# 第一章

# 统计学的基本概念(1)

例　小儿麻痹症是20世纪五十年代的一种流行病，对于一种疫苗的有效性检验，收集了20万儿童并随机分成两组：实验组和对照组．试验结果是对照组中有138个儿童受到感染，而实验组中则有56个受到感染．试根据这组数据判断疫苗的有效性．

统计假设检验的结果表明138与56的差异是高度显著的，因此该疫苗统计有效．

# 本章要点

一、直方图

二、总体与样本

三、经验分布函数

四、统计量

五、三大常用分布

六、抽样分布

# §1.1　直方图

## 一、直方图的做法：例1.1

直方图近似描绘了连续型随机变量的概率密度函数.

统计学主要研究如何以有效的方法收集、整理与分析带有随机性影响的数据，从而对所考察的问题作出推断和预测，进一步为采取某种决策提供理论依据和建议.

# §1.2　总体与样本

我们把研究对象的全体称为总体，组成总体的每个成员称为个体. (以例1.1为对照)

特指：

研究对象的某项数量指标的全体称为总体，组成总体的每个成员的该项数量指标称为个体.

# 总体分布

总体指标是一个随机变量，记为 $X : X \sim f(x,\theta)$

当总体 $X$ 是离散型随机变量时，定义总体分布并记为

$f(x,\theta) \triangleq P(X = x)$ ，即为总体 $X$ 的概率函数.

当总体 $X$ 是连续型随机变量时，定义总体分布为

$f(x,\theta) \triangleq f_X(x)$ ，即为总体 $X$ 的概率密度函数.

总体分布的记号一致但指向不同!

而 $F_X(x) = F(x) \triangleq P(X \le x),\quad x \in R$ 称为总体 $X$ 的

分布函数.

例1　设总体 $X \sim B(1, p)$，试写出总体分布 $f(x, p)$.

解　$f(x, p) \triangleq P(X = x) = p^x(1-p)^{1-x}, x = 0, 1.$

例2　设总体 $X \sim P(\lambda)$，试写出总体分布 $f(x, \lambda)$.

解　$f(x, \lambda) \triangleq P(X = x) = \dfrac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, x = 0, 1, 2, \cdots.$

例3　设总体 $X \sim R(0,\theta)$, 写出总体分布 $f(x,\theta)$.

解

$$f(x,\theta) \triangleq f_X(x) = \begin{cases} \dfrac{1}{\theta}, & 0 < x < \theta, \\ 0, & \text{其余}. \end{cases}$$

例4　设总体 $X \sim N(\mu,\sigma^2)$, 写出总体分布 $f(x,\mu,\sigma^2)$.

解

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

# 样本

在统计学中，总体分布往往是未知的，有时虽然总体分布的类型已知，但分布中含有未知参数，如所作直方图中数据，从直方图判断食品重量服从正态分布，但是均值 $\mu$ 和方差 $\sigma^2$ 这两个参数是未知的.

我们需要一些观测数据来求得 $\mu$ 和 $\sigma^2$，即进行试验. 但是考虑到经济因素，考虑到有些试验又是破坏性试验，如测试灯泡的寿命等，在这种情形下，就要考虑数据的选取方式.

我们希望从客观存在的总体中按一定原则选取一些个体、即抽样，通过对这些个体作观察或测试来推断关于总体分布的某些量，例如总体 $X$ 的均值、方差、中位数等.

这些选取的个体便称为取自总体的一个样本，这些个体的观测值称为样本观测值.

在抽样前，样本的观测值是不确定的. 为了体现其随机性，在统计学中把样本记作 $(X_1, X_2, \cdots, X_n)$. 事实上是一个 **n** 维随机向量. 抽样后通过试验或观测得到的数值称为样本观测值，记作 $(x_1, x_2, \cdots, x_n)$，是一组数据或看成 **n** 维空间中的一个点. 称 $n$ 为样本大小，或样本容量.

在实际工作中，我们通常把看到的一堆数据称为样本．但是严格来讲：样本不是数据而是一组随机变量.

数据是抽样完成以后得到的一次观测值，而统计学则研究如何利用样本的信息在抽样前就某个问题制定“方针政策”．

# 简单随机样本

从总体中抽取样本的方法有很多种，主要并且常用的就是所谓的简单随机抽样方法，即有放回重复独立的抽取，这样得到的样本便称之为简单随机样本.

由抽样方式即可知，简单随机样本具有以下两个特点：

1. 独立性：随机变量 $X_1, X_2, \cdots, X_n$ 是相互独立的；
2. 代表性：每个个体 $X_i$ 的分布都和总体分布相同；

即 $X_i \sim f(x_i, \theta), i = 1, 2, \cdots, n.$

问题：如何求样本的联合分布即联合概率密度函数或联合概率函数?

设总体 $X$ 为离散型随机变量，$X \sim f(x,\theta) \triangleq P(X=x)$.

则样本 $(X_1, X_2, \cdots, X_n)$ 的联合概率函数定义为:

$$
\begin{aligned}
f^*(x_1,x_2,\cdots,x_n,\theta) &\triangleq P(X_1=x_1, X_2=x_2,\cdots,X_n=x_n) \\
&= P(X_1=x_1)P(X_2=x_2)\cdots P(X_n=x_n) \\
&= f(x_1,\theta)f(x_2,\theta)\cdots f(x_n,\theta) \\
&= \Pi_{i=1}^{n} f(x_i,\theta), \qquad x_i \in R,\ i=1,\cdots,n.
\end{aligned}
$$

**例5** 设总体 $X \sim B(1, p)$, $(X_1, X_2, \cdots, X_n)$是取自该总体的一个样本, 试写出样本 $(X_1, X_2, \cdots, X_n)$ 的联合概率函数.

**解**

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

$$= \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, \quad x_i = 0, 1,\ i = 1, 2, \cdots, n.$$

例6 设总体 $X \sim P(\lambda)$, $(X_1, X_2, \cdots, X_n)$为取自该总体的一个样本，写出样本$(X_1, X_2, \cdots, X_n)$的联合概率函数.

解 $f^*(x_1, x_2, \cdots, x_n, \lambda) = \prod_{i=1}^{n} f(x_i, \lambda) = \prod_{i=1}^{n} P(X_i = x_i)$

$$= e^{-\lambda}\frac{\lambda^{x_1}}{x_1!}\cdots e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} = e^{-n\lambda}\cdot\frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1!x_2!\cdots x_n!}$$

$$x_i = 0, 1, 2, \cdots, \quad i = 1, 2, \cdots, n.$$

设总体 $X$ 为连续型随机变量，密度函数为 $X \sim f(x,\theta)$，则样本 $(X_1, X_2, \cdots, X_n)$ 的联合密度函数为:

$$f^*(x_1, x_2, \cdots, x_n, \theta) \triangleq f_{(X_1, X_2 \cdots, X_n)}(x_1, x_2, \cdots, x_n)$$

$$= f(x_1,\theta)\cdot f(x_2,\theta)\cdots f(x_n,\theta)$$

$$= \prod_{i=1}^{n} f(x_i,\theta)$$

例7　设总体 $X \sim R(0,\theta)$, $(X_1, X_2, \cdots, X_n)$ 为取自该总体的一个样本，写出样本 $(X_1, X_2, \cdots, X_n)$ 的联合密度函数.

解　因总体 $X \sim R(0,\theta)$，因此相应的密度函数为

$$f(x,\theta)=\begin{cases}\dfrac{1}{\theta}, & 0<x<\theta, \\ 0, & \text{其余}.\end{cases}$$

因此，样本的联合密度函数为

$$f^*(x_1,x_2,\cdots,x_n,\theta)=\begin{cases}\dfrac{1}{\theta^n}, & 0<x_i<\theta, i=1,2,\cdots n, \\ 0, & \text{其余}.\end{cases}$$

例8　设总体 $X \sim N(\mu,\sigma^2)$, $(X_1, X_2, \cdots, X_n)$ 为取自该总体的一个样本，写出 $(X_1, X_2, \cdots, X_n)$ 的联合密度函数.

解　因 $X \sim N(\mu,\sigma^2)$，相应的密度函数为

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

因此，样本的联合密度函数为

$$f^*(x_1, x_2, \cdots, x_n, \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\cdot\sigma} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

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

$$-\infty < x_i < +\infty, \quad i = 1, 2, \cdots, n.$$

例9　设$(X_1,X_2,\cdots,X_n)$是取自总体 $X$ 的一个样本，$X$ 的概率密度函数为

$$f(x,\theta)=\begin{cases}\dfrac{2x}{\theta^2}, & 0<x<\theta,\\ 0, & \text{其余}.\end{cases}$$

试写出 $(X_1,X_2,\cdots,X_n)$ 的联合密度函数.

解　样本的联合密度函数为

$$f^*(x_1,x_2,\cdots,x_n,\theta)=\begin{cases}\dfrac{2^n x_1x_2\cdots x_n}{\theta^{2n}}, & 0<x_i<\theta, i=1,2,\cdots,n\\ 0, & \text{其余}.\end{cases}$$

例10　设 $(X_1, X_2, \cdots, X_n)$ 是取自总体 $X$ 的一个样本，$X$ 的概率函数为

$$f(x, p) = P(X = x) = p(1-p)^{x-1}, x = 1, 2, \cdots$$

试写出样本 $(X_1, X_2, \cdots, X_n)$ 的联合概率函数.

解　样本的联合概率函数为

$$f^*(x_1, x_2, \cdots, x_n, p) = p^n (1-p)^{\sum_{i=1}^{n} x_i - n}$$

$$x_i = 1, 2, \cdots, i = 1, 2, \cdots, n$$

# §1.3　经验分布函数

设$(X_1, X_2, \cdots, X_n)$是取自总体 $X$ 的一个样本，$(x_1, \cdots, x_n)$是样本观测值，定义函数

$$F_n(x) = \frac{1}{n} \cdot \{X_1, \cdots, X_n \text{中小于或等于} x \text{的个数}\},$$

$$-\infty < x < +\infty$$

则称该函数为随机变量(或总体) $X$ 的经验分布函数.

相应地，称函数

$$\tilde{F}_n(x) = \frac{1}{n} \cdot \{x_1, \cdots, x_n \text{ 中小于或等于 } x \text{ 的个数}\},$$

$$-\infty < x < +\infty$$

为经验分布函数的观测值.

例11　设有一组样本观测值 ( 5, 3, 7, 5, 4 )，试写出经验分布函数的观测值.

解(1)　由定义　　**（5,3,7,5,4）**

$$\tilde{F}_5(x)=\frac{1}{5}\cdot\{x_1,\cdots,x_5 \text{ 中小于或等于} x \text{ 的个数}\}$$

$$=\begin{cases} 0, & x<3, \\ \frac{1}{5}, & 3\le x<4, \\ \frac{2}{5}, & 4\le x<5, \\ \frac{4}{5}, & 5\le x<7, \\ 1, & x\ge 7. \end{cases}$$

解(2)　先写出样本观测值的频率分布：　　（5,3,7,5,4）

| 样本观测值 | 3 | 4 | 5 | 7 |
|---|---|---|---|---|
| 频率 | $\frac{1}{5}$ | $\frac{1}{5}$ | $\frac{2}{5}$ | $\frac{1}{5}$ |

把它看成是某个离散型随机变量的概率函数，那么与它对应的分布函数便是经验分布函数的观测值. (略)

# 经验分布函数和总体分布函数的联系

**定理1.1** 设 $(X_1, X_2, \cdots, X_n)$ 是取自总体 $X$ 的一个样本, $F(x)$ 为总体分布函数，则对任意实数 $x \in (-\infty, +\infty)$ 以及任意的 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P(|F_n(x) - F(x)| \geq \varepsilon) = 0, \quad -\infty < x < +\infty$$

即 $F_n(x) \xrightarrow{P} F(x)$, 当 $n \to \infty$ 时.

证明主要步骤：(1)对任意实数 $x\in(-\infty,+\infty)$，设

$$Y_i=\begin{cases}1, & X_i\le x,\\ 0, & X_i>x.\end{cases}\quad i=1,2,\cdots,n$$

，则

$Y_i\sim B(1,p)$，且 $p=P(X_i\le x)=P(X\le x)=F(x)$;

(2)记 $\overline{Y}=\dfrac{1}{n}\sum_{i=1}^{n}Y_i$, 则 $\overline{Y}=F_n(x), E(\overline{Y})=p$;

(3)由大数定律 $\overline{Y}\xrightarrow{P}p$，此即 $F_n(x)\xrightarrow{P}F(x)$.

或 $\lim\limits_{n\to\infty}P(\left|F_n(x)-F(x)\right|\ge\varepsilon)=0.$