统计学概论

续本达

清华大学 工程物理系

2024-11-04

复习

指数分布族

指数分布族

一族的分布，它们的概率密度函数或分布律记为 $f(x|\vec{\theta})$ ，如果有

$$f(x|\vec{\theta}) = \exp\left[\sum_{i=1}^{s}\eta_i(\vec{\theta})T_i(x) - B(\vec{\theta})\right]h(x)$$

的形式，则称为 s 维指数分布族 。

如果把 $\eta_i(\vec{\theta})$ 当作自变量反解 θ，那么形式有进一步的化简。

指数分布族标准形式

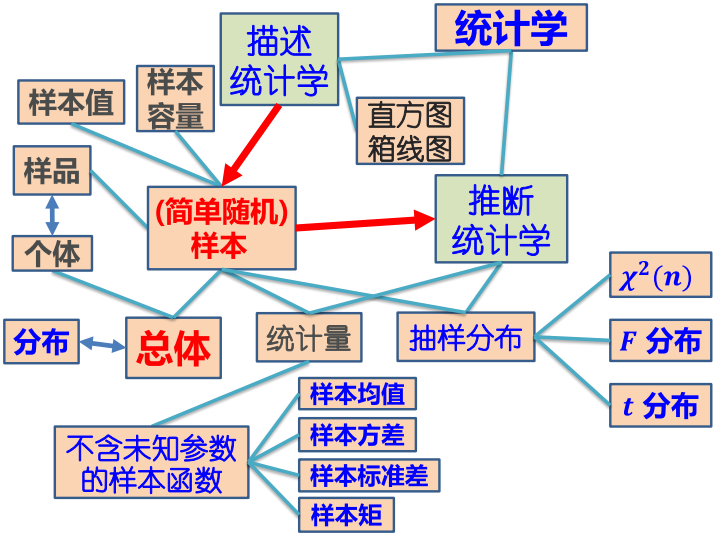
$$f(x|\vec{\eta}) = \exp\left[\sum_{i=1}^{s}\eta_i T_i(x) - A(\vec{\eta})\right]h(x)$$

其中 $\eta_i = \eta_i(\vec{\theta})$, $A[\vec{\eta}(\vec{\theta})] = B(\vec{\theta})$。

数据

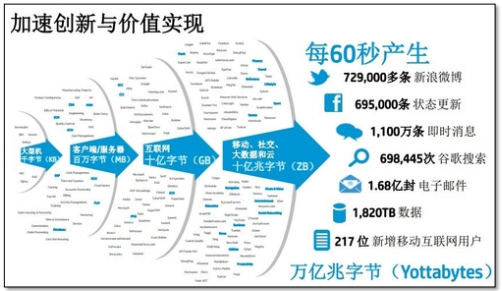
复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

概念地图



复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

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…

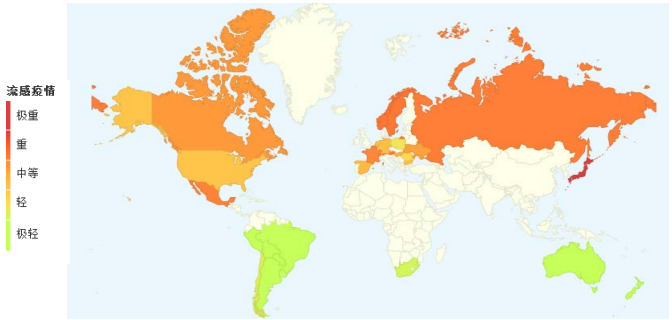
国际数据公司IDC报告：2011 年全球被创建和被复制的数据总量超过1.8 ZB，且增长趋势遵循新摩尔定律(全球数据量大约每两年翻一番)。

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

定义

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

统计学定义的宽泛性

统计学的定义，表明统计学包罗万象，甚至包含所有实证科学的步骤。

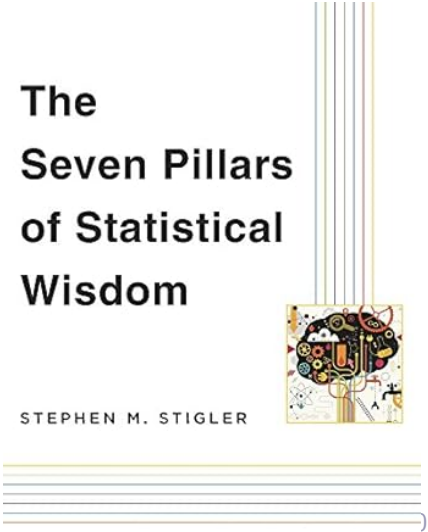
对所有学科的渗透

The best thing about being a statistician is that you get to play in everyone's backyard – John Tukey

同样，开展实验物理学研究，需要掌握统计学，甚至成为统计学家。

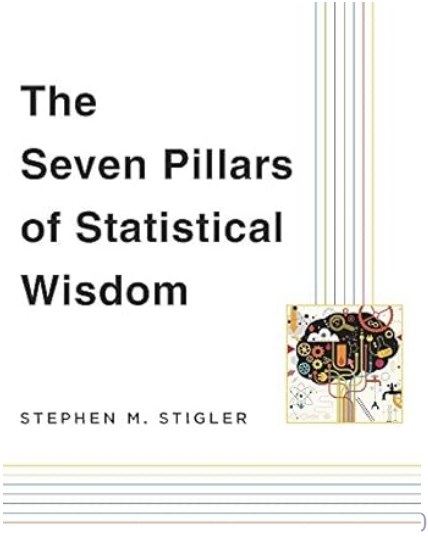
统计学的七大准则

1. Aggregation 概括
   削减信息让人获得更多信息。
2. Information Measurement 信息度量
   根号 n 准则。
3. Likelihood 似然与概率论
4. Intercomparison 完备性
   以统计学的方法论独立于科学领域完成分析。
5. Regression 回归
6. Design 实验设计
7. Residual 分离已知与未知



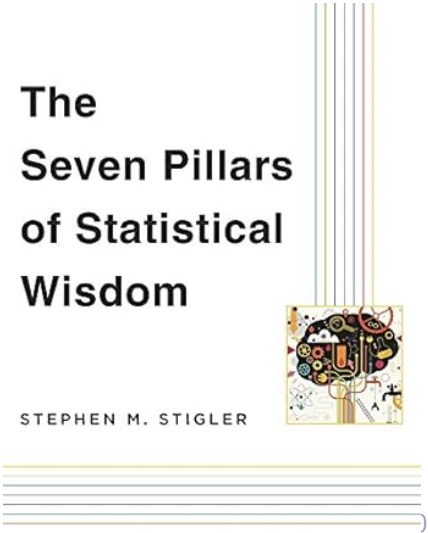
统计学的七大准则

1. Aggregation 概括
   削减信息让人获得更多信息。
2. Information Measurement 信息度量
   根号 n 准则。
3. Likelihood 似然与概率论
4. Intercomparison 完备性
   以统计学的方法论独立于科学领域完成分析。
5. Regression 回归
6. Design 实验设计
7. Residual 分离已知与未知



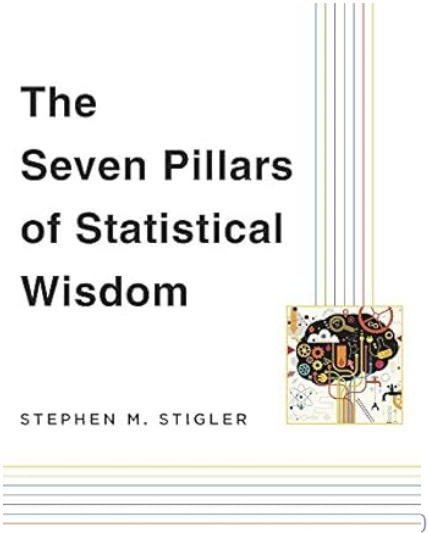
统计学的七大准则

1. Aggregation 概括
   削减信息让人获得更多信息。
2. Information Measurement 信息度量
   根号 n 准则。
3. Likelihood 似然与概率论
4. Intercomparison 完备性
   以统计学的方法论独立于科学领域完成分析。
5. Regression 回归
6. Design 实验设计
7. Residual 分离已知与未知



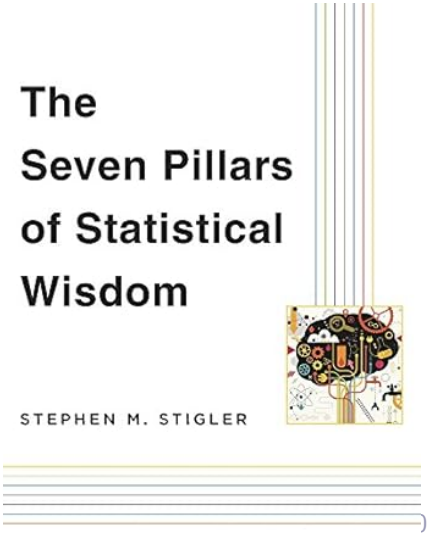
统计学的七大准则

1. Aggregation 概括
   削减信息让人获得更多信息。
2. Information Measurement 信息度量
   根号 n 准则。
3. Likelihood 似然与概率论
4. Intercomparison 完备性
   以统计学的方法论独立于科学领域完成分析。
5. Regression 回归
6. Design 实验设计
7. Residual 分离已知与未知



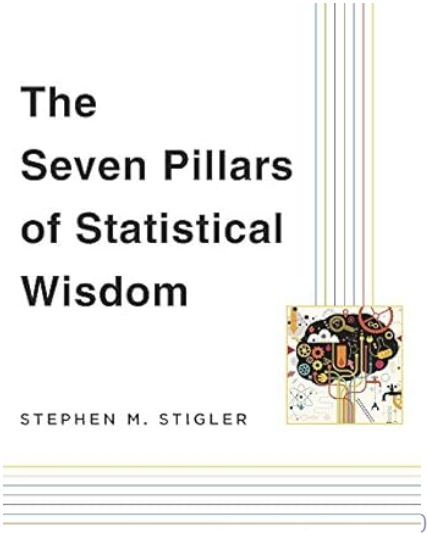
统计学的七大准则

1. Aggregation 概括
   削减信息让人获得更多信息。
2. Information Measurement 信息度量
   根号 n 准则。
3. Likelihood 似然与概率论
4. Intercomparison 完备性
   以统计学的方法论独立于科学领域完成分析。
5. Regression 回归
6. Design 实验设计
7. Residual 分离已知与未知



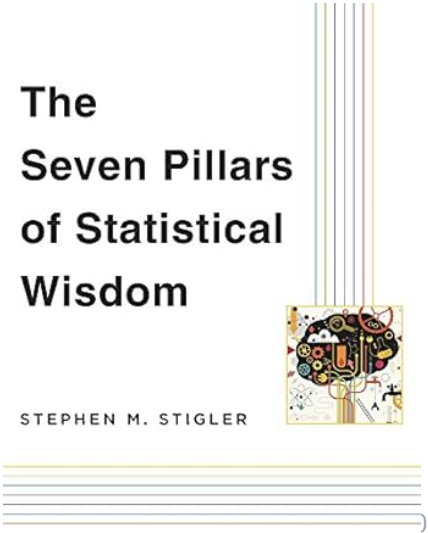
统计学的七大准则

1. Aggregation 概括
   削减信息让人获得更多信息。
2. Information Measurement 信息度量
   根号 n 准则。
3. Likelihood 似然与概率论
4. Intercomparison 完备性
   以统计学的方法论独立于科学领域完成分析。
5. Regression 回归
6. Design 实验设计
7. Residual 分离已知与未知



统计学的七大准则

1. Aggregation 概括
   削减信息让人获得更多信息。
2. Information Measurement 信息度量
   根号 n 准则。
3. Likelihood 似然与概率论
4. Intercomparison 完备性
   以统计学的方法论独立于科学领域完成分析。
5. Regression 回归
6. Design 实验设计
7. Residual 分离已知与未知



描述统计学 descriptive statistics

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

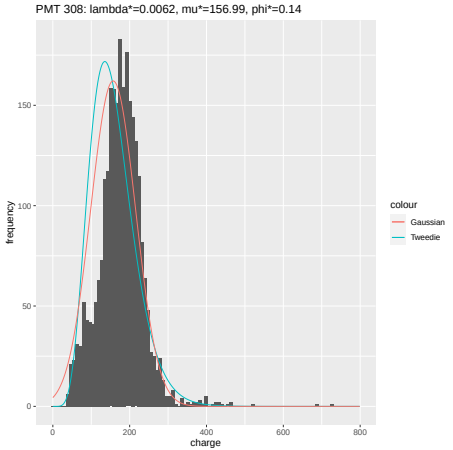
- 对随机现象进行观测、试验，以取得有代表性的观测值
- 收集、加工数据，并用图形、表格和数值方法来汇总数据的统计学

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



描述统计学 descriptive statistics

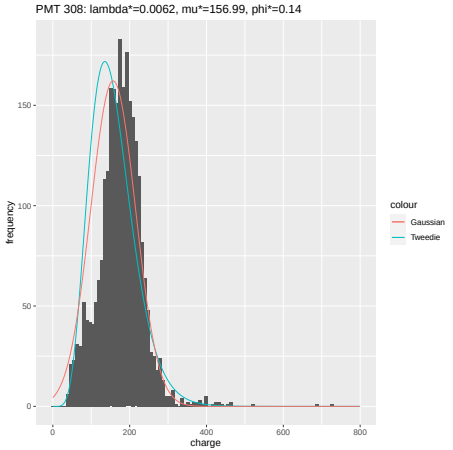
- 对随机现象进行观测、试验，以取得有代表性的观测值
- 收集、加工数据，并用图形、表格和数值方法来汇总数据的统计学

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



描述统计学 descriptive statistics

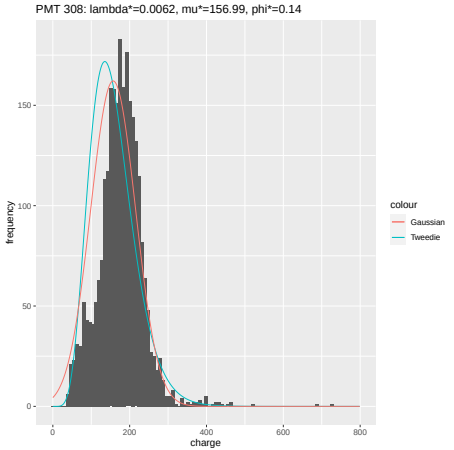
- 对随机现象进行观测、试验，以取得有代表性的观测值
- 收集、加工数据，并用图形、表格和数值方法来汇总数据的统计学

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



R 语言的图表

- R 语言积累了多种对概括数据用的图表：
  - 核心制图系统
  - ggplot 系统
- 让人快速理解数据。

直方图

数据的展示

- 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

数据的展示

- 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

例：伊特拉斯坎人的脑容量

```
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
```

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

stem-leaf 的字符型总结

```
stem(a_etruscan) # 字符的总结
```

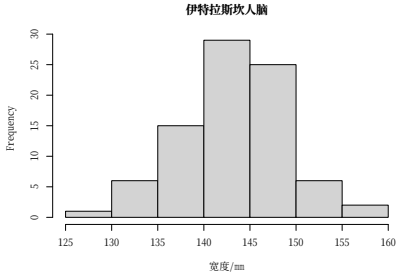
```
The decimal point is 1 digit(s) to the right of the |

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

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

画直方图

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")
```

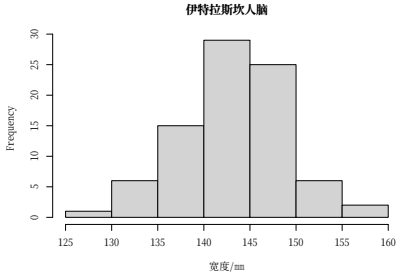


复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

画直方图

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")
```

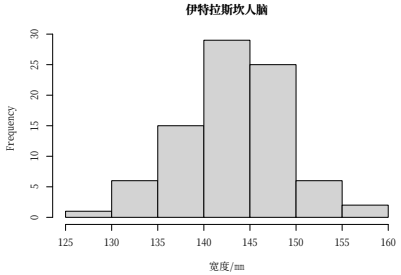


画直方图

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

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")
```

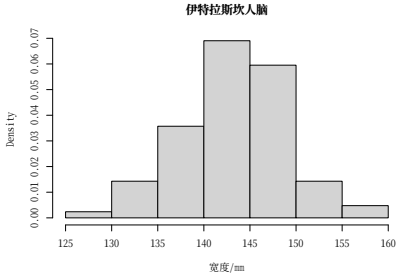


频数 vs. 频率直方图

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

- 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。

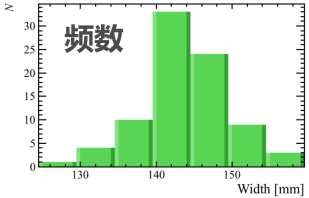
```
hist(a_etruscan, freq=FALSE, main="伊特拉斯坎人脑", xlab="宽度/mm")
```

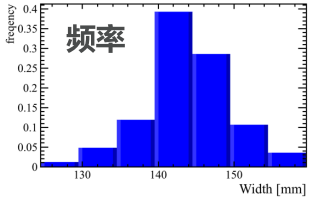


直方图的 4 种画法

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。

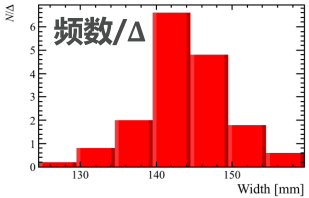
频数累加等于样本总量n

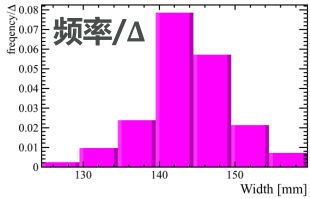




频率累加等于1

面积累加等于样本总量n





面积累加等于1

箱线图

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

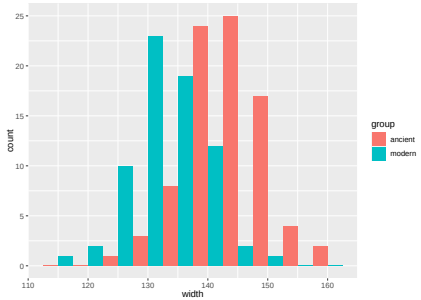
两种伊特拉斯坎人

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

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

叠叠画

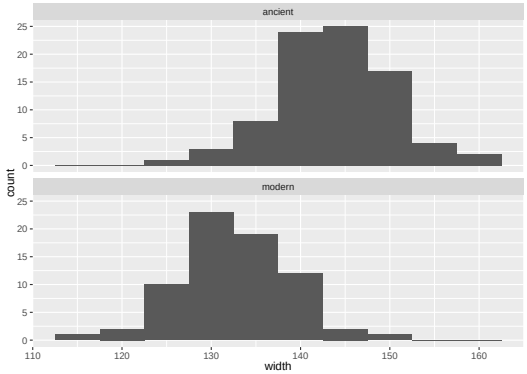
```
library(ggplot2)
p <- ggplot(etruscan, aes(x=width, fill=group)) + geom_histogram(binwidth=5, position = "dodge")
print(p)
```



排排画

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

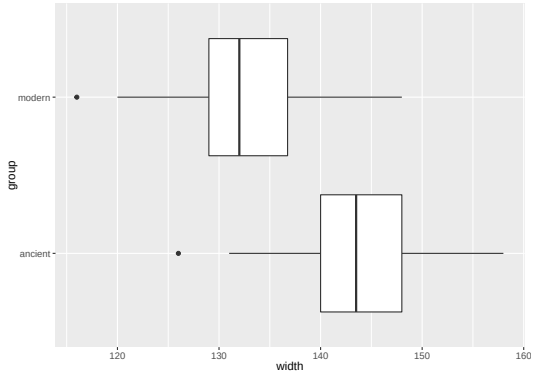
```
p <- ggplot(etruscan, aes(x=width)) + geom_histogram(binwidth=5) + facet_wrap(~group, nrow=2)
print(p)
```



箱线图

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量

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



分位数

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 。

样本分位数

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 。

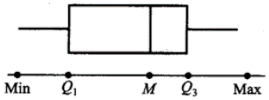
样本分位数：设有容量为 n 的样本观察值 $(x_1, x_2, \cdots, x_n)$，样本 p 分位数 $(0 < p < 1)$ 记为 x_p，它具有以下性质：

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。

箱线图的要素

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：

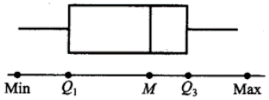
1. 画一水平数轴，轴上标上 Min, Q_1, M, Q_3, Max。在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于 Q_1、Q_3 的上方。在 M 点上方箱子内部画一条垂直线段。
1. 自箱子左侧引一条水平线直至最小值 Min；在同一水平高度自箱子右侧引一条水平线直至最大值。



箱线图的要素

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：

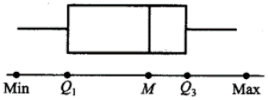
1. 画一水平数轴，轴上标上 Min, Q_1, M, Q_3, Max。在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于 Q_1、Q_3 的上方。在 M 点上方箱子内部画一条垂直线段。
2. 自箱子左侧引一条水平线直至最小值 Min；在同一水平高度自箱子右侧引一条水平线直至最大值。



箱线图的要素

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 Min, Q_1, M, Q_3, 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。



推断统计学

推断统计学

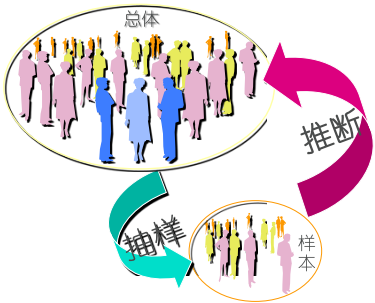
- 对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性
- 用样本数据对总体的某些特征进行估计和检验的统计学

内容

1. 参数估计
2. 假设检验

目的

- 对总体特征作出推断
  - 总体：随机变量



推断统计学

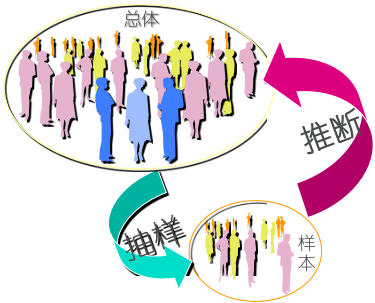
- 对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性
- 用样本数据对总体的某些特征进行估计和检验的统计学

内容

1. 参数估计
2. 假设检验

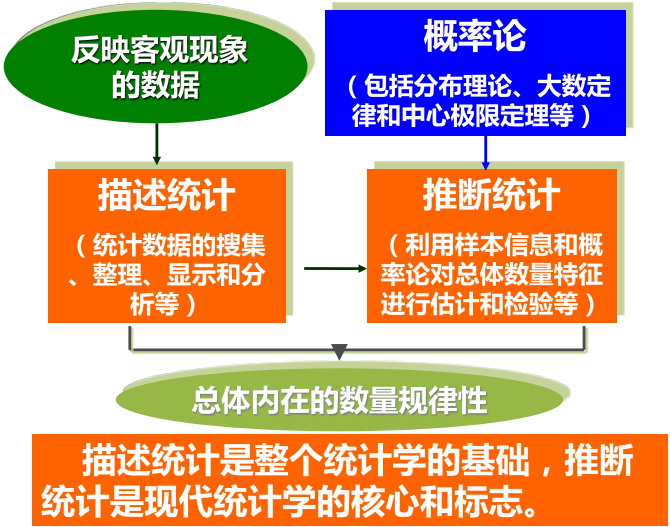
目的

- 对总体特征作出推断
  - 总体：随机变量



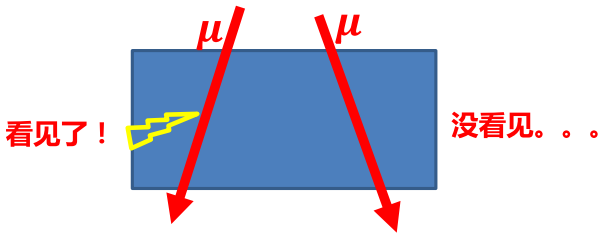
描述与推断相辅相成

复习
数据
统计学
直方图
箱线图
推断统计学
随机样本
统计量



随机样本

探测器效率：搭建一个宇宙线缪子（$\mu^{\pm}$）探测器。要求探测效率 $\epsilon > 95\,\%$.



探测器效率：

$$\epsilon = \frac{n_{\text{obs}}}{n_{\text{tot}}}$$

总体

定义 (总体)

研究对象全体元素组成的集合
所研究的对象的某个 (或某些) 数量指标的全体，它是一个随机变量 (或多维随机变量)，记为 X。可以有限，也可以无限。

X 的分布函数和数字特征称为总体的分布函数和数字特征.

三层含义

1. 研究对象的全体
2. 数据
3. 分布

总体

定义 (总体)

研究对象全体元素组成的集合
所研究的对象的某个 (或某些) 数量指标的全体，它是一个随机变量 (或多维随机变量)，记为 X。可以有限，也可以无限。

X 的分布函数和数字特征称为总体的分布函数和数字特征.
三层含义

1. 研究对象的全体
2. 数据
3. 分布

定义

个体

组成总体的每一个元素，即总体的每个数量指标，可看作随机变量 X 的某个取值。用 X_i 表示。

样本

从总体中抽取的部分个体.

- 用 $(X_1, X_2, \cdots, X_n)$ 表示， n 为样本容量。
- 称 $(x_1, x_2, \cdots, x_n)$ 为总体 X 的一个容量为 n 的样本观测值.

样本空间

样本所有可能取值的集合.

定义

个体

组成总体的每一个元素，即总体的每个数量指标，可看作随机变量 X 的某个取值。用 X_i 表示。

样本

从总体中抽取的部分个体.

- 用 $(X_1, X_2, \cdots, X_n)$ 表示， n 为样本容量。
- 称 $(x_1, x_2, \cdots, x_n)$ 为总体 X 的一个容量为 n 的样本观测值.

样本空间

样本所有可能取值的集合.

定义

个体

组成总体的每一个元素，即总体的每个数量指标，可看作随机变量 X 的某个取值。用 X_i 表示。

样本

从总体中抽取的部分个体.

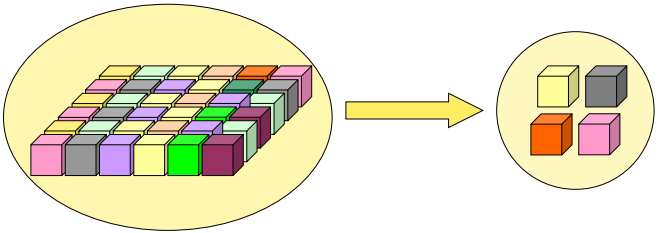
- 用 $(X_1, X_2, \cdots, X_n)$ 表示， n 为样本容量。
- 称 $(x_1, x_2, \cdots, x_n)$ 为总体 X 的一个容量为 n 的样本观测值.

样本空间

样本所有可能取值的集合.

样本的生成

从总体中抽取，并作为总体代表的一部分总体单位的集合体。



- 样本取自总体，不唯一

简单随机样本

若总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 满足:

1. $X_1, X_2, \cdots, X_n$ 与 X 有相同的分布
2. $X_1, X_2, \cdots, X_n$ 相互独立

则称 $(X_1, X_2, \cdots, X_n)$ 为 简单随机样本 .

抽样方法

- 对有限总体，放回抽样所得到的样本为简单随机样本。
- 如果放回抽样不方便, 常用不放回抽样代替，条件是 $N/n \geq 10$。N 为总体中个体总数， n 为样本容量。

简单随机样本

若总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 满足:

1. $X_1, X_2, \cdots, X_n$ 与 X 有相同的分布
2. $X_1, X_2, \cdots, X_n$ 相互独立

则称 $(X_1, X_2, \cdots, X_n)$ 为 简单随机样本 .

抽样方法

- 对有限总体，放回抽样所得到的样本为简单随机样本。
- 如果放回抽样不方便, 常用不放回抽样代替，条件是 $N/n \geq 10$。N 为总体中个体总数， n 为样本容量。

统计量

定义

样本 $(X_1, X_2, \cdots, X_n)$ 的不含有未知参数的连续函数 $g(X_1, X_2, \cdots, X_n)$ 称为 统计量 。

1. 利用样本的函数进行统计推断
2. 样本是随机变量，统计量也是随机变量。

例

考虑 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ ，如果参数 μ, σ^2 已知，则它是统计量，否则不是。

定义

样本 $(X_1, X_2, \cdots, X_n)$ 的不含有未知参数的连续函数 $g(X_1, X_2, \cdots, X_n)$ 称为 统计量 。

1. 利用样本的函数进行统计推断
2. 样本是随机变量，统计量也是随机变量。

例

考虑 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ ，如果参数 μ, σ^2 已知，则它是统计量，否则不是。

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本, 分别称下列统计量为

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, 例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本, 分别称下列统计量为

样本均值 $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

样本方差 $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$, 例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本, 分别称下列统计量为

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, 例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本, 分别称下列统计量为

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$，例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$

常见统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本, 分别称下列统计量为

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

样本标准差 $\sqrt{S^2}$

样本 k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$，例如 $A_1 = \overline{X}$

样本 k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \equiv S_n^2$$