第五节 · 统计与统计量

司继春

上海对外经贸大学统计与信息学院

在这一节中我们将讨论统计学的一些基本概念，这些概念是我们后面学习统计学理论的基础。我们首先介绍统计学中总体、样本和模型的概念，进而介绍统计量的概念及性质。

1 统计的基本概念

1.1 统计学中的数据

统计学是一门关于数据的学科，所有的统计方法都是围绕着数据展开的，因而我们从数据的分类入手，介绍统计学的一些基本概念。

现实生活中碰到的数据是多种多样的，针对同一个个体，我们可以通过很多特征对其进行刻画。比如，对于一个人来说，其性别、年龄、身高等都是其个人的特征；而对于一家企业来说，其所有权性质、企业年龄、注册资本等也是其特征。我们通常把这些描述个体的特征称为变量（Variable）。然而注意到，这些变量的性质并不一样。比如我们可以比较两个人的身高、年龄的大小，然而我们却不能比较性别的大小。因而尽管我们经常把数据全都编码为数值型（比如男 =1, 女 =0 等），然而这些数值的大小并不是都有意义。根据数据度量的层次，一般可以将数据分为以下三类：

1. **分类变量**（**Categorical variable**）：指数据仅仅用于区分类别，而数据没有数值上的意义，比如性别、企业注册类型等。

2. **顺序变量**（**Ordinal variable**）：指数据的值不仅仅用于区分类别，还可以用于排序。比如奖学金等级（一二三等），空气污染等级（重度污染，轻度污染，良好）等。

3. **数值变量**（**Numerical variable**）：指不仅仅数据的排序有意义，而且数据值的差是有意义的。通常又可以将数值型数据分为离散变量和连续变量，前者如次数、人数、年龄等，后者如温度、长度、金额等等。

当然，数据的分类方法并不唯一。比如有的分类方法将数据分为定类数据、定序数据、定距数据和定比数据。而还有一些数据是复合类型，比如对于**截尾数**

据（**Censored data**），就结合了顺序变量和数值变量的特点。针对不同类型的数据，使用的统计方法经常有很大的差别。

而根据时间和个体进行划分，我们经常使用的数据一般有两种最基本的数据类型：**横截面数据**（**Cross-sectional data**）与**时间序列数据**（**Time series data**）。

其中，横截面数据，或者简称截面数据，指同一时间点或者时间段，对不同主体的某些变量进行观测。比如在实验中，对于某一次实验，不同的实验对象的不同观察指标组成的数据即横截面数据。再比如，在调查数据中，很多家庭的多个变量组成的数据也是横截面数据。横截面数据只有个体上的差异而没有时间上的差别。一般我们用 N 记为数据中个体的个数。

而时间序列数据是对于一个或者多个变量在不同时间上的观测。比如 2000 年到现在我国每个季度的 GDP 即时间序列数据。再比如 2000 年到现在我国每个季度的货币供给 M0、M1、M2 也是时间序列数据。一般我们用 T 记为数据中时间的长度。时间序列数据只有时间上的差异而不存在个体上的差异。

除这两种外，还有这两种类型数据的合并数据，如常用的**面板数据**（**Panel data**）或者**纵向数据**（**Longitudinal data**）、**重复截面数据**（**Repeated cross-sectional data**）等等。其中面板数据指的是同时观测多个个体，而对于每个个体，在不同时间段对某些变量进行观测。比如，单独看上海市从 2000 年到现在每年的 GDP 是时间序列数据，然而如果我们可以观察到全国每个省从 2000 年到现在每年的 GDP，那么就是面板数据。面板数据既有时间上的信息又有不同个体的信息，我们一般把 $N \gg T$ 的面板数据称为长面板数据。

1.2 统计模型

在获得数据之后，我们需要对这些数据建立概率模型。一般而言，一个典型的统计学问题可以如下描述：进行一次或者一系列的随机试验，并且从这些随机试验中收集了一些数据，而统计的任务就是使用这些数据提取我们希望得到的信息，得到一些结论。在统计学中，我们希望得到的结论是多种多样的，比如在机器学习中，最关心的结论是预测是否正确，而在计量经济学中，很多时候识别因果关系则是最希望得到的结论。一般来说，统计方法可以分为描述性统计方法和推断统计。

针对已经有的数据，首先可以进行一些**描述性统计**（**descriptive statistics**）的工作。描述性统计通过表和图的形式对数据加以展示，常用的描述性统计量一般包括均值、标准差、最大值、最小值、中位数、四分位数等。描述性统计经常作为初步的研究，研究者可以通过描述性统计对数据的分布情况做初步的了解，检查数据中可能有的错误，帮助研究者发现数据中特有的现象等等。

描述性统计虽然重要，但是只能作为初步的观察，而更进一步的结论则需要借助**统计推断**（**Statistical inference**）来实现。统计推断通过概率建模的方法，用已知的样本对未知的总体进行推断，一般包含着**参数估计**（**Estimation**）、**假设检验**（**Hypothesis testing**）等内容。因而在这里，理解总体和样本的概

念是非常重要的。

在统计推断理论中，数据集 $\{x_i, i=1,...,N\}$ 被视为是概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中一系列随机变量（向量）$\{X_i, i=1,...,N\}$ 的实现，概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中的概率函数 $\mathscr{P}$ 则被称为**总体**（**Population**）。

在统计推断中，我们经常将数据集 $\{x_i, i=1,...,N\}$ 建模为样本空间 $\Omega = \mathbb{R}^n$ 中随机试验的一个实现。如果随机变量 $\{X_i, i=1,...,N\}$ 中 X_i 是相互独立的，且具有相同的分布函数，那么我们称 $\{X_i, i=1,...,N\}$ 为**独立同分布的**（**independent and identically distributed**，**i.i.d**）。如果 $\{X_i, i=1,...,N\}$ 是来自于 $\prod_{i=1}^{N}(\mathbb{R}, \mathscr{B}, P)$ 的一组随机变量，且独立同分布，我们称 $\{X_i, i=1,...,N\}$ 为**随机样本**（**Random sample**），其中总体为 P，而随机变量的个数 N 称为**样本量**（**Sample size**）。

由于随机样本其分布都相同且相互独立，因而总体 P 由 X_i 的边缘分布 $F_X(\cdot)$ 确定，其联合分布可以写为：

$$F(x) = \prod_{i=1}^{N} F_{X_i}(x_i) = \prod_{i=1}^{N} F_X(x_i)$$

其中第一个等号使用了独立的假设，而第二个等号使用了同分布的假设。

例 1. 如果我们希望使用统计手段调查一条生产线上的次品率，我们经常会对这个产品线上的产品进行抽样。如果我们抽取了 N 个样本 $\{X_i, i=1,...,N\}$，假设 $X_i = 1$ 为次品，否则为 0，X_i 独立同分布，那么这里 $\Omega = \{0,1\}^N$，而总体 P 为一个伯努利分布 $P(X_i = 1) = p$。我们的目的即希望通过样本 $\{X_i, i=1,...,N\}$ 对总体 P 做出推断。由于每个 X_i 其概率质量函数为：

$$P(X_i = x) = p^x (1-p)^{1-x}, x = 0, 1$$

如果假设 X_i 为独立同分布，那么样本 $\{X_i, i=1,...,N\}$ 的联合密度函数为：

$$P(x) = \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i} = p^{N_1}(1-p)^{N-N_1}$$

其中 $N_1 = \sum_{i=1}^{N} x_i$。比如，如果我们观察到数据 $x = (0,1,0,0,1)'$，那么观察到这组数据的概率为：

$$P(x) = p^2(1-p)^3$$

例 2. 有的时候我们会对某些测量感兴趣（如体温、距离、长度等），如果我们对其进行 N 次观测，假设每次观测误差都是独立同分布的，记每次观测为 X_i，那么这里 $\Omega = \mathbb{R}^N$，而其联合分布函数：

$$F(x) = \prod_{i=1}^{N} F_{X_i}(x_i) = \prod_{i=1}^{N} F_X(x_i) = \prod_{i=1}^{N} P(X_i \leq x_i)$$

其中 $F_X(\cdot)$ 为每次测量的边缘分布。在这里，我们的总体 P 与边缘分布 $F_X(\cdot)$ 是等价的。

而在一些问题中，我们所关注的个体是有限的。比如如果我们只对一批产品的合格率感兴趣，或者当我们关注全国人民的收入分布时，全国人民是一个有限的集合。然而在这些情况下，调查每一个个体经常是不现实的，所以我们需要在所关注的个体中找到一个子集进行研究。当然，对所有关注的个体进行调查也是有可能的，比如**普查**（**Census**），包括人口普查、经济普查等。

一般的，如果令 $\mathcal{P} = \{y_1, y_2, ..., y_M\}$ 为我们所关心的全体，而全体不能一一进行调查，我们通常对其一个子集 $S \subset \mathcal{P}$ 进行调查。因而这里就涉及到我们如何从 $\mathcal{P}$ 中挑选出子集 S，即**抽样**（**Sampling**）问题。

抽样方法分为两种：**概率抽样**（**Probability sampling**）和**非概率抽样**（**Nonprobability sampling**）。其中概率抽样指每个个体都有正的概率被抽中，且该概率已知（或者可以被计算）。概率抽样的统计性质良好，因而我们下面主要集中在概率抽样的条件下进行讨论。

一个最简单的概率抽样方法即从 M 个元素中等可能的不放回地抽取（**sampling without replacement**）N 个样本 $\{X_i, i = 1, ..., N\}$，即**简单随机抽样**（**simple random sampling**）。由于在这种情况下，样本的分布由 $\mathcal{P}$ 和每个个体被抽中的概率决定，因而我们通常将 $\mathcal{P}$ 称之为总体。

除了简单随机抽样之外，综合考虑调查成本和其他因素，还有其他的抽样方法，如：

1. 系统抽样（Systematic sampling）：指先根据一定规则对个体排序，再根据一定的规则选取样本。比如对一个班的学生进行抽样，可以抽取学号尾数为 1 的所有个体。

2. 分层抽样(Stratified sampling)：如果总体分成不同的层级，或者**层**(**strata**)，那么可以在每个层中进行抽样，再将每个层中抽取的总体进行汇总。比如如果需要抽取全国的样本，可以在每个城市单独抽样然后汇总。

3. 整群抽样（Cluster sampling）：先对所有个体分组，再抽取组别，进而调查被抽中的组的所有个体。比如想要在全校学生中抽样，可以抽取班级，再调查被抽中的班级的所有学生。

4. 多阶段抽样（Multi-stage sampling）：先对所有个体分组，抽取分组，再在每个组内部抽样。与分层抽样的差别是，分层抽样要求每个组内都进行抽样，而多阶段抽样只对抽中的组进行抽样。比如可以对全国所有城市中抽取 20 个城市，再在这 20 个城市中进行抽样。

5. 面板抽样（Panel sampling）：指先随机抽取样本，之后隔一段时间对同一批个体进行反复的调查，可以获得面板数据。

注意以上抽样方法并不能一定保证每个个体被抽中的概率相等。

在这里需要特别注意的是，在有限总体的情况下，我们虽然可以认为样本 X_i 是同分布的，但是经常样本未必是独立的，因而样本的联合分布不一定可以写成样本边缘分布的乘积的形式。比如，如果我们关心全国的家庭总消费，不同地区可能有不同的消费和物价水平，因而每个家庭的消费可能满足如下形式:

$$C_{ri} = \alpha_r + \epsilon_{ri}$$

其中 r 代表地区，而 i 代表每个家庭。如果假设 ϵ_{ri} 为独立同分布的，且 $\mathrm{Cov}(\epsilon_i, \alpha_r) = 0$，那么对于同一地区的不同个体，$\mathrm{Cov}(C_{ir}, C_{jr}) = \mathrm{Cov}(\alpha_r + \epsilon_i, \alpha_r + \epsilon_j) = \mathrm{Var}(\alpha_r)$，因而消费水平可能不是独立的，除非 $\alpha_r = \alpha$ 为所有地区都相等的常数。

以上定义了总体，即一个概率函数 P，然而现实中总体 P 是不能被观测的，而统计推断的任务就是使用可以观测的样本 $\{X_i\}$ 对未知的总体进行推断。而所谓的**统计模型**（**Statistical model**）即通过对总体 P 做一系列的假设，简化问题并对总体进行推断。

一般来说，统计模型分为**参数模型**（**parametric model**）和**非参数模型**（**nonparametric model**），以及介于两者之间的**半参数模型**（**semi-parametric model**）。

参数模型即假设总体 P 属于某一个参数族 $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$，其中 $\Theta \subset \mathbb{R}^d$，且一旦 θ 确定了，那么 P_θ 为一个确定的概率函数。其中 Θ 被称为参数空间（parameter space），而 d 称为参数空间的维数。

对于参数族 $\{P_\theta, \theta \in \Theta\}$，如果当 $\theta_1 \neq \theta_2$ 时，必然有 $P_{\theta_1} \neq P_{\theta_2}$，那么我们称其为**可识别的**（**identifiable**）。如果参数族不可识别，意味着存在多于一个 θ 代表了同一个概率函数，或者模型的解不唯一。因而一般我们要求参数模型必须为可识别的。

例 3. 在例 (2) 中，我们可以假设每一次测量 $X_i \sim N(\mu, \sigma^2)$，其中 μ 为测量的真实值（如真实体温、长度），而 σ^2 代表每次测量可能的误差大小，且假设误差服从正态分布。在这里，我们假设总体 $P \in \{N(\mu, \sigma^2)\}$，参数空间为 $\Theta = \mathbb{R} \times \mathbb{R}^+$。只要 μ 和 σ^2 确定了，那么总体 P 也就确定了。

相反，如果对总体 P 不做任何参数上的假定，我们称其为非参数模型。此类模型并不假设 P 属于某一个参数族，而是对 P 做了其他的假定，如对分布函数的连续型、光滑性、对称性等做出假定。

2 统计量及其抽样分布

在统计理论中，所有的统计方法都是通过**统计量**（**Statistic**）实现的。一般的，对于概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中的一组样本 $\{X_i, i = 1, 2, ..., N\}, X = [X_1, ..., X_N]'$，统计量即样本的一个不依赖于其具体实现的函数 $T(X)$。由于 X 为随机向量，因而统计量 $t = T(X)$ 作为随机向量的函数仍然是概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的随

机变量，所以统计量 s 同样具有期望、方差、分布等随机变量所具有的特征。其中统计量的分布称为**抽样分布**（**sampling distribution**），而统计量的标准差则被成为**标准误**（**standard error**）。

统计量是所有统计方法的基本工具，在不同的统计方法中有不同的称谓，如在描述性统计中，统计量一般称之为**描述性统计量**（**descriptive statistic**）；在参数估计中，我们称其为**估计量**（**estimator**）；而在假设检验中，我们称其为**检验统计量**（**test statistic**）。

这里需要注意参数 θ 和统计量 t 的差别。参数是总体的特征，因而是不可观测的，而估计量是样本的函数，由于样本是可以观测的，因而估计量也必须是可以计算可以观测的。统计推断的目标即使用有限的样本，通过计算样本统计量，对总体的参数做出推断。

为了符号统一，在不引起歧义的情况下，接下来我们将不区分样本 $\{X_i\}$ 及其实现 $\{x_i, i = 1, 2, ..., N\}$，即当我们使用 $\{x_i\}$ 时，仍然代表来自于总体的一组样本，即概率空间中的一系列随机变量。

接下来我们以样本均值和样本方差为例，介绍一些简单的统计量的抽样分布。

2.1 样本均值

样本均值（Mean）是对数据平均水平的最常用的度量，其定义为：

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = L'x$$

其中 $x = (x_1, x_2, ..., x_N)'$，$L = \frac{1}{N}\iota$。注意样本均值使得样本均方误差最小化，即：

$$\bar{x} = \arg\min_{c} \frac{1}{N}\sum_{i=1}^{N} (x_i - c)^2$$

如果假设 $\{x_i\}$ 为独立同分布的样本，且 $\mathbb{E}(x_i) = \mu$，$\mathrm{Var}(x_i) = \sigma^2$（或记为 $x_i \sim (\mu, \sigma^2)\ i.i.d$），那么我们有：

$$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(x_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu$$

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(x_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{\sigma^2}{N}$$

更进一步，如果 $\{x_i\}$ 来自于正态总体且独立同分布，即：$x_i \sim N(\mu, \sigma^2)\ i.i.d$，

或者等价的记为 $x \sim N\left(\mu\iota, \sigma^2 I\right)$，那么可以得到 $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$，或者：

$$\sqrt{N}\frac{\bar{x}-\mu}{\sigma} = \frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{N}}} \sim N\left(0,1\right) \tag{1}$$

即正态总体的样本均值服从正态分布。

2.2 样本方差

样本方差是数据离散程度最常用的度量，其定义为：

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2 = \frac{1}{N-1}x'M_0x$$

其中 $M_0 = I - P_0 = I - \frac{1}{N}\iota\iota'$。相应的，样本标准差定义为：$s = \sqrt{s^2}$。由于：

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i^2+\bar{x}^2-2\bar{x}x_i\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2+\bar{x}^2-\frac{1}{N}\sum_{i=1}^{N}2\bar{x}x_i\\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2+\bar{x}^2-2\bar{x}\cdot\frac{1}{N}\sum_{i=1}^{N}x_i\\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2+\bar{x}^2-2\bar{x}^2\\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2-\bar{x}^2
\end{aligned}$$

从而样本方差可以写为：

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

即样本均值可以写为样本平方的均值减去样本均值的平方，乘以 $\frac{N}{N-1}$。注意到在计算样本均值时，我们使用 N 做为分母，然而在计算样本方差时，我们使用 $N-1$ 作为分母，这是由于使用 $N-1$ 做分母可以保证样本方差的期望 $\mathbb{E}s^2 = \sigma^2$，

即如果假设 $x_i \sim \left(\mu, \sigma^2\right)$ $i.i.d$，那么：

$$\begin{aligned}
\mathbb{E}s^2 &= \frac{N}{N-1}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2\right) \\
&= \frac{N}{N-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(x_i^2\right) - \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2\right] \\
&= \frac{N}{N-1}\left[\mu^2+\sigma^2 - \frac{1}{N^2}\mathbb{E}\left(\sum_{i=1}^{N}x_i^2 + 2\sum_{1\le i<j\le N}x_ix_j\right)\right] \\
&= \frac{N}{N-1}\left[\mu^2+\sigma^2 - \frac{1}{N^2}\mathbb{E}\left(\sum_{i=1}^{N}x_i^2\right) - \frac{2}{N^2}\mathbb{E}\left(\sum_{1\le i<j\le N}x_ix_j\right)\right] \\
&= \frac{N}{N-1}\left[\mu^2+\sigma^2 - \frac{1}{N}\left(\mu^2+\sigma^2\right) - \frac{2}{N^2}\frac{N^2-N}{2}\mu^2\right] \\
&= \frac{N}{N-1}\left[\mu^2+\sigma^2 - \frac{1}{N}\left(\mu^2+\sigma^2\right) - \frac{N-1}{N}\mu^2\right] \\
&= \sigma^2
\end{aligned}$$

而如果更进一步，假设 $x \sim N\left(\mu\iota, \sigma^2 I\right)$，由于 $M_0\mu\iota = \mu\left(I - \frac{1}{N}\iota\iota'\right)\iota = \mu\left(\iota - \frac{1}{N}\iota\iota'\iota\right) = 0$，那么

$$\frac{(N-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}x'M_0x = \left(\frac{x-\mu}{\sigma}\right)'M_0\left(\frac{x-\mu}{\sigma}\right)$$

其中 $\frac{x-\mu}{\sigma} \sim N(0, I)$，且 M_0 为投影矩阵，$\mathrm{tr}\left(M_0\right) = \mathrm{tr}\left(I - \frac{1}{N}\iota\iota'\right) = \mathrm{tr}\left(I\right) - \frac{1}{N}\mathrm{tr}\left(\iota\iota'\right) = N - \frac{1}{N}\mathrm{tr}\left(\iota'\iota\right) = N-1$，因而可以得到：

$$\frac{(N-1)s^2}{\sigma^2} \sim \chi^2(N-1)$$

即正态总体的样本方差标准化之后服从卡方分布，且自由度为 $N-1$。注意这一结论是在正态总体且独立同分布的假定下得到的。

在以下程序中，我们对样本方差的分布进行了模拟：

代码 1: Stata 样本方差的模拟

```
cap program drop std_normal_var
program define std_normal_var, rclass
  version 12
  syntax [,obs(integer 10) mu(real 0) sigma(real 1)]
  drop _all
  set obs `obs'
  tempvar x
```

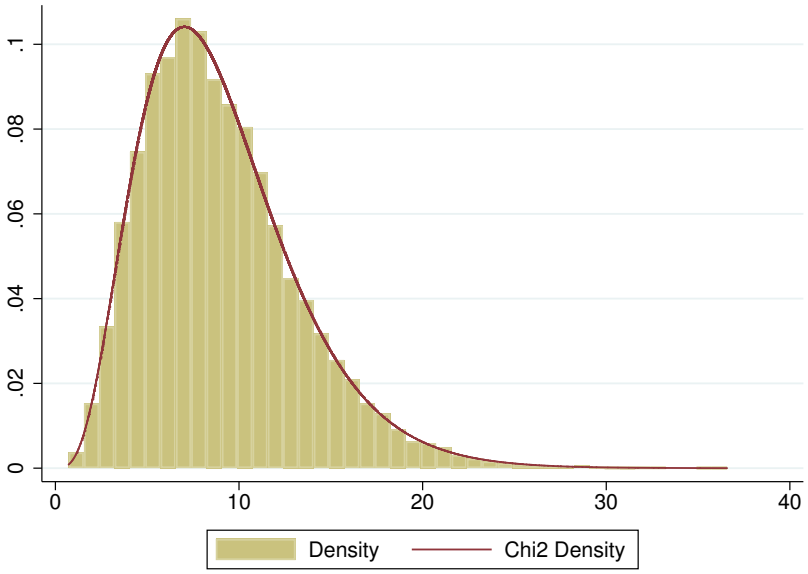


图 1: 样本方差的模拟

```
    gen `x'=`sigma'*(rnormal()+`mu')
    su `x'
    return scalar std_var=(`obs'-1)*r(Var)/`sigma'
end

simulate std_var=r(std_var), reps(20000): std_normal_var


sort std_var
gen chi2_den=chi2den(9,std_var)
label variable chi2_den "Chi2 Density"
twoway (hist std_var, density) ///
       (line chi2_den std_var) ///
       , graphr(fcolor(white) color(white)) xtitle("")
graph export standard_normal_var.eps, replace
```

我们首先定义了一个程序（program），其作用是给定 $\mu = mu$ 和 $\sigma = sigma$，生成 $N = obs(=10)$ 个正态分布样本，计算样本方差 s^2，并返回 std_var 为 $\frac{(N-1)s^2}{\sigma^2}$。接下来，我们使用 simulate 命令将以上过程重复 20000 次，最终得到了 20000 个正态分布样本方差的分布。

图 (1) 展示了程序结果，其中直方图为模拟的 20000 次 str_var 的直方图，而红色的线表示 $\chi^2(9)$ 的密度函数。可见两者之间差距非常小。

如果我们将式 (1) 中的总体方差 σ^2 替换为样本方差 s^2，即：

$$\begin{aligned}\sqrt{N}\frac{\bar{x}-\mu}{s} &= \sqrt{N}\frac{\bar{x}-\mu}{\sqrt{s^2}} \\ &= \sqrt{N}\frac{L'x-\mu}{\sqrt{\frac{1}{N-1}x'M_0x}} \\ &= \sqrt{N}\frac{L'(x-\mu)}{\sqrt{\frac{1}{N-1}x'M_0x}} \\ &= \sqrt{N}\frac{L'\left(\frac{x-\mu}{\sigma}\right)}{\sqrt{\frac{1}{N-1}\frac{1}{\sigma^2}x'M_0x}}\end{aligned}$$

由于 $\frac{x-\mu}{\sigma} \sim N(0,I)$，因而 $\sqrt{N}L'\frac{x-\mu}{\sigma} \sim N(0,NL'L) = N(0,1)$，而分母上 $\frac{1}{\sigma^2}x'M_0x \sim \chi^2_{N-1}$，且 $LM = \frac{1}{N}\iota\left(I-\frac{1}{N}\iota\iota'\right) = 0$（分子与分母独立，即 $\bar{X}$ 与 s^2 是独立的[1]），因而：

$$\sqrt{N}\frac{\bar{x}-\mu}{s} = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}} \sim t(N-1) \tag{2}$$

由于 t_{N-1} 分布在 $N \to \infty$ 时趋向于标准正态分布，因而当样本足够大时，上述分布趋向于正态分布。对于样本 $\{x_i\}$，如果我们对每个样本都减去样本均值，再除以样本标准差，即：

$$x_i^s = \frac{x_i-\bar{x}}{s}$$

我们称这个过程为**标准化**（**standardrize**），标准化之后的数据 $\{x_i^s\}$ 其样本均值为 0，而样本方差为 1。注意由于：

$$\frac{\bar{x}-\mu}{s} = \frac{1}{N}\sum_{i=1}^{N}\frac{x_i-\mu}{s} = \frac{1}{N}\sum_{i=1}^{N}x_i^s$$

因而式 (2) 表明，标准化之后的样本均值乘以 $\sqrt{N}$ 服从 t 分布，且其自由度为 $N-1$。仍然，这一结论只有在独立同分布的正态总体下才成立。

在以下程序中，我们分别对 $\sqrt{N}\frac{\bar{x}-\mu}{\sigma}$ 和 $\sqrt{N}\frac{\bar{x}-\mu}{s}$ 的分布进行了模拟：

代码 2: Stata 样本标准化均值的模拟

```
clear
set more off
cap program drop std_normal_mean
program define std_normal_mean, rclass
  version 12
  syntax [,obs(integer 10) mu(real 0) sigma(real 1)]
  drop _all
```

[1]注意这一结论的前提是总体为正态分布。

```
  set obs `obs'
  tempvar x
  gen `x'=`sigma'*(rnormal()+`mu')
  su `x'
  return scalar mean1=sqrt(`obs')*(r(mean)-`mu')/`sigma'
  return scalar mean2=sqrt(`obs')*(r(mean)-`mu')/r(sd)
end
local N=8
local df=`N'-1
simulate std_mean=r(mean1) sample_mean=r(mean2),/*
*/    reps(20000): std_normal_mean, obs(`N')

// standardised with sigma
sort std_mean
gen normal_den_std_mean=normalden(std_mean)
label variable normal_den_std_mean "Std Normal Density"
gen student_den_std_mean=tden(`df',std_mean)
label variable student_den_std_mean "t(`df') Density"
twoway (hist std_mean, density) ///
     (line normal_den_std_mean std_mean) ///
     (line student_den_std_mean std_mean) ///
     , graphr(fcolor(white) color(white)) xtitle("") ///
     saving(standard_normal_mean_1, replace)

// standardised with std deviation
sort sample_mean
gen normal_den_std_mean2=normalden(sample_mean)
label variable normal_den_std_mean2 "Std Normal Density"
gen t_den_std_sample_mean=tden(`df',sample_mean)
label variable t_den_std_sample_mean "t(`df') Density"
twoway (hist sample_mean, density) ///
     (line normal_den_std_mean2 sample_mean) ///
     (line t_den_std_sample_mean sample_mean) ///
     , graphr(fcolor(white) color(white)) xtitle("") ///
     saving(standard_normal_mean_2, replace)

graph combine standard_normal_mean_1.gph /*
  */  standard_normal_mean_2.gph, ///
      graphr(fcolor(white) color(white)) col(2)
graph export standard_normal_mean.eps, replace
```

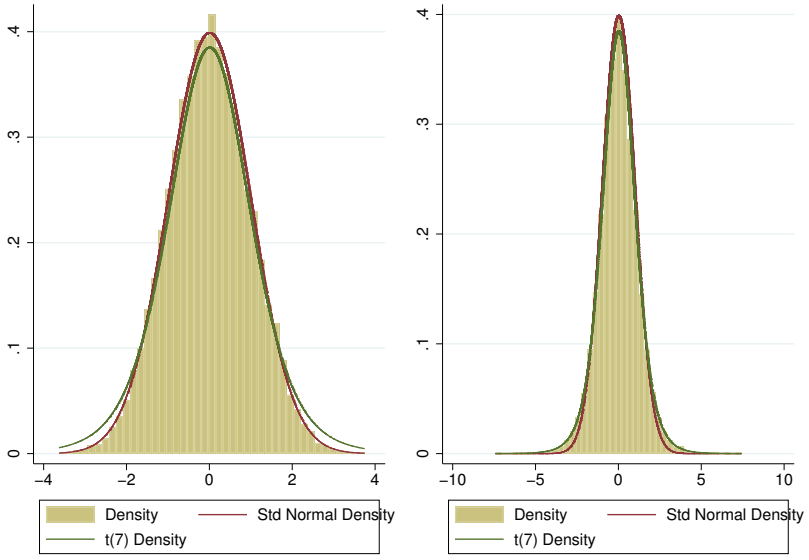


图 2: 样本方差的模拟

在以上程序中，我们同样给定 $\mu = mu$ 和 $\sigma = sigma$，生成 $N = obs(=8)$ 个正态分布样本，计算样本方差 s^2，并返回 std_mean 为 $\sqrt{N}\frac{\bar{x}-\mu}{\sigma}$ 而返回回 std_sample_mean 为 $\sqrt{N}\frac{\bar{x}-\mu}{s}$。按照理论，以上两个统计量分别服从标准正态分布和 t 分布。我们使用 simulate 命令将以上过程重复 20000 次，最终得到了 20000 个正态分布的以两种方法标准化以后的样本均值。

图 (1) 展示了程序结果，其中左图为 20000 次计算 $\sqrt{N}\frac{\bar{x}-\mu}{\sigma}$ 的直方图，右图为 20000 次计算 $\sqrt{N}\frac{\bar{x}-\mu}{s}$ 的直方图，红线为标准正态分布的密度函数，绿线为 $t(7)$ 分布的密度函数。可以发现，左图使用正态分布拟合更好，而右图用 $t(7)$ 分布拟合更好，符合我们理论的预期。实际上，左图中，$t(7)$ 分布高估了 $\sqrt{N}\frac{\bar{x}-\mu}{\sigma}$ 的尾巴厚度，而在右图中，正太分布低估了 $\sqrt{N}\frac{\bar{x}-\mu}{s}$ 的尾巴厚度。此外，由于 $t\,(N-1)$ 随着 $N \to \infty$，因而当样本量设置的大一点时，两个分布是很难区分开的，读者可以自行进行实验。

2.3 分位数与次序统计量

上一节中我们介绍了使用样本均值度量平均水平，使用样本方差度量数据的离散程度。而除了平均值，**中位数**（**median**）是平均水平的另外一种度量方法。

对于一个总体 P，如果其分布函数为 $F\,(x)$，那么中位数定义为 $F^{-1}\left(\frac{1}{2}\right)$。例如，由于对称性，正态分布 $N\left(\mu,\sigma^2\right)$ 的中位数为 μ。注意中位数是以下最小

化问题的解：

$$\min_c \mathbb{E}\left|X-c\right| \tag{3}$$

为证明以上结论，注意当以上目标函数取最小值时，一阶条件意味着：

$$\begin{aligned}
0 &= \frac{\partial \mathbb{E}\left|X-c\right|}{\partial c} \\
&= \frac{\partial \int_{\mathbb{R}} \left|x-c\right| dF\left(x\right)}{\partial c} \\
&= \int_{\mathbb{R}} \frac{\partial \left|x-c\right|}{\partial c} dF\left(x\right) \\
&= \int_c^{\infty} \left(-1\right) dF\left(x\right) + \int_{-\infty}^{c} 1 dF\left(x\right) \\
&= -\left[1-F\left(c\right)\right] + \left[F\left(c\right)-0\right] \\
&= -1+2F\left(c\right)
\end{aligned}$$

因而当 $c=F^{-1}\left(\frac{1}{2}\right)$ 时，上式成立。

对于一组样本 $\{x_1, x_2, ..., x_N\}$，记最小的样本为 $x_{(1)}$，第二小的样本为 $x_{(2)}$，以此类推，最大的样本记为 $x_{(N)}$。我们称 $x_{(n)}$ 为**次序统计量**（**order statistics**）。**样本中位数**（**sample median**，记为 M）定义为：

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n\text{为奇数} \\ \frac{\left(x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right)}{2} & n\text{为偶数} \end{cases}$$

注意实际上中位数是以下最小化问题的解：

$$\min_c \frac{1}{N}\sum_{i=1}^{N}\left|x_i-c\right|$$

以上最小化问题是式 (3) 的样本等价形式。

更进一步，我们还可以定义其他的**分位数**（**quantiles**）。q 分位数（q-Quantiles）即将实轴分为概率相等的 q 部分。$q-1$ 个分位数值将实轴分为 q 个概率相等的部分。比如，四分位数（quartiles，记为 Q），即 $Q_1 = F^{-1}\left(0.25\right), Q_2 = F^{-1}\left(0.5\right), Q_3 = F^{-1}\left(0.75\right)$。此外，百分位数（percentiles，记为 P），即 $P_p = F^{-1}\left(\frac{p}{100}\right), p=1,2,...,99$。

如果令：

$$\psi_q\left(x\right) = \begin{cases} qx & x>0 \\ \left(q-1\right)x & x\leq 0 \end{cases}$$

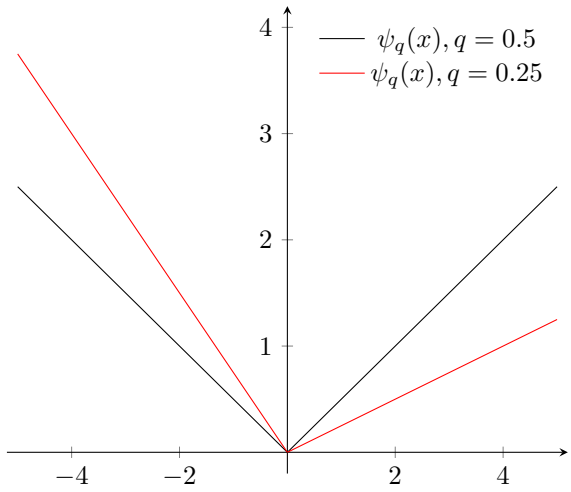


图 3: $\psi_q(x)$ 函数示意图

其中 $0 < q < 1$，那么可以证明，$F^{-1}(q)$ 是以下最小化问题的解：

$$\min_c \mathbb{E}\psi_q(X - c)$$

类似的，对于 $0 < q < 1$，令 $\{Nq\}$ 代表 Nq 的四舍五入，那么样本的分位数即 $x_{(\{Nq\})}$。同样，样本分位数也是以下最小化问题的解：

$$\min_c \frac{1}{N}\sum_{i=1}^{N}\psi_q(x_i - c)$$

以上我们使用次序统计量 $x_{(n)}$ 定义了中位数、分位数，接下来我们讨论次序统计量的分布问题。对于次序统计量，我们有如下定理：

定理 1. 如果 $x_{(1)}, x_{(2)}, ..., x_{(N)}$ 为独立同分布随机样本 $\{X_i, 1 \le i \le N\}$ 的次序统计量，且总体分布函数为 $F(x)$，密度函数为 $f(x)$，那么次序统计量 $x_{(n)}$ 的密度函数为：

$$f_{x_{(n)}}(x) = \frac{N!}{(n-1)!\,(N-n)!} f(x)\,[F(x)]^{n-1}\,[1 - F(x)]^{N-n}$$

证明. 可以首先计算 $x_{(n)}$ 的分布函数 $F_{x_{(n)}}(x) = P\left(x_{(n)} \le x\right)$，密度函数即其分布函数的导数。现在令 Y 为小于等于 x 的样本数，即 $Y = \#\{x_i \le x\}$，那么 $Y \sim Bi(N, F(x))$。而 $x_{(n)} \le x$ 等价于 $Y \ge n$，因而：

$$F_{x_{(n)}} = P(Y \ge n) = \sum_{k=n}^{N}\begin{pmatrix} N \\ k \end{pmatrix}[F(x)]^k\,[1 - F(x)]^{N-k}$$

对以上分布函数求导可得结论。 □

此外，我们还可以计算不同次序统计量之间的联合分布。所有次序统计量的联合密度函数为：

$$f_{x_{(1)}\ldots x_{(n)}}(x_1,\ldots,x_n)=\begin{cases}N!\prod_{i=1}^{N}f(x_i) & -\infty<x_1<\ldots<x_n<\infty\\ 0 & otherwise\end{cases}$$

其中 $N!$ 是由于对于一个随机样本，有 $N!$ 中情况可以得到 $-\infty<x_1<\ldots<x_n<\infty$ 的实现。比如样本 $\{x_1=1,x_2=2\}$ 和样本 $\{x_1=2,x_2=1\}$ 都可以产生相同的次序统计量，因而有 2! 中情况。根据以上的密度函数，可以得到任意两个次序统计量 $\left(x_{(i)},x_{(j)}\right),1\leq i<j\leq N$ 的联合密度函数为：

$$f_{x_{(i)},x_{(j)}}(u,v)=\begin{cases}\frac{N![F(u)]^{i-1}[F(v)-F(u)]^{j-i-1}[1-F(v)]^{N-j}f(u)f(v)}{(i-1)!(j-i-1)!(N-j)!} & u<v\\ 0 & otherwise\end{cases}$$

3 描述性统计

正如前文所述，统计方法包括描述性统计和统计推断两部分。虽然统计推断是数理统计的核心，然而在数据分析之前，做好描述性统计是非常重要的。

描述性统计即使用图和表的形式对数据的分布特征进行度量。描述性统计可以帮助研究者初步掌握数据的分布情况，并发现数据中潜在的问题，比如异常值的存在、数据的不一致性等。此外，描述性统计给研究者提供了一些需要解释的现象，比如研究人员发现城市的大小、姓氏的分布等都服从幂律（Power law），该如何解释这些分布规律成为了一些学科的研究热点。最后，描述性统计的结果有利于对统计推断结果的理解，如回归分析中，回归结果经常需要与描述性统计相配合才能得到回归系数的影响大小。

下面我们分描述性统计量和统计图表两部分介绍描述性统计的初步知识。

3.1 描述性统计量

针对已有的数据，可以使用一些样本统计量对数据进行描述。而针对不同的数据类型，使用的描述性统计量也不尽相同，比如，对于分类数据，就不能计算其均值和中位数等。下面我们将分别对这些描述性统计量做简要介绍。

3.1.1 位置度量

位置度量（measure of location）是一组数据中心的测量。常用用于位置度量的描述性统计量有众数、中位数、平均数等。

1. **众数（mode）**：适用于分类数据和顺序数据，指样本数量最多的类别。如一个班中男生 30 人，女生 20 人，那么众数即男生。

2. **中位数**（**median**）：适用于顺序数据和数值型数据，但是不适用于分类数据。对于顺序数据，如一等奖 3 人、二等奖 5 人、三等奖 7 人，那么中位数即为二等奖。中位数具有不易受异常值影响的优点。此外，中位数对于单调变换有不变性，比如一组数据 $\{x_i\}$ 的中位数为 M，那么经过单调变换之后的数据，比如 $\{\ln(x_i)\}$ 的中位数为 $\ln(M)$。

3. **平均数**（**mean**）：仅适用于数值型数据。与中位数相比，平均数比较容易受到异常值的影响，然而其具有非常良好的性质，因而是最常使用的平均水平的度量。

3.1.2 离散程度度量

常见的度量平均水平的描述性统计量有：异众比率、四分位差、极差、标准差、离散系数等。

1. **异众比率**（**variation ratio**）：适用于分类数据和顺序数据，指样本中非众数组的比例。

2. **四分位差**（**quartile deviation**）、**极差**（**range**）：适用于顺序数据和数值型数据，其中四分位差定义为 $Q_3 - Q_1$，而极差定义为 $x_{(N)} - x_{(1)}$。同样，这两个变量不太容易受到极端值的影响。

3. **标准差**（**standard deviation**）：适用于数值型数据，是离散程度的最常用的度量。

4. **离散系数**（**coefficient of variation**）：适用于数值型数据，定义为样本标准差与样本均值的比值：
$$v = \frac{s}{\bar{x}}$$
其优点是消除了单位以及平均水平大小的影响。

3.1.3 偏度系数与峰度系数

偏度系数度量了数据分布的对称性，而峰度系数则度量了分布的厚尾性。常用的偏度系数有：

1. **样本偏度系数**（**sample skewness**）：是偏度系数的最常用度量，定义为：
$$b_1 = \frac{N^2}{(N-1)(N-2)} \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{s^3}$$
当样本偏度系数大于 0 时为右偏，小于 0 时为左偏。

2. **非参数偏度**（**nonparametric skew**）：定义为：
$$b_2 = \frac{\bar{x} - M}{s}$$

即当样本均值大于中位数时为右偏，小于时为左偏。

当分布对称时，两个偏度系数都等于 0。注意两种定义下的偏度系数可能会出现符号相反的情况，即样本偏度系数大于 0 并不一定代表均值大于中位数。

而峰度系数度量了数据分布的厚尾特性（tailedness），其定义为：

$$k = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4}{s^4}$$

由于正态分布的峰度系数为 3，因而应用中经常将峰度系数减 3 处理，即定义：

$$k = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4}{s^4} - 3$$

需要注意的是，虽然峰度系数中文名称似乎与分布的峰有关，然而其度量的是分布的尾巴的厚度。虽然一般情况下，厚尾伴随着尖峰，然而情况并不总是如此。

3.1.4 相关系数

我们经常会关心两组数据 $\{(x_i, y_i), i = 1, ..., N\}$ 的相关性，此时需要使用相关系数。常用的相关系数包括：

1. **Pearson 相关系数（Pearson correlation coefficient）**：是最常用的相关系数，简称样本相关系数（sample correlation coefficient），其计算公式为：

$$\rho = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{(N-1)\, s_x s_y} = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}}{\frac{N-1}{N} s_x s_y} = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}}\sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}}}$$

   注意样本相关系数只能度量变量之间的线性相关性。

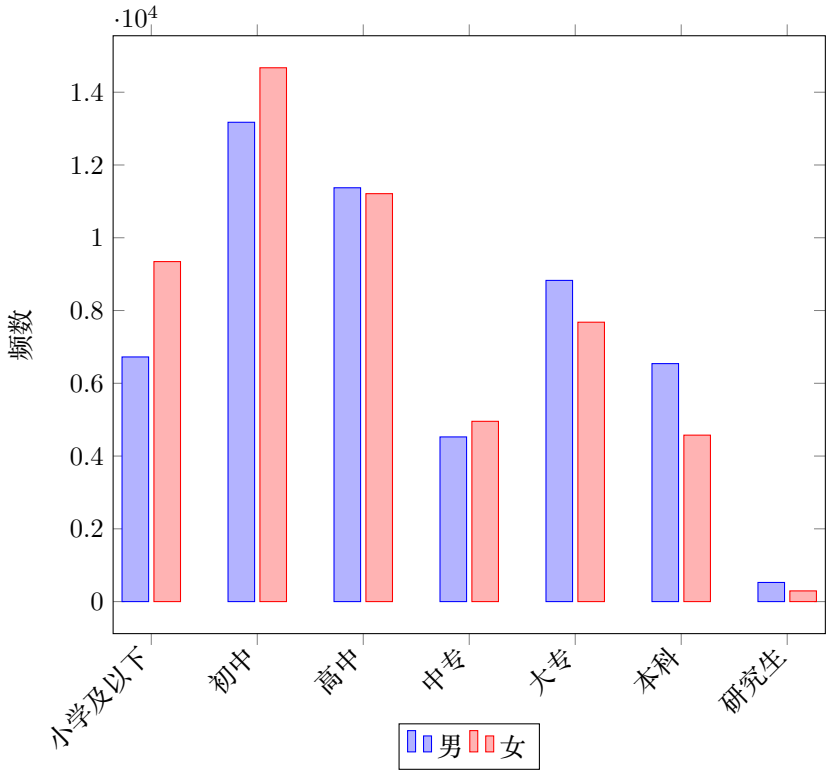
2. **Spearman 秩相关系数（Spearman's rank correlation coefficient）**：即数据排序的相关系数，度量了两个变量单调变换的相关性。如果记 $r(x_i)$ 为 x_i 在样本中的排序，$r(y_i)$ 为 y_i 在样本中的排序，那么秩相关系数被定义为 $r(x_i)$ 与 $r(y_i)$ 的样本相关系数。

例 4. 一组数据：$\{(1,1), (2,4), (3,9)\}$，可以得到 $\bar{x} = 2, \bar{y} = \frac{14}{3}, s_x = 1, s_y = \frac{7}{\sqrt{3}}$，其样本相关系数为：

$$\rho = \frac{(1 + 8 + 27) - 3 \times 2 \times \frac{14}{3}}{2 \times 1 \times \frac{7}{\sqrt{3}}} \approx 0.9897$$

而两列数据的排序分别为：$\{(1,1), (2,2), (3,3)\}$，因而其秩相关系数为：

$$r = \frac{(1 + 4 + 9) - 3 \times 2 \times 2}{2 \times 1 \times 1} = 1$$



数据来源：2009 年中国城镇住户调查

图 4: 条形图示例：我国人口教育程度分布

即两组数据存在着完全负向的单调关系，然而并不存在完全单调的负向线性关系。

3.2 数据的图表展示

虽然以上描述性统计量从各个方面对数据进行了描述，然而这些统计量仍然不够直观。而图、表可以以更加直观的方式呈现数据。下面我们分别介绍一些简单的数据图表。

3.2.1 统计图

图形可以非常直观的展示数据的分布等情况。以下介绍几种最为常见的统计表：

1. 条形图（bar chart）：为一个二维图，其中 x 轴为分类数据或者顺序数据，纵轴可以为频数、频率或者其他数值型数据等。当纵轴为为频数时，条形图的高度表示频数，而条形图的宽度没有意义。

2. 饼图（pie chart）：使用圆形及扇形角度表示比例，一般用于展示分类数据

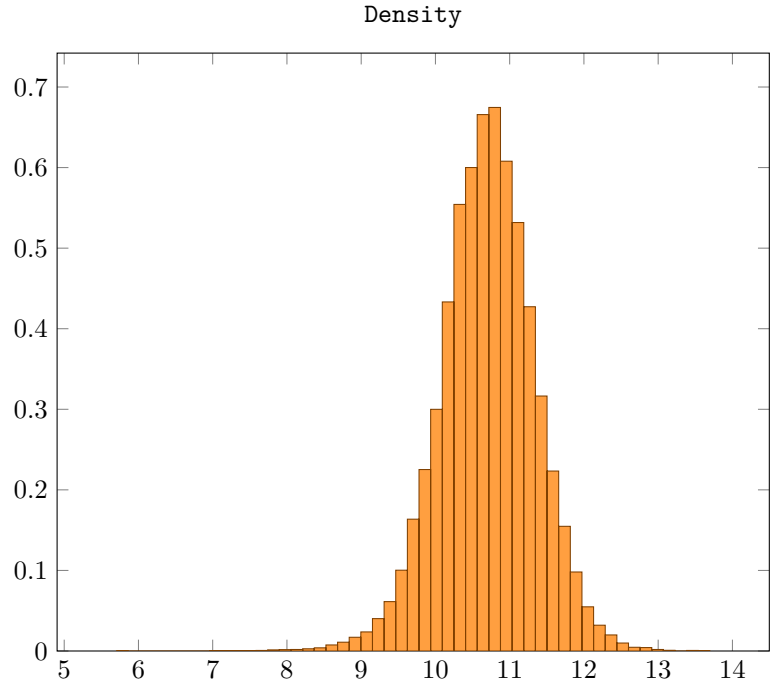


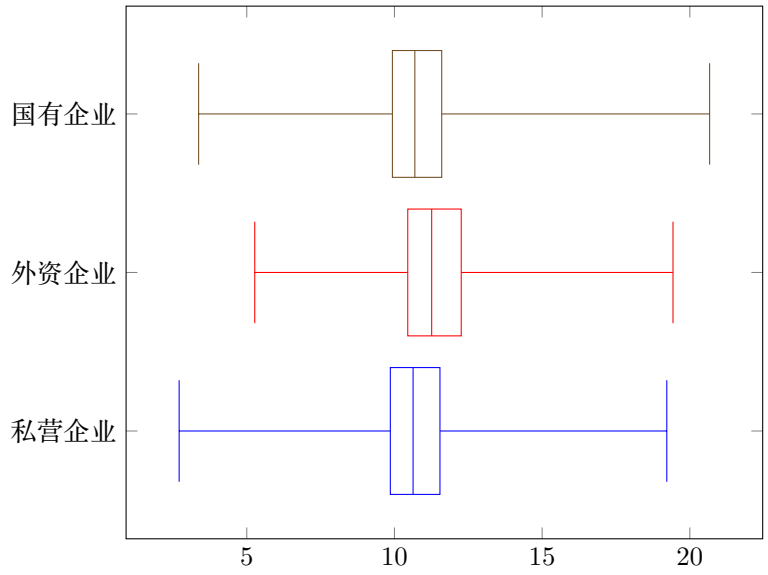
图 5: 直方图示例：我国家庭收入对数的分布

的比例。

3. 直方图 (histogram)：用来表示数值型数据密度分布的一种图，使用矩形**面积**代表落在某一区间内概率的大小，是数值型数据最常用的分布图形。注意直方图与条形图的差别，条形图针对分类数据，其宽度没有意义，只有长度有意义，而直方图每个柱形的宽度代表区间大小，其面积代表落入该区间的概率；条形图针对分类数据，因而不同类别之间矩形是分开的，而直方图代表的是连续的区间，因而矩形之间是紧密相连的。

4. 箱线图 (box chart)：即将数据的最小值、最大值、中位数、四分位数 (Q_1, Q_3) 画在图中，两个四分位数作为「箱子」，并在中间表示中位数，再将最大值、最小值与箱子连接所做的图 (如图 (6) 所示)。箱线图可以比较直观的观察数据的平均水平、离散程度以及偏态。

5. 散点图 (scatter diagram)：一种二维图，x 轴与 y 轴分别表示一个变量，将数据的 $x - y$ 组合以散点的形式画在图上，表示两个变量之间的相关关系。例如，图 (7) 展示了 2011 年我国人口大于 300 万的城市的人口数与其人口数排名之间的关系，可以看到除了某些异常点之外，两个变量之间有着近似的线性关系。此外，散点的颜色、大小等也可以表示第三维数据，如气泡图等。



数据来源：2013 年中国工业企业数据库

图 6: 箱线图示例：分所有制企业规模（ln (总资产)）

6. 线图（line plot）：横轴为某一序列（如时间），纵轴为数值型数据，一般用来描述时间序列数据。

在画图时需要特别注意图的坐标轴。例如，为了便于比较，一般条形图、直方图、线图的横轴应该以 0 开始，不以 0 为起始的图经常给人以误导。

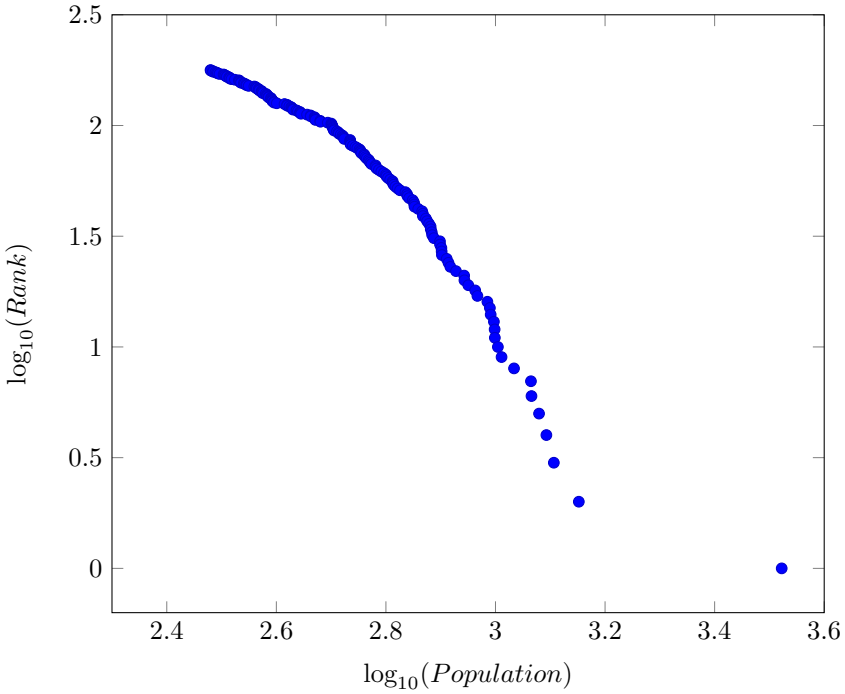
除了以上介绍的这些图之外，还有很多其他类型的图，在这里我们不一一赘述。

3.2.2 统计表

除了图之外，统计表格也是经常使用的描述性统计工具。一般来说，一张统计表应该包含表头、行标题、列标题、数值、附注等部分。此外，在格式上，统计表格一般要遵循一定的规范，如表格除了上下两条横线用粗线之外，其他线一般用细线；统计表格一般两边不封口。表 (1) 展示了一张典型的描述性统计表格。

4 充分统计量

以上我们介绍了统计量的概念。我们知道，在参数模型中，我们假设总体 P 属于某一个参数族 $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$，其中 $\theta \in \Theta$ 为参数。在获得了一个样本 $x = (x_1, ..., x_N)$ 之后，我们通常会使用一些统计量 $T(x)$ 来描述总体 P 的分布，一般而言，这些统计量的个数小于样本数量。一个很自然的问题是，这些统计量 $T(x)$ 在何种程度上代表了样本 x 所包含的信息呢？是不是存在有限个统



数据来源：2011 年《中国城市统计年鉴》

图 7: 散点图示例：我国城市人口排名与人口数的关系

表 1: 描述性统计表示例

变量	(1) 样本量	(2) 均值	(3) 标准差	(4) 最小值	(5) 最大值
总人口	286	439.5	312.0	19.50	3,330
人均生产总值	286	38,812	24,247	6,457	163,014
第一产业比重	286	13.06	8.130	0.0600	48.64
第二产业比重	286	51.96	10.49	17.02	89.34
第三产业比重	286	34.98	9.056	10.15	76.07

注：数据来源：2011 年《中国城市统计年鉴》

计量使得这些统计量能够完全代表样本的信息呢？这里我们需要所谓的**充分统计量**（**Sufficient statistics**）的概念。

定义 1. （充分统计量）若 $x = (x_1, ..., x_N)$ 为来自于未知总体 $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ 的一组样本，如果给定一组统计量 $T(x)$，样本的条件分布 $f(x|T(x) = t)$ 不依赖于 θ，那么我们称 $T(x)$ 为充分统计量。

以上定义意味着，如果我们计算得到了充分统计量 $T(x)$，那么样本 x 不包含除了 $T(x)$ 之外的关于 θ 的任何信息，或者说，充分统计量 $T(x)$ 包含了使用样本 x 对总体 P（或者等价的，θ）进行推断所需要的所有信息。这也就意味着，如果两个样本 x^1 和 x^2，有 $T\left(x^1\right) = T\left(x^2\right)$，那么我们对于 θ 的所有推断应该是等价的，而不管 $x^1 = x^2$ 是否成立。

例 5. 假设 $x_i \sim Ber(p)$ $i.i.d$，$x = (x_1, ..., x_N)$，那么样本的联合分布为：

$$f(x = \tilde{x}) = \prod_{i=1}^{N} p^{\tilde{x}_i} (1-p)^{1-\tilde{x}_i} = p^{\sum_{i=1}^{N} \tilde{x}_i} (1-p)^{N-\sum_{i=1}^{N} \tilde{x}_i} \tag{4}$$

如果记统计量 $N_1(x) = \sum_{i=1}^{N} x_i$，那么 $N_1(x) \sim Bi(N, p)$。由于 $f(x|N_1(x)) = \frac{f(x, N_1(x))}{f_{N_1}(N_1(x))}$，因而我们需要计算 $f(x, N_1(x))$ 的联合分布。可得：

$$f(x = \tilde{x}, N_1(x) = n) = \begin{cases} p^n (1-p)^{N-n} & if \ \sum_{i=1}^{N} \tilde{x}_i = n \\ 0 & if \ \sum_{i=1}^{N} \tilde{x}_i \neq n \end{cases}$$

因而：

$$\begin{aligned} f(x|N_1(x) = n) &= \begin{cases} \frac{p^n (1-p)^{N-n}}{\begin{pmatrix} N \\ n \end{pmatrix} p^n (1-p)^{N-n}} & if \ \sum_{i=1}^{N} x_i = n \\ 0 & if \ \sum_{i=1}^{N} x_i \neq n \end{cases} \\ &= \begin{cases} \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} & if \ \sum_{i=1}^{N} x_i = n \\ 0 & if \ \sum_{i=1}^{N} x_i \neq n \end{cases} \end{aligned}$$

注意以上条件分布并不依赖于未知参数 p，因而 $N_1(x)$ 是 p 的一个充分统计量。

以上我们通过观察猜测的方式找到了伯努利总体的充分统计量，然而该过程比较繁琐。实际上，在寻找充分统计量时，我们有如下简单的定理可以使用：

定理 2. （因子分解定理）若 $x = (x_1, ..., x_N)$ 为来自于未知总体 $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ 的一组样本，令 $f(x|\theta)$ 样本的联合概率密度函数。统计量 $T(x)$

为充分统计量的充要条件是存在函数 $g(t|\theta)$ 和 $h(x)$，使得：

$$f(x|\theta) = g(T(x)|\theta) \cdot h(x)$$

例如，通过观察式 (4) 可以发现，$f(x|p) = p^{N_1(x)}(1-p)^{N-N_1(x)}$，因而 $N_1(x)$ 是其充分统计量。

例 6. 假设 $x_i \sim U(0,\theta)$ $i.i.d$，那么样本 x 的联合密度函数为：

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^{N}\left[\frac{1}{\theta}1_{(0,\theta)}(x_i)\right] \\ &= \frac{1}{\theta^N}\prod_{i=1}^{N}1_{(0,\theta)}(x_i) \\ &= \frac{1}{\theta^N}1\left\{\max_i x_i \le \theta\right\} \end{aligned}$$

因而 $T(x) = \max_i\{x_i\}$ 即其充分统计量。

例 7. 若 $x_i \sim P(\lambda)$ $i.i.d$，那么样本 x 的联合密度函数为：

$$\begin{aligned} f(x|\lambda) &= \prod_{i=1}^{N}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda} \\ &= e^{-N\lambda}\lambda^{\sum_{i=1}^{N}x_i}\prod_{i=1}^{N}\frac{1}{x_i!} \end{aligned}$$

其中可以令 $h(x) = \prod_{i=1}^{N}\frac{1}{x_i!}$，$T(x) = \frac{1}{N}\sum_{i=1}^{N}x_i$，$g(T(x)|\lambda) = e^{-N\lambda}\lambda^{NT(x)}$，因而 $T(x) = \frac{1}{N}\sum_{i=1}^{N}x_i$ 是泊松分布的充分统计量。

例 8. 假设 $x_i \sim N\left(\mu,\sigma^2\right)$ $i.i.d$，那么样本 x 的联合密度函数为：

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right\} \\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i^2+\mu^2-2\mu x_i\right)\right\} \\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}x_i^2+N\mu^2-2\mu\sum_{i=1}^{N}x_i\right)\right\} \\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{N}{2\sigma^2}\left(\overline{x^2}+\mu^2-2\mu\bar{x}\right)\right\} \end{aligned}$$

因而 $T(x) = \left(\bar{x},\overline{x^2}\right)'$ 是正态分布的充分统计量。由于 $s^2 = \frac{N}{N-1}\left(\overline{x^2}-\bar{x}^2\right)$，因

而 $T'(x)=\left(\bar{x},s^2\right)$ 也是正态分布的充分统计量。

回忆指数分布族的定义，我们发现以上两例中分布都属于指数分布族。实际上，根据指数分布族的定义：

$$f(x_i|\theta)=h(x_i)\cdot\exp\left\{\sum_{k=1}^{K}\left[\eta_k(\theta)\cdot T_k(x_i)\right]-B(\theta)\right\}$$

那么在独立同分布的假定下，样本 $x=(x_1,...,x_N)$ 的联合分布为：

$$f(x|\theta)=\left[\prod_{i=1}^{N}h(x_i)\right]\cdot\exp\left\{\sum_{k=1}^{K}\left[\eta_k(\theta)\cdot\sum_{i=1}^{N}T_k(x_i)\right]-NB(\theta)\right\}$$

因而 $T(x)=\left[\sum_{i=1}^{N}T_1(x_i),...,\sum_{i=1}^{N}T_K(x_i)\right]$ 为其充分统计量。

例 9. 对于伯努利分布，其分布函数：

$$\begin{aligned}f(x_i|\lambda)&=\exp\{x_i\ln p+(1-x_i)\ln(1-p)\}\\&=\exp\left\{x_i\ln\frac{p}{1-p}+\ln(1-p)\right\}\end{aligned}$$

因而 $T(x)=\sum_{i=1}^{N}x_i$ 为其充分统计量。

进一步，在得到充分统计量之后，我们会关心充分统计量的分布。对于指数分布族，我们有如下定理：

定理 3. 若 $x=(x_1,...,x_N)$ 为来自于未知总体 $P\in\mathcal{P}=\{P_\theta:\theta\in\Theta\}$ 的一组独立同分布的样本，若 $\{P_\theta:\theta\in\Theta\}$ 为指数分布族，即其密度函数为：

$$f(x_i|\theta)=h(x_i)\cdot\exp\left\{\sum_{k=1}^{K}\left[\eta_k(\theta)\cdot T_k(x_i)\right]-B(\theta)\right\}$$

那么 $T(x)=\left[\sum_{i=1}^{N}T_1(x_i),...,\sum_{i=1}^{N}T_K(x_i)\right]$ 为其充分统计量。若集合 $\{(\eta_1(\theta),...,\eta_K(\theta)),\theta\in\Theta\}$ 包含了 $\mathbb{R}^K$ 中的一个开集，那么 $T(x)$ 的联合密度函数同样属于指数分布族，且具有如下形式：

$$f_T(t_1,...,t_K|\theta)=H(t_1,...,t_K)\exp\left\{\sum_{k=1}^{K}\left[\eta_k(\theta)\cdot t_k\right]-NB(\theta)\right\}$$

注意集合 $\{(\eta_1(\theta),...,\eta_K(\theta)),\theta\in\Theta\}$ 包含了 $\mathbb{R}^K$ 中的一个开集的假设排除了诸如 $N\left(\mu,\mu^2\right)$ 这样的指数分布族。

例 10. 例 (9) 中，我们得到伯努利分布的一个充分统计量为 $T(x)=\sum_{i=1}^{N}x_i$。

根据上述定理，充分统计量 $T(x)$ 的密度函数为：

$$\begin{aligned} f_T(t) &= H(t)\exp\left\{t\ln\frac{p}{1-p}+N\ln(1-p)\right\} \\ &= H(t)\exp\{t\ln p - t\ln(1-p)+N\ln(1-p)\} \\ &= H(t)\exp\{t\ln p+(N-t)\ln(1-p)\} \\ &= H(t)\,p^t(1-p)^{N-t} \end{aligned}$$

实际上，我们知道 $T(x)$ 服从二项分布，即 $T(x)\sim Bi(N,p)$，可以验证，以上的形式当 $H(t)=\begin{pmatrix} N \\ t \end{pmatrix}$ 时，上述密度函数即得到了二项分布，验证了以上定理。

实际上，对于某一个总体，可能有不止一组充分统计量。比如，次序统计量 $T(x)=\left(x_{(1)},...,x_{(N)}\right)$ 包含了样本所有的信息，因而一定是充分统计量，然而这样的充分统计量并没有达到数据压缩的目的。因而自然的想法是，在所有的充分统计量中，我们是不是可以找到一组最少的充分统计量。为此，我们定义**最小充分统计量**（**minimal sufficient statistics**）的概念：

定义 2. 一个充分统计量 $T(x)$，如果对于任何其他的充分统计量 $T'(x)$，$T(x)$ 都是 $T'(x)$ 的函数，那么我们称 $T(x)$ 为一个最小充分统计量。

以上定义意味着，$T(x)$ 本身就是充分统计量，而且 $T(x)$ 可以由其他的任何充分统计量 $T'(x)$ 计算得到，因而从这个意义上说，$T(x)$ 比 $T'(x)$ 要「小」。例如，对于泊松分布，我们已经找到其一个充分统计量为 $T(x)=\sum_{i=1}^{N}x_i$，而次序统计量也是其充分统计量，我们可以使用次序统计量计算出 $T(x)$，但是不能通过 $T(x)$ 计算出次序统计量，因而 $T(x)$ 是比次序统计量要「小」的充分统计量。

以下定理可以帮助我们寻找最小充分统计量：

定理 4. 对于任意的两组来自于未知总体 $P\in\mathcal{P}=\{P_\theta:\theta\in\Theta\}$ 的样本 $x=(x_1,...,x_N)$，$y=(y_1,...,y_N)$，如果存在统计量 $T(x)$，使得当两个样本联合密度函数可以写为 $f(x|\theta)=f(y|\theta)\,\phi(x,y)$ 的形式时，必然有 $T(x)=T(y)$，那么 $T(x)$ 为 θ 的最小充分统计量。

例 11. 假设 $x_i\sim N\left(\mu,\sigma^2\right)$ $i.i.d$，由于密度函数在任何一个点处都大于 0，那么任意两个样本 x 和 y 联合密度函数的比例 $\frac{f(x|\theta)}{f(y|\theta)}$ 为：

$$\begin{aligned} \frac{f(x|\theta)}{f(y|\theta)} &= \frac{(2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{N}{2\sigma^2}\left(\overline{x^2}+\mu^2-2\mu\bar{x}\right)\right\}}{(2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{N}{2\sigma^2}\left(\overline{y^2}+\mu^2-2\mu\bar{y}\right)\right\}} \\ &= \exp\left\{-\frac{N}{2\sigma^2}\left[\left(\overline{x^2}-\overline{y^2}\right)-2\mu\left(\bar{x}-\bar{y}\right)\right]\right\} \end{aligned}$$

如果上述比例与 θ 无关则必然有 $\overline{x^2}=\overline{y^2},\bar{x}=\bar{y}$，所以 $T(x)=\left(\bar{x},\overline{x^2}\right)$ 为正态分布的最小充分统计量。

实际上，最小充分统计量并非只有一组，比如由于 $s^2=\frac{N}{N-1}\left(\overline{x^2}-\bar{x}^2\right)$，因而 $\left(\bar{x},s^2\right)$ 是 $\left(\bar{x},\overline{x^2}\right)$ 的函数，同时反过来 $\left(\bar{x},\overline{x^2}\right)$ 也是 $\left(\bar{x},s^2\right)$ 的函数，所以 $\left(\bar{x},s^2\right)$ 也是正态分布的最小充分统计量。如果找到了一组最小充分统计量，那么这组最小充分统计量的所有一一映射都是最小充分统计量。

习题

练习 1. 等式 $\mathbb{E}s=\sigma$ 是否成立？如果成立，请证明，如果不成立，请指出其大小关系。

练习 2. （编程题）使用 CFPS 数据中的「cfps_family_econ.dta」数据集，分别计算：

1. 每个家庭的房贷支出占家庭总收入的比例（s）；
2. 每个家庭房贷支出比例的均值（$\bar{s}$）；
3. 每个家庭房贷的均值（$\bar{m}$）及收入的均值（$\bar{y}$）；
4. 比较 $\bar{s}$ 与 $\bar{m}/\bar{y}$ 的大小，为什么会出现这种情况？从中你有何种启发？

练习 3. 证明 $F^{-1}(q)$ 是以下最小化问题的解：

$$\min_c \mathbb{E}\psi_q(X-c)$$

练习 4. 求以下分布的充分统计量：

1. 泊松分布
2. 指数分布
3. 正态分布
4. Beta 分布

练习 5. （编程题）模拟 1000 次来自于正态总体 $N\left(\mu,\sigma^2\right)$ 的样本均值，样本量分别设为 $N=(5,10,100)$，观察其抽样分布是否为正态分布。

参考文献

[1] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[2] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.