条水平线直至最小值 Min ；在同一水平高度自箱子右侧引一条水平线直至最大值。



复习

数据

描述统计学

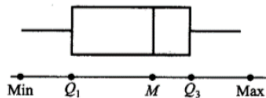
直方图

箱线图

推断统计学

箱线图是基于数据最小值 Min 、最大值 Max 、三个常用分位数 Q_1, M, Q_3 ，用箱子和线画出的图形。做法如下：

- ① 画一水平数轴，轴上标上 $\text{Min}, Q_1, M, Q_3, \text{Max}$ 。在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于 Q_1, Q_3 的上方。在 M 点上方箱子内部画一条垂直线段。
- ② 自箱子左侧引一条水平线直至最小值 Min ；在同一水平高度自箱子右侧引一条水平线直至最大值。



复习

数据

描述统计学

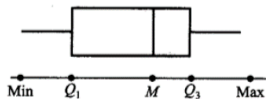
直方图

箱线图

推断统计学

箱线图是基于数据最小值 Min 、最大值 Max 、三个常用分位数 Q_1, M, Q_3 ，用箱子和线画出的图形。做法如下：

- ① 画一水平数轴，轴上标上 $\text{Min}, Q_1, M, Q_3, \text{Max}$ 。在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于 Q_1, Q_3 的上方。在 M 点上方箱子内部画一条垂直线段。
- ① 自箱子左侧引一条水平线直至最小值 Min ；在同一水平高度自箱子右侧引一条水平线直至最大值。



统计学概论

续本达

复习

数据

描述统计学

直方图

箱线图

推断统计学

推断统计学

复习

数据

描述统计学

直方图

箱线图

推断统计学

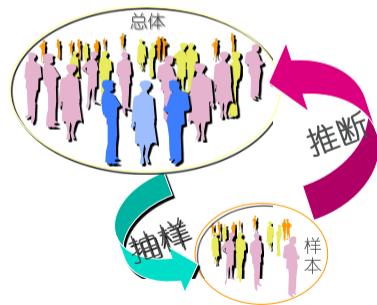
- 对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性
- 用样本数据对总体的某些特征进行估计和检验的统计学

内容

- ① 参数估计
- ② 假设检验

目的

- 对总体特征作出推断



复习

数据

描述统计学

直方图

箱线图

推断统计学

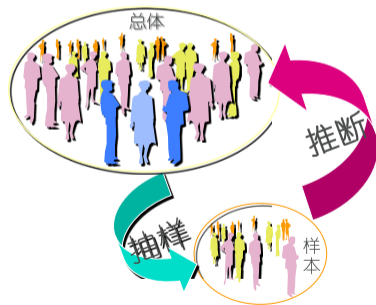
- 对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性
- 用样本数据对总体的某些特征进行估计和检验的统计学

内容

- ① 参数估计
- ② 假设检验

目的

- 对总体特征作出推断



复习

数据

描述统计学

直方图

箱线图

推断统计学

