统计学与统计量

司继春

[1]上海对外经贸大学

2023年10月

概览

统计学中的数据

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；
- 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）。

数据的种类

根据数据度量的层次，一般可以将数据分为以下三类：

- 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、省份等。
- 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。
- 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。

分类变量数据和顺序变量数据又可以合称为定性数据（qualitative data），而相应的数值变量数据为定量数据（quantitative data）。

根据个体和时间进行划分

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）**：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。
  - 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。
- **时间序列数据（time series data）**：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。
  - 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。

根据个体和时间进行划分

除这两种外，还有这两种类型数据的合并数据，如：

- **面板数据（panel data）或者纵向数据（longitudinal data）**：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$N \gg T$的面板数据称为短面板（short panel）数据。
- **重复截面数据（repeated cross-sectional data）**：与面板数据类似，不同的是不同时期所观测到的个体是不同的
  - 比如每两年进行一次调查，但是每次调查的对象都不相同，那么这就不构成面板数据，而是重复截面数据。

统计方法

一般来说，统计方法可以分为描述性统计方法和推断统计。

- 描述性统计（descriptive statistics）：通过表和图的形式对数据加以展示，其基本工具为描述性统计量，用以描述数据的分布特征。
  - 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。
- 统计推断（statistical inference）：通过概率建模的方法，使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）等方法。

总体和样本

- 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）**。
- 而其中$\boldsymbol{x} = [x_1, ..., x_N]'$为从总体中进行抽样的一个**样本（sample）**，其中的x_i为一个**观测（observation）**。
- 在接下来的符号中，如无必要我们将不区分随机变量X_i及其实现x_i，从而x_i即指代随机变量X_i。此外，我们用粗体的$\boldsymbol{x}$代表样本。
- 总体可以区分为**有限总体（finite population）**和**无限总体（infinite population）**。其中：
  - 无限总体指潜在的有无数多个个体，或者无法确定潜在研究对象的个数
  - 比如宇宙中的行星、一条生产线在生命周期内所生产的所有产品等等。

有限总体和抽样

而在另一些问题中，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。

- 对所有关注的个体进行调查也是有可能的，这种方法称为普查（census）
  - 我国有人口普查、经济普查等。
- 一般的，如果令$\mathcal{P} = \{y_1, y_2, \ldots, y_M\}$为我们所关心的全体，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。
  - S中元素的个数：样本量（sample size）

抽样方法

抽样方法分为两种：

- **概率抽样（probability sampling）**：根据随机性原则进行抽样，总体中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。
- **非概率抽样（nonprobability sampling）**：总体中有些个体被抽中的概率为0，或者被抽中的概率无法被确定。
  - 由于非概率抽样的样本可能是有系统性偏差的，所以很难代表总体。
  - 常用的非概率抽样包括方便抽样、自愿样本、滚雪球抽样、配额抽样等等。
  - 由于非概率抽样的统计性质难以分析，在进行调查时应尽量避免使用非概率抽样

抽样框

抽样框（sampling frame），即一个可以识别和区分抽样单位的名单。

- 最简单的情况下，如果我们需要对一个学校的学生进行抽样，我们可以把该学校包含姓名、学号的名单当做抽样框。
- 而有的时候直接对个人进行抽样比较困难时，我们也可能将个体所属的单位进行抽样
  - 比如如果要在全市的居民中进行抽样，全市所有居民的名单是非常难以获得的，我们可以转而列出该城市的所有街道、乡镇作为抽样框，进而在街道、乡镇进行进一步抽样。

简单随机抽样

最简单的抽样方法即从总体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）

- 优点：简单随机抽样的样本被抽中的概率是等可能的，所以统计分析非常方便，接下来如果没有特殊假定，我们都假设样本来自于简单随机抽样。
- 缺点：成本高，甚至不可能。

其他抽样方法

- 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。
- 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。
- 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。

其他抽样方法

- 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。
- 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。

统计推断

- 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。

统计模型

- 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。
- 统计模型分为
  - 参数模型（parametric model）
  - 非参数模型（nonparametric model）
  - 半参数模型（semi-parametric model）

参数模型

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。

参数模型

测量问题

我们可以假设每一次测量$x_i \sim N\left(\mu, \sigma^2\right)$，其中$\mu$为测量的真实值（如真实体温、长度），而$\sigma^2$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$P \in \left\{N\left(\mu, \sigma^2\right)\right\}$，参数空间为$\Theta = \mathbb{R} \times \mathbb{R}^+$。只要$\mu$和$\sigma^2$确定了，那么总体$P$也就确定了。

数据生成过程

数据生成过程（data generating process, DGP），即使用概率语言描述现实中的数据是如何被“生成”的。

- 特别是对于多维的随即向量$x_i \in \mathbb{R}^k$，通常直接建模x_i的联合分布是相对复杂的，我们通常会使用条件分布等工具对数据进行建模。

数据生成过程

数据生成过程

上一章中我们对同时对到达银行的人数和办理外汇业务的人数(N, M)进行了建模，如果我们可以观察到样本$\{(n_i, m_i), i = 1, ..., N\}$，那么

$$M|N \sim Bi(N, p), N \sim P(\lambda)$$

实际上建模了样本(n_i, m_i)被“生成”出来的概率模型，即数据生成过程。

数据生成过程

数据生成过程

高斯混合模型中，通常我们只能观察到X的样本$\{x_i\}$，而不能观察到相应的D的实现d_i，以下模型：

$$\begin{cases} X|D=1 \sim N\left(\mu_1, \sigma_1^2\right) \\ X|D=0 \sim N\left(\mu_0, \sigma_0^2\right) \\ D \sim Ber\left(p\right) \end{cases}$$

同样建模了x_i的数据生成过程，而其中由于变量D不可观测，我们通常称其为潜变量（latent variable）。

统计量

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$\{x_i, i = 1, 2, ..., N\}, \boldsymbol{x} = [x_1, ..., x_N]'$，统计量即样本的一个不依赖于其具体实现的函数$T(\boldsymbol{x})$。

- 由于$\boldsymbol{x}$为随机向量，因而统计量$t = T(\boldsymbol{x})$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：
  - 统计量的分布称为**抽样分布**（sampling distribution）
  - 统计量的标准差则被成为**标准误**（standard error）

统计量示例：样本均值

样本均值

统计学中一个最简单、最常用的统计量即样本均值，即对于一组随机样本$(x_1, ..., x_N)$，样本均值被定义为：

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

注意到样本均值$\bar{x}$作为一个统计量，也是一个随机变量，其随机性来源于抽样的随机性：由于我们只有有限样本，而样本是从总体中随机抽取的，从而即使样本量相同，两个不同的样本计算得到的$\bar{x}$也一定是不相同的。

统计量示例：样本均值

样本均值

比如，如果假设$[x_1, ..., x_N]'$为随机样本，来自于正态总体：$x_i \sim N\left(\mu, \sigma^2\right)$，那么样本$[x_1, ..., x_N]'$可以看作是从正态分布$N\left(\mu, \sigma^2\right)$中抽取的N个随机数，比如在Stata中，我们可以使用如下代码模拟该过程：

```
clear
set obs 100
gen x=rnormal(1,2)
```

即使用随机数生成函数rnormal()生成了100个期望为1，方差为4的随机样本。我们可以使用su命令计算其均值：

```
su x
```

统计量示例：样本均值

样本均值

如果我们把以上过程重复10000遍，就可以得到10000个样本均值，我们使用以下代码完成以上过程：

代码 1: 样本均值的抽样分布

```
// statistic_mean.do
clear all
set seed 19880505
frame create means m
forvalues i=1/10000{
    clear
    set obs 100
    gen x=rnormal(1,2)
    qui: su x
    frame post means (r(mean))
}
frame change means
su m
```

统计量示例：样本均值

样本均值

- 由于重复次数足够多，我们大体可以认为以下直方图即样本均值$\bar{x}$的分布，即抽样分布。
- 根据$\bar{x}$的描述性统计结果，可以得到：

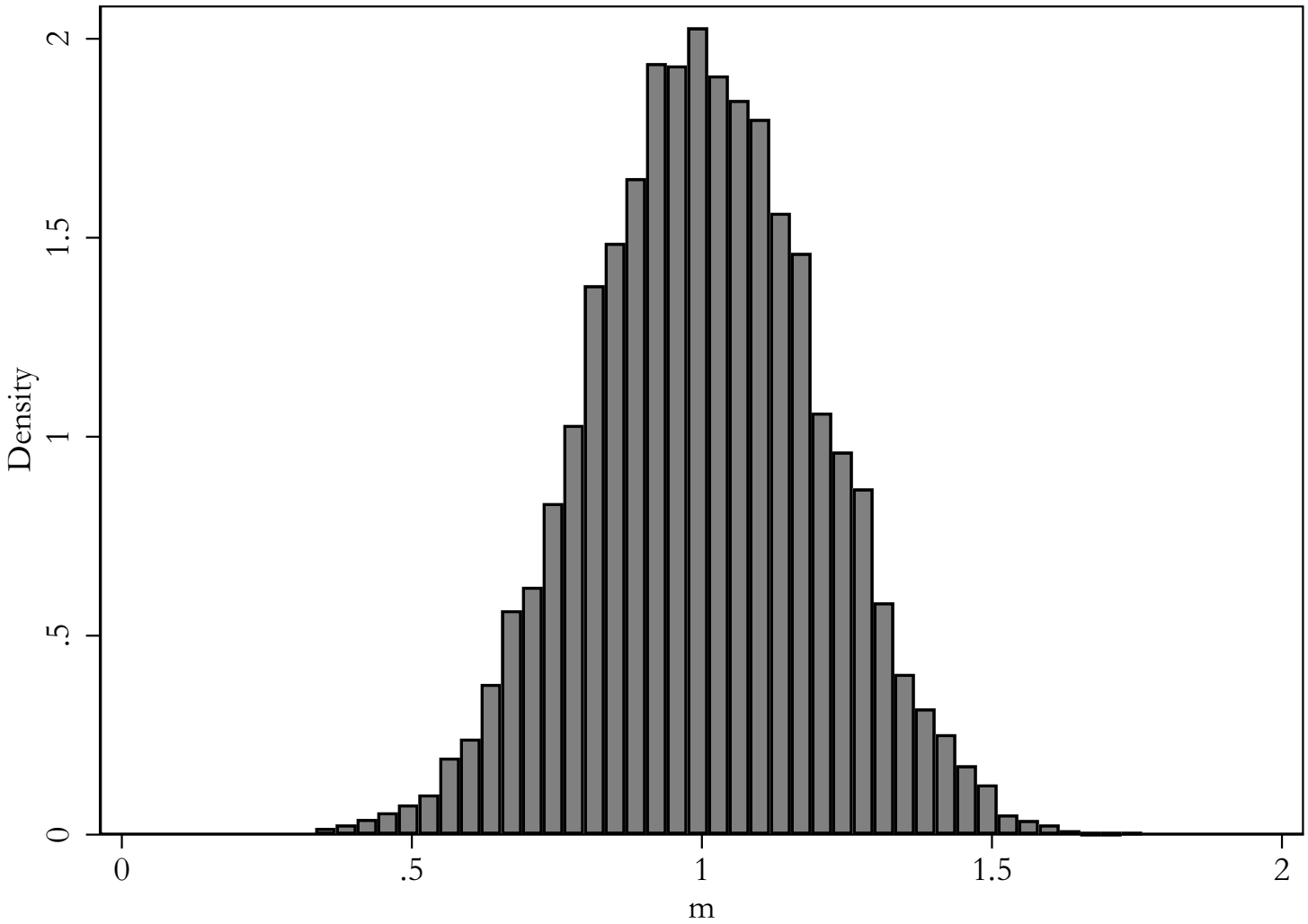
$$\begin{cases} \mathbb{E}(\bar{x}) \approx 1 \\ \mathbb{V}(\bar{x}) \approx 0.04 = \frac{4}{100} \end{cases}$$

其中我们称$\sqrt{0.04} = 0.2$，即$\bar{x}$的标准差，为标准误。

  - 注意x_i的标准差为2，而$\bar{x}$的标准差（标准误）为0.2，两者是不同的。

- 此外，在实际应用中，我们只能看到一个样本，从而抽样分布、标准误等都是无法通过以上重复抽样的方法获得的，我们需要使用统计理论计算以上的结论，这也是我们学习统计理论的重要任务。

统计量示例：样本均值



统计量

- 所有统计方法都借助于统计量来实现：
  - 描述性统计：描述性统计量（descriptive statistic）
  - 推断统计：
    - 参数估计：估计量（estimator）
    - 假设检验：检验统计量（test statistic）

统计量与参数

- 参数是总体的特征
  - 总体不可观测，从而参数不可观测
- 统计量是样本的特征
  - 统计量必须可以使用样本计算出来
  - 样本可观测，从而统计量可观测

统计学的目的：使用样本推断总体，即使用统计量推断参数。

抽样误差和非抽样误差

所有的统计模型都会存在误差，我们一般将误差分为两种：

- **抽样误差（sampling error）**：由于我们不能观察到总体，而只能观察到样本，而样本具有随机性，从而所得到的统计量与真实参数之间是由误差的，这种误差被称为抽样误差。
  - 只要我们观察到的是样本而非总体，抽样误差就是不可避免的
- **非抽样误差（non-sampling error）**：所有其他的误差，包括各种的系统性误差等。
  - 一份关于收入的调查中可能对于极其富有的人群是无法调查的，从而依据该调查计算的平均收入是低估的，这种误差是不会随着样本量的增大而减少的，也非常难以计算出非抽样误差的大小。

抽样误差

抽样误差

如果总体是全校学生的身高，而我们无法观察到总体，只能从全校学生中抽取出一部分同学观察其身高，然而，抽样的过程是随机的，两次策略相同的抽样过程也会导致抽出的样本是不相同的，从而计算得到的平均身高与真实的全体学生的平均身高总是有差异的，这个差异仅仅是由于抽样导致的，从而称为抽样误差。

抽样误差

- 抽样误差虽然不可避免，然而是可以计算的
  - 抽样分布、标准误等工具就是计算抽样误差的重要工具。
- 抽样误差与样本量紧密相关，样本量变大时抽样误差也会降低
  - 理想情况下也可以通过增加样本量的方式降低抽样误差。

抽样误差

抽样误差

在上例中，我们可以想象从M个学生的总体中抽取出N个学生的样本。如果（想象）重复抽样过程R次，我们就会得到R个样本，进而计算出R个平均数。这R个平均数并不会完全相等，而是随机的，从而会产生一个分布，从而我们可以计算抽样误差的大概范围。

随着$R \to \infty$，这个分布就会变成抽样分布。而现实中，我们无法获得R次抽样，只有一次抽样，从而只能计算出一个平均数，为了获得抽样分布，我们就必须使用统计理论推导出一定条件下抽样分布的形式，从而计算抽样误差。

随机样本的联合分布

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：

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

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。同样，联合密度函数为：

$$f(\boldsymbol{x}) = \prod_{i=1}^{N} f_{x_i}(x_i) = \prod_{i=1}^{N} f_x(x_i)$$

随机样本的联合分布

伯努利分布

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$\{x_i, i = 1, ..., N\}$，记$x_i = 1$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$P(x_i = 1) = p$。

随机样本的联合分布

伯努利分布

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:

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

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: $P(\boldsymbol{x}) = p^2 (1-p)^3$

随机样本的联合分布

测量问题

测量问题中，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$\Omega = \mathbb{R}^N$，由于假设$x_i \sim N\left(\mu, \sigma^2\right)$，所以：

$$f\left(\boldsymbol{x}\right) = \prod_{i=1}^{N} f_x\left(x_i\right) = \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left(x_i - \mu\right)^2}{2\sigma^2} \right\} \right]$$

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：

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

注意样本均值使得样本均方误差最小化，即：

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

样本均值

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

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

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

从而标准误

$$s.e.(\bar{x})=\frac{\sigma}{\sqrt{N}}$$

正态总体样本均值的抽样分布

更进一步，如果$\{x_i\}$来自于正态总体且独立同分布，
即：$x_i \sim N\left(\mu, \sigma^2\right)\ i.i.d$，由于联合正态分布中各分量的和仍然是正态分布，因而可以得到

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

或者：

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

即正态总体的样本均值服从正态分布。

样本均值

样本均值的模拟

在24页的例子中，由于样本来自于$N(1,4)$的总体，从而根据以上结论，有：

$$\mathbb{E}(\bar{x}) = 1$$

$$\mathbb{V}(\bar{x}) = \frac{4}{100} = 0.04$$

与我们模拟的结果相差无几。此外，由于总体是正态总体，从图中也可以看到\bar{x}的抽样分布也是接近正态的。

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：

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

相应的，样本标准差定义为：

$$s = \sqrt{s^2}$$

样本方差

由于:

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

从而样本方差可以写为:

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

样本方差

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：

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

样本方差的抽样分布

正态总体的样本方差标准化之后服从卡方分布：

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

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。注意：

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

对比在以上假定下：

$$\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \chi^2\,(N)$$

（why？）

自由度

直观理解：

- 由于：

$$\sum_{i=1}^{N}\frac{x_i-\bar{x}}{\sigma}=\frac{1}{\sigma}\sum_{i=1}^{N}(x_i-\bar{x})=0$$

而：

$$\sum_{i=1}^{N}\frac{x_i-\mu}{\sigma}\neq 0$$

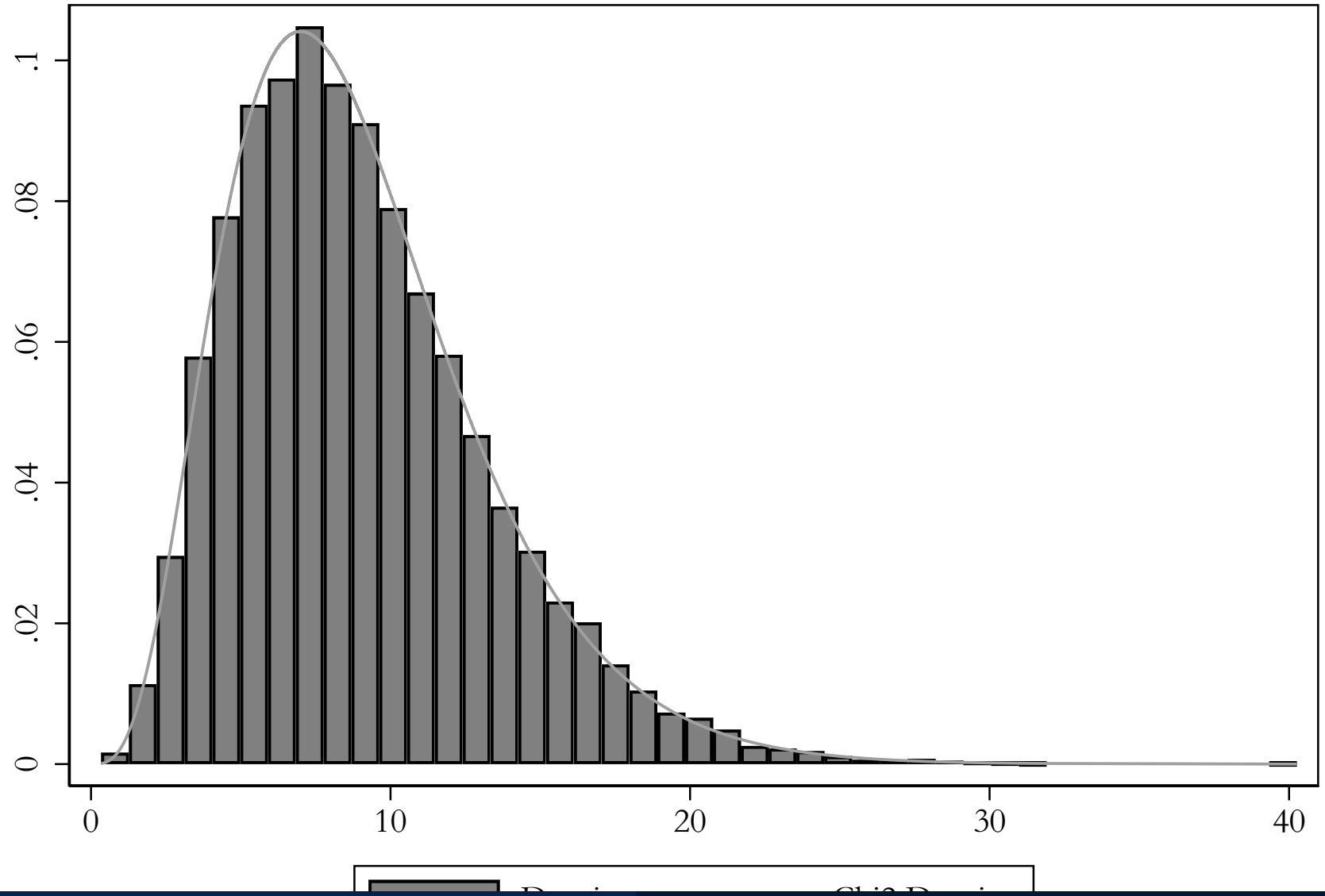
- $\left\{\frac{x_i-\bar{x}}{\sigma},i=1,...,N\right\}$中任何一个都可以被其他$N-1$个表示出，比如：

$$\frac{x_N-\bar{x}}{\sigma}=-\sum_{i=1}^{N-1}\frac{x_i-\bar{x}}{\sigma}$$

所以只有$N-1$个$\frac{x_i-\bar{x}}{\sigma}$是“自由”的

样本方差的抽样分布

模拟结果（standard_normal_var.do）：



正态总体的标准化

对于正态总体，有：

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

如果我们将上式中的总体方差σ^2替换为样本方差s^2，即：

$$\begin{aligned}\sqrt{N}\frac{\bar{x}-\mu}{s}&=\sqrt{N}\frac{\bar{x}-\mu}{\sqrt{s^2}}\\&=\sqrt{N}\frac{\frac{\bar{x}-\mu}{\sigma}}{\sqrt{s^2/\sigma^2}}\\&=\frac{\frac{\bar{x}-\mu}{\sqrt{\sigma^2/N}}}{\sqrt{\frac{(N-1)s^2}{\sigma^2}/(N-1)}}\quad\dashrightarrow\frac{N(0,1)}{\sqrt{\frac{\chi^2(N-1)}{N-1}}}\end{aligned}$$

正态总体标准化的抽样分布

然而上式不代表：

$$\sqrt{N}\frac{\bar{x}-\mu}{s}\sim t\left(N-1\right)$$

- 原因：分子分母独立性还没有验证！
- 可以证明分子和分母是独立的，从而以上结论成立
- 同样注意：这一结论是在正态总体且独立同分布的假定下得到的

标准化

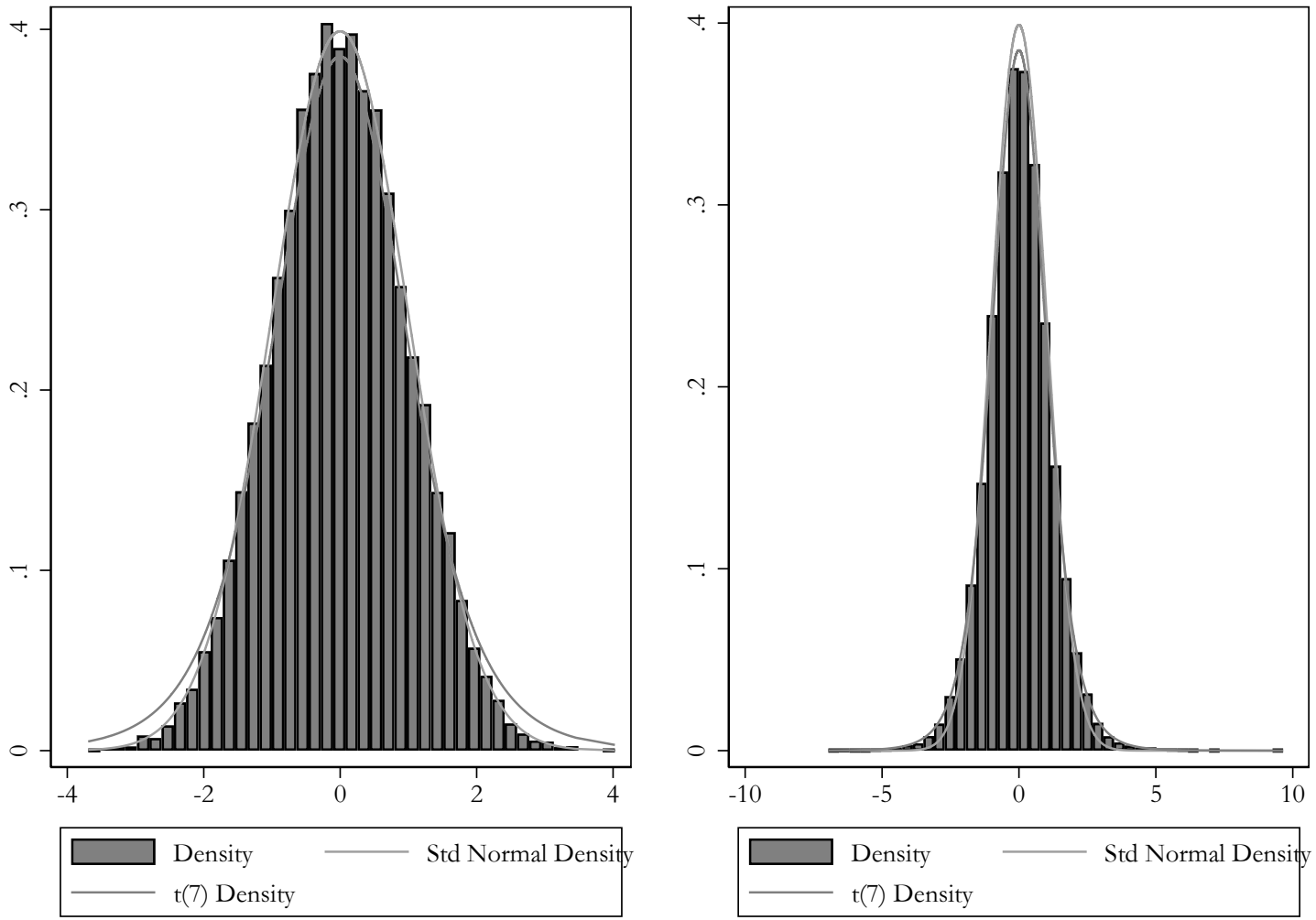
- 对于样本$\{x_i\}$，如果我们对每个样本都减去样本均值，再除以样本标准差，即：

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

  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（请验证）。

抽样分布的模拟

模拟结果（standard_normal_mean.do）：



中位数

- 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：

$$F^{-1}\left(\frac{1}{2}\right)$$

  - 如果随机从F中抽取一个随机数，一般的可能性在中位数左边，一般的可能性在中位数右边
  - 例如，由于对称性，正态分布$N\left(\mu,\sigma^2\right)$的中位数为$\mu$。

中位数

注意中位数是以下最小化问题的解：

$$\min_{c} \mathbb{E}\left|X-c\right|$$

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：

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

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

中位数

- 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：

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

分位数

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。

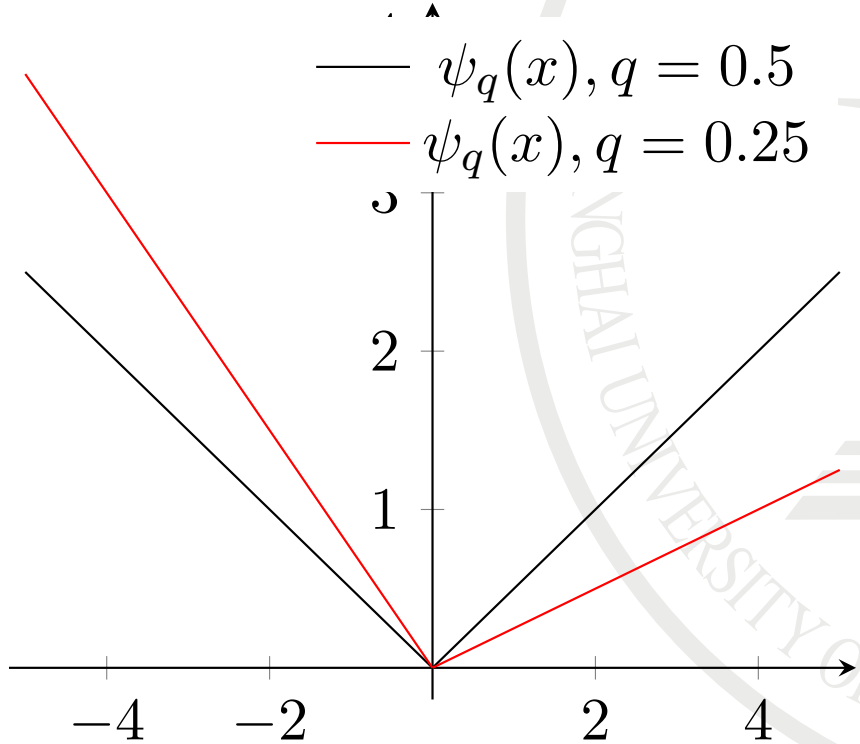
- 四分位数（quartiles，记为Q），即$Q_1 = F^{-1}(0.25), Q_2 = F^{-1}(0.5), Q_3 = F^{-1}(0.75)$，如果从$F$中抽取随机数，会有25%的小于$Q_1$，75%的小于$Q_3$。
- 百分位数（percentiles，记为P），即$P_p = F^{-1}\left(\frac{p}{100}\right), p = 1, 2, ..., 99$。

分位数

令:

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

其中$0 < q < 1$



样本分位数

- $F^{-1}(q)$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：

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

样本分位数的抽样分布

定理

如果$x_{(1)}, x_{(2)}, ..., x_{(N)}$为独立同分布随机样本$\{x_i, 1 \leq i \leq N\}$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：

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

样本分位数的抽样分布

- 令Y为小于等于x的样本数，即$Y = \#\{x_i \leq x\}$，那么$Y \sim Bi(N, F(x))$。
- 而$x_{(n)} \leq x$等价于$Y \geq n$，因而：

$$F_{x_{(n)}} = P(Y \geq n) = \sum_{k=n}^{N} \binom{N}{k} [F(x)]^k [1 - F(x)]^{N-k}$$

对以上分布函数求导可得结论。

作业

- 5.1
- 5.2
- 5.3