概率论与数理统计

崔文泉 (wqcui@ust.edu.cn)

2021 Autumn

第一章　基本概念

目录

§1.1 引言

Case study 1: Who Are Those Speedy Drivers?

在Penn State University 作了一个调查，被调查者要回答他们开车的最大速度？随机采访了87位男士和102位女士，得到数据如下：(单位: mph)

```
> male
 110 109  90 140 105 150 120 110 110  90 115  95 145 140 110 105 85  95 100
 115 124  95 100 125 140  85 120 115 105 125 102  85 120 110 120 115 94 125
  80  85 140 120  92 130 125 110  90 110 110  95  95 110 105  80 100 110 130
 105 120  90 100 105 100 120 100 100  80 100 120 105  60 125 120 100 115  95
 110 101  80 112 120 110 115 125  55  90 105
> female
  80  75  83  80 100 100  90  75  95  85  90  85  90  90 120  85  100 120 75
  85  80  70  85 110  85  75 105  95  75  70  90  70  82  85 100  90  75  90
 110  80  80 110 110  95  75 130  95 110 110  80  90 105  90 110  75  100 90
 110  85  90  80  80  85  50  80  90 100  80  80  80  95 100  90  100  95 80
  80  50  88  90  90  85  70  90  30  85  85  87  85  90  85  75  90 102  80
 100  80  95  90  80  95 110
```

§1.1 引言

从这些数据中我们能了解到什么呢? 开车最快速度和性别有关系吗? 这些数据服从正态分布吗? 简单的数据总结得到

	male	Female
Min. :	55.0	30.0
1st Qu.:	95.0	80.0
Median :	110.0	89.0
Mean :	107.4	88.4
3rd Qu.:	120.0	95.0
Max. :	150.0	130.0

i.e. 显然, 有一半的男士开车的最快速度≥110, 有3/4 的人最快速度≥ 95, 而开车最快的速度为150, 最慢的速度为55. 对女士而言, 有一半的人开车的最快速度≥89, 有3/4的人的最快速度≥ 80, 而开车最快的速度为130,最慢的速度为30.

§1.1 引言

进一步，我们还以对这些数据的分布有如下了解

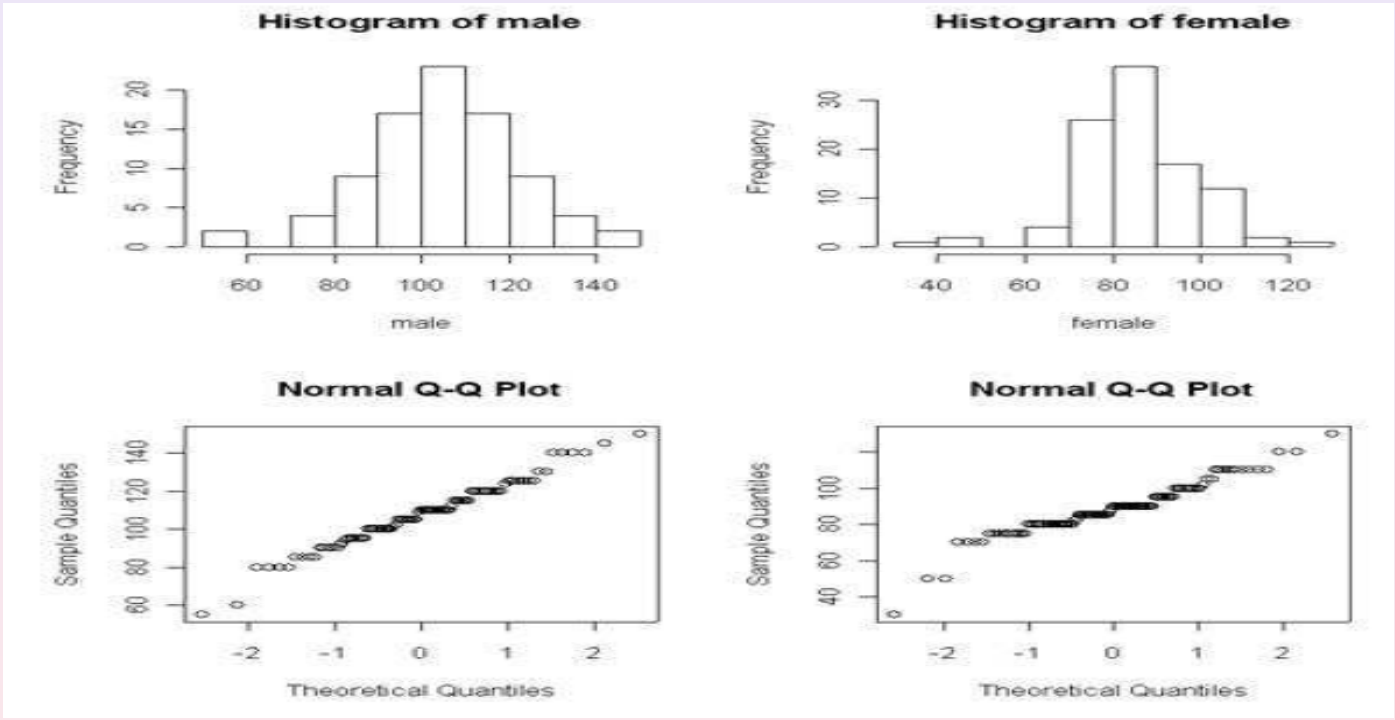


图: 直方图和正态Q-Q图

从这些分析我们可以认为这些数据是服从某个正态分布的。

§1.1 引言

Case study 2: 在卢瑟福试验中,每隔一段时间观察一次由某种铀所放射的到达计数器的粒子数, 共观察100次, 得到结果如下:

i	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
v_i	1	5	16	17	26	11	9	9	2	1	2	1	0

其中v_i表示观察到i个粒子的次数。有理论知识认为放射粒子数服从Poisson分布，试问是否真是这样？

Case 1, Case 2 反映了统计的两个方面：描述性统计(Descriptive Statistics) 和推断统计(Statistical Inference)。

像**Case 1**那样简单的总结数据，我们称为描述性统计。而像**Case 2** 那样，利用分布的性质对问题作出推断我们称为推断统计。概率论在推断统计中起着极其重要的作用。

§1.1 引言

这两个例子也反映了统计是如何处理一个实际问题的：

实际问题与建模——收集数据与统计分析——作出统计推断

因此，我们也可以这样定义数理统计学：

定义1.1.1 数理统计学是一门使用概率论和数学的方法，研究怎样收集带有随机误差的数据，并在设定的模型下，对这种数据进行分析，以对所研究的问题作出推断的一门学科。

§ 1.2.1总体，样本

在数理统计学里，根据针对问题的不同，我们把一个统计问题分解为这两个问题：

- 参数估计:利用样本对总体中未知的参数做出推断的问题
- 假设检验:利用样本对一个假设问题作出推断

在统计学里，有一些专门的术语来描述一个统计问题。我们来介绍一些常见的术语和一个问题的统计描述。

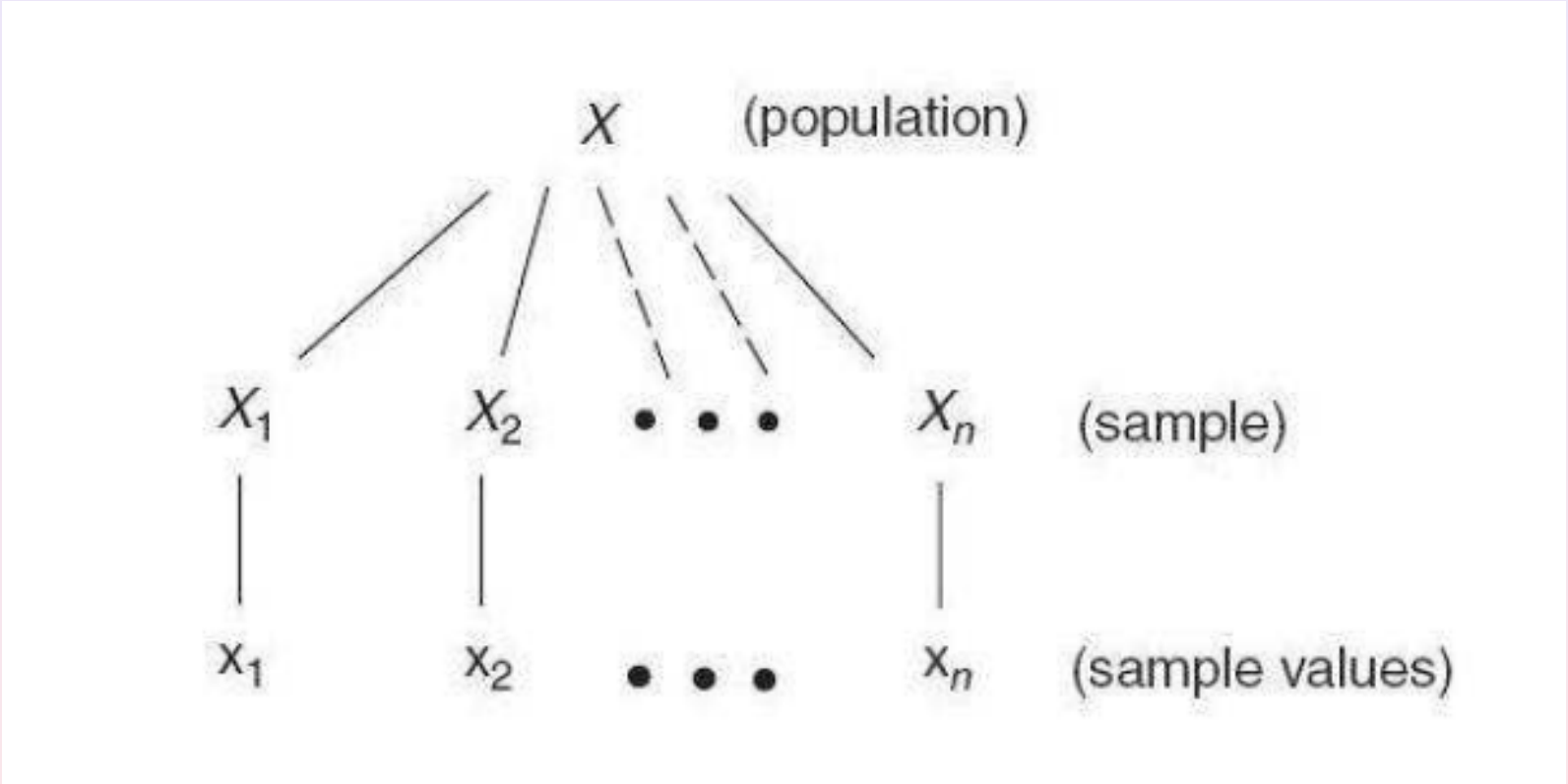
- 总体
- 样本
- 统计量

§1.2.1总体，样本

直观上讲，总体就是所考察对象的全体。比如**Case 1**, 我们要考察Penn State University 男士和女士开车的最快速度, 因此总体就是该校所有人。总体的每一个基本单元称为一个个体，从总体中抽出的一部分个体组成一个样本，总体包含个体的数目称为总体容量，样本包含个体的数目成为样本容量或者样本大小。

抽象地说, 我们关心的是这些人开车的最快速度的分布, 即如果我们用X表示任意一个人开车的最快速度，则X是一个随机变量，而对一个特定的人，他(她)的最快速度只是X的一个值。了解一个随机变量等价于了解它的分布，因此，**总体是一个分布**。从总体中抽取一个个体就是做一次随机试验，而抽取样本容量为n的一个样本，就是做n次随机试验，记为$X_1, \cdots, X_n$。而试验得到的值$x_1, \cdots, x_n$则称为该样本的观察值。如下图所示:

§1.2.1总体，样本



当试验是独立重复的进行时，则称样本$X_1, \cdots, X_n$为简单样本。以后我们若无特殊说明，所说的样本都是指简单样本。

§1.2.1总体，样本

当总体为某个确定的分布F时, 则也称该总体为F总体。比如从正态分布里抽样时, 则称为正态总体；从指数分布中抽取样本时, 则称为指数总体等等。

定义1.2.2 若用$r.v.X$表示所研究对象的某一指标，则总体即为$r.v.X$(的分布)。从此总体中抽取的n个随机变量$X_1, \cdots, X_n$称为样本，而样本$X_1, \cdots, X_n$的值$x_1, \cdots, x_n$称为样本的观察值。

设总体X有概率函数(离散型即为分布律，连续场合下即为概率密度)$f(x)$，则在简单样本情形下，样本$X_1, \cdots, X_n$的联合分布为

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

§1.2.2统计量

只依赖于样本的量称为统计量。比如设$X_1, \cdots, X_n$为从总体$F_\theta(x)$中抽取的一个样本，其中θ为未知的参数，则$\sum_{i=1}^{n} X_i$为一个统计量，而$\sum_{i=1}^{n} X_i - \theta$ 就不是统计量。

统计量的作用在于集中有用的信息，降低数据的维数。

- 常见的统计量

以下我们设$X_1, \cdots, X_n$为样本。

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

§1.2.2统计量

3. 次序统计量$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$

3-1. 样本中位数

$$m = \begin{cases} X_{(\frac{n+1}{2})}, & n\ is\ odd \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}], & n\ is\ even \end{cases}$$

3-2. 样本$p(0 < p < 1)$分位数$X_{[(n+1)p]}$, 此处$[a]$表示不超过a的最大整数.

3-3. 样本极大值和样本极小值: $X_{(n)}$和$X_{(1)}$

3-4. 极差: $X_{(n)} - X_{(1)}$

4. 样本k阶矩

4-1. 样本k阶原点矩$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

4-2. 样本k阶中心矩$m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$

§ 1.2.3充分统计量

当从数据出发进行统计推断时，我们需要对数据的信息进行提炼。即需要构造合适的统计量，一个理想的统计量是完全包含了样本的信息，没有损失任何样本包含的有关参数的信息。换句话说，只要算出了这个统计量的值，就算把原来的样本都丢掉了，也没有任何损失。这种统计量我们称为**充分统计量**。

§1.2.3充分统计量

定义1.2.3：设$T(X)$为一统计量，$X=(X_1,\cdots,X_n)$为从总体$F_\theta(x)$里抽取的样本，θ为参数。如果

$$P(X_1\le x_1,\cdots,X_n\le x_n|T(X)=t)\text{ 与参数 }\theta\text{ 无关}$$

则称$T(X)$为参数θ的一个充分统计量.

充分性原则 在存在充分统计量的情形下，所有的统计推断都可以基于充分统计量进行.

定理1.2.1：设样本$X=(X_1,\cdots,X_n)$的概率函数$f_\theta(x_1,\cdots,x_n)$依赖于参数θ，$T=T(X)$为一统计量，则T为参数θ的充分统计量的充要条件为$f_\theta(x_1,\cdots,x_n)$可以分解为

$$f_\theta(x_1,\cdots,x_n)=g_\theta(T(x_1,\cdots,x_n))h(x_1,\cdots,x_n)$$

其中h仅与$x=(x_1,\cdots,x_n)$有关。

§1.2.3充分统计量

例题1.2.1：设$X = (X_1, \cdots, X_n)$为从$0-1$分布中抽取的简单样本，则$T(X) = \sum_{i=1}^{n} X_i$为充分统计量。

§1.2.3充分统计量

例题1.2.1：设$X=(X_1,\cdots,X_n)$为从$0-1$分布中抽取的简单样本，则$T(X)=\sum_{i=1}^n X_i$为充分统计量。

证：记$T=T(X)$，按定义只要证明下列条件概率与θ无关。当$\sum_{i=1}^n X_i=t_0$时有

$$P(X_1=x_1,\cdots,X_n=x_n|T=t_0)$$

$$=\frac{P(X_1=x_1,\cdots,X_n=x_n,T=t_0)}{P(T=t_0)}$$

$$=\frac{P(X_1=x_1,\cdots,X_n=t_0-\sum_{i=1}^{n-1}x_i)}{P(T=t_0)}$$

§ 1.2.3充分统计量

$$= \frac{\theta^{t_0}(1-\theta)^{n-t_0}}{\binom{n}{t_0}\theta^{t_0}(1-\theta)^{n-t_0}} = \frac{1}{\binom{n}{t_0}}$$

因此有

$$P(X_1 = x_1, \cdots, X_n = x_n | T = t_0) = \begin{cases} \dfrac{1}{\binom{n}{t_0}} & \sum_{i=1}^n x_i = t_0 \\ 0 & \sum_{i=1}^n x_i \neq t_0 \end{cases}$$

上述条件概率与θ无关，因此$T(X) = \sum_{i=1}^n X_i$为θ的充分统计量。

§1.2.3充分统计量

例题1.2.2 常见总体下的充分统计量(设样本为$X = (X_1, \cdots, X_n)$)

[1] 二项分布$B(n,p)$ 参数p的充分统计量为$\sum_{i=1}^{n} X_i$

[2] 均匀分布$U(0,\theta)$ 参数θ的充分统计量为$X_{(n)}$

[3] 指数分布$Exp(\lambda)$ 参数λ的充分统计量为$\sum_{i=1}^{n} X_i$

[4] 正态分布$N(\mu,\sigma^2)$ 参数(μ,σ^2)的充分统计量为$\bar{X}, S^2$

§ 1.3.1 χ^2, t, F分布

在数理统计学里，有三个非常重要的分布:

1. χ^2 分布

定义1.3.4 ： 设$X_1, \cdots, X_n$为相互独立且具有共同的分布(i.i.d) $N(0,1)$的随机变量，则称$X = \sum_{i=1}^{n} X_i^2$的分布为自由度是n的χ^2分布，记为$X \sim \chi_n^2$。X的pdf:

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{x/2} I(x > 0).$$

§ 1.3.1 χ^2, t, F分布

$X \sim \chi_n^2$的性质：

1. $EX = n, D(X) = 2n$
2. 关于参数n具有再生性，即若$Y \sim \chi_m^2$且与X相互独立，则$X + Y \sim \chi_{n+m}^2$.

§ 1.3.1 χ^2, t, F分布

2. t 分布 (Student t 分布)

定义1.3.5：设随机变量$X \sim N(0,1)$, $Y \sim \chi^2_n$且X和Y相互独立，令

$$T = \frac{X}{\sqrt{\frac{1}{n}Y}}$$

则称T的分布为自由度是n的t分布，记为$T \sim t_n$。

可以计算出T的概率密度为

$$g(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad t \in R.$$

t_n的性质:

1. t分布关于$t=0$对称；
2. $\lim_{n\to\infty} g(t) = \phi(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$.

§ 1.3.1 χ^2, t, F分布

3. F **分布**

定义1.3.6 ： 设随机变量X, Y相互独立而且分别服从χ^2_n和χ^2_m，令

$$Z = \frac{1}{n}X \Big/ \frac{1}{m}Y$$

则称Z的分布为自由度是n和m(第一自由度是n，第二自由度是m)的F分布，记为$F \sim F(n, m)$. 同样，可以得到$F(n, m)$具有概率密度：

$$p(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} m^{m/2} n^{n/2} x^{m/2-1} (n + mx)^{-\frac{n+m}{2}} I(0 < t < \infty).$$

§ 1.3.1 χ^2，t，F分布

分位数

定义1.3.7：设随机变量X的分布函数为F，$0 < \alpha < 1$，称数x_α为随机变量X的(上)α分位数，如果

$$1 - F(x_\alpha) = P(X \geq x_\alpha) = \alpha$$

记$F(n, m)$的上α分位数为$F_\alpha(n, m)$，则有$F_\alpha(n, m) = F^{-1}_{1-\alpha}(m, n)$。

对标准正态分布，χ^2和t分布，其上α分位数分别记为$u_\alpha, \chi^2_n(\alpha), t_n(\alpha)$。

§ 1.3.2 正态总体下$\bar{X}$与S^2的分布

定理1.3.2 ： 设$X_1, \cdots, X_n$为从正态总体$N(\mu, \sigma^2)$中抽取的样本，$\bar{X}$与S^2分别为样本均值和样本方差，则

(1) $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

(2) $(n-1)S^2 \Big/ \sigma^2 \sim \chi^2_{n-1}$

(3) $\bar{X}$ 与 S^2 相互独立

(4) $\sqrt{n}(\bar{X} - \mu) \Big/ S \sim t_{n-1}$

§ 1.3.2 正态总体下$\bar{X}$与S^2的分布

定理1.3.3：设$X_1, \cdots, X_n$为从正态总体$N(\mu_1, \sigma_1^2)$中抽取的样本，$Y_1, \cdots, Y_m$为从正态总体$N(\mu_2, \sigma_2^2)$中抽取的样本, 而且两组样本独立，用$\bar{X}, S_X^2, \bar{Y}, S_Y^2$分别表示两组样本的样本均值和样本方差，则

(1) $$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$$

(2) $$\frac{S_X^2}{S_Y^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n-1, m-1)$$

(3) 当$\sigma_1^2 = \sigma_2^2$,有

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{nm}{(n+m)(n+m-2)}[(n-1)S_X^2 + (m-1)S_Y^2]}} \sim \chi^2_{n+m-2}$$

§1.4总结

数据在使用前要注意其收集的合法性(主要的是设计好的试验，感兴趣可以参看参考文献[4])。在合法的数据下，才能展开统计推断工作。

在给定统计模型假设的前提下，一个统计推断问题可以按照如下的步骤进行：

1. 寻求用于统计推断的统计量(在存在充分统计量的情形下使用充分统计量)；
2. 统计量的分布；
3. 基于该统计量和统计推断方法作出推断；
4. 根据统计推断结果对问题作出解释。

统计三大分布及正态总体下样本均值和样本方差的分布，在我们后面的学习中占着重要的地位和应用。