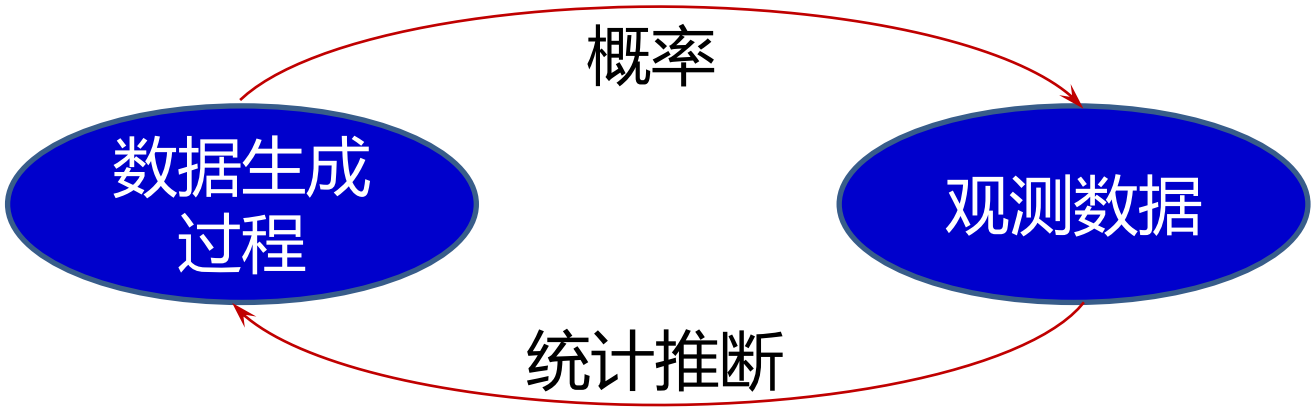
第六章 样本及抽样分布

§0 从概率到数理统计
§1 随机样本
§2 直方图、分位数与箱线图
§3 抽样分布

§0 从概率到数理统计

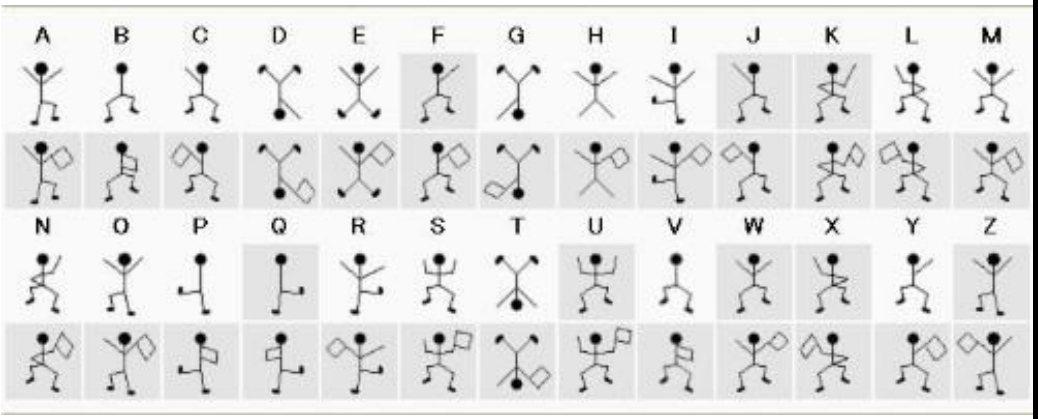


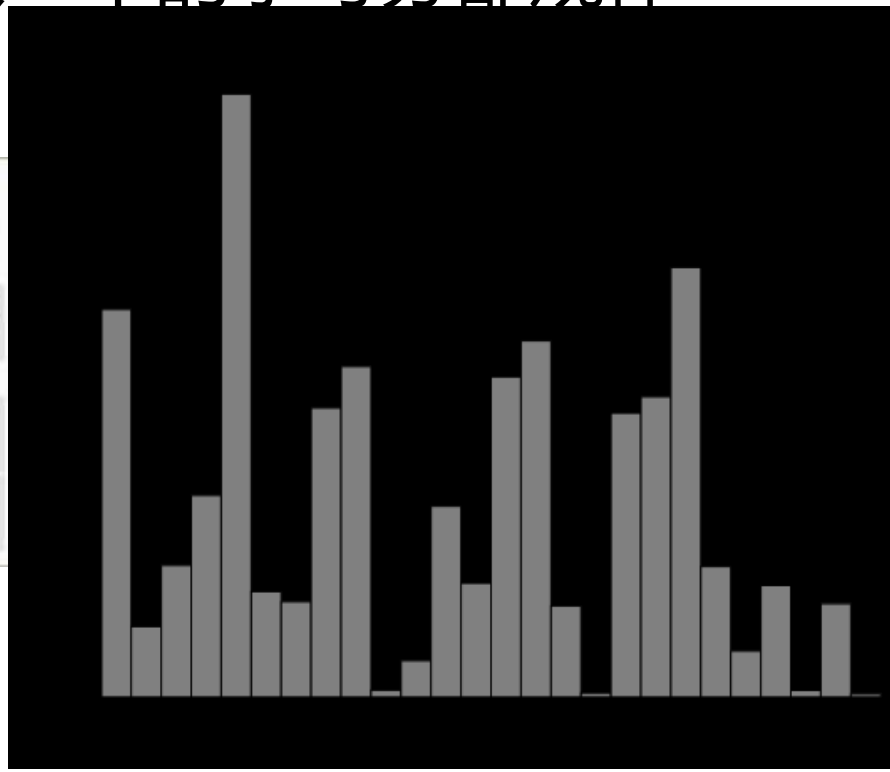
数理统计的研究内容

- 怎样有效地收集、整理和分析带有随机性的数据
- 如何量化不确定性
- 怎样对所考察的问题作出推断或预测
- 为决策和行动提供依据和建议

数理统计——解释现象背后的规律

- 大量随机现象必然呈现出它的规律性，因而从理论上讲，只要对随机现象进行足够多次观察，被研究的随机现象的规律性一定能清楚地呈现出来
  - 福尔摩斯“跳舞的小人”中的字母分部规律





数理统计——解释现象背后的规律

- 大量随机现象必然呈现出它的规律性，因而从理论上讲，只要对随机现象进行足够多次观察，被研究的随机现象的规律性一定能清楚地呈现出来
  - 福尔摩斯“跳舞的小人”中的字母分部规律
  - “模仿游戏”中“电报中固定位置密文的固定含义”Wetter(天气)“
  - 信号的盲检测

数理统计——解释现象背后的规律

- 大量随机现象必然呈现出它的规律性，因而从理论上讲，只要对随机现象进行足够多次观察，被研究的随机现象的规律性一定能清楚地呈现出来
- 只对随机现象进行次数不多的观察试验——只是局部观察资料
  - 数理统计的任务就是研究怎样有效地收集、整理、分析所获得的有限的资料
  - 对所研究的问题，尽可能地作出精确而可靠的结论

数理统计的基本方法

两个基本方法

- 参数估计
  - 根据数据,用一些方法对分布的未知参数进行**估计**
- 假设检验
  - 根据数据,用一些方法对分布的未知参数进行**检验**

统计推断的两个重要基础

- 收集数据——从总体$X \sim F(x)$抽取样本$X_1, X_2, \ldots, X_n$
- 加工整理数据——统计量

§1 随机样本

一、基本概念

1. 总体——试验的全部可能的观察值称为总体
   - 一个总体对应一个随机变量
   - 对总体的研究就是对一个随机变量X的研究
   - X的分布函数和数字特征就称为总体的分布函数和数字特征
   - 总体可以是具体事物的集合，如一批产品，也可以是关于事物的度量数据集合，如长度测量
2. 个体——总体中的每个可能观察值称为个体
   - 个体是随机试验的观察值
   - 总体中的个体，应当有共同的可观察的特征。该特征与研究目的有关

总体	个体	特征
一批产品	每件产品	等级
一批灯泡	每个灯泡	寿命
一年的日空气质量	每天的日空气质量	PM2.5
数轴上某一线段	线段中每一点	坐标
一批彩票	每张彩票	号码
全国人口寿命	每个人的寿命	寿命

3. 容量——总体中包含个体的数量称为总体容量

有限总体与无限总体

容量有限的总体称为有限总体，容量无限的称为无限总体

某工厂10月份生产的灯泡寿命所组成的总体中, 个体的总数就是10月份生产的灯泡数，这是一个有限总体

该工厂生产的所有灯泡寿命所组成的总体是一个无限总体, 它包括以往生产和今后生产的灯泡寿命

当有限总体包含的个体的总数很大时, 可近似地将它看成是无限总体

4. 总体分布——我们把数量指标取不同数值的比率叫做总体分布

例1. 在2000名大学一年级学生的年龄中, 年龄指标值为“15”，“16”，“17”，“18”，“19”，“20”的依次有9，21，132，1207，588，43 名

总体就是数集 {15, 16, 17, 18, 19, 20}

总体分布为：

年龄	15	16	17	18	19	20
比率	$\frac{9}{2000}$	$\frac{21}{2000}$	$\frac{132}{2000}$	$\frac{1207}{2000}$	$\frac{588}{2000}$	$\frac{43}{2000}$

5. 样本

- 设X是具有分布函数F的随机变量，若$X_1, X_2, \ldots, \mathrm{X}_n$是具有统一分布函数$F$、相互独立的随机变量，则称$X_1, X_2, \ldots, \mathrm{X}_n$ 为从分布函数F (或总体F、或总体X)得到的容量为n的**简单随机样本**，简称样本
- 观察值$x_1, x_2, \ldots, x_n$称为样本值，又称为X的n个独立的观察值
- 样本中所含个体的个数，称为样本容量

二、样本选取

选取样本的目的——从样本的特征对总体特征做出估计和推断。因此，抽样必须尽可能多地反映总体的特征。

选取样本是需要考虑的因素：

(1)独立性：抽样时互不影响。

(2)代表性：样本的分布与总体相同

抽样方式：

(1)不重复抽样(不放回)

(2)重复抽样(放回)

简单随机抽样

- 获得简单随机样本的抽样方法称为简单随机抽样
- 一般地，对于有限总体采用放回抽样就能得到简单随机样本；当总体的个数比样本容量大的多时，可将不放回抽样近似地当成放回抽样

简单随机抽样

由定义，若$X_1, X_2, \ldots, X_n$为F的一个样本，则$X_1, X_2, \ldots, X_n$的联合分布函数为

$$F^*(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F(x_i)$$

若设X的分布律为$P\{X = x\} = p(x)$，则$(X_1, X_2, \ldots, X_n)$的联合分布律为

$$P\{X_1 = x_1, \ldots, X_n = x_n\} = \prod_{i=1}^{n} p(x_i)$$

又若X具有概率密度f，$X_1, X_2, \ldots, X_n$的联合概率密度为

$$f^*(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

例2. 设总体X服从参数为$\lambda(\lambda>0)$的指数分布，$(X_1,X_2,\dots,X_n)$是来自总体的样本，求样本$(X_1,X_2,\dots,X_n)$的概率密度。

解：总体X的概率密度为

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

因为$X_1,X_2,\dots,X_n$相互独立，且与X有相同的分布，所以$(X_1,X_2,\dots,X_n)$的概率密度为

$$f_n(x_1,x_2,\dots,x_n)=\prod_{i=1}^{n}f(x_i)$$

$$=\begin{cases}\lambda^n\, e^{-\lambda\sum_{i=1}^{n}x_i}, & x_i>0\\ 0, & \text{其他}\end{cases}$$

例3. 设总体X服从两点分布 $B(1,p)$ ，其中 $0<p<1$ ，$(X_1,X_2,\dots,X_n)$是来自总体的样本，求样本 $(X_1,X_2,\dots,X_n)$的分布律。

解：总体X的分布率为

$$P\{X=i\}=p^i(1-p)^{1-i},\qquad (i=0,1)$$

所以$(X_1,X_2,\dots,X_n)$的分布律为

$$P\{X_1=x_1,X_2=x_2,\dots,X_n=x_n\}$$
$$=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

其中，$x_1,x_2,\dots,x_n$在集合{0,1}中取值。

例4. 若$X_1, X_2, \dots, X_n$是参数为λ的泊松分布总体X的样本，求$(X_1, X_2, \dots, X_n)$的联合分布律。

解：总体X的分布律为

$$p(x) = P\{X = x\} = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0,1,2, \dots$$

所以$(X_1, X_2, \dots, X_n)$的联合分布率为

$$P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} = \prod_{i=1}^{n} p(x_i)$$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} e^{-n\lambda},$$

$$x_i = 0,1, \dots, i = 1,2, \dots, n$$

例5. 若$X_1, X_2, \dots, X_n$是总体$X \sim N(\mu, \sigma^2)$的样本，求$(X_1, X_2, \dots, X_n)$的联合概率密度。

解：总体X的概率密度为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$$

于是$(X_1, X_2, \dots, X_n)$的联合概率密度

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(x_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}},$$

$$-\infty < x_i < +\infty, i = 1,2, \dots, n$$

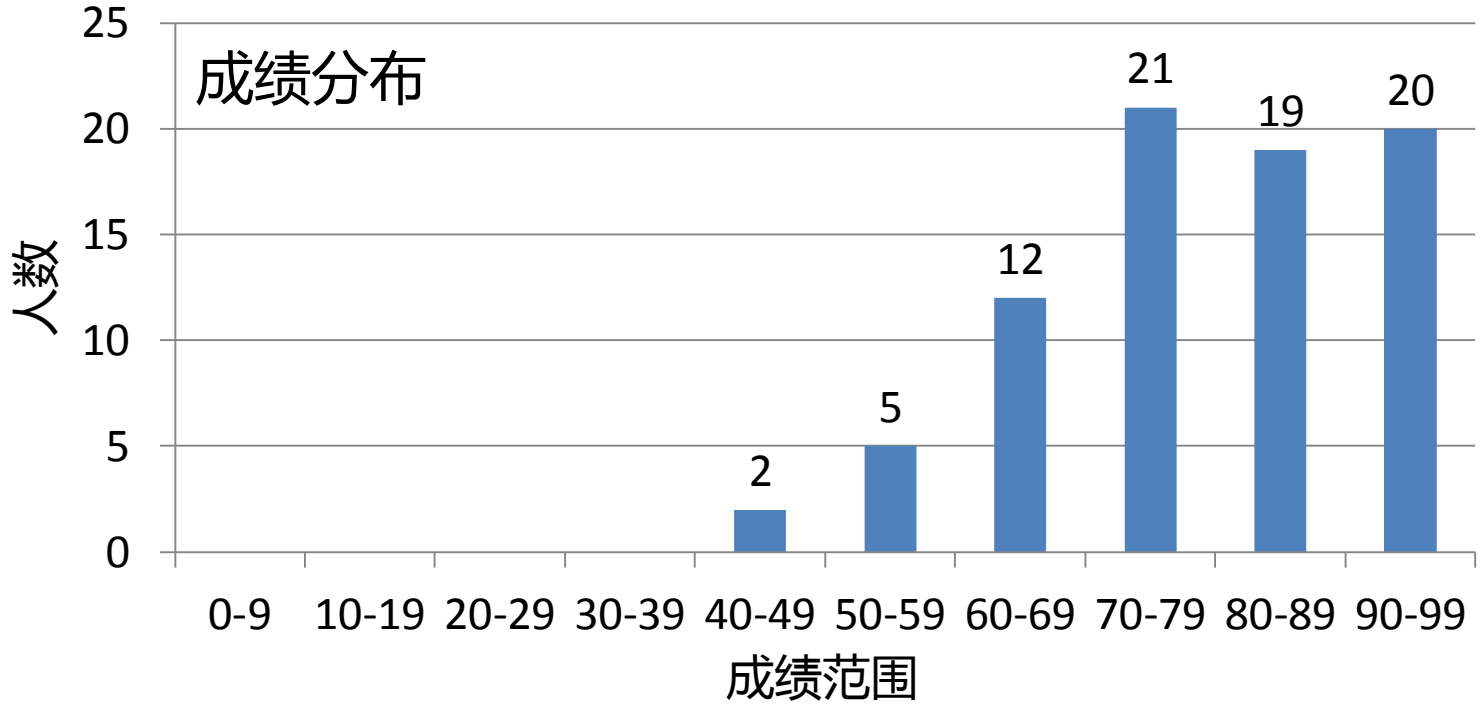
概念对比——概率、统计

概率	数理统计
样本空间	总体
随机事件	样本
样本点	个体

§2 直方图、分位数与箱线图

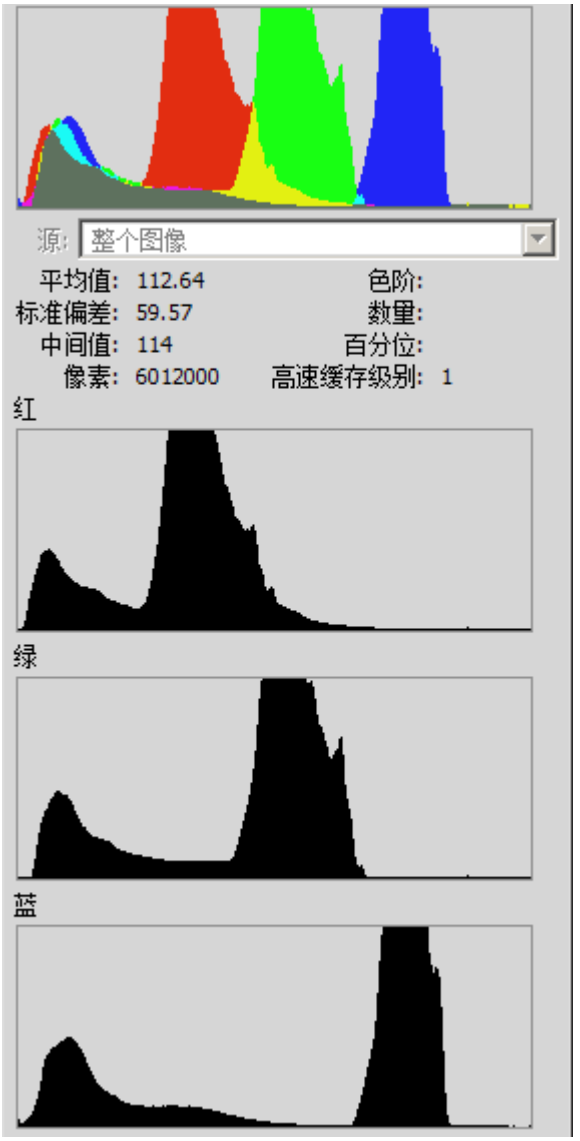
一、直方图

- 数值分布的一种图形化表示手段，可以很好地表达概率分布，便于从总体上把握分布
- 由Karl Pearson在1895年首次引入

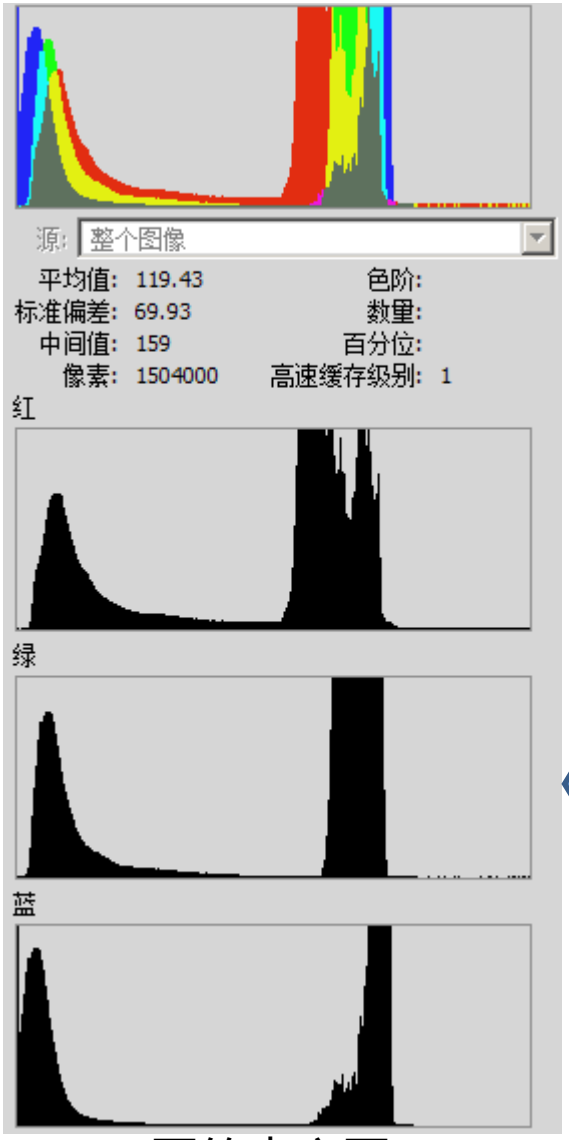


在图像中的应用

Size: ~2000* 3000



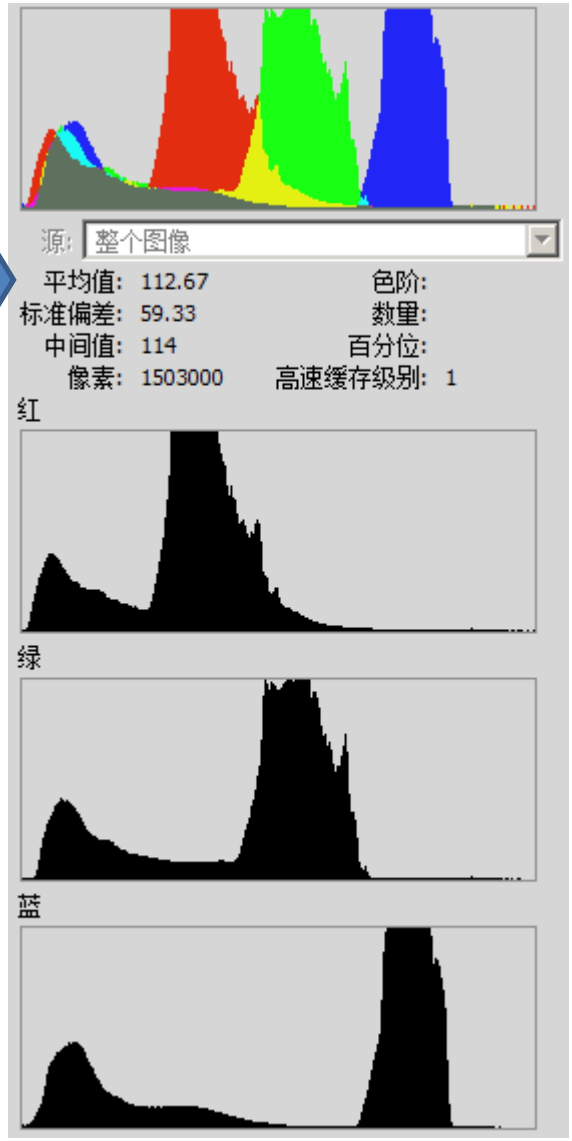
在图像中的应用



下图的直方图

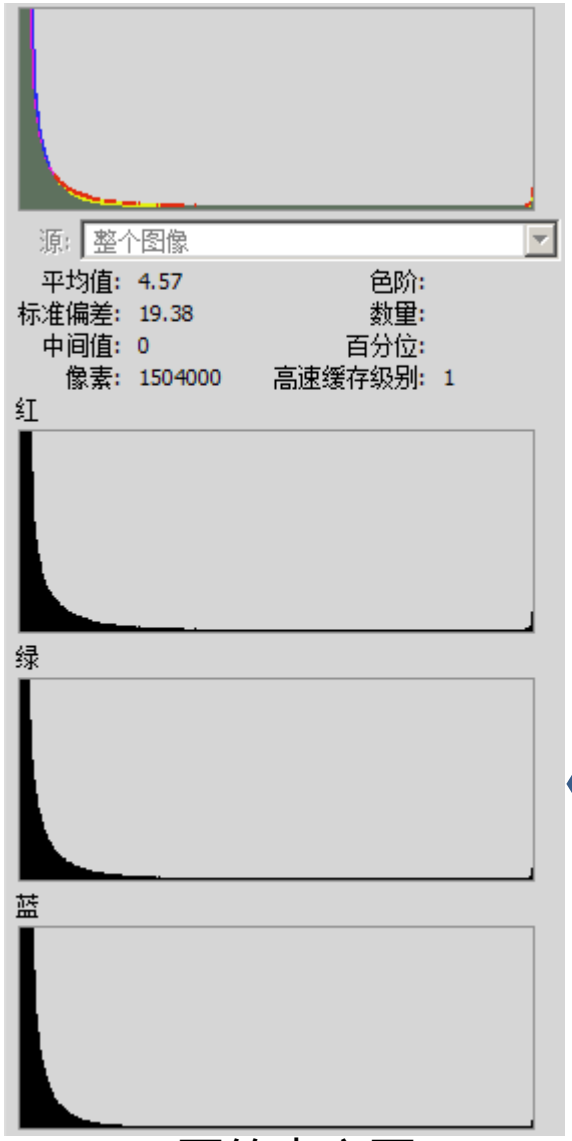
Size: ~1000* 1500

Size: ~1000* 1500



上图的直方图

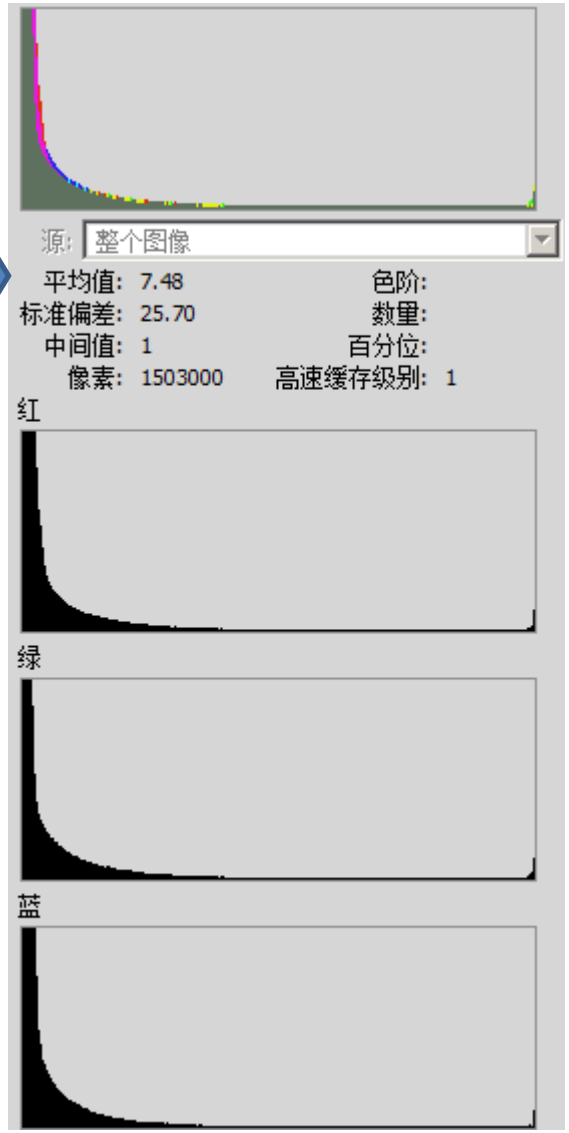
在图像中的应用



下图的直方图

Sobel 垂直边缘

Sobel 垂直边缘



上图的直方图

分位数

设有容量为n的样本观察值$x_1, x_2, \dots, x_n$，样本的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

将$x_1, x_2, \dots, x_n$排序，保证$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$

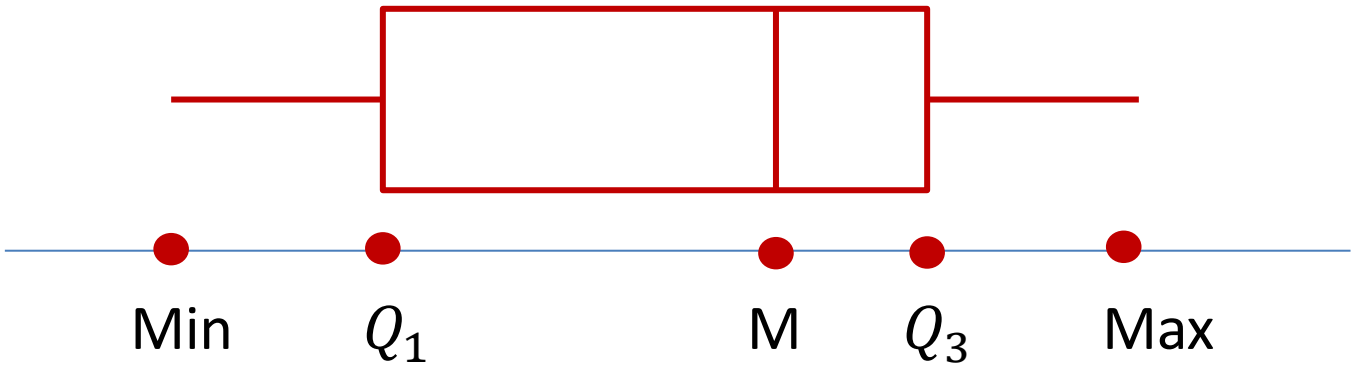
$$x_p = \begin{cases} x_{(\lceil np \rceil)}, & np\text{不是整数} \\ \frac{1}{2}\left[x_{(np)} + x_{(np+1)}\right], & np\text{为整数} \end{cases}$$

$p=0.5$时$x_{0.5}$记为Q_2或M，称为样本中位数，即

$$x_{0.5}=\begin{cases} x_{\left(\left\lceil\frac{n}{2}\right\rceil\right)}, & n\text{是奇数} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right], & n\text{是偶数} \end{cases}$$

$x_{0.25}$记为Q_1，称为**第一四分位数**，$x_{0.75}$记为Q_3，称为**第三四分位数**

箱线图由最小值Min、第一四分位数Q_1、样本中位数M、第三四分位数Q_3和最大值Max构成



四分位数间距——第一四分位数Q_1与第三四分位数Q_3的间距，记为IQR (Interquartile Range)

如果数据小于$Q_1 - 1.5IQR$或大于$Q_3 + 1.5IQR$，而认为异常，此时的箱线图就需要做出修正，排除异常值

例6. 下面是21个病人住院的时间(天)，试画出修正的箱线图

1, 2, 3, 3, 4, 4, 5, 6, 6, 7, 7, 9, 9, 10, 12, 12, 13, 15, 18, 23, 55

解： M=7 ($\lceil 21 * 0.5 \rceil = 11th$)，

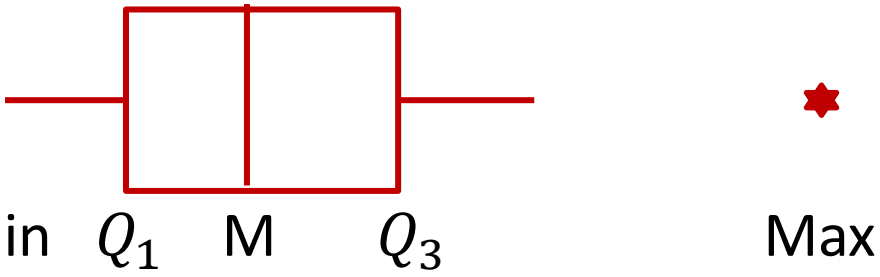
$Q_1 = 4$ ($\lceil 21 * 0.25 \rceil = 6th$)

$Q_3 = 12$ ($\lceil 21 * 0.75 \rceil = 16th$)

Min=1， Max=55， $IQR = 8$，

$Q_1 - 1.5IQR = -8, Q_3 + 1.5IQR = 24$

于是“55”是Outlier，修正最大值



§3 抽样分布

统计推断的两个重要基础

收集数据——从总体$X \sim F(x)$抽取样本$X_1, X_2, \ldots, X_n$

加工整理数据——统计量

1.统计量

定义(统计量)：若$X_1, X_2, \dots, X_n$是来自总体X 的一个样本， $g(X_1, X_2, \dots, X_n)$是$X_1, X_2, \dots, X_n$的函数，若g中不含任何未知参数，则称$g(X_1, X_2, \dots, X_n)$是一统计量。

关于统计量的一些理解：

- 统计量是由随机变量组成的随机样本的函数，不含任何未知参数
- 注意：统计量是随机变量，
- 于是$x_1, x_2, \dots, x_n$是相应于样本$X_1, X_2, \dots, X_n$的样本值，而称$g(x_1, x_2, \dots, x_n)$是$g(X_1, X_2, \dots, X_n)$的观察值

例7. 设$X_1, X_2, \ldots, X_n$为来自总体$X \sim N(\mu, \sigma^2)$的一个样本，其中μ未知，σ^2已知，问下列随机变量中那些是统计量。

$\frac{X_1 + X_2}{2}$；

$\frac{X_1 + \cdots + X_n}{n} - \mu$；

$\frac{(X_1 + X_2)^2}{\sigma^2}$

常用统计量

样本均值

反映了总体均值的信息

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

它反映了总体方差的信息

样本方差

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$$

样本标准差

$$S=\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)}$$

常用统计量

它反映了总体k 阶矩的信息

样本k阶原点矩

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad k = 1,2,\ldots$$

样本k阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k, \qquad k = 1,2,\ldots$$

它反映了总体k 阶中心矩的信息

上述统计量与数字特征的差异

- 统计量是随机变量
- 数字特征是常数

对应观察值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \,, b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \,, \qquad k = 1,2,\ldots$$

若总体X的k阶矩$E(X^k)$记为μ_k存在，则当$n \to \infty$时，

$A_k \xrightarrow{P} \mu_k, k = 1,2, \ldots$

证明：$X_1, X_2, \ldots, X_n$独立且与X同分布，

于是$X_1^k, X_2^k, \ldots, X_n^k$独立且与$X^k$同分布，

从而有

$$E(X_1^k) = E(X_2^k) = \cdots, = E(X_n^k) = \mu_k$$

由Khinchin大数定理

$$A_k \xrightarrow{P} \mu_k, k = 1,2, \ldots$$

进一步，由依概率收敛的序列性质知道，若g为连续函数，则有

$$g(A_1, A_2, \ldots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \ldots, \mu_k)$$

矩估计法的基础

概念对比——概率、统计

概率	数理统计	计算机科学(PR&ML)
样本空间	总体	
随机事件	样本	训练样本集
样本点	个体	单个样本
方差	方差	类内距离
	估计	学习
协方差	协方差	特征

2. 经验分布函数

——与总体分布函数F(x)相对应的统计量

设$X_1, X_2, \dots, X_n$是总体F的一个样本。用$S(x)$，$-\infty < x < +\infty$表示$X_1, X_2, \dots, X_n$中不大于x的随机变量个数，定义经验分布函数为：

$$F_{n(x)} = \frac{1}{n} S(x), -\infty < x < +\infty$$

例8. 设总体F具有一个样本值1, 2, 3，经验分布函数$F_3(x)$的观察值

$$F_3(x)=\begin{cases}0, & \text{若}x<1,\\ \frac{1}{3}, & \text{若}1\le x<2\\ \frac{2}{3}, & \text{若}2\le x<3\\ 1, & \text{若}x\ge 3\end{cases}$$

例9. 设总体F具有一个样本值1, 1, 2，经验分布函数$F_3(x)$的观察值

$$F_3(x)=\begin{cases}0, & \text{若}x<1,\\ \frac{2}{3}, & \text{若}1\le x<2\\ 1, & \text{若}x\ge 2\end{cases}$$

一般地，设$x_1, x_2, \dots, x_n$是总体F的一个样本值，先将$x_1, x_2, \dots, x_n$按自小到大排序，并重新编号满足：

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

则经验分布函数$F_n(x)$的观察值为：

$$F_n(x) = \begin{cases} 0, & \text{若} x < x_{(1)} \\ \dfrac{k}{n}, & \text{若} x_{(k)} \le x < x_{(k+1)} \\ 1, & \text{若} x \ge x_{(n)} \end{cases}$$

对于经验分布函数$F_n(x)$，格里汶科1933年证明了:

对一切实数x，当$n \to \infty$时，$F_n(x)$以概率1一致收敛于分布函数$F(x)$，即：

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |F_n(x) - F(x)| = 0\right\} = 1$$