描述性统计

司继春

[1]上海对外经贸大学

2024年10月

概览

描述性统计

描述性统计即使用描述性统计量对数据的分布特征进行度量，通常使用图和表的形式对数据的特征进行展示。

- 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。
- 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），为何会出现这种现象?
- 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。

描述性统计

一般而言，我们对数据的描述性统计会从位置度量、离散程度度量、偏度、峰度等不同的角度进行：

- **位置度量**（**measure of location**）是是对数据中心的测量
- **离散程度度量**（**measure of dispersion**）则是对数据的变异性进行度量
- **偏度**和**峰度**则是对分布的尾部厚度进行度量。

对于不同性质的数据，其使用的统计量也是不尽相同的

定性数据的描述性统计

- 对于定性数据（包括分类数据和顺序数据），由于其取值不能直接参与算术运算，从而样本平均数、样本方差等都是无法计算的。
- 对于定性数据，一个常用的进行描述性统计的方法是制作**频数表**（**frequency table**）。频数表中一般包含每个分类的
  - 频数
  - 频率
  - 累积频率

频数表

最高学历的频数表

在cfps_adult.dta文件中，te4变量为最高学历，我们可以使用如下tabulate命令制作频数表：

```
1 tab te4
```

其中tab为tabulate命令的缩写。或者，使用outreg2命令将其导出到文件中：

```
1 outreg2 te4 using te4tabulate.tex, cross side replace
```

频数表

最高学历的频数表

	(1)	(2)
te4	Freq	Percent
-1	36***	(0.0969)
-8	33,797***	(90.98)
1	373***	(1.004)
2	668***	(1.798)
3	1,293***	(3.481)
4	549***	(1.478)
5	234***	(0.630)
6	188***	(0.506)
7	9***	(0.0242)
Total	37147	

定性数据的位置度量

- 对于定性数据，其位置度量通常使用**众数（mode）**来度量。
- 众数即样本数量最多的组别
  - 如一个班中男生30人，女生20人，那么众数即男生。众数的计算完全可以通过频数表进行。

最高学历的众数

接上例，如果我们在样本中剔除不适用、不知道的样本，再制作频数表：

```
1 tab te4 if te4>=0
```

可以得到频数最多的组为“初中”，其频率有39.02%。

频数表

- 对于顺序数据，由于其可以比较大小，也可以计算其中位数

最高学历的中位数

接上例，在剔除了不适用、不知道的样本后，最高学历变量是按照文盲、小学、初中、高中、大专、本科、硕士的顺序排序的，注意到根据累积频率，小学及以下学历占31.41%，而初中及以下占70.43%，从而学历的中位数应该在初中组。

定性数据的离散程度度量

- 而对于离散程度，相对于众数，可以使用**异众比率（variation ratio）**度量，即非众数组的比例。

最高学历的异众比率

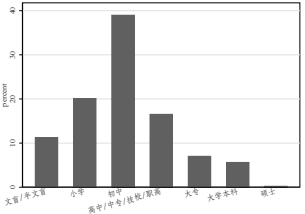
比如在上例中，我们已经知道在剔除了不适用、不知道后，众数组的比例为39.02%，那么异众比率为1-39.02%=60.98%。

- 该比率越大，意味着样本在不同的组别更加分散，从而离散程度也越大。

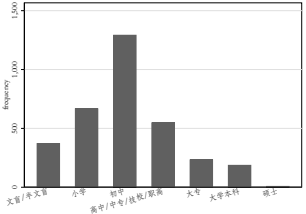
条形图

- 对于定性数据，通常可以使用**条形图（bar chart）**以及**饼图（pie chart）**等进行可视化展示。
- 条形图主要用来展示不同类别之间某些变量之间的比较，比如频率、频数以及某些其他变量的均值等等。
- 条形图制作时既可以是水平的也可以是竖直的，在制作时可以根据需求进行调整。

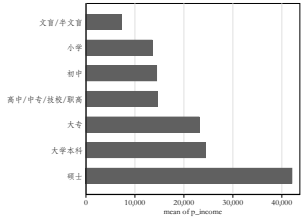
条形图


percent
0
10
20
30
40
文盲/半文盲
小学
初中
高中/中专/技校/职高
大专
大学本科
硕士

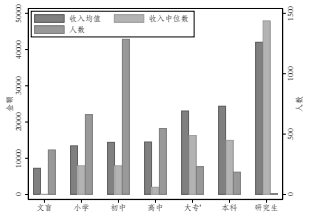
(a)


frequency
0
50
100
150
文盲/半文盲
小学
初中
高中/中专/技校/职高
大专
大学本科
硕士

(b)


文盲/半文盲
小学
初中
高中/中专/技校/职高
大专
大学本科
硕士
0
10,000
20,000
30,000
40,000
mean of pc_income

(c)


收入均值
收入中位数
人数
金额
人数
0
10000
20000
30000
40000
50000
0
50
100
150
文盲
小学
初中
高中
大专
本科
研究生

(d)

条形图的绘制

最高学历的频数、百分比条形图

如果需要描述最高学历的频率的条形图，可以使用：

```
graph bar if te4>=0, over(te4, label(angle(15)))
```

而如果需要描述最高学历的频数，可以使用：

```
graph bar (count) if te4>=0, over(te4, label(angle(15)))
```

此外，还可以根据最高学历分组，表示不同组别的平均收入：

```
graph hbar (mean) p_income if te4>=0, over(te4)
```

注意上面使用了hbar而不是bar绘制了水平的条形图

条形图的绘制

表示多个变量

如果需要如上图(d)那样，将多个变量同时画在一张图上，可以使用twoway bar命令。其中，“twoway”表示该图是一张“双向图”，可以简单理解为一张“二维”图，即有横纵坐标的图。该命令需要首先将数据整理成分类别的加总变量，该步骤一般可以使用collapse完成:

```
1 drop if te4<0
2 collapse (count) p_income (mean) mean_income = p_income (median)
     median_income = p_income, by(te4)
```

条形图的绘制

表示多个变量

然后，使用twoway命令可以将多张图画在一张图上，twoway命令后面可以接多个括号，每个括号都是一张twoway图：

```
gen te4_left=te4-0.2 // 左边柱形的横坐标
gen te4_right=te4+0.2 // 右边柱形的横坐标
twoway (bar mean_income te4_left, barw(0.2)) (bar median_income
    te4, barw(0.2)) (bar p_income te4_right, yaxis(2) barw(0.2)),
    xlabel(1 "文盲" 2 "小学" 3 "初中" 4 "高中" 5 "大专'" 6 "本科" 7 "研
    究生") ytitle("金额", axis(1)) ytitle("人
    数", axis(2)) legend(pos(11) ring(0) label(1 "收入均
    值") label(2 "收入中位数") lab(3 "人数"))
```

条形图的绘制

注意:

1. 在以上图表中，纵坐标总是从0开始的，这是一个绘图中的好习惯
    1. 如果不这样画，图中每个柱形的高度差就会失真，在视觉上会夸大不同分类之间的差距，给人不正确的印象。
2. 上图(d)使用了双坐标轴，尽管这一做法比较常见，然而实际使用中也需要额外注意
    1. 从属于不同坐标轴的柱形高度不能进行比较，然而视觉上却给人一种可以比较的错觉，而这种错觉有时会带来一些误解，在实际使用中类似双坐标轴的做法也需要谨慎对待。

饼图

而饼图的用法与条形图类似，不过饼图更加强调变量所占总的比例，而非其绝对数值。在Stata中：

- 可以使用“graph pie”命令制作饼图。
- 除了简单的饼图外，还有：
  - 环形图（doughnut chart）
  - 太阳图（sunburst chart）等各种变形。

  篇幅所限，我们不再赘述。

位置度量

- 位置度量一组数据中心的测量，对于定量数据而言，无论是前面介绍的众数、中位数，还是平均数，都可以应用在定量数据中。
- 而对于定量数据而言，最为常用的位置度量标准是**算数平均数（arithmetic mean）**，即样本均值。
- 此外，样本中位数也是定量数据位置度量的常用统计量。

位置度量

算数平均数

在CFPS数据中，如果我们要计算平均身高，可以使用summarize命令（简写为su）：

```
use datasets/cfps_adult.dta, clear
su qp101 if qp101>=0
```

注意在以上命令的结果中除了汇报了均值以外，还汇报了样本量、标准差、最小值、最大值等，并没有汇报中位数。为了计算中位数，需要在su命令中加入detail选项（简写为de）：

```
su qp101 if qp101>=0, de
```

样本均值 v.s. 中位数

- 与样本均值相比，样本中位数具有不容易受异常值影响的特点。
  - 比如，如果我们在一个40人的班级中计算平均身高，一个身高为220厘米的将会极大拉高均值，如果除去这个人的平均身高为165厘米，那么这个高个将会拉高平均值：$\frac{39\times165+220}{40}-165=1.375$厘米。
  - 而于此同时，对于中位数而言，不管这个人是220厘米还是190厘米，中位数都不会随着他的身高而改变。

样本均值 v.s. 中位数

然而为什么平常应用中还是更多使用均值呢?

- 首先是，通常而言均值更加容易计算。为了计算中位数，我们必须知道比较详细的样本数据。然而计算均值时我们只需要知道一组数据的总和以及样本量就可以了。
  - 比如，如果我们知道了全国的总工资收入和全国的人口，两者相除就是平均收入，然而根据这些信息我们无法计算收入的中位数。
  - 特别是工资收入的数据可能是把每个企业的工资支出全部加起来，此时就没有详细的个人数据，那么无论如何都无法计算出中位数了。
- 其次是，样本均值的统计性质更加简单，包括标准误、抽样分布等都容易计算和推导，这单我们将在后续的学习中更进一步了解。

样本均值 v.s. 中位数

- 此外，样本均值在某些时候的解释性更强。
  - 比如，对于配对样本，样本均值有个非常好的性质：$\overline{x_1 \pm x_2} = \bar{x}_1 \pm \bar{x}_2$例如，如果$x_1$代表一个家庭中丈夫的收入，$x_2$代表妻子的收入，那么$\overline{x_1 - x_2}$即先计算丈夫与妻子收入的差，再把收入差求平均，而$\bar{x}_1 - \bar{x}_2$则代表所有丈夫收入的均值和所有妻子收入的均值之间的差异，这也就意味着比较丈夫和妻子之间的平均收入是有意义的：$\bar{x}_1 - \bar{x}_2$可以被解释为是丈夫和妻子收入差异的平均数。
  - 而对于差的中位数并不等于中位数的差，记$M_1 - M_2$为所有丈夫收入的中维数与所有妻子的中位数之间的差距，然而中位数收入的丈夫和中位数收入的妻子可能来自于不同的家庭，此时$M_1 - M_2$的解释是比较困难的。

样本均值 v.s. 中位数

中位数的优点：

- 不容易受异常值影响，从而很多时候相比于平均数，中位数能更好反映“中间”水平
  - 全国家庭中位数收入总是小于平均收入，中位数资产总是小于平均资产
- 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$\ln(M)$。

几何平均数

除了以上的算数平均数之外，还有几何平均数、调和平均数等很多种平均数。

- **几何平均数（geometric mean）**的定义为:

$$GM = \sqrt[N]{\prod_{i=1}^{N} x_i} = \exp\left\{\frac{1}{N}\sum_{i=1}^{N}\ln\left(x_i\right)\right\}$$

根据定义，几何平均数要求x_i的取值范围为$(0, +\infty)$。

几何平均数

- 几何平均数最常见的用途是用在增长率的计算上。如果即x_t为某个变量在时间t的取值，那么其增长率可以定义为：

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1$$

而从$t = 0$到$t = T$期间的增长率为：

$$r_T = \frac{x_T - x_0}{x_0} = \frac{x_T}{x_0} - 1$$

记$GM\left(1 + r_t\right)$为$1 + r_t$的几何平均数，根据定义：

$$GM\left(1 + r_t\right) = \sqrt[T]{\prod_{t=1}^{T} \frac{x_t}{x_{t-1}}} = \sqrt[T]{\frac{x_1}{x_0}\frac{x_2}{x_1} \cdots \frac{x_T}{x_{T-1}}} = \sqrt[T]{\frac{x_T}{x_0}}$$

从而：

$$x_T = x_0 \times \left[GM\left(1 + r_t\right)\right]^T$$

从而$GM\left(1 + r_t\right) - 1$可以看作是平均增长率。

几何平均数

沪深300指数的平均收益率

我们使用hs300index.dta，即沪深300指数的数据，计算了几何平均数：

代码 1: 几何平均数的计算

```
// geometric_mean.do
use datasets/hs300index.dta, clear
// 按照时间排序，用行号作为新的时间
sort day
gen t=_n
tsset t
// 计算增长率
gen r=clsindex/L.clsindex
gen log_r=log(r)
su log_r
local geo_mean=exp(r(mean))
di "几何平均数=`geo_mean'"
di "第一天沪深指数300=" clsindex[1]
di "最后一天沪深指数300=" clsindex[_N]
```

调和平均数

- **调和平均数（harmonic mean）**：

$$HM = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}} = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{x_i}\right]^{-1}$$

- 调和平均数通常用于建模一些比率或者比值，比如速度（距离除以时间）、密度（质量除以体积）或者金融中的市盈率（股票价格除以每股收益）等。

调和平均数

速度的调和平均数

如果一辆小汽车以$80\mathrm{km/h}$开了$200\mathrm{km}$，又以$160\mathrm{km/h}$开了$200\mathrm{km}$，那么总共开了$400\mathrm{km}$，总共花了$200/80 + 200/160 = {}^{15}/_{4}\mathrm{h}$，实际平均速度为：$400\mathrm{km}/\left({}^{15}/_{4}\mathrm{h}\right) = {}^{320}/_{3}\mathrm{km/h}$。如果使用算数平均数，平均速度为$120\mathrm{km/h}$，而如果使用调和平均数，得到的结果为：

$$\frac{2}{\frac{1}{80} + \frac{1}{160}} = \frac{320}{3}$$

得到了正确的答案。

广义平均数

- 观察以上几何平均数和调和平均数，其形式是类似的
- 一般的，利用任何一个可逆的函数$f(\cdot)$，我们都可以定义一个新的平均数：

$$f^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}f(x_i)\right\}$$

- 特别是，如果我们取$f(x)=x^p$，那么平均数可以定义为：

$$\left\{\frac{1}{N}\sum_{i=1}^{N}x_i^p\right\}^{\frac{1}{p}}$$

我们就得到了**广义平均数（generalized means，或者Hölder mean）**。

- 当$p=1$时，以上定义即算数平均数；
- $p=-1$时，即调和平均数；
- 当$p\to 0$时，即几何平均数。

众数的计算

- 对于定量数据，特别是服从连续型分布的数据而言，众数无法通过定性数据中的方法计算，因为理论上对于连续型分布的随机变量，其相等的概率应该为0。
- 为了定义众数，通常可以将数据的取值范围分为m个长度相同的区间，取样本量最多的区间的组中值为该组数据的众数。
- 以上众数的计算方法非常依赖于组别的数量，从统计学理论上讲，“最优”组别个数的选择依赖于样本的数量：样本量越大，可以选取更多的组别个数，该部分理论可以参考直方图的带宽选择。

众数的计算

身高的众数

- 对于身高数据，我们首先使用如下公式将数据分为m组：

$$g_i = \left\lfloor m \times \frac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\} + \epsilon} \right\rfloor$$

其中$\lfloor x \rfloor$代表比x小的最大整数，比如$\lfloor 4.5 \rfloor = 4$，ϵ为一个非常小的整数，为了使得上式中的分式的取值范围为$[0, 1)$而非$[0, 1]$。

- 接下来，我们只要统计出每个分组的样本量，记样本量最大的组别为g^*，那么众数可以计算为：

$$mode = \min\{x_i\} + (\max\{x_i\} - \min\{x_i\} + \epsilon) \times \frac{g^* + 0.5}{m}$$

descriptive_mode.do实现了该过程

离散程度的度量

与位置度量相对应，离散程度度量通常可以使用标准差、四分位差、极差等指标。

- 其中，**标准差（standard deviation）**即样本标准差。
  - 我们一般使用标准差而非方差度量离散程度
  - 虽然标准差的平方即方差，但是方差的单位与均值是不同的，从而方差无法和均值直接进行比较，但是标准差可以。

标准差

身高的标准差

- 如果我们要度量身高的离散程度，身高的单位为厘米，那么身高的均值的单位仍然为厘米，而身高的方差的单位为厘米^2，而标准差的单位仍然是厘米，所以我们可以将均值和标准差进行比较。
- 比如，在身高的例子中，使用su命令已经计算得出标准差为8.26厘米，那么我们可以说一个身高为180厘米的人高出平均水平
  约$(180 - 164.143)\,/8.26 \approx 1.92$个标准差。

标准差

- 在上例中，我们直接将身高差异于标准差进行比较，由于单位相同，所以这种比较是合理的，而且更容易直观上理解差异的大小。
- 比如，对于正态分布而言，我们知道在均值左右各1.96个标准差即包含了95的样本：

$$P\left(|x-\mu|<1.96\sigma\right)=95\%$$

那么我们大约能判断出身高为180厘米的人应该比大约97.5%的人身高都要高了，当然前提是身高近似服从正态分布，所以1.92个标准差已经是很大的差距了。

离散系数

- 借鉴以上“去量纲”的思想，可以使用标准差进一步定义**离散系数（coefficient of variation）**：

$$v = \frac{s}{\bar{x}}$$

即标准差除以均值

- 由于标准差和均值的单位相同，从而离散系数的单位为1，可以对任意的两组数据进行比较。
- 针对单位不同，或者平均水平相差较大的两组数据，如果需要比较离散程度，那么使用离散系数是更加严谨的方法。
  - 一个简单的例子是，如果需要比较老虎的体重的离散程度和猫的体重的离散程度，那么用标准差显然是不现实的：老虎体重相差10斤是几乎可以忽略不计的，而两只猫如果相差10斤那就差别非常悬殊了。

离散系数

身高的离散系数

使用CFPS数据我们使用如下代码计算了男性和女性的体重的离散系数:

```
use datasets/cfps_adult.dta, clear
drop if qp102<0
su qp102 if cfps_gender==1
di r(sd)/r(mean)
su qp102 if cfps_gender==0
di r(sd)/r(mean)
```

计算得到男性体重的标准差为22.10斤，女性的为18.72斤，两者之间差了3.38斤，然而如果计算离散系数，男性为0.168，女性为0.166，实际上相差非常小。

四分位差和极差

- 除此之外，**四分位差（quartile deviation）**、**极差（range）**也经常被用于度量离散程度。
  - 其中四分位差定义为上四分位数与下四分位数之间的差异，即$Q_3 - Q_1$；
  - 而极差定义为$x_{(N)} - x_{(1)}$，即数据中最大值与最小值之间的差异。
  - 其中，由于四分位数不容易受到异常值的影响，从而四分位差也不太容易受到异常值的影响。

四分位差和极差的计算

在身高的例子中，根据加入detail选项的su命令的结果，可以计算出四分位差为170-159=11厘米，而极差为216-80=136厘米。

偏度系数

- 度量偏度最常用的统计量是**样本偏度系数（sample skewness）**，定义为:

$$b_1 = \frac{N^2}{(N-1)(N-2)} \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{s^3}$$

以上定义也就是随机变量偏度系数的样本对应，其中左边的$\frac{N^2}{(N-1)(N-2)}$是对自由度进行适当补偿

- 与随机变量的偏度类似，当样本偏度系数大于0时为右偏，小于0时为左偏，当分布对称时，偏度系数为0。

偏度系数的计算

在cfps_family_econ.dta数据中，使用带detail选项的su命令计算家庭收入（fincome1）的偏度系数，大约为24.31，表现出了很强的右偏特征，意味着在分布的右侧出现了拖尾，即有数量不多但是家庭收入非常之高的家庭。

非参数偏度

- 一般的，对于一个右偏的分布，均值大于中位数，比如上例中平均收入为55534.86而中位数收入为38200。
  - 这是由于右偏分布容易在分布的右侧出现极端值，而均值容易受到极端值的影响，而中位数不容易受到极端值的影响。
- 反之，左偏分布的均值一般小于中位数。
- 为此，我们还可以定义**非参数偏度（nonparametric skew）**：

$$b_2 = \frac{\bar{x} - M}{s}$$

即当样本均值大于中位数时为右偏，小于时为左偏。

- 样本偏度系数和非参数偏度系数符号有可能是不同的！（虽然少见）

非参数偏度

非参数偏度和样本偏度的符号问题

数据集rr000005.dta中记录了股票世纪星源（代码：000005）的股票价格以及收益率。我们可以使用带detai选项的su命令计算其偏度系数，计算结果样本偏度系数为0.916，为右偏，然而样本均值为-0.0009，中位数为0，从而非参数偏度为-0.0454，出现了背离的情况，虽然样本均值与中位数之间的差异非常之小。

峰度系数

- 而与随机变量的峰度类似，**样本峰度系数（sample kurtosis）**度量了数据分布的**厚尾特性（tailedness）**，其定义为：

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

- 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**超额峰度系数（excess kurtosis）**：

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

- 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（尖峰厚尾），然而情况并不总是如此。

峰度系数

- 此外需要注意的是，峰度系数计算的是经过标准化之后的样本的四次方的均值，从而在计算时已经剔除了离散程度。
- 如果某组数据的峰度大于3，意味着这组数据出现极端值（可能在左边、右边或者两边）的可能性比正态分布要大，但是这种现象并不是方差能够解释的。

峰度系数

峰度系数的计算

同样使用数据集rr000005.dta，使用带detai选项的su命令计算其峰度系数，约为8.89，这意味着虽然该股票的收益率近乎是没有偏向的，但是峰度系数大于3。这也意味着，如果使用正态分布建模该股票的收益率，那么就会低估尾部事件，即出现暴涨、暴跌的可能性。实际上金融中资产价格的峰度系数通常都是要大于3的，所以使用正态分布建模资产收益率是不可行的，这也是金融计量经济学研究的重要内容之一。

权重的使用

- 为了使得样本对总体的代表性更好，我们可能需要在做统计分析时对数据进行加权。
  - 比如，在很多的抽样调查中，为了得到比较好的统计性质，通常抽样并非等概率的进行，而是针对某一群体，比如低收入家庭等，进行有重点地抽样。
  - 此时，如果我们希望获得收入的总体均值，简单的计算样本平均会导致平均收入的低估。

权重的使用

- 为了解决这一问题，通常我们会使用**Horvitz-Thompson估计量**，也就是使用每个样本被抽中的概率π_i的倒数$1/\pi_i$作为权重，计算加权平均：

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

- 如果在上式中，令$x_i = 1$，为了使得$\bar{x}^w = 1$，需要：

$$M = \sum_{i=1}^{N}\frac{1}{\pi_i}$$

而M即为总体中个体的数量。

- 概率的倒数可以简单理解为一个样本代表了总体中多少个个体，
  - 比如，如果一个样本被抽中的概率为万分之一，那么一个样本大概代表了一万个个体，自然概率倒数之和即为总体中个体的数量。

逆概率加权

- 可以证明，如果使用以上的加权方法计算平均数，那么：

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

其中$\frac{1}{M}\sum_{i=1}^{M}x_i$即为（有限）总体的均值。换句话说，Horvitz-Thompson估计量是无偏估计量。

- 将Horvitz-Thompson估计量重新整理：

$$\bar{x}^{w}=\frac{1}{M}\sum_{i=1}^{N}\frac{1}{\pi_i}x_i=\frac{\sum_{i=1}^{N}\frac{1}{\pi_i}x_i}{\sum_{i=1}^{N}\frac{1}{\pi_i}}=\sum_{i=1}^{N}\left(\frac{\frac{1}{\pi_i}}{\sum_{i=1}^{N}\frac{1}{\pi_i}}\right)x_i\triangleq\sum_{i=1}^{N}w_i x_i$$

其中$\sum w_i=1$。以上估计量也称为**加权平均（weighted average）**，其中权重为概率的倒数，因而通常也被称为**逆概率加权（inverse probability weighting）**。

- 在实际的调查数据中，权重通常会使用“一个个体代表了多少个个体”这种形式给出。

数据中的加权

CHFS中的权重

在中国家庭金融调查（China Household Finance Survey）的数据（chfs_ind.dta）中，如果不使用权重，我们可以使用命令

```
1 su labor_inc
```

来计算未加权的平均劳动收入。

数据中的加权

CHFS中的权重

然而实际上，该调查在进行调查时进行了抽样设计，在数据集中，swgt变量指明了数据中1个人代表了总体中的多少人。在Stata中，有四种加权方式可供使用：

- fweight：频数权重，即如果我们观察到m个一模一样的观测，那么我们可以将这m个一模一样的观测合并为一个观测，并设其权重为m。fweight必须设定一个整数型变量作为权重，不接受小数。
- aweight：分析权重，适用于加总的数据，比如我们使用的数据为每个省份的平均值，那么可以用省份的人口作为权重。权重在使用时会默认规范化所有的权重之和为N：$\sum_{i=1}^{N} w_i = N$。
- pweight：抽样权重，适用于抽样数据，权重为每个个体被抽中的概率的倒数。
- iweight：重要性权重，Stata内部处理方法与aweight类似，区别在于使用iweight不会做规范化，适用于出于其他目的的加权。

数据中的加权

CHFS中的权重

因而我们可以使用如下命令计算加权平均:

```
1 su labor_inc [aw=swgt]
```

注意到经过加权后的均值为30185.8，大于未经加权的均值28426.33。

加总数据中的权重

- 加权平均的另一个应用是在加总的数据中。
- 比如如果我们有每个城市的平均收入，为了计算全国的平均收入，我们可以按照如下计算：

$$\bar{x} = \frac{\sum_{c=1}^{C}(\bar{x}_c \times p_c)}{\sum_{c=1}^{C} p_c} = \sum_{c=1}^{C}\left(\frac{p_c}{\sum_{c=1}^{C} p_c} \times \bar{x}_c\right) \triangleq \sum_{c=1}^{C}(w_c \times p_c)$$

  其中权重w_c为每个城市人口p_c占全国人口的比例。
- 实际上，以上计算方法也是一种逆概率加权：
  - 由于对于城市数据而言，每个城市只有1条数据，从而$1/p_c$代表了每个城市c中一个个体被抽中的概率，从而根据逆概率加权的思想，权重应该为$1/(1/p_c)=p_c$，即使用人口数量进行加权。

加总数据中的权重

全国人均公共图书册数

如果我们需要计算2010年全国人均公共图书册数，使用citydata.dta中的城市数据，我们分别计算了使用人口加权和不加权两种不同的均值：

```
use datasets/citydata.dta, clear
keep if year==2010
su v210
su v210 [aw=v4]
```

广义平均数的加权

- 此外，除了算数平均数，以上的广义平均数，包括调和平均数、几何平均数等，都可以使用加权的版本：

$$\left\{\frac{1}{N}\sum_{i=1}^{N} w_i x_i^p\right\}^{\frac{1}{p}}$$

其中$\sum w_i = 1$。

- 权重应该用什么？（习题）

分布图

- 对于定量数据，我们可以使用很多种图表展示其分布的形状，如：
  - 直方图
  - 箱线图
  - 经验分布函数
  - QQ图

经验分布函数

- 我们知道，累积分布函数可以描述一个随机变量是所有特征
- 最直观的方法即使用样本数据计算出累积分布函数，即**经验分布函数（empirical distribution function）**
- 回忆累积分布函数的定义为：$F(x) = P(X \leq x)$即小于x的概率
- 我们可以把概率换为样本中可以计算的比例，即对于某个x：

$$\hat{F}(x) = \frac{\#\{x_i \leq x\}}{N} = \frac{\sum_{i=1}^{N} 1\{x_i \leq x\}}{N}$$

其中#{}代表满足大括号中条件的样本数量，除以样本量就得到了小于x的比例。

经验分布函数

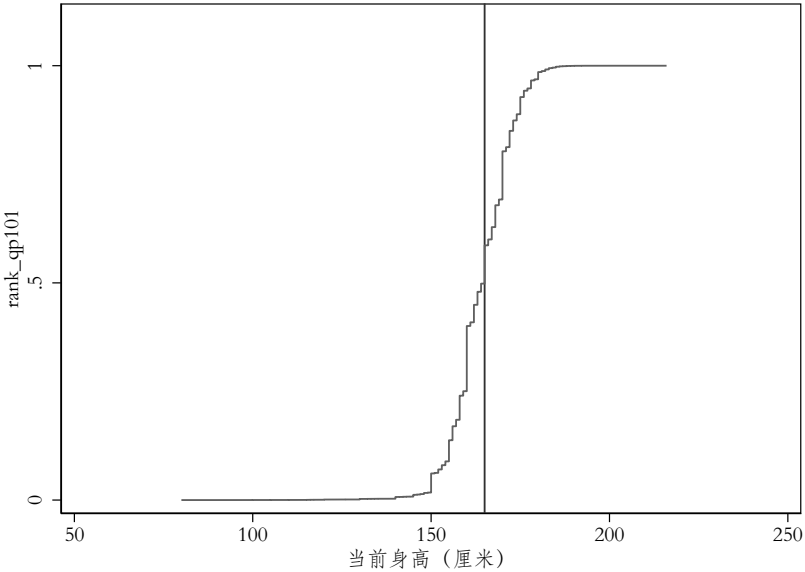
身高的经验分布函数

我们使用如下代码画出CFPS数据中身高的经验分布函数:

代码 2: 经验分布函数

```
// empirical_distribution.do
clear
use datasets/cfps_adult.dta
drop if qp101<0
// 排序
sort qp101
// 生成排序序号
// 注意中有缺失值，这里不删除缺失值，但是在计算时将其排除qp101
gen notmissing_qp101=qp101~=.
qui: su notmissing_qp101
gen rank_qp101=_n/r(sum)
// 画图
qui: su qp101, de
local median_qp101=r(p50)
```

经验分布函数



直方图

- 虽然经验分布函数能够表示分布的所有特征，然而很多时候并不直观
- 密度函数往往比分布函数更为直观，而为了观察密度函数一个更加常用的方法是使用**直方图（histogram）**。
- 在直方图中，可以首先把数据的取值范围划分为很多不同的组，比如最常用的方法是把最小值到最大值的区间划分为长度相同的R个分组，那么：

$$\frac{x_{(N)} - x_{(1)}}{R} \triangleq 2h$$

即为每个小组的长度，我们称之为组距，而每个组的中心称为组中值。

直方图中的分组

在CFPS数据的身高数据中，最小值为80，最大值为216，分为20组，那么每个组的组距就应该为$(216-80)/20=6.8$，分组为：

$$(80, 86.8], (86.8, 93.6], ..., (209.2, 216]$$

其中第一组的组中值为83.4，第二组的组中值为90.2，以此类推。

直方图

- 接下来，根据落入到每个组的频数，除以样本量N，就可以计算出每个组的频率，接下来就可以使用类似于条形图的形式，以每个组的起始作为宽度，以频率、频数等作为高速，制作条形图，就得到了直方图。
- 值得注意的是，直方图的高度有几种不同的选择：
  - 直接使用频数；
  - 使用频率，从而所有组别的频率相加为1；
  - 使用密度，即将频率再除以组距$2h$作为高度，从而线下面积为1。

直方图

身高的直方图

在Stata中，可以使用“histogram ”命令绘制直方图，比如对于以上身高数据：

```
use datasets/cfps_adult.dta, clear
hist qp101 if qp101>=0, bin(40) fraction
```

其中bin(40)选项表示直方图需要有40个条形；fraction选项指高度使用频率。

直方图

- 以上介绍的是每个组别的组距都相同的情况。
- 如果存在组别的组距不同，此时应该保证每个条形的面积之比应该等于每个组别的频数之比。
- 从这点来看：
  - 由于直方图的条形宽度有意义，所以**直方图是使用面积，而非长度表示概率**；
  - 而条形图由于条形的宽度没有意义，所以只有长度有意义，而面积是没有意义的。
- 此外：由于条形图针对分类变量，而直方图针对数值型变量，而数值型变量多数都是连续的，所以直方图的条形之间通常是紧挨真的，而不像条形图是分开的。

密度函数图

- 直方图已经很直观，但是因为其进行了分组，也损失了很多信息，那么我们能不能把密度函数直接画出来呢?
- 我们知道分布函数的导数等于密度函数，我们不妨从刚刚的经验分布函数出发。
- 回忆密度函数的定义

$$f(x)=\frac{dF(x)}{dx}=\lim_{h\to 0}\frac{F(x+h)-F(x-h)}{2h}$$

- 定义中要求$h\to 0$，但是对于样本，由于我们没有无穷多的样本，所以一般我们会选择一个正的h
- 我们把h称作**窗宽（bandwidth）**。

核密度函数

- 为了获得密度函数的估计，带入经验分布函数的定义，有

$$\hat{f}(x)=\frac{\hat{F}(x+h)-\hat{F}(x-h)}{2h}=\frac{\#\{x-h<x_i\leq x+h\}}{2Nh}$$

即落入区间$(x-h,x+h]$区间的样本量除以$2Nh$。

- 根据导数的定义，h应该越小越好，然而如果使用很小的h，会导致区间$(x-h,x+h]$中样本量非常少甚至没有
- 所以在实际操作中，应该存在一个最好的h
  - 如果有大量样本，应该允许更小的h
  - 如果样本量不大，h应该选的略微大一点。

核密度函数

- 现在，如果对于每一个$x \in \mathbb{R}$都把$\hat{f}(x)$计算出来，那么我们就有了密度函数的估计。
- 实际上，我们可以将以上的估计进行推广，注意到如果定义函数

$$K_0(x) = \frac{1\{|x| < 1\}}{2}$$

那么密度函数估计式可以写为

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} K_0\left(\frac{x-x_i}{h}\right)}{Nh}$$

核密度函数

密度函数需要满足线下面积为1，我们可以检查

$$
\begin{aligned}
\int_{\mathbb{R}} \hat{f}(x)\,dx &= \int_{\mathbb{R}} \frac{\sum_{i=1}^{N} K_0\left(\frac{x-x_i}{h}\right)}{Nh}dx \\
&= \frac{1}{Nh}\int_{\mathbb{R}} \sum_{i=1}^{N} K_0\left(\frac{x-x_i}{h}\right)dx \\
&= \frac{1}{Nh}\sum_{i=1}^{N}\int_{\mathbb{R}} K_0\left(\frac{x-x_i}{h}\right)dx \\
&= \frac{1}{Nh}\sum_{i=1}^{N}\int_{\mathbb{R}} K_0(z)\,d(zh+x_i) \\
&= \frac{1}{Nh}\sum_{i=1}^{N} h\int_{\mathbb{R}} K_0(z)\,dz \\
&= \frac{1}{N}\sum_{i=1}^{N}\int_{\mathbb{R}} K_0(z)\,dz
\end{aligned}
$$

因而只要$\int_{\mathbb{R}} K_0(z)\,dz = 1$，那么就有$\int_{\mathbb{R}} \hat{f}(x)\,dx = 1$

核密度函数

- 然而上述定义的密度函数仍然有其缺点，即由于$K_0(x)$有不连续、不可导的点，所以导致$\hat{f}(x)$也有不可导的点。
- 为了克服这一缺点，我们可以选取其他连续、光滑的函数，即“**核函数**”（**kernel function**）。我们要求核函数$K(x)$满足如下几个要求：
  1. $\int_{\mathbb{R}} K(x)\,dx = 1$；
  2. $K(x) \geq 0$；
  3. $K(x) = K(-x)$，即关于y轴对称。
- 检查以上要求，$K_0(x)$都满足，我们称其为矩形（rectangular）核函数

核密度估计

实际上任意对称的密度函数都满足如上要求，所以我们可以选择一些连续可导的密度函数作为核密度函数，比如：

- 高斯（Gaussian）核函数，即标准正态分布的密度函数：$K(x)=\phi(x)$
- 三角（triangular）核函数：$K(x)=1\{|x|<1\}\cdot(1-|x|)$
- Epanechnikov核函数：$K(x)=1\left\{|x|<\sqrt{5}\right\}\cdot\left[\frac{3}{4\sqrt{5}}\left(1-0.2x^2\right)\right]$
- Epan2核函数：$K(x)=1\{|x|<1\}\cdot\left[\frac{3}{4}\left(1-x^2\right)\right]$
- 余弦（cosine）核函数：$K(x)=1\{|x|<1/2\}\cdot[1+\cos(2\pi x)]$

实践中，核函数的选取对核密度估计的影响并不大，但是窗宽的选择影响会比较大。

核密度函数

身高的核密度函数

在Stata中，可以使用“kdensity”命令绘制该密度函数图，该命令默认使用了Epanechnikov核函数。比如对于以上身高数据：

```
use datasets/cfps_adult.dta, clear
kdensity qp101 if qp101>=0
```

即绘制了身高的核密度函数，该命令会自动选择一个最优的h，如果没有特殊需要，可以不手动制定h，如果需要指定h，可以使用“bwidth()”选项。

直方图与核密度函数

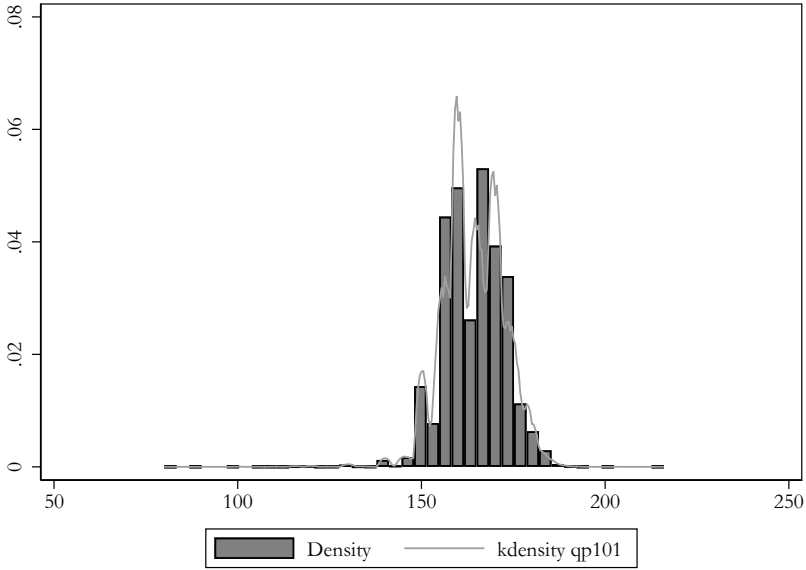
身高的直方图与核密度函数

我们可以使用如下代码绘制了CFPS中身高数据的直方图和核密度函数的估计图:

代码 3: 直方图和核密度函数

```
// histogram_kdensity.do
clear
use datasets/cfps_adult.dta
drop if qp101<0
twoway (hist qp101,bin(40) density) (kdensity qp101)
graph export hist_kdens.pdf, replace
```

直方图与核密度函数



箱线图

- 直方图或者核密度函数虽然也可以表示数据的所有信息，但是很多时候仍然不够直观
- 一个更加简洁的方法是绘制**箱线图（box chart）**。
- 在箱线图中，从小到大有五条线：最小值（L）、下25%分位数（Q_1）、中位数（M）、75%分位数（Q_3）、最大值（U）
- 将Q_1和Q_3两边连接起来作为「箱子」，再将最大值、最小值与箱子连接起来，就得到了箱线图。
- 其中，最小值L和最大值U有时也会如下设定：

$$U = Q_3 + 1.5\,(Q_3 - Q_1)$$
$$L = Q_1 - 1.5\,(Q_3 - Q_1)$$

如果按照此设定，超过U、低于L的样本都被视作异常值，单独标注在图中。

箱线图

身高的箱线图

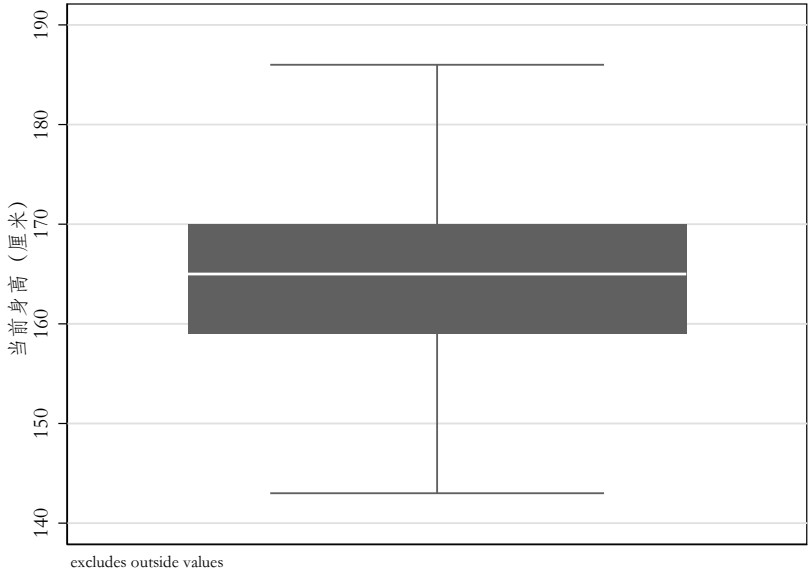
我们使用如下代码分别绘制了身高的箱线图以及分性别的身高的箱线图:

代码 4: 箱线图

```
// box_plot.do
clear
use datasets/cfps_adult.dta
drop if qp101<0
graph box qp101, noout
graph export box_plot.pdf, replace
graph hbox qp101, over(cfps_gender)
graph export box_plot_by_sex.pdf, replace
```

其中noout选项用于控制不显示异常值，over选项用于分组，box命令为竖直的箱线图，而hbox为水平的箱线图

箱线图



箱线图

- 箱线图可以比较直观的观察数据的平均水平、离散程度甚至偏度、峰度。比如：
  - 通过中位数的比较就可以比较出两组数据的位置度量；
  - 而根据箱子的长度（宽度）即四分位差，就可以比较两组数据的离散程度；
  - 而观察U,L与四分位数之间的距离以及异常值的情况可以看出数据的偏态，四分位数与U, L间距较远的为尾巴比较厚的一边；
  - 而观察两边尾巴的厚度可以比较峰度。
- 与经验分布函数、核密度函数、直方图相比，虽然箱线图只保留了分布的一部分信息，但是这些信息更加直观、简洁。

相关性

- 当存在成对的两组变量时，我们经常会关注两个变量之间的相关性。
- 在多元随机变量的介绍中，我们已经介绍而来总体的协方差、相关系数等概念
- 现在我们从样本的角度出发，介绍如何度量两组数据的相关性。

相关系数

- 对于两组数据$\{(x_i, y_i), i = 1, ..., N\}$，为了度量其相关性，一个简单的想法是计算总体相关系数：

$$\rho = \frac{\mathbb{C}(x, y)}{\sqrt{\mathbb{V}(x)}\sqrt{\mathbb{V}(y)}} = \frac{\mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y)}{\sqrt{\mathbb{E}\left[(x - \mathbb{E}x)^2\right]}\sqrt{\mathbb{E}\left[(y - \mathbb{E}y)^2\right]}}$$

- 将以上公式中的所有期望都换成平均：

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

- 以上统计量称为**样本相关系数（sample correlation coefficient）**，或者**Pearson相关系数（Pearson correlation coefficient）**，有时也称为简单相关系数。
- 注意在以上公式中，计算x和y的方差时，使用N做分母，而非$N-1$做分母，这是为了保证样本相关系数能够取到± 1。

简单相关系数

- 正如总体的相关系数一样，如果样本相关系数为正，则意味着x的增加伴随着y的增加，我们称其为正相关；反之，则称为负相关。
- 如果相关系数为0，则意味着两者没有**线性**相关性；
- 而如果样本相关系数为± 1，则意味着样本存在着完美的线性关系：

$$y_i = \alpha + \beta x_i$$

其中当$\beta > 0$时，样本相关系数为1，当$\beta < 0$时，样本相关系数为-1。

简单相关系数

身高和体重的相关系数

在Stata中，可以使用corr命令计算变量之间的相关系数，该命令会计算出所给出的所有变量的相关系数矩阵，比如代码：

```
drop if qp101<0
drop if qp102<0
corr qp101 qp102 cfps_gender
```

秩相关系数

- 样本相关系数作为总体相关系数的样本等价，只能度量两个变量之间的线性相关性。
- 为了解决这个问题，还可以定义很多其他的相关系数。其中，一个比较常用的是**Spearman秩相关系数（Spearman’s rank correlation coefficient）**，即数据排序的相关系数。
- 其计算方法如下：
  1. 分别对样本$\{x_i\}$和和$\{y_i\}$排序；
  2. 记录每个样本的序号，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，从而$r\left(x_{(n)}\right)=n$；
  3. 计算$r(x_i)$与$r(y_i)$之间的样本相关系数，即秩相关系数。

秩相关系数

- 由于该相关系数计算的是数据的排序序号的相关系数，所以该相关系数可以度量任意的单调的相关性，即：
  - 如果$y_i = f(x_i)$其中$f(\cdot)$为单调递增的函数，那么计算得到的x_i和y_i之间的秩相关系数就等于1
  - 如果$f(\cdot)$为单调递减的函数，那么计算得到的x_i和y_i之间的秩相关系数就等于-1。
- 也因为如此，Spearman秩相关系数具有单调不变性的有点，即如果$f(\cdot)$为单调递增的函数，那么x_i, y_i的秩相关系数与$f(x_i), y_i$、$x_i, f(y_i)$以及$f(x_i), f(y_i)$的秩相关系数都是相等的。

秩相关系数的计算

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：

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

而两列数据的排序序号分别为：$\{(1,1),(2,2),(3,3)\}$，因而其秩相关系数为：$r=\frac{(1+4+9)-3\times 2\times 2}{2\times 1\times 1}=1$即两组数据存在着完全正向的单调关系，然而并不存在完全的正向线性关系。

秩相关系数

Stata中秩相关系数计算

身高和体重之间的关系可能并不是线性的，为此可以继续计算秩相关系数，比如如下代码：

```
drop if qp101<0
drop if qp102<0
sort qp101
gen qp101_rank=_n
sort qp102
gen qp102_rank=_n
corr qp101_rank qp102_rank
```

即先排序，获得排序序号后计算排序序号的相关系数，得到两者的秩相关系数为0.684。

秩相关系数

Stata中秩相关系数计算

或者，可以直接使用spearman命令计算秩相关系数：

```
1 spearman qp101 qp102
```

注意以上两种方法计算结果不同，是由于样本中存在大量身高相同的人，如果使用第一种计算方法，这些身高相同的人的序号是不一样的（见习题中的解决方法），导致了第一种计算方法的误差。为了验证Spearman秩相关系数的单调不变性，我们对两个变量取对数，重新计算：

```
1 gen log_qp101=log(qp101)
2 gen log_qp102=log(qp102)
3 spearman qp101 qp102 log_qp101 log_qp102
```

可以看到Spearman秩相关系数对于单调变换是保持不变的。

相关性图

- 为了表达相关性，最常用的图表是**散点图（scatter diagram）**，即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。
- 此外，还有**线图（line plot）**，一般用于描述时间序列随时间的变化（横轴为时间），或者x轴与y轴之间的某个函数关系。

散点图与拟合图

身高和体重的关系

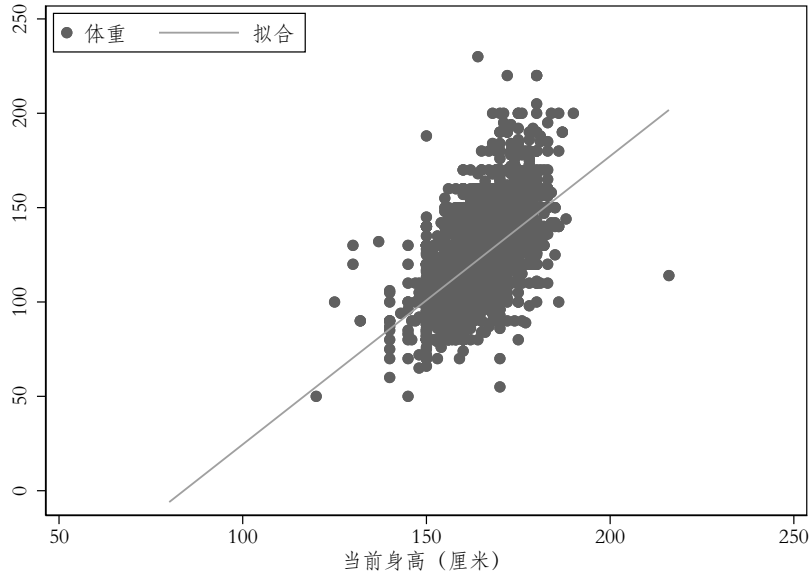
我们使用CFPS数据绘制身高和体重之间关系的散点图:

代码 5: 散点图：身高体重

```
// scatter_weight_height.do
clear
use datasets/cfps_adult.dta
drop if qp101<0
drop if qp102<0
twoway (scatter qp102 qp101 if mod(_n,10)==0) (lfit qp102 qp101),
    legend(pos(11) ring(0) label(1 "体重") label(2 "拟合"))
graph export scatter_weight_height.pdf, replace
```

- 为了避免数据量太大影响美观，我们只绘制了10%的样本;
- lfit命令可以绘制一个x轴和y轴变量的一个线性拟合关系的图，即一个线图，其原理主要使用的是一元线性回归

散点图与拟合图



散点图与拟合图

Zipf’s law

- 统计学家发现，在自然和社会中的很多数量都服从齐夫定律（Zipf’s law）
  - 如城市的大小、姓氏的分布、文本中单词的分布等等。
- 该定律可以通过一个对数-对数图观察到
  - 其中横轴为将样本按照从大到小的顺序排序，将其序号取对数；
  - 纵轴为频数的对数，
  - 两者之间应该大约是一个线性关系。
- 比如，对于城市的大小，我们可以先按照城市的人口从大到小排序，将序号取对数放在横轴；再将人口取对数放在纵轴。

散点图与拟合图

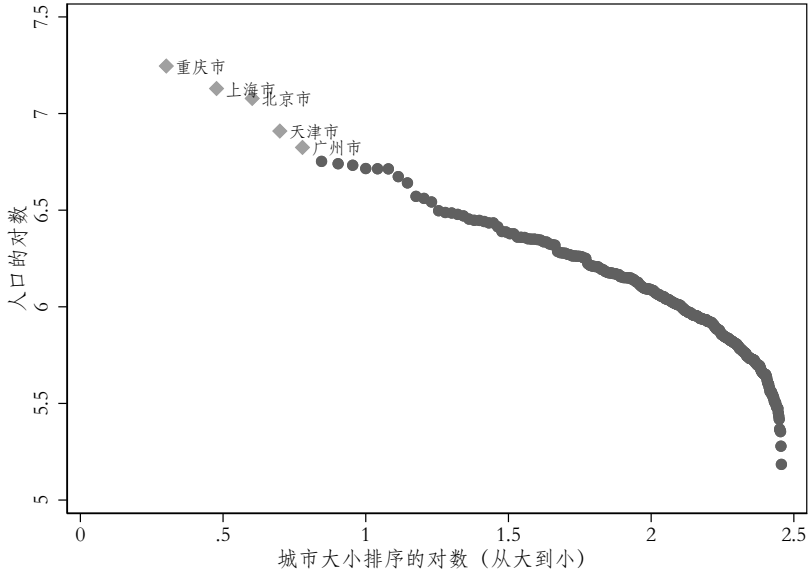
Zipf’s law

以下代码使用我国《城市统计年鉴》2011年的数据绘制了上述的散点图，其中人口数据使用了市辖区的人口:

代码 6: 齐夫定律

```
// zipf_law.do
clear
use datasets/citydata.dta
keep if Year==2011
sort v87
gen rank=_N-_n+1
gen log_rank=log10(rank)
gen log_pop=log10(v87*10000) // 数据的单位为万人
twoway (scatter log_pop log_rank if _n<_N-5) (scatter log_pop
    log_rank if _n>=_N-5, mlabel(City)), legend(off) xtitle("城市大
    小排序的对数（从大到小）") ytitle("人口的对数")
graph export zipf_law.pdf, replace
```

散点图与拟合图


重庆市
上海市
北京市
天津市
广州市
人口的对数
城市大小排序的对数（从大到小）
7.5
7
6.5
6
5.5
5
0
.5
1
1.5
2
2.5

气泡图

- 此外，散点的颜色、大小等也可以表示第三维数据，如**气泡图（bubble chart）**等，
- 这样就不仅仅可以表达两个变量之间的相关性，通过额外的颜色、散点大小等还可以表达更多维度的相关性。

气泡图

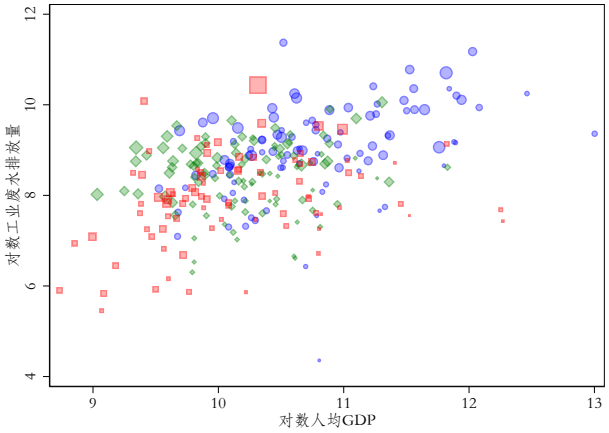
气泡图示例

使用citydata.dta中的数据，我们使用如下代码绘制了人均GDP和工业废水排放量之间关系的气泡图，其中气泡大小代表城市人口，而颜色则代表东、中、西部：

代码 7: 气泡图

```
// bubble_chart.do
clear
use datasets/citydata.dta
keep if Year==2011
// 省份代码、判断东中西部
gen prov=floor(CityCode/10000)
gen east=inlist(prov,11,12,13,21,31,32,33,35,37,44,46)
gen middle=inlist(prov,14,22,23,34,36,41,43,42)
gen west=inlist(prov,15,45,50,51,52,53,54,61,62,63,64,65)
// 产生变量
gen log_gdp_per_capita=log(v84)-log(v86)
gen log_pop=log(v86)
gen log_waste_water=log(v266)
```

气泡图



注：1. 数据来源：2011 年《中国城市统计年鉴》；
2. 图中散点大小表示城市人口，红色代表西部，绿色代表中部，蓝色代表东部。

Figure: 气泡图

统计图的规范

一般而言，统计图表的制作和报告需要满足一定的规范。比如：

- 在绘制统计图，特别是直方图、线图、条形图等图形时，为了便于比较，纵坐标轴都需要从0开始，否则错误的比例非常容易给人以误导。
- 此外，在文章中，一般需要对统计图编码标注，并写明图标题，如“图1：人口分布直方图”等
- 一般图标题应该在图的下方。
- 如果有必要，可以在图的下方标明注释，一般是对图的制作和内容的解释。

统计表的规范

而对于统计表格，同样有比较统一的规范

- 一张统计表应该包含表头、行标题、列标题、数值、附注等部分。
- 表头，即表的标题，在文章中一般要进行编号，与统计图不一样的是，表头一般在表格的上方。
- 统计表一般除了上下两条横线用粗线（或者双横线，有时也会使用单横线）之外，其他线一般用细线；统计表格一般两边不封口。
- 如有必要，可以在表的下方标明注释，一般包括对表的解释以及数据来源等。

统计表

Table: 描述性统计

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
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》

描述性统计表

- 在论文中，一般描述性统计表格的构成内容都是类似的，包括:
  - 变量名
  - 样本量
  - 均值
  - 标准差
  - 最小值
  - 最大值等。
- 当然，为了文章的内容服务，描述性统计表格可以根据需要增加或者删除某些变量，或者分组别进行描述性统计等
- 在Stata中，可以使用“outreg2”等命令导出该表格。

描述性统计表

Stata中的描述性统计表

为了获得上表的结果，可以使用数据集中的citydata.dta，并结合使用outreg2命令：

```
use datasets/citydata.dta
keep if Year==2011
outreg2 using sum_city.tex, sum(log) keep(v4 v92 v94 v96)
```

其中sum(log)代表要输出描述性统计表格，sum_city.tex为文件名，如果使用LaTeX就使用.tex作为后缀名，否则可以使用.doc作为扩展名以导出Word格式；keep()选项指我们要导出描述性统计的变量的名字。

练习题

- 6.1
- 6.3
- 6.4
- 6.8