第五章 统计量及其分布

Instructor: 郝壮

haozhuang@buaa.edu.cn
School of Economics and Management
Beihang University

March 1, 2022

概率论与数理统计

- 概率论：研究随机现象
  - 随机变量及其概率分布全面地描述了随机现象的统计规律性
  - 在概率论中, 通常假定概率分布是已知的, 一切计算及其推理均基于这个已知的分布进行
- 然而, 当我们研究并解决实际问题时, 情况往往并非如此.

例5.0.1 某公司要采购一批产品, 每件产品不是合格品就是不合格品, 不合格品率为未知参数 p.

由此, 若从该批产品中随机抽取一件, 用x表示这一批产品的不合格数, 不难看出x服从一个二点分布$b(1,p)$, 但分布中的参数p是不知道的.

例5.0.1 某公司要采购一批产品, 每件产品不是合格品就是不合格品, 不合格品率为未知参数 p.

由此, 若从该批产品中随机抽取一件, 用x表示这一批产品的不合格数, 不难看出x服从一个二点分布$b(1,p)$, 但分布中的参数p是不知道的.

一些问题:

- p的大小如何;
- p大概落在什么范围内;
- 能否认为p 满足设定要求 (如$p \leqslant 0.05$).

- 数理统计的研究内容：
  - 研究如何更合理, 更有效地抽取样本, 从而获得观测数据和资料的方法.
  - 如何利用一定的数据资料, 对所关心的问题, 得出尽可能精确且可靠的统计结论 **(即统计推断问题, 本学期重点)**.

- 数理统计的研究内容：
  - 研究如何更合理, 更有效地抽取样本, 从而获得观测数据和资料的方法.
  - 如何利用一定的数据资料, 对所关心的问题, 得出尽可能精确且可靠的统计结论 **(即统计推断问题, 本学期重点)**.
- **统计推断问题：**
  - **样本及抽样分布**
  - **参数估计：** 点估计, 区间估计, 估计的优良性质等
  - **假设检验：** 基本思想, 单(双)正态总体参数假设检验, 其他分布参数的假设检验, 拟合优度检验等
  - **方差分析**(多组总体位置参数的假设检验问题)

上述问题对应着本学期的以下学习内容：

- §1 第五章 统计量及其分布 (2周)
- §2 第六章 参数估计 (3周)
- §3 第七章 假设检验 (2周)
- §4 第八章 方差分析 (1周)

第五章　统计量及其分布

- §5.1 总体与样本
- §5.2 样本数据的整理与显示
- §5.3 统计量及其分布
- §5.4 三大抽样分布
- §5.5 充分统计量 (略)

§5.1 总体与样本 (population and sample)

案例： 若规定灯泡寿命低于1000小时者为次品, 如何确定次品率? 由于灯泡寿命试验是破坏性试验, 不可能把整批灯泡逐一检测, 只能抽取一部分灯泡作为样本进行检验, 以样本的信息来推断总体的信息.

定义： 在统计学中把研究对象的全体称为**总体**(population), 构成总体的每个成员称为**个体**(individual).

总体的三层含义：

- 总体是研究对象的全体 (有限总体, 无限总体)
- 总体是数据
- 总体是分布(本质是一个分布, 其数量指标为服从该分布的随机变量)

总体的含义举例

例: 考察某厂的产品质量, 以0记合格品, 以1记不合格品,

则**总体** $=$ {该厂生产的全部合格品与不合格品} $=$
{由0或1组成的一堆数}

若以p 表示这堆数中1的比例(不合格品率), 则该总体可由一个二点分布表示:

X	0	1
P	$1-p$	p

比如: 两个生产同类产品的工厂的产品的总体分布:

X	0	1
p	0.983	0.017

X	0	1
p	0.915	0.085

§5.1 总体与样本(population and sample)

总体参数： 描述总体的特征, 要调查的指标.

例： 一卷磁带上的伤痕数X服从泊松分布$P(\lambda)$, 但分布的参数λ却未知. 显然, λ的大小决定了一批产品的质量.

本例中： 总体分布的类型明确, 但总体含有未知参数λ. 统计的任务: 确定λ, 即确定最终的总体分布.

重要的总体参数如：

- 总体均值 μ
- 总体方差 σ^2
- 总体标准差 σ

5.1.2 样本

定义： 为了了解总体的分布, 从总体中随机地抽取n个个体, 记其指标值为$x_1, x_2, \ldots, x_n$, 则$x_1, x_2, \ldots, x_n$称为总体的一个**样本**(sample), n称为**样本容量**(sample size), 样本中的个体称为**样品**(sample).

思考： 为什么要抽样?

5.1.2 样本

定义： 为了了解总体的分布, 从总体中随机地抽取n个个体, 记其指标值为$x_1, x_2, \ldots, x_n$, 则$x_1, x_2, \ldots, x_n$称为总体的**一个样本**(sample), n称为**样本容量**(sample size), 样本中的个体称为**样品**(sample).

思考： 为什么要抽样?

普查(Census)的代价：

- 费用昂贵
- 时间过长
- 毁坏性实验
- 无法普查：观测值是无穷个

5.1.2 样本

样本具有两重性:

- 一方面, 由于样本是从总体中随机抽取的, 抽取前(或抽取后观测前)无法预知它们的数值, 因此, 样本是**随机变量**, 可以用大写字母 $X_1, X_2, \ldots, X_n$表示;
- 另一方面, 样本在抽取以后经观测就有确定的观测值, 因此, 样本又是**一组数值**, 也可以用小写字母 $x_1, x_2, \ldots, x_n$ 表示.

简单起见, 无论是样本还是其观测值, 本书中样本均用 $x_1, x_2, \ldots, x_n$ 表示, 学习中应从上下文中加以区别其代表的是随机变量还是观测值.

例5.1.3

例5.1.3 啤酒厂生产的瓶装啤酒规定净含量为640克. 由于随机性, 事实上不可能使得所有的啤酒净含量均为640克. 现从某厂生产的啤酒中随机抽取10瓶测定其净含量, 得到如下结果:

$$641, 635, 640, 637, 642, 638, 645, 643, 639, 640$$

- 这是一个容量为10的样本的观测值(observation), 对应的总体为该厂生产的瓶装啤酒的净含量.
- 这样的样本称为**完全样本**(complete sample, 相对于**分组样本**, categorical sample).

例5.1.4 考察某厂生产的某种电子元件的寿命, 选了100只进行寿命试验, 得到如下数据:

寿命范围	元件数	寿命范围	元件数	寿命范围	元件数
(0 24]	4	(192 216]	6	(384 408]	4
(24 48]	8	(216 240]	3	(408 432]	4
(48 72]	6	(240 264]	3	(432 456]	1
(72 96]	5	(264 288]	5	(456 480]	2
(96 120]	3	(288 312]	5	(480 504]	2
(120 144]	4	(312 336]	3	(504 528]	3
(144 168]	5	(336 360]	5	(528 552]	1
(168 192]	4	(360 384]	1	>552	13

- 表中的样本观测值没有具体的数值, 只有一个范围, 这样的样本称为**分组样本**.

样本的要求：简单随机样本

要使得推断可靠, 对样本就有要求, 使样本能很好地代表总体. 通常有如下两个**简单随机抽样**要求(Simple random sampling)：

- **独立性:** 样本中每一样品的取值不影响其它样品的取值: $x_1, x_2, \ldots, x_n$ 相互独立.
- **随机性:** 总体中每一个个体都有同等机会被选入样本: x_i 与总体X有相同的分布.

用以上简单随机抽样方法得到的样本称为**简单随机样本**, 也简称**样本**.

- 于是, 样本 $x_1, x_2, \ldots, x_n$ 可以看成是独立同分布(independently identically distributed, iid) 的随机变量, 其共同分布即为总体分布.

设总体X具有分布函数$F(x)$, $x_1, x_2, \ldots, x_n$ 为取自该总体的容量为n的样本, 则样本联合分布函数为

$$F\left(x_1, \ldots, x_n\right) = \prod_{i=1}^{n} F\left(x_i\right)$$

样本的联合密度函数:

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

(以后会学到, 以上两式是MLE的理论基础.)

注意

- 总体分为有限总体与无限总体.
- 实际中总体中的个体数大多是有限的. 当个体数充分大时, 将有限总体看作无限总体是一种合理的抽象.
- 对无限总体, 独立性容易实现, 思考随机性被违反的情况? 会导致哪些问题?
- 对有限总体, 只要总体所含个体数很大, 特别是与样本量相比很大, 则独立性也可基本得到满足(下面用例子说明).

例5.1.5

例5.1.5 设有一批产品共N个, 需要进行抽样检验以了解其不合格品率p. 现从中采取不放回抽样抽出2个产品, 这时, 第二次抽到不合格品的概率依赖于第一次抽到的是否是不合格品.

如果第一次抽到不合格品,
则$P(x_2 = 1|x_1 = 1) = (Np - 1)/(N - 1)$.

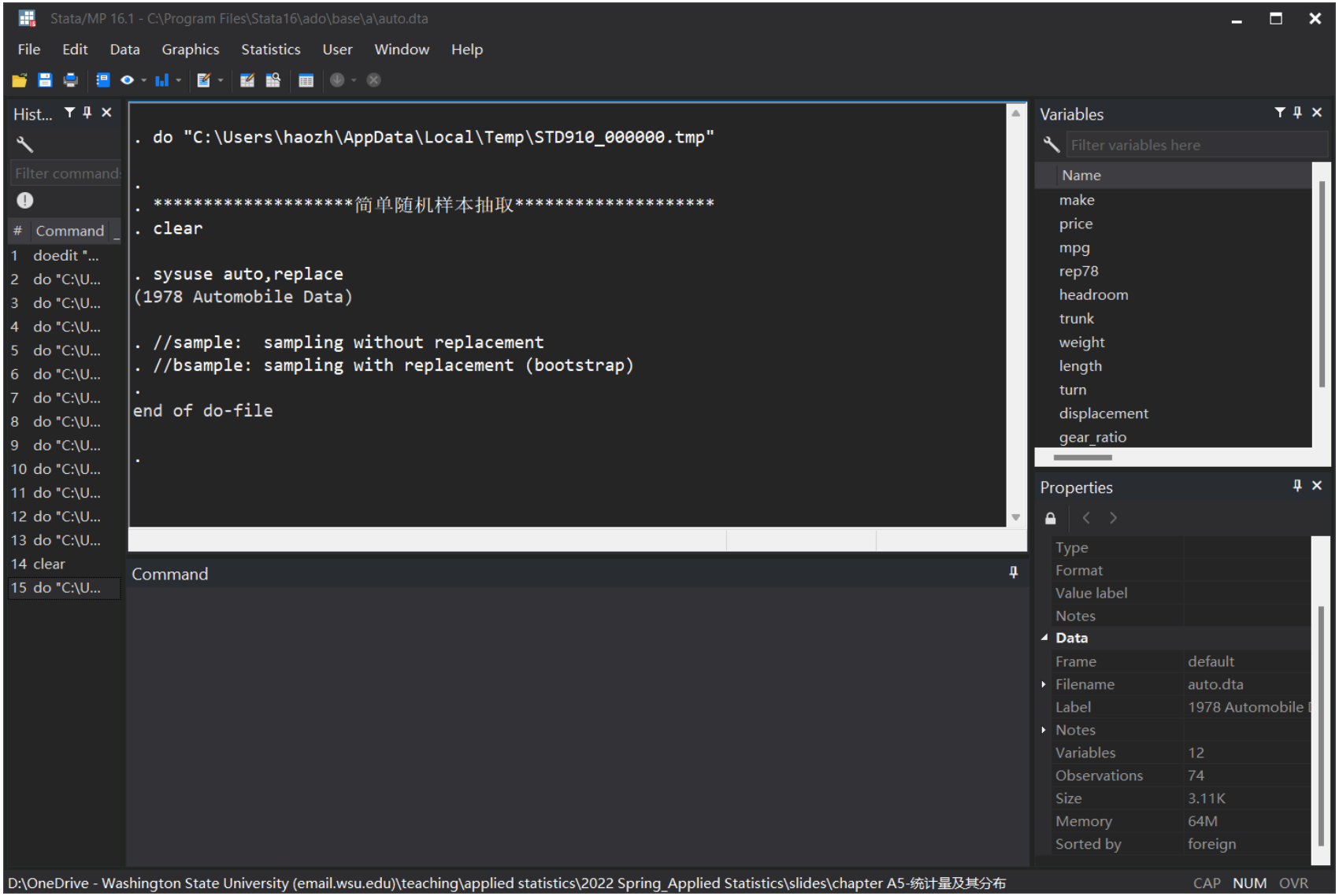
而若第一次抽到的是合格品, 则第二次抽到不合格品的概率为$P(x_2 = 1|x_1 = 0) = (Np)/(N - 1)$.

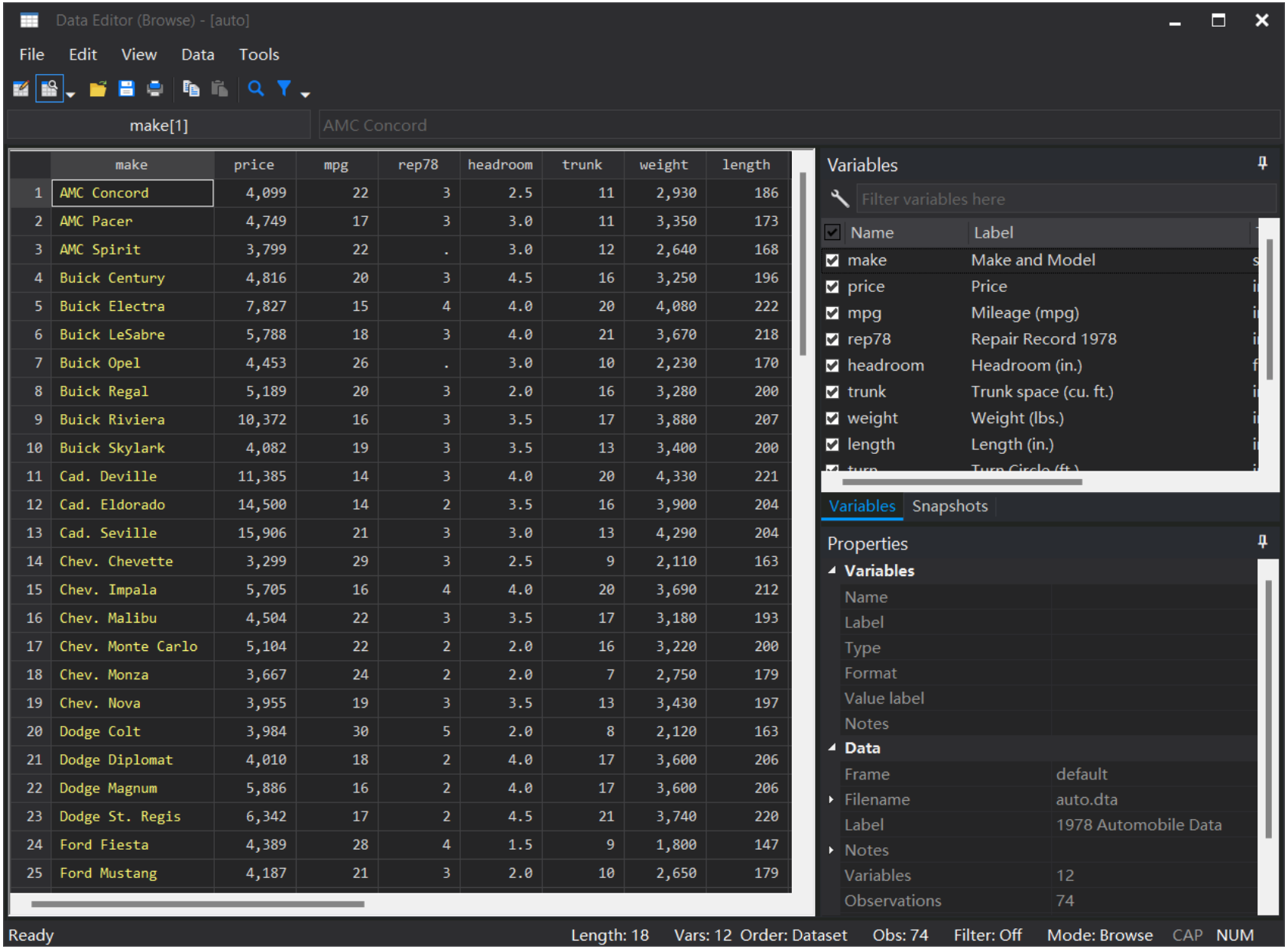
显然, 如此得到的样本不是简单随机样本. 但是, 当N很大时, 我们可以看到上述两种情形的概率都近似等于p. 所以当N 很大, 而n不大(一个经验法则是 $n/N \leqslant 0.1$)时可以把该样本近似地看成简单随机样本.

简单随机样本: STATA实现

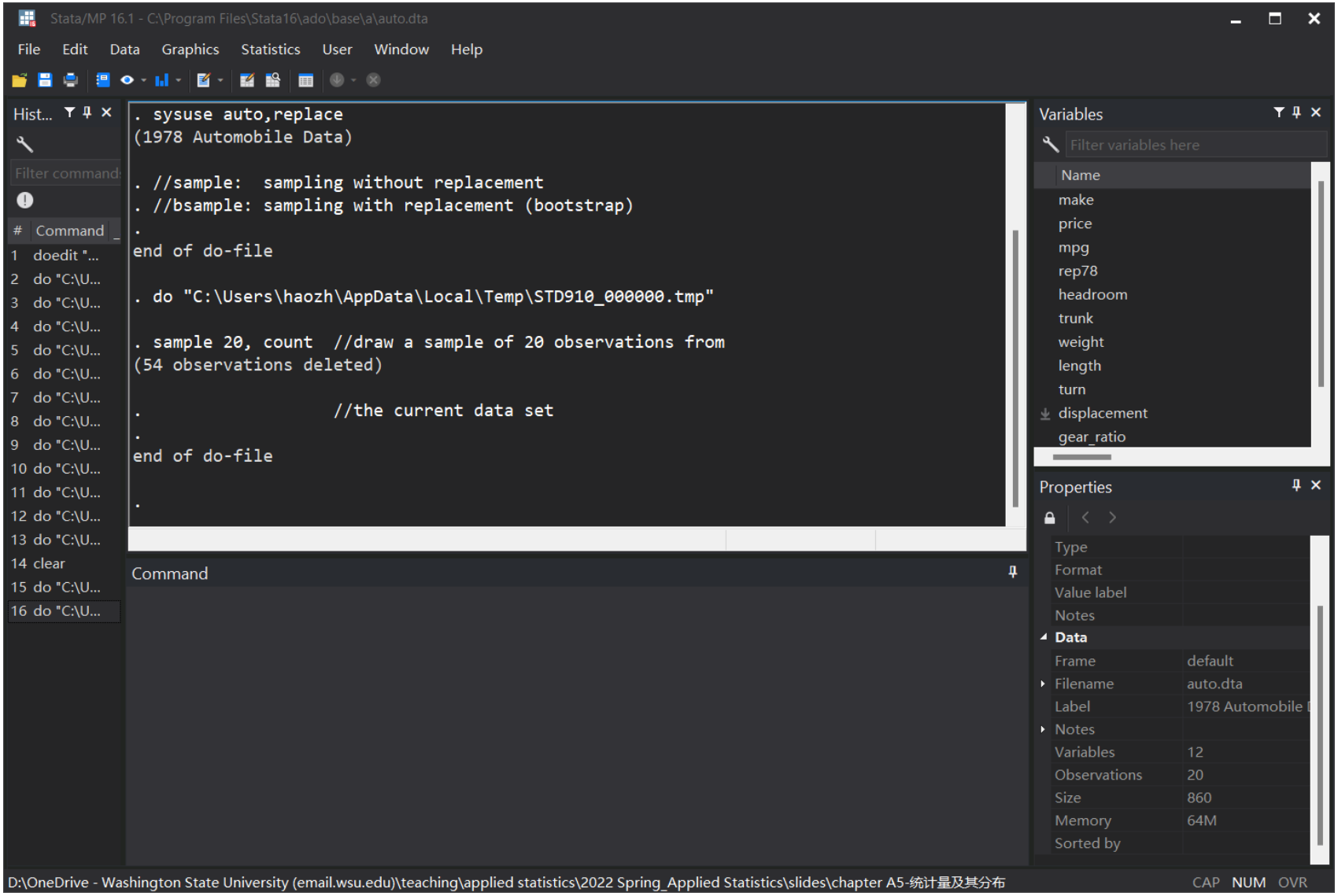
```
sysuse auto,replace
//sample:  sampling without replacement
//bsample: sampling with replacement (bootstrap)

sample 20, count  //draw a sample of 20 observation
                  //the current data set
sample 20         //draw a sample of 20%  observati
```

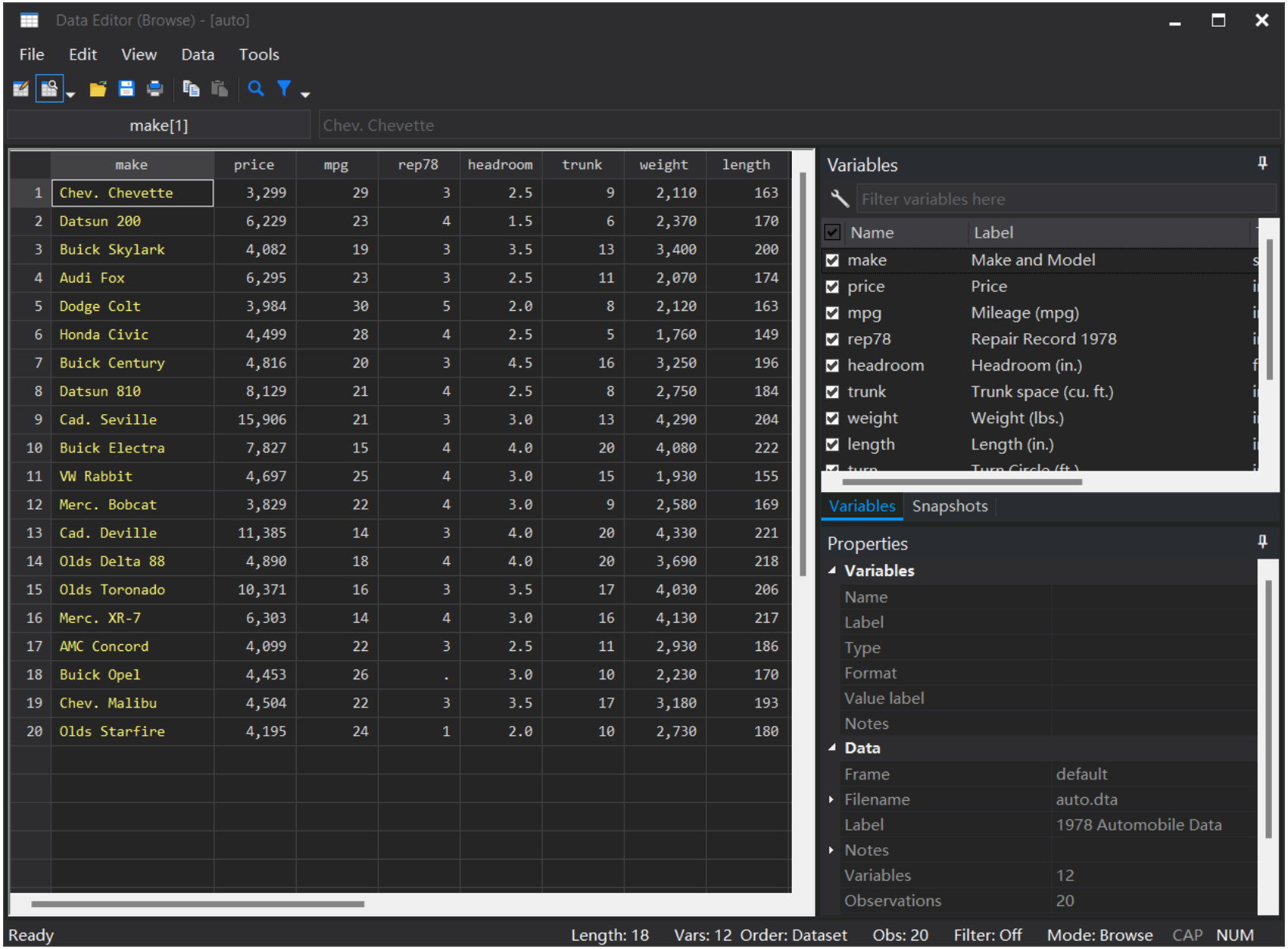
注意：该STATA实现仅为示意. 实际上, 利用的数据auto.dta已经是一个从总体中抽样的样本了.

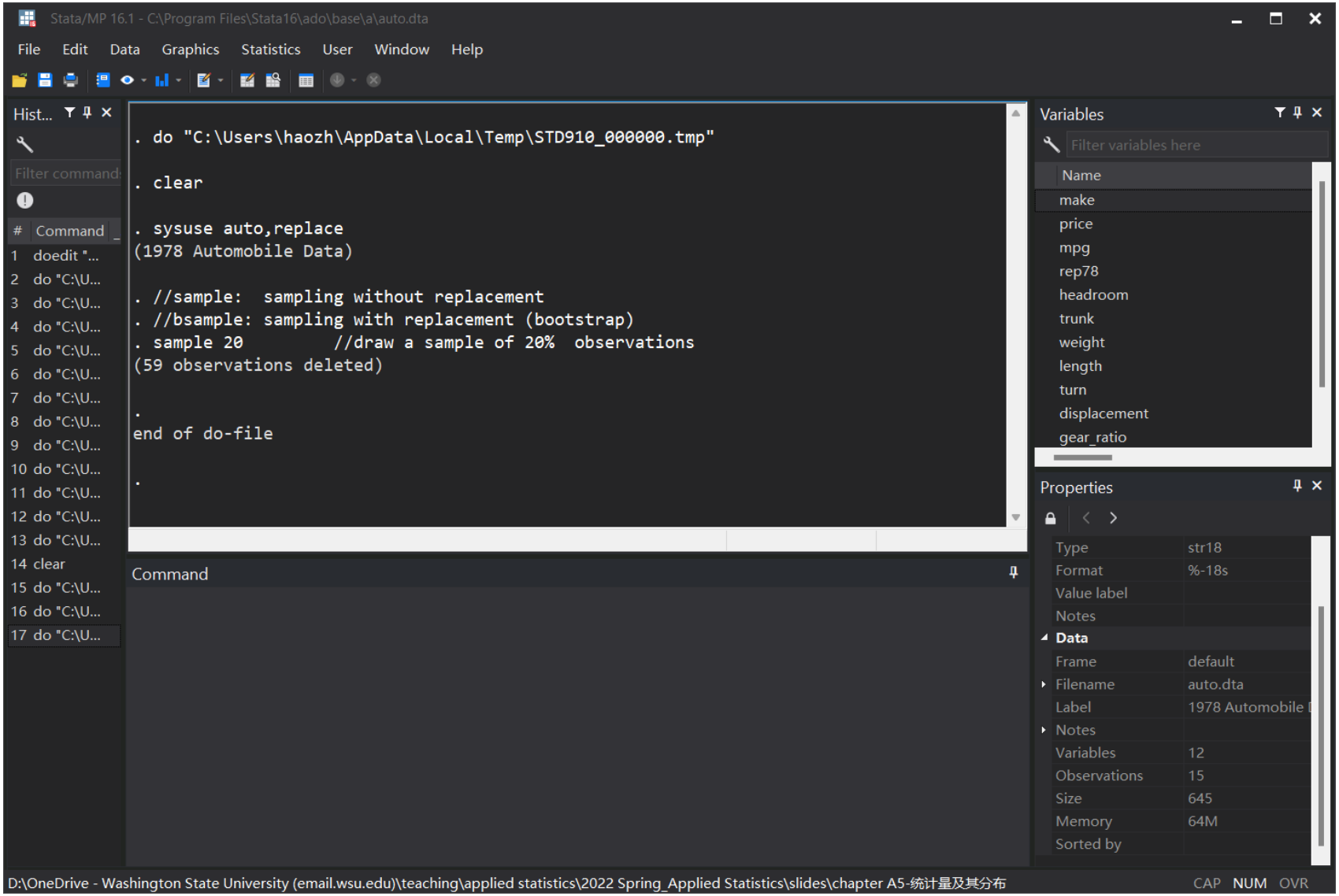

Stata/MP 16.1 - C:\Program Files\Stata16\ado\base\a\auto.dta
File Edit Data Graphics Statistics User Window Help
Hist...
Filter command
Command
1 doedit "...
2 do "C:\U...
3 do "C:\U...
4 do "C:\U...
5 do "C:\U...
6 do "C:\U...
7 do "C:\U...
8 do "C:\U...
9 do "C:\U...
10 do "C:\U...
11 do "C:\U...
12 do "C:\U...
13 do "C:\U...
14 clear
15 do "C:\U...
. do "C:\Users\haozh\AppData\Local\Temp\STD910_000000.tmp"
.
. **********简单随机样本抽取**********
. clear
. sysuse auto,replace
(1978 Automobile Data)
. //sample: sampling without replacement
. //bsample: sampling with replacement (bootstrap)
.
end of do-file
.
Command
Variables
Filter variables here
Name
make
price
mpg
rep78
headroom
trunk
weight
length
turn
displacement
gear_ratio
Properties
Type
Format
Value label
Notes
Data
Frame default
Filename auto.dta
Label 1978 Automobile
Notes
Variables 12
Observations 74
Size 3.11K
Memory 64M
Sorted by foreign
D:\OneDrive - Washington State University (email.wsu.edu)\teaching\applied statistics\2022 Spring_Applied Statistics\slides\chapter A5-统计量及其分布
CAP NUM OVR

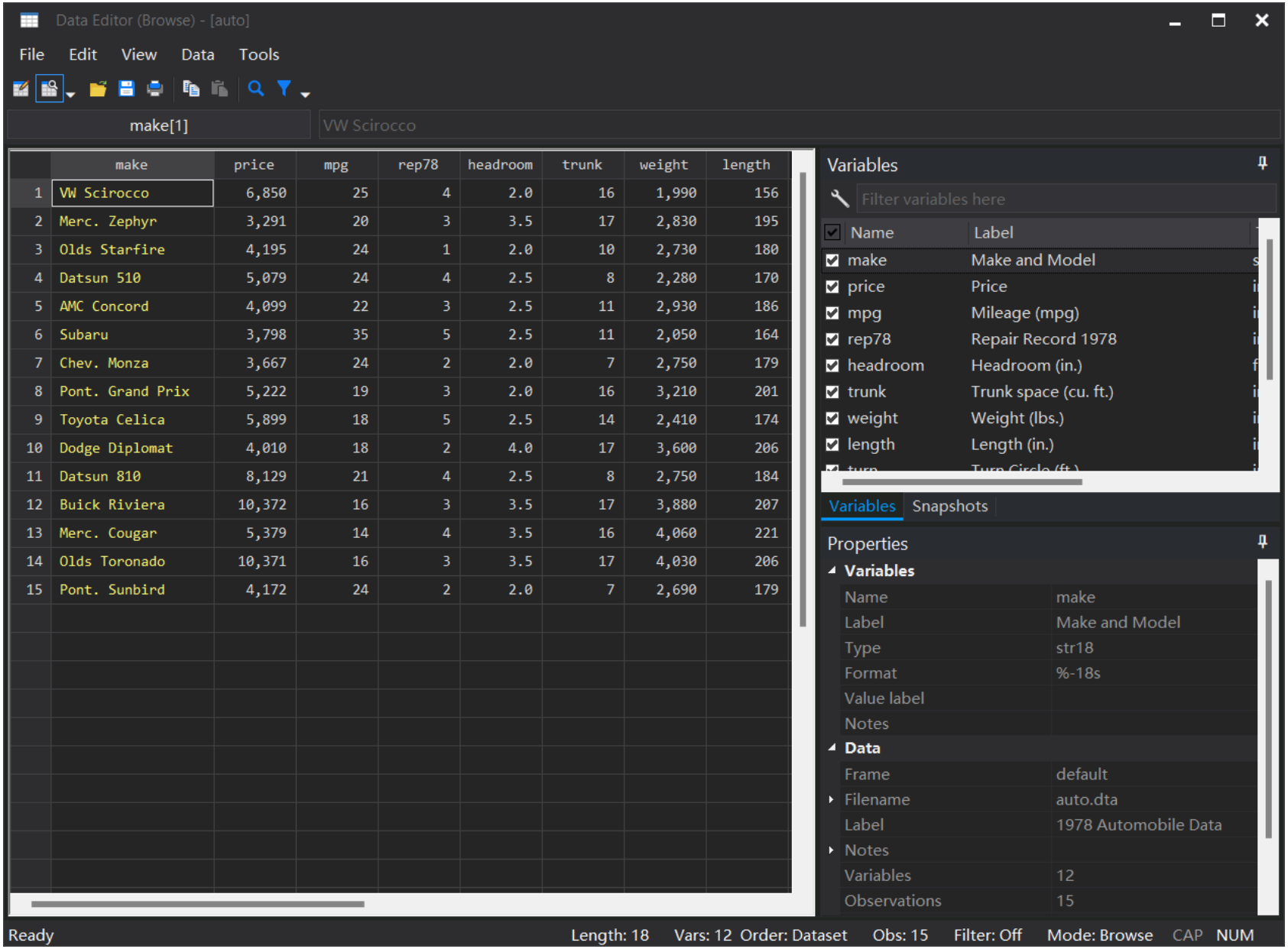


	make	price	mpg	rep78	headroom	trunk	weight	length
1	AMC Concord	4,099	22	3	2.5	11	2,930	186
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168
4	Buick Century	4,816	20	3	4.5	16	3,250	196
5	Buick Electra	7,827	15	4	4.0	20	4,080	222
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218
7	Buick Opel	4,453	26	.	3.0	10	2,230	170
8	Buick Regal	5,189	20	3	2.0	16	3,280	200
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147
25	Ford Mustang	4,187	21	3	2.0	10	2,650	179


Stata/MP 16.1 - C:\Program Files\Stata16\ado\base\a\auto.dta
File Edit Data Graphics Statistics User Window Help
Hist...
Filter command
Command
1 doedit "...
2 do "C:\U...
3 do "C:\U...
4 do "C:\U...
5 do "C:\U...
6 do "C:\U...
7 do "C:\U...
8 do "C:\U...
9 do "C:\U...
10 do "C:\U...
11 do "C:\U...
12 do "C:\U...
13 do "C:\U...
14 clear
15 do "C:\U...
16 do "C:\U...
. sysuse auto,replace
(1978 Automobile Data)
. //sample: sampling without replacement
. //bsample: sampling with replacement (bootstrap)
.
end of do-file
. do "C:\Users\haozh\AppData\Local\Temp\STD910_000000.tmp"
. sample 20, count //draw a sample of 20 observations from
(54 observations deleted)
. //the current data set
.
end of do-file
.
Command
Variables
Filter variables here
Name
make
price
mpg
rep78
headroom
trunk
weight
length
turn
displacement
gear_ratio
Properties
Type
Format
Value label
Notes
Data
Frame default
Filename auto.dta
Label 1978 Automobile
Notes
Variables 12
Observations 20
Size 860
Memory 64M
Sorted by
D:\OneDrive - Washington State University (email.wsu.edu)\teaching\applied statistics\2022 Spring_Applied Statistics\slides\chapter A5-统计量及其分布
CAP NUM OVR


Data Editor (Browse) - [auto]
File Edit View Data Tools
make[1]
Chev. Chevette
make price mpg rep78 headroom trunk weight length
1 Chev. Chevette 3,299 29 3 2.5 9 2,110 163
2 Datsun 200 6,229 23 4 1.5 6 2,370 170
3 Buick Skylark 4,082 19 3 3.5 13 3,400 200
4 Audi Fox 6,295 23 3 2.5 11 2,070 174
5 Dodge Colt 3,984 30 5 2.0 8 2,120 163
6 Honda Civic 4,499 28 4 2.5 5 1,760 149
7 Buick Century 4,816 20 3 4.5 16 3,250 196
8 Datsun 810 8,129 21 4 2.5 8 2,750 184
9 Cad. Seville 15,906 21 3 3.0 13 4,290 204
10 Buick Electra 7,827 15 4 4.0 20 4,080 222
11 VW Rabbit 4,697 25 4 3.0 15 1,930 155
12 Merc. Bobcat 3,829 22 4 3.0 9 2,580 169
13 Cad. Deville 11,385 14 3 4.0 20 4,330 221
14 Olds Delta 88 4,890 18 4 4.0 20 3,690 218
15 Olds Toronado 10,371 16 3 3.5 17 4,030 206
16 Merc. XR-7 6,303 14 4 3.0 16 4,130 217
17 AMC Concord 4,099 22 3 2.5 11 2,930 186
18 Buick Opel 4,453 26 . 3.0 10 2,230 170
19 Chev. Malibu 4,504 22 3 3.5 17 3,180 193
20 Olds Starfire 4,195 24 1 2.0 10 2,730 180
Variables
Filter variables here
Name Label
make Make and Model
price Price
mpg Mileage (mpg)
rep78 Repair Record 1978
headroom Headroom (in.)
trunk Trunk space (cu. ft.)
weight Weight (lbs.)
length Length (in.)
Variables Snapshots
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Frame default
Filename auto.dta
Label 1978 Automobile Data
Notes
Variables 12
Observations 20
Ready
Length: 18 Vars: 12 Order: Dataset Obs: 20 Filter: Off Mode: Browse CAP NUM


Stata/MP 16.1 - C:\Program Files\Stata16\ado\base\a\auto.dta
File Edit Data Graphics Statistics User Window Help
Hist...
Filter command
Command
1 doedit "...
2 do "C:\U...
3 do "C:\U...
4 do "C:\U...
5 do "C:\U...
6 do "C:\U...
7 do "C:\U...
8 do "C:\U...
9 do "C:\U...
10 do "C:\U...
11 do "C:\U...
12 do "C:\U...
13 do "C:\U...
14 clear
15 do "C:\U...
16 do "C:\U...
17 do "C:\U...
. do "C:\Users\haozh\AppData\Local\Temp\STD910_000000.tmp"
. clear
. sysuse auto,replace
(1978 Automobile Data)
. //sample: sampling without replacement
. //bsample: sampling with replacement (bootstrap)
. sample 20 //draw a sample of 20% observations
(59 observations deleted)
.
end of do-file
.
Command
Variables
Filter variables here
Name
make
price
mpg
rep78
headroom
trunk
weight
length
turn
displacement
gear_ratio
Properties
Type str18
Format %-18s
Value label
Notes
Data
Frame default
Filename auto.dta
Label 1978 Automobile
Notes
Variables 12
Observations 15
Size 645
Memory 64M
Sorted by
D:\OneDrive - Washington State University (email.wsu.edu)\teaching\applied statistics\2022 Spring_Applied Statistics\slides\chapter A5-统计量及其分布
CAP NUM OVR


Data Editor (Browse) - [auto]
File Edit View Data Tools
make[1]
VW Scirocco
make price mpg rep78 headroom trunk weight length
1 VW Scirocco 6,850 25 4 2.0 16 1,990 156
2 Merc. Zephyr 3,291 20 3 3.5 17 2,830 195
3 Olds Starfire 4,195 24 1 2.0 10 2,730 180
4 Datsun 510 5,079 24 4 2.5 8 2,280 170
5 AMC Concord 4,099 22 3 2.5 11 2,930 186
6 Subaru 3,798 35 5 2.5 11 2,050 164
7 Chev. Monza 3,667 24 2 2.0 7 2,750 179
8 Pont. Grand Prix 5,222 19 3 2.0 16 3,210 201
9 Toyota Celica 5,899 18 5 2.5 14 2,410 174
10 Dodge Diplomat 4,010 18 2 4.0 17 3,600 206
11 Datsun 810 8,129 21 4 2.5 8 2,750 184
12 Buick Riviera 10,372 16 3 3.5 17 3,880 207
13 Merc. Cougar 5,379 14 4 3.5 16 4,060 221
14 Olds Toronado 10,371 16 3 3.5 17 4,030 206
15 Pont. Sunbird 4,172 24 2 2.0 7 2,690 179
Variables
Filter variables here
Name Label
make Make and Model
price Price
mpg Mileage (mpg)
rep78 Repair Record 1978
headroom Headroom (in.)
trunk Trunk space (cu. ft.)
weight Weight (lbs.)
length Length (in.)
Variables Snapshots
Properties
Variables
Name make
Label Make and Model
Type str18
Format %-18s
Value label
Notes
Data
Frame default
Filename auto.dta
Label 1978 Automobile Data
Notes
Variables 12
Observations 15
Ready
Length: 18 Vars: 12 Order: Dataset Obs: 15 Filter: Off Mode: Browse CAP NUM

5.2.1 经验分布函数

设 $x_1, x_2, \ldots, x_n$ 是取自总体分布函数为$F(x)$的样本, 若将样本观测值由小到大进行排列, 为$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, 则称 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 为**有序样本**(ordered sample).

用有序样本定义如下函数

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ k/n, & x_{(k)} \le x < x_{(k+1)}, \\ 1, & x_{(n)} \le x \end{cases} \qquad k = 1, 2, \ldots, n-1$$

则$F_n(x)$是一非减右连续函数, 且满足$F_n(-\infty) = 0$ 和 $F_n(+\infty) = 1$.

由此可见, $F_n(x)$是一个分布函数, 并称$F_n(x)$**经验分布函数**.

例5.2.1 某食品厂生产听装饮料, 现从生产线上随机抽取5听饮料, 称得其净重(单位:克)

$$351, 347, 355, 344, 351.$$

这是一个容量为5的样本, 经排序可得有序样本:

- $x_{(1)} = 344, x_{(2)} = 347, x_{(3)} = 351, x_{(4)} = 351, x_{(5)} = 355$

其经验分布函数为

$$F_n(x) = \begin{cases} 0, & x < 344 \\ 0.2, & 344 \le x < 347 \\ 0.4, & 347 \le x < 351 \\ 0.8, & 351 \le x < 355 \\ 1, & x \ge 355 \end{cases}$$

一般化地, 对任意给定的实数x, $F_n(x)$是样本中事件$\{x_i \leqslant x\}$(观测样本中的任意一个样本点x_i, 它的观测值小于x)的**频率**.

当n固定时, $F_n(x)$是样本的函数, 是一个**随机变量**. 若对任意给定的实数x, 定义指示函数(indicator funciton)

$$I_i(x) = \begin{cases} 1, & x_i \leqslant x \\ 0, & x_i > x \end{cases}$$

则由经验分布定义, 对任意给定的实数x,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_i(x)$$

则由经验分布定义, 对任意给定的实数x,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_i(x)$$

注意到$I_i(x)$独立同分布$b(1, F(x))$, (为什么? 因为x_i服从$F(x)$, $x_i \leqslant x$的概率是$F(x)$, 由$I_i(x)$的定义即可得$I_i(x) \sim b(1, F(x))$, 所以$E(I_i(x)) = F(x)$

则由伯努里大数定律: **只要 n 相当大, $F_n(x)$依概率收敛于$F(x)$.**

格里文科定理

由以上分析, 进一步不加证明的给出格里纹科定理.

定理5.2.1 (格里文科定理) 设$x_1, x_2, \ldots, x_n$是取自总体分布函数为$F(x)$的样本, $F_n(x)$ 是其经验分布函数, 当$n \to \infty$时, 有对任意x,

$$P\left\{\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \to 0\right\} = 1$$

(sup：上确界), 即$F_n(x)$对于所有x点点收敛到$F(x)$.

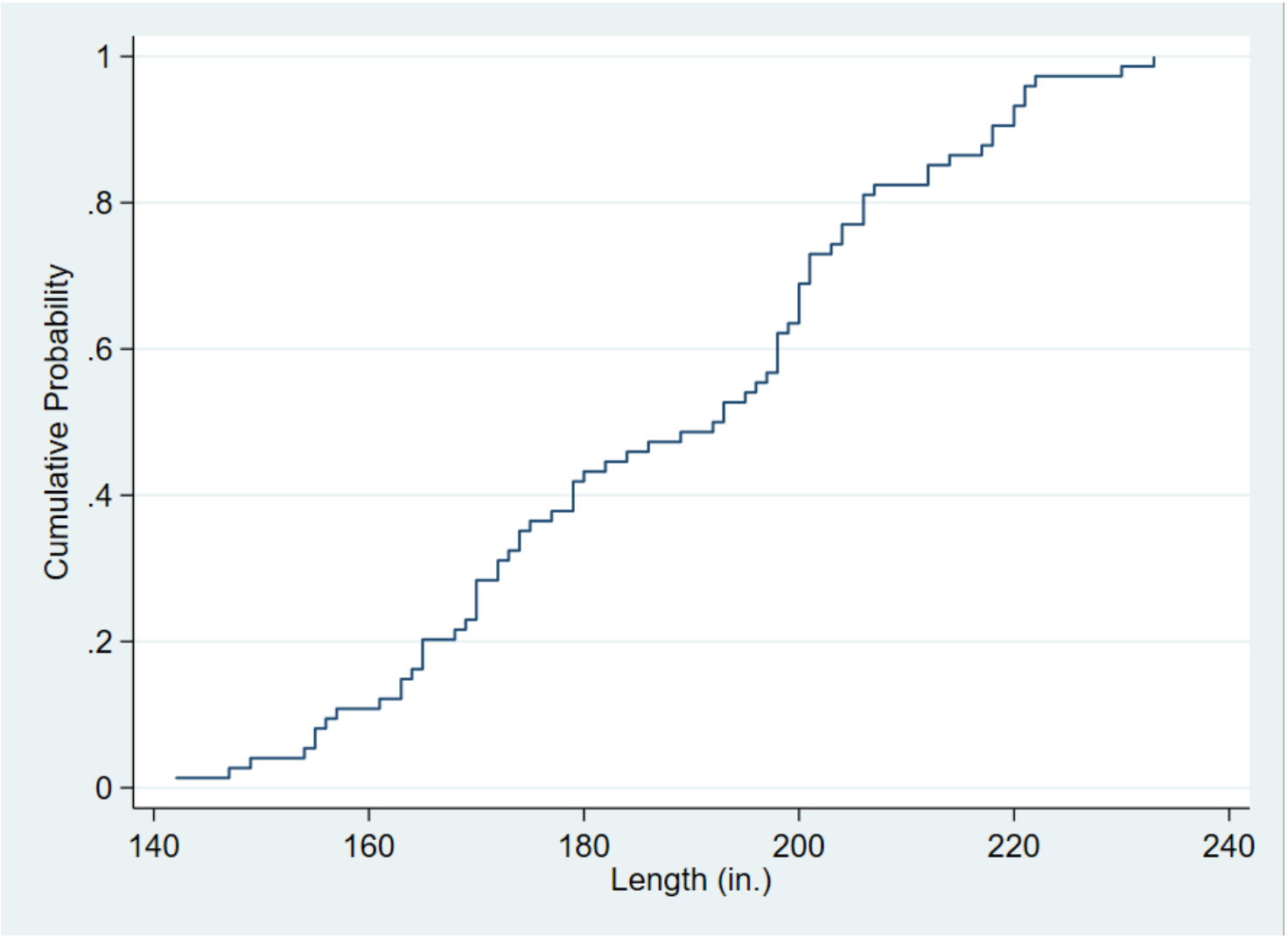
格里文科定理表明：当n相当大时, 经验分布函数是总体分布函数$F(x)$的一个良好的近似.

经典的统计学中一切统计推断都以样本为依据, 其理由就在于此.

用Stata给出有序样本, 做经验分布函数图

```
ssc install cdfplot, replace
help cdfplot
sysuse auto,replace
keep length foreign
//produce ordered sample
sort length

// produce empirical distribution plot
sum length
    cdfplot length
    cdfplot length, normal
    cdfplot length, by(foreign)
    cdfplot length, by(foreign) norm saving(mygraph,replace)
```

cdfplot length



5.2.2 频数-频率分布表(frequency-relative frequency table)

样本数据的整理是统计研究的基础, 整理数据的最常用方法之一是给出其**频数分布表**或**频率分布表**.

例5.2.2 为研究某厂工人生产某种产品的能力, 随机调查了20位工人某天生产的该种产品的数量, 数据如下

160	196	164	148
175	178	166	181
161	168	166	162
156	170	157	162

对这20个数据(样本)进行整理,具体步骤如下:

- (1) 对样本进行分组：作为一般性的原则, 组数通常在5-20个;
- (2) 确定每组组距：近似公式为组距d = (最大观测值-最小观测值)/组数;
- (3) 确定每组组限：各组区间端点为$a_0, a_1 = a_0 + d, a_2 = a_0 + 2d, \ldots, a_k = a_0 + kd$, 形成如下的分组区间$(a_0, a_1], (a_1, a_2], \ldots, (a_{k-1}, a_k]$ 其中a_0 略小于最小观测值, a_k略大于最大观测值.
- (4) 统计样本数据落入每个区间的个数——频数, 并列出其频数频率分布表.

表5.2.1 例5.2.2 的频数频率分布表

组序	分组区间	组中值	频数	频率	累计频率(%)
1	(147, 157]	152	4	0.20	20
2	(157, 167]	162	8	0.40	60
3	(167, 177]	172	5	0.25	85
4	(177, 187]	182	2	0.10	95
5	(187, 197]	192	1	0.05	100
合计			20	1	

5.2.3 样本数据的图形显示

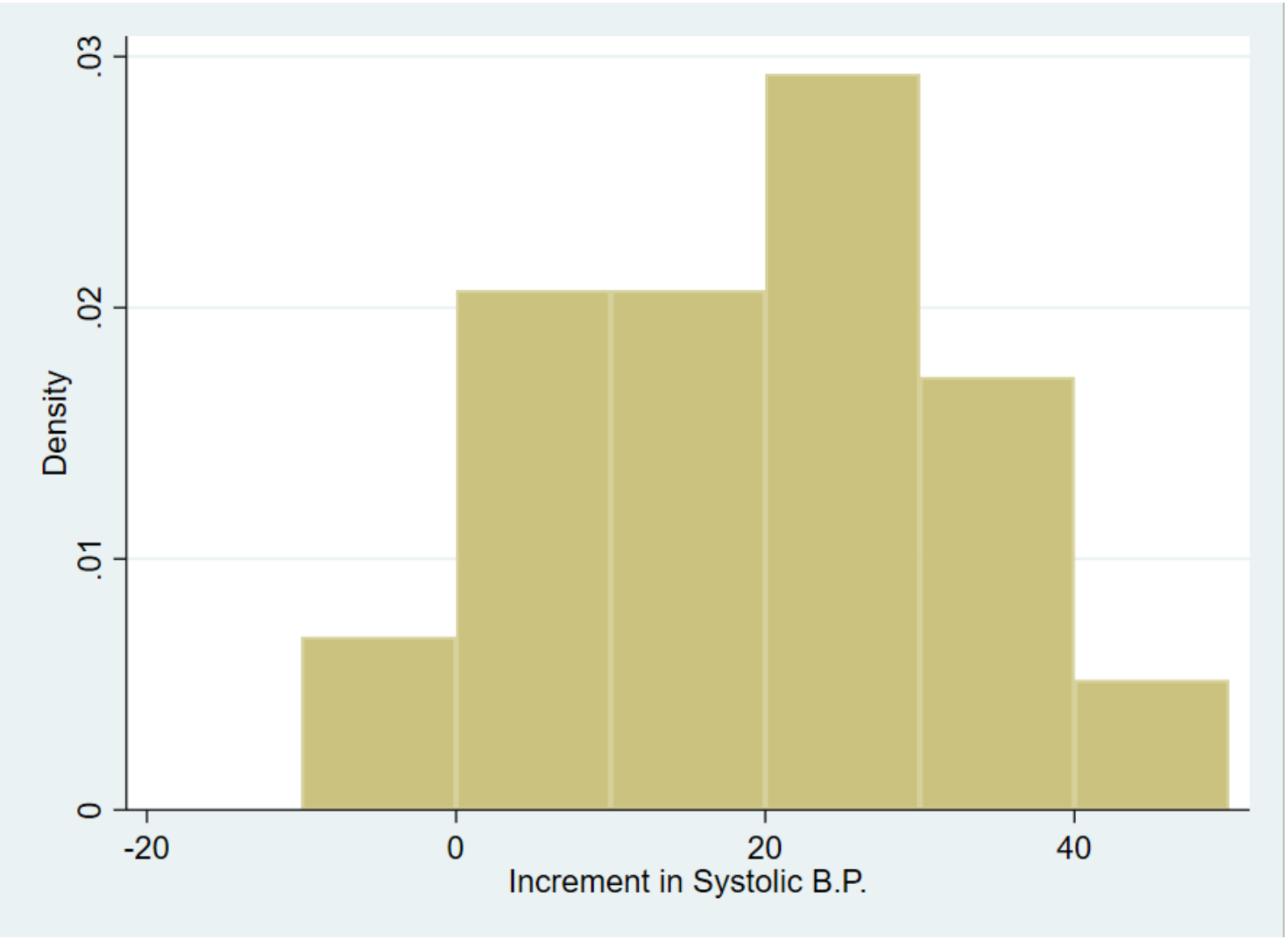
一. 直方图(histogram)

- 直方图是频数分布的图形表示, 它的横坐标表示所关心变量的取值区间.
- 纵坐标有三种表示方法：频数(frequency), 频率(fraction), 或是密度(即频率/组距, density), 它可使得诸长条矩形面积和为1.
- 三种直方图的差别仅在于纵轴刻度的选择, 直方图本身并无变化.

用stata做直方图

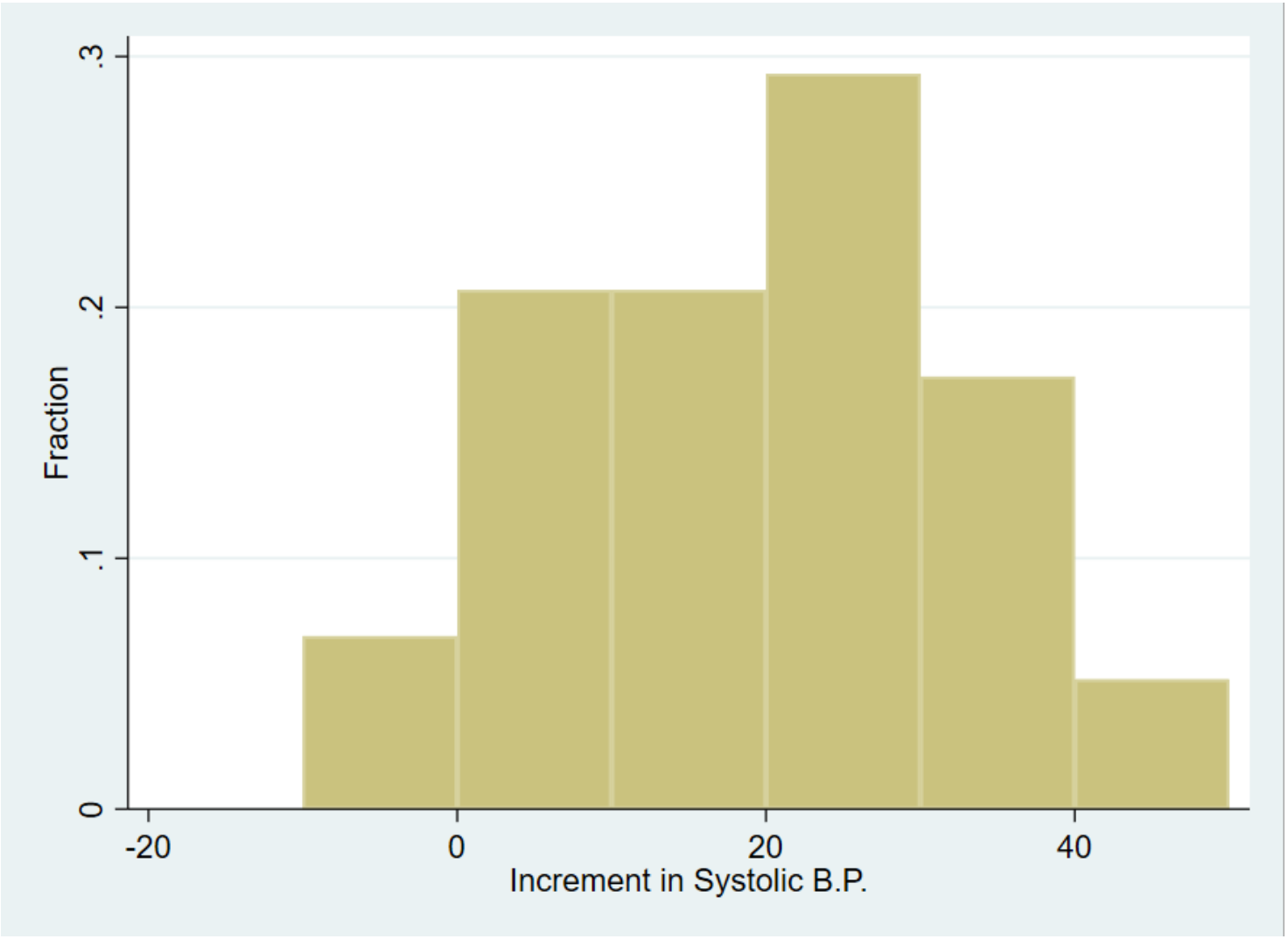
```
help hist
webuse systolic

hist drug, discrete
hist drug, discrete frequency
hist systolic
hist systolic, bins(10)
hist systolic, width(10) start(-10) //default: 密度(即 频率/组距, density
hist systolic, width(10) start(-10) fraction //频率(fraction)
hist systolic, width(10) start(-10) frequency //频率(fraction)
```

hist systolic, width(10) start(-10) //default: 密度(即频率/组距, density)

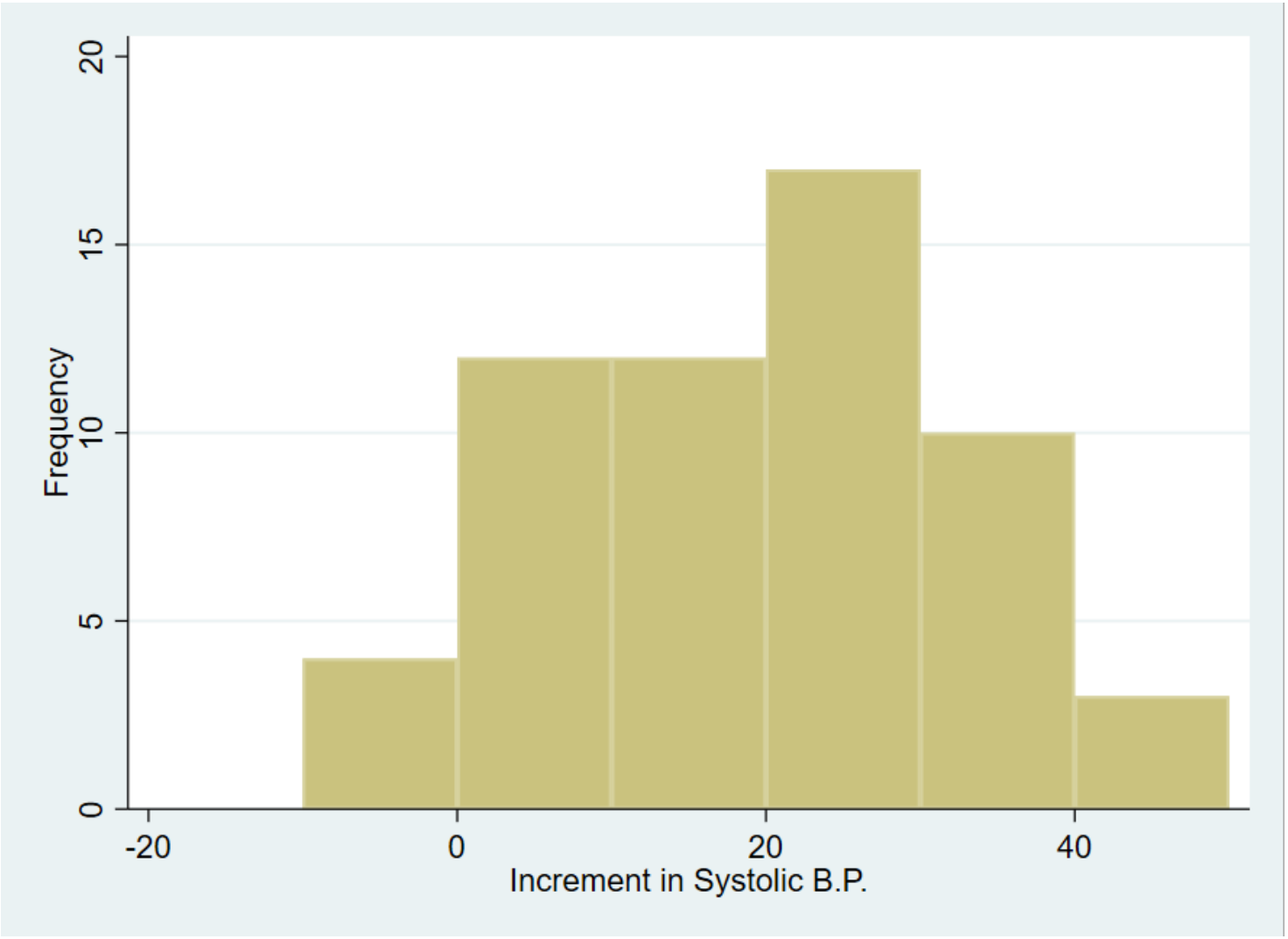


cdfplot length

hist systolic, width(10) start(-10) fraction //频率(fraction)



hist systolic, width(10) start(-10) frequency //频率(fraction)



5.2.3 样本数据的图形显示

二. 茎叶图 把每一个数值分为两部分, 前面一部分(百位和十位)称为茎, 后面部分(个位)称为叶, 然后画一条竖线, 在竖线的左侧写上茎, 右侧写上叶, 就形成了茎叶图. 如:

数值	分开	茎	和	叶
112	11 | 2	11	和	2

例5.2.3 某公司对应聘人员进行能力测试, 测试成绩总分为150分. 下面是50位应聘人员的测试成绩(已经过排序):

64	67	70	72	74	76	76	79	80	81
82	82	83	85	86	88	91	91	92	93
93	93	95	95	95	97	97	99	100	100
102	104	106	106	107	108	108	112	112	114
116	118	119	119	122	123	125	126	128	133

- 我们用这批数据给出一个茎叶图, 见下页.

图5.2.3　测试成绩的茎叶图

6	4	7								
7	0	2	4	6	6	9				
8	0	1	2	2	3	5	6	8		
9	0	1	2	3	3	5	6	6	7	9
9	1	1	2	3	3	3	5	6	6	7	9
10	0	0	2	4	6	6	7	8	8	
11	2	2	4	6	8	9	9			
12	2	3	5	6	8					
13	3									

- 在要比较两组样本时, 可画出它们的**背靠背的茎叶图**. 见教材图5.2.4.
- 注意：茎叶图保留数据中全部信息. 当样本量较大, 数据很分散, 横跨二, 三个数量级时, 茎叶图并不适用.

用stata做茎叶图

```
clear
webuse systolic
stem systolic , lines(1)
stem systolic if drug==1, lines(1)
stem systolic if drug==2, lines(1)

. stem systolic , lines(1)

Stem-and-leaf plot for systolic (Increment in Systo

 -0* | 6532
  0* | 111334577999
  1* | 122233556699
  2* | 11222344555667889
  3* | 1123334466
  4* | 224
```

更多描述性统计在下半学期会介绍!

如：饼状图 (stata: graph pie)

5.3.1 统计量与抽样分布

- 当人们需要从样本获得对总体各种参数的认识时, 最好的方法是构造样本的函数, 不同的函数反映总体的不同特征.

定义5.3.1 设 $x_1, x_2, \ldots, x_n$ 为取自某总体的样本, 若样本函数$T = T(x_1, x_2, \ldots, x_n)$中**不含有任何未知参数**, 则称$T$为**统计量**(statistic).

统计量的分布称为**抽样分布**(sample distribution).

按照这一定义：若 $x_1, x_2, \ldots, x_n$为样本, 则 $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$以及经验分布函数$F_n(x)$都是统计量. 而当μ, σ未知时, $x_1 - \mu$, x_1/σ 等均不是统计量.

- 尽管统计量不依赖于未知参数, 但是它的分布一般是依赖于未知参数的.
- 下面介绍一些常见的统计量及其抽样分布.

5.3.2 样本均值(sample mean)及其抽样分布(sample distribution)

定义5.3.2: 设 $x_1, x_2, \ldots, x_n$为取自某总体的样本, 其算术平均值称为**样本均值**, 一般用$\bar{x}$表示, 即

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 思考：在分组样本场合, 样本均值如何计算?
- 二者结果相同吗?

5.3.2 样本均值(sample mean)及其抽样分布(sample distribution)

定义5.3.2: 设 $x_1, x_2, \ldots, x_n$为取自某总体的样本, 其算术平均值称为**样本均值**, 一般用$\bar{x}$表示, 即

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 思考：在分组样本场合, 样本均值如何计算?
- 二者结果相同吗? (一般不同, 但近似见例5.3.1)

样本均值的基本性质

定理5.3.1: 若把样本中的数据与样本均值之差称为偏差(deviation), 则样本所有**偏差之和**(sum of deviations)为0, 即$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=0$.

证明:

$$\sum\left(x_i-\bar{x}\right)=\sum x_i-n\bar{x}=\sum x_i-n\cdot\frac{\sum x_i}{n}=0$$

样本均值的基本性质

定理5.3.2: 数据观测值与均值的**偏差平方和**(sum of squared deviations)最小, 即在形如$(x_i - c)^2$的函数中, $\sum (x_i - \bar{x})^2$最小, 其中c为任意给定常数.

证明: 对任意的c有

$$\begin{aligned}\sum (x_i - c)^2 &= \sum (x_i - \bar{x} + \bar{x} - c)^2 \\ &= \sum (x_i - \bar{x})^2 + n(\bar{x} - c)^2 + 2\sum (x_i - \bar{x})(\bar{x} - c) \\ &= \sum (x_i - \bar{x})^2 + n(\bar{x} - c)^2 \geqslant \sum (x_i - \bar{x})^2\end{aligned}$$

样本均值的抽样分布

定理5.3.3 设$x_1, x_2, \ldots, x_n$是来自某个总体的样本, $\bar{x}$为样本均值.

- (1) 若总体分布为$N(\mu, \sigma^2)$, 则 $\bar{x}$ 的精确分布为$N(\mu, \sigma^2/n)$;
- (2) 若总体分布未知或不是正态分布, 但$E(x) = \mu, Var(x) = \sigma^2$, 则$n$ 较大时$\bar{x}$的**渐近分布**为$N(\mu, \sigma^2/n)$,常记为$\bar{x} \sim AN\left(\mu, \sigma^2/n\right)$ (asymptotically normal)

这里**渐近分布**是指n较大时的近似分布. 下面给出证明.

样本均值的抽样分布

(1) 若总体分布为$N(\mu, \sigma^2)$, 则 $\bar{x}$ 的精确分布为$N(\mu, \sigma^2/n)$;

证明： 由卷积公式： $\sum_{i=1}^{n} x_i \sim N\left(n\mu, n\sigma^2\right)$, 所以 $\bar{x} \sim N(\mu, \sigma^2/n)$

(2) 若总体分布未知或不是正态分布, 但 $E(x) = \mu, Var(x) = \sigma^2$,则$n$ 较大时 $\bar{x}$ 的**渐近分布**为$N(\mu, \sigma^2/n)$.

证明： 由中心极限定理

$$\frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma\sqrt{n}} = \sqrt{n}(\bar{x} - \mu)/\sigma \xrightarrow{L} N(0, 1)$$

拓展练习： 利用STATA软件验证总体分别为正态分布, 均匀分布, 指数分布时样本均值的抽样分布 (示意代码传至课程中心)

复习：独立同分布下的中心极限定理

定理4.4.1 林德贝格-勒维中心极限定理 (Lindeberg-Levy CLT): 设$\{X_n\}$为独立同分布随机变量序列, X_i的数学期望为μ, $E(X_i)=\mu$, 方差为$\sigma^2>0$, $Var(X_i)=\sigma^2$, 记

$$Y_n^*=\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$$

则对任意实数y, Y_n^* 的分布弱收敛于标准正态分布, 即

$$\lim_{n\to\infty}P\left(Y_n^*\leqslant y\right)=\lim_{n\to\infty}P\left\{\frac{\sum_{i=1}^{n}X_i-n\mu}{\sigma\sqrt{n}}\leq y\right\}$$

$$=\Phi(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}e^{-\frac{t^2}{2}}dt$$

即$\{Y_n^*\}$依分布收敛于标准正态分布的随机变量:

$Y_n^*\xrightarrow{L}X\sim N(0,1)$.

5.3.3 样本方差与样本标准差

定义5.3.3 设$x_1, x_2, \ldots, x_n$是来自某个总体的样本, 则它关于样本均值$\bar{x}$的平均偏差平方和(average sum of the squared deviations)

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

称为**样本方差**(sample variance), 其算术平方根$s_n = \sqrt{s_n^2}$称为**样本标准差**.

在n 不大时, 常用$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$作为样本方差, 其算术平方根$s = \sqrt{s^2}$也称为**样本标准差**.

- $(n-1)$是自由度调整.

在这个定义中, $\sum_{i=1}^{n}(x_i-\bar{x})^2$称为偏差平方和, $(n-1)$称为偏差平方和的自由度. 其含义是：在$\bar{x}$确定后, n 个偏差$x_1-\bar{x}, x_2-\bar{x}, \ldots, x_n-\bar{x}$ 中只有$n-1$个数据可以自由变动, 而第n个则不能自由取值.

样本偏差平方和有三个不同的表达式：

$$\sum(x_i-\bar{x})^2=\sum x_i^2-\frac{(\sum x_i)^2}{n}=\sum x_i^2-n\bar{x}^2$$

它们都可用来计算样本方差.

- 思考：分组样本如何计算样本方差?

样本均值的数学期望和方差, 以及样本方差的数学期望都不依赖于总体的分布形式.

定理5.3.4 设总体 X 具有二阶矩, 即 $E(x)=\mu, Var(x)=\sigma^2<\infty$, $x_1, x_2, \ldots, x_n$ 为从该总体得到的样本, $\bar{x}$和s^2 分别是样本均值和样本方差, 则

$$E(\bar{x})=\mu, \quad \mathrm{Var}(\bar{x})=\sigma^2/n, \quad E\left(s^2\right)=\sigma^2$$

(后面会学到**无论总体的分布如何**, 样本均值是总体期望的无偏估计, 样本方差是总体方差的无偏估计)

证明:

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

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

证明(续)：

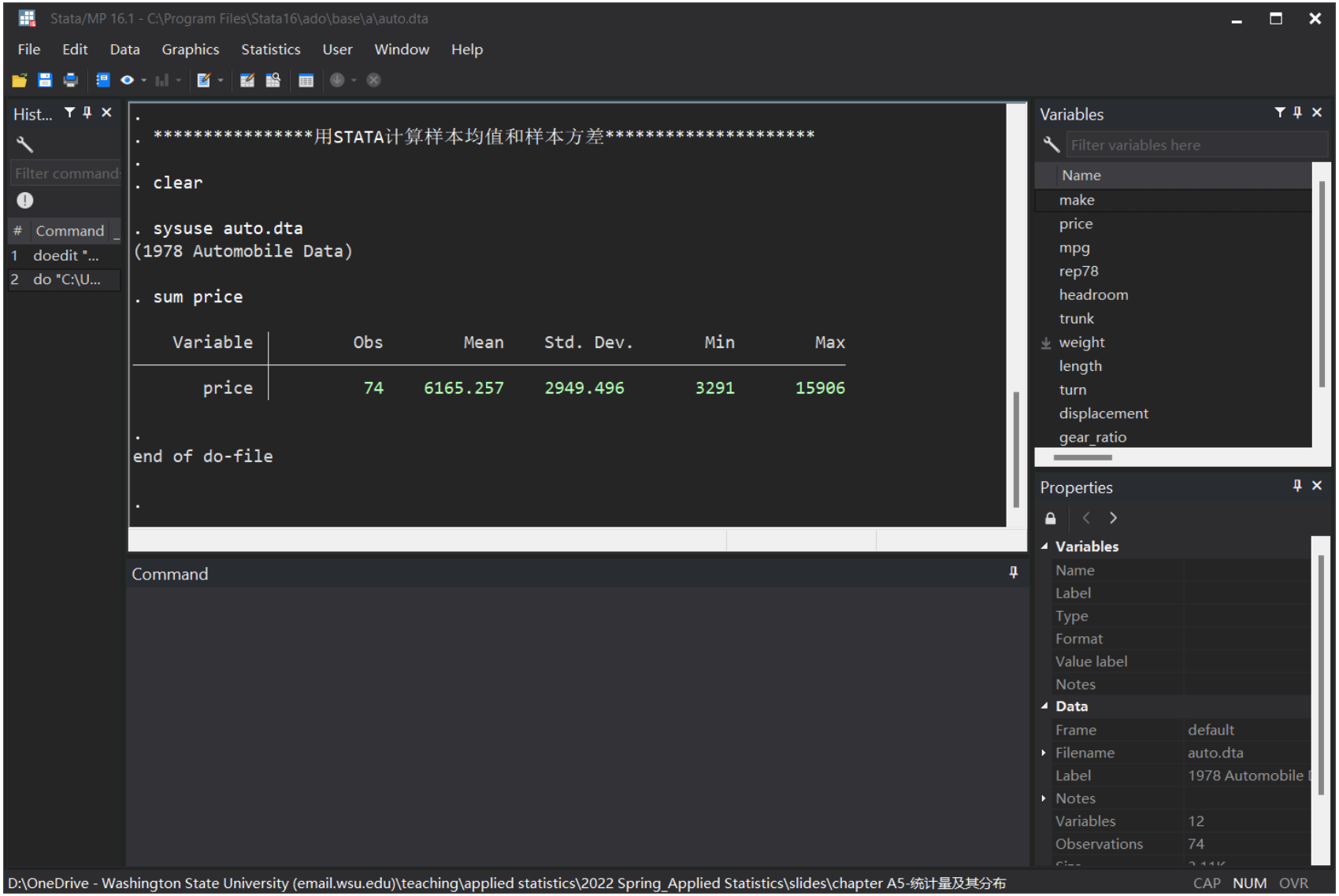
$$
\begin{aligned}
E\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right) &= E\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)\\
&= nEx_i^2-nE\left(\bar{x}^2\right)\\
&= n\left(\left(Ex_i\right)^2+Var\left(x_i\right)\right)-n\left(\left(E\bar{x}\right)^2+\mathrm{Var}(\bar{x})\right)\\
&= n\left(\mu^2+\sigma^2\right)-n\left(\mu^2+\sigma^2/n\right)\\
&= (n-1)\sigma^2
\end{aligned}
$$

两边同除$n-1$得证.

思考： $E(s)=\sigma$是否成立?

用STATA计算样本均值和样本方差

```
clear
sysuse auto.dta
sum price
```


Stata/MP 16.1 - C:\Program Files\Stata16\ado\base\a\auto.dta
File Edit Data Graphics Statistics User Window Help
Hist...
Filter command
Command
1 doedit "...
2 do "C:\U...
. **********用STATA计算样本均值和样本方差**********
. clear
. sysuse auto.dta
(1978 Automobile Data)
. sum price
Variable | Obs Mean Std. Dev. Min Max
price | 74 6165.257 2949.496 3291 15906
end of do-file
Command
Variables
Filter variables here
Name
make
price
mpg
rep78
headroom
trunk
weight
length
turn
displacement
gear_ratio
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Frame default
Filename auto.dta
Label 1978 Automobile
Notes
Variables 12
Observations 74
D:\OneDrive - Washington State University (email.wsu.edu)\teaching\applied statistics\2022 Spring_Applied Statistics\slides\chapter A5-统计量及其分布
CAP NUM OVR

5.3.4 样本矩及其函数

样本均值和样本方差的更一般的推广是样本矩, 这是一类常见的统计量.

定义5.3.4 $a_k = \frac{1}{n}\sum_{i=1}^{n}(x_i^k)$ **称为样本** k **阶原点矩**. 样本一阶原点矩就是样本均值.

$b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$ **称为样本**k**阶中心矩**. 样本二阶中心矩就是样本方差.

当总体关于分布中心对称时, 我们用$\bar{x}$和s刻画样本特征很有代表性, 而当其不对称时, 需要一些刻画分布形状的统计量, 如样本偏度和样本峰度, 它们都是样本中心矩的函数.

定义5.3.5, 5.3.6:

- $\hat{\beta}_S = b_3/b_2^{3/2}$ 称为**样本偏度**,
- $\hat{\beta}_k = b_4/b_2^2 - 3$ 称为**样本峰度**.

样本偏度$\hat{\beta}_S$反映了总体分布密度曲线的对称性信息. 样本峰度$\hat{\beta}_k$反映了总体分布密度曲线在其峰值附近的陡峭程度.

5.3.5 次序统计量(order statistics)及其分布

另一类常见的统计量是次序统计量.

定义5.3.7: 设 $x_1, x_2, \ldots, x_n$是取自总体X的样本, $x_{(i)}$ 称为该样本的第i个次序统计量, 它的取值是将样本观测值由小到大排列后得到的第 i 个观测值. 其中$x_{(1)} = min\{x_1, x_2, \ldots, x_n\}$称为该样本的**最小次序统计量**, 称$x_{(n)} = max\{x_1, x_2, \ldots, x_n\}$为该样本的**最大次序统计量**.

- 在一个样本中, $x_1, x_2, \ldots, x_n$是独立同分布的, 而次序统计量$(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ 则既不独立, 分布也不相同, 见下面例子.

例 5.3.6

例 5.3.6 设总体 X 的分布为仅取 $0, 1, 2$ 的离散均匀分布, 分布列为

x	0	1	2
p	$1/3$	$1/3$	$1/3$

现从中抽取容量为 3 的样本, 其一切可能取值有 $3^3 = 27$ 种, 现将它们列在下表左侧, 其右侧是相应的次序统计量观测值.

表 5.3.6 例 5.3.6 中样本取值及其次序统计量取值

x_1	x_2	x_3	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	x_1	x_2	x_3	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$
0	0	0	0	0	0	1	2	0	0	1	2
0	0	1	0	0	1	2	1	0	0	1	2
0	1	0	0	0	1	0	2	2	0	2	2
1	0	0	0	0	1	2	0	2	0	2	2
0	0	2	0	0	2	2	2	0	0	2	2
0	2	0	0	0	2	1	1	2	1	1	2
2	0	0	0	0	2	1	2	1	1	1	2
0	1	1	0	1	1	2	1	1	1	1	2
1	0	1	0	1	1	1	2	2	1	2	2
0	1	2	0	1	2	2	2	1	1	2	2
0	2	1	0	1	2	1	1	1	1	1	1
1	0	2	0	1	2	2	2	2	2	2	2
2	0	1	0	1	2						

由于样本取上述每一组观测值的概率相同, 都为 $1/27$, 由此可给出 $x_{(1)}, x_{(2)}, x_{(3)}$ 的分布列如下:

$x_{(1)}$	0	1	2
p	$\frac{19}{27}$	$\frac{7}{27}$	$\frac{1}{27}$

$x_{(2)}$	0	1	2
p	$\frac{7}{27}$	$\frac{13}{27}$	$\frac{7}{27}$

$x_{(3)}$	0	1	2
p	$\frac{1}{27}$	$\frac{7}{27}$	$\frac{19}{27}$

所以三个次序统计量的分布是不相同的.

进一步,给出两个次序统计量的联合分布,如, $x_{(1)}$ 和 $x_{(2)}$ 的联合分布列为

$x_{(1)}$ \ $x_{(2)}$	0	1	2
0	$7/27$	$9/27$	$3/27$
1	0	$4/27$	$3/27$
2	0	0	$1/27$

因为 $P\left(x_{(1)}=0\right)P\left(x_{(2)}=0\right)=\frac{19}{27}\times\frac{7}{27}$, 而 $P\left(x_{(1)}=0,x_{(2)}=0\right)=\frac{7}{27}$,两者不等.

由此看出$x_{(1)}$ 和 $x_{(2)}$ 不独立.

5.3.5 次序统计量及其分布

单个次序统计量的分布

定理5.3.5 设总体X的密度函数为$p(x)$, 分布函数为$F(x)$, $x_1, x_2, \ldots, x_n$为样本, 则第k个次序统计量$x_{(k)}$的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(1-F(x))^{n-k}p(x)$$

- 特别地, 若总体分布为$U(0,\theta)$, 则$x_{(n)}$ 的分布密度函数为

$$p(y) = ny^{n-1}/\theta^n, \quad y < \theta$$

5.3.6 样本分位数与样本中位数

样本中位数(median)也是一个很常见的统计量, 它也是次序统计量的函数, 通常如下定义:

$$m_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, n\text{为奇数} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), n\text{为偶数} \end{cases}$$

更一般地, 样本p分位数(quantile)m_p可如下定义:

$$m_p = \begin{cases} x_{([np+1])}, np\text{不为整数} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), np\text{为整数} \end{cases}$$

- "[ ]"表示取整数
- 样本分位数和中位数永远存在
- 样本中位数相比样本均值更加稳健: 受极值影响小

用STATA计算上述统计量

```
clear
sysuse auto.dta
sum price, de
```

Price

Percentiles		Smallest		
1%	3291	3291		
5%	3748	3299		
10%	3895	3667	Obs	74
25%	4195	3748	Sum of Wgt.	74
50%	5006.5		Mean	6165.257
		Largest	Std. Dev.	2949.496
75%	6342	13466		
90%	11385	13594	Variance	8699526
95%	13466	14500	Skewness	1.653434
99%	15906	15906	Kurtosis	4.819188

样本分位数和中位数的渐近分布

定理5.3.7 设总体密度函数为$p(x)$, x_p为总体p分位数, $p(x)$在x_p处连续且 $p(x_p) > 0$, 则当$n \to \infty$时样本 p 分位数m_p 的渐近分布为

$$m_p \dot{\sim} N\left(x_p, \frac{p(1-p)}{n \cdot p^2(x_p)}\right)$$

特别, 对样本中位数, 当$n \to \infty$时近似地有

$$m_{0.5} \dot{\sim} N\left(x_{0.5}, \frac{1}{4n \cdot p^2(x_{0.5})}\right)$$

证明略.

5.3.7 五数概括与箱线图

次序统计量的应用之一是**五数概括**与**箱线图**. 在得到有序样本后, 容易计算如下五个值:

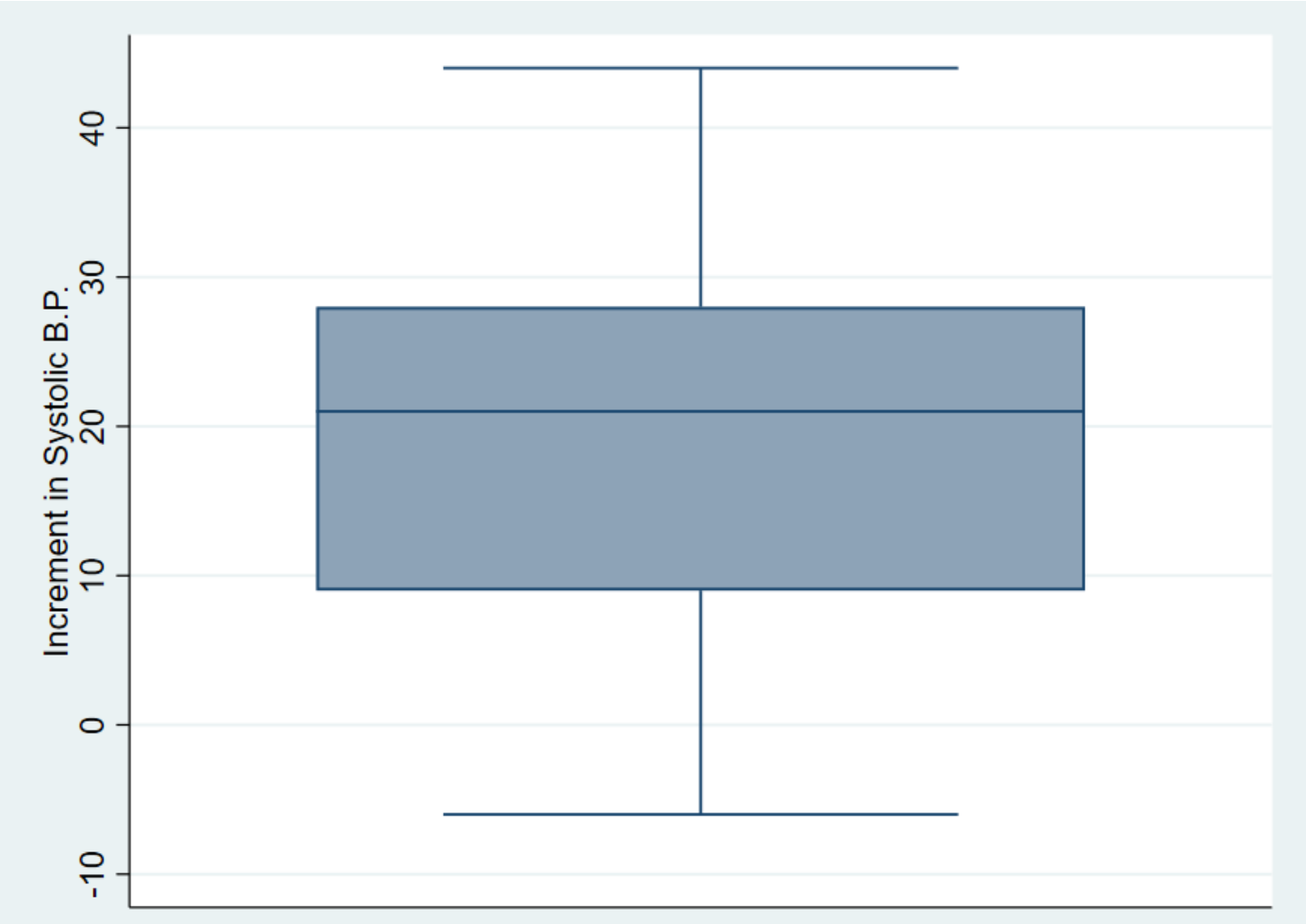
- 最小观测值 $x_{min} = x_{(1)}$, 最大观测值 $x_{max} = x_{(n)}$,
- 中位数 $m_{0.5}$,
- 第一4分位数 $Q_1 = m_{0.25}$, 第三4分位数 $Q_3 = m_{0.75}$.

所谓**五数概括**就是指用这五个数: $x_{min}, Q_1, m_{0.5}, Q_3, x_{max}$来大致描述一批数据的轮廓.

箱线图(box and whisker plot, box plot): 五数概括的图形表示: 箱子+线段. 箱子从第一4分位数到第三4分位数, 中间含有中位数, 线段向两侧延伸到最小和最大观测值.

用stata做箱线图

```
webuse systolic
graph box systolic
```



§5.4 三大抽样分布

有很多统计推断是基于正态分布的假设的, 以标准正态变量为基石而构造的三个著名统计量在实际中有广泛的应用, 这是因为这三个统计量不仅有明确背景, 而且其抽样分布的密度函数有明确表达式, 它们被称为统计中的“三大抽样分布”

- χ^2分布(卡方分布)
- F分布
- t分布

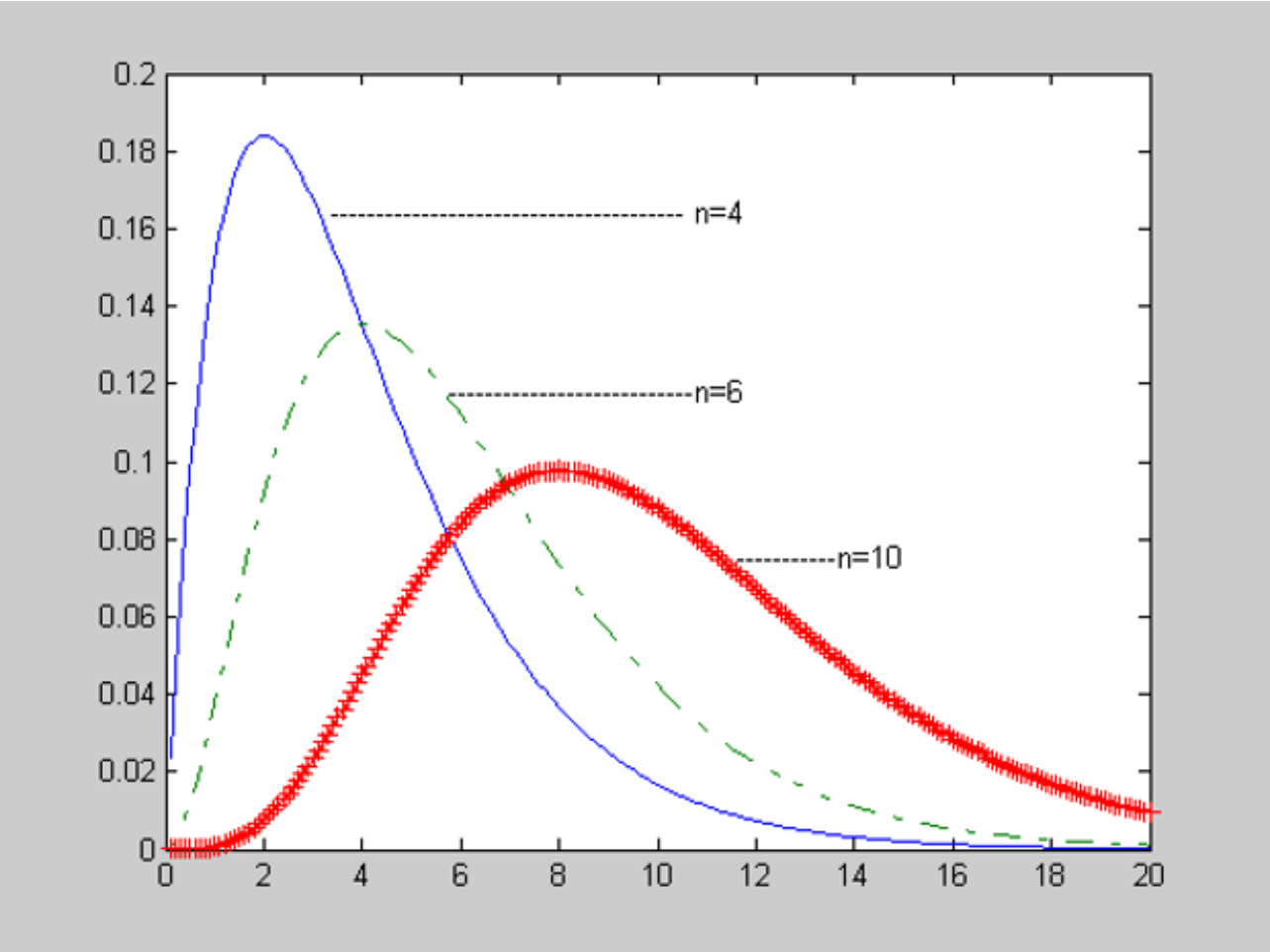
5.4.1 χ^2分布(卡方分布)

定义5.4.1 设 $X_1, X_2, \ldots, X_n$ 独立同分布于标准正态分布$N(0,1)$, 则$\chi^2 = X_1^2 + \ldots + X_n^2$的分布称为自由度为$n$的$\chi^2$分布, 记为 $\chi^2 \sim \chi^2(n)$.

- χ^2分布的密度函数(推导见教材):

$$p(y) = \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n/2}} y^{n/2-1} \mathrm{e}^{-\frac{y}{2}} \quad (y > 0)$$

  其中 $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$ (伽玛函数).
- 若 $\chi^2 \sim \chi^2(n)$, 则$E\left(\chi^2\right) = n, Var\left(\chi^2\right) = 2n$. (推导见教材2.5常用连续分布p104-105)
- 可应用于单个正态总体方差检验.



- 该密度函数的图像是一只取非负值的偏态分布

- 当随机变量 $\chi^2 \sim \chi^2(n)$时, 对给定 $\alpha(0 < \alpha < 1)$, 称满足$P\left(\chi^2 \leq \chi^2_{1-\alpha}(n)\right) = 1 - \alpha$的$\chi^2_{1-\alpha}(n)$是自由度为$n$的卡方分布的$1 - \alpha$分位数.
- 分位数 $1 - \alpha$ 可以从附表3中查到.

用STATA计算$\chi^2(n)$的分位数$\chi^2_{1-\alpha}(n)$

```
help function
help invchi2// the inverse of cumulative chi-square distribution
// chi2(): if chi2(df,x) = p,  then invchi2(df,p) = x

//calculating the 0.005 percentile of chi(5)
di invchi2(5,0.005)

0.4117
```

课堂练习: 查附表3, 验证$\chi^2(5)$的0.005分位数$\chi^2_{0.005}(5) = 0.4117$.

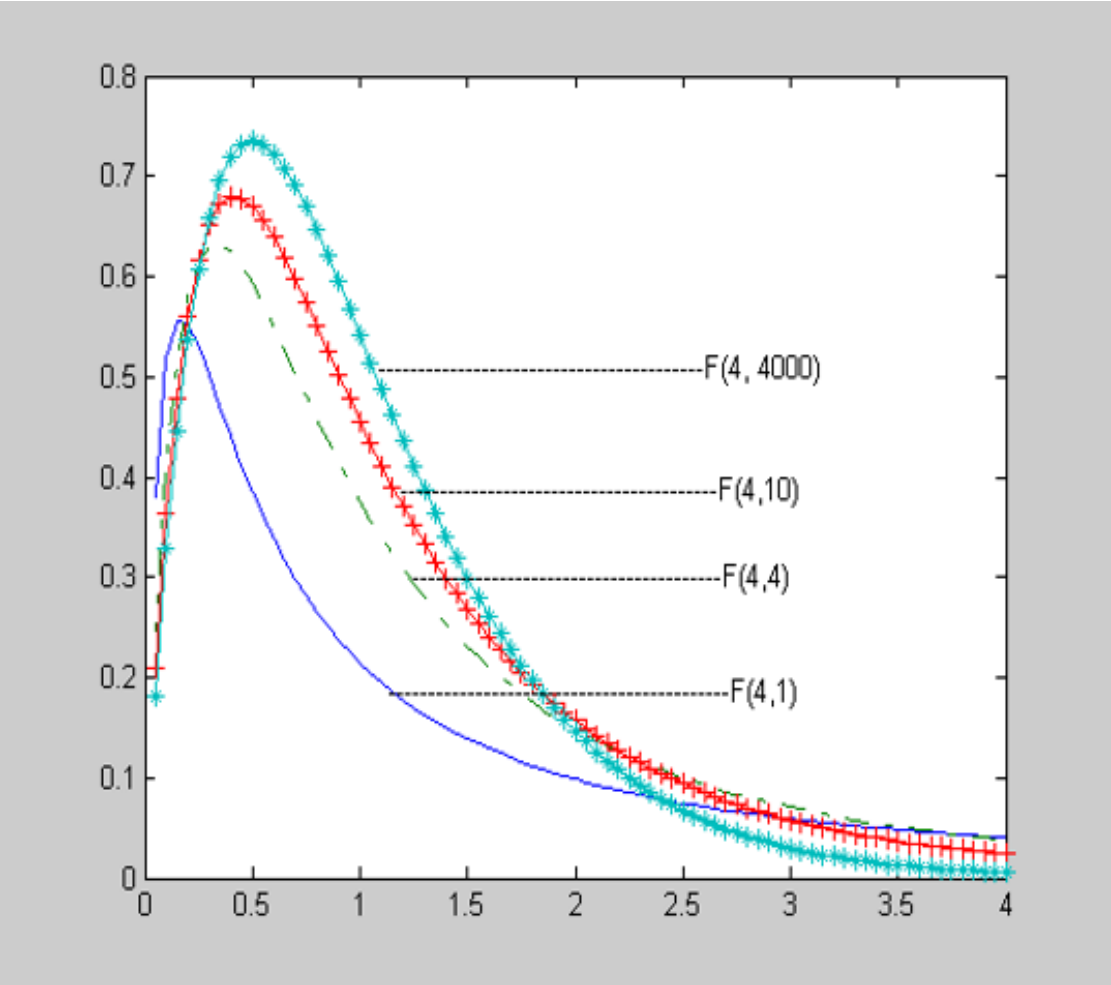
5.4.2 F 分布

定义5.4.2 设$X_1 \sim \chi^2(m), X_2 \sim \chi^2(n)$, X_1与X_2独立, 则称$F = (X_1/m)/(X_2/n)$的分布是自由度为m与 n的F分布, 记为$F \sim F(m,n)$, 其中m 称为分子自由度, n称为分母自由度.

- F分布的密度函数(推导见教材):

$$p(y) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} y^{\frac{m}{2}-1}\left(1+\frac{m}{n}y\right)^{-\frac{m+n}{2}}$$

- 可应用于两正态总体方差比的检验; 可应用于回归模型线性关系检验



- 该密度函数的图象也是一只取非负值的偏态分布

5.4.2 F 分布

由F分布的构造知

- 若$F \sim F(m,n)$, 则$1/F \sim F(n,m)$

且

- $F_{\alpha}(n,m) = \dfrac{1}{F_{1-\alpha}(m,n)}$.

- 当随机变量$F \sim F(m,n)$ 时, 对给定 $\alpha(0 < \alpha < 1)$, 称满足 $P(F \leqslant F_{1-\alpha}(m,n)) = 1-\alpha$的$F_{1-\alpha}(m,n)$ 是自由度为m 与 n的F 分布的$1-\alpha$分位数 (附表5).

用STATA计算$F_{1-\alpha}(m,n)$分位数

```
help function
help invF// the inverse cumulative F distribution:
//   if F(df1,df2,f) = p, then invF(df1,df2,p) = f
// 计算自由度为(10, 5)的F分布的0.95分位数
di invF(10, 5, 0.95)

4.7350631
```

课堂练习: 查附表5, 找到自由度为(10, 5)的F分布的0.95分位数

5.4.3 t 分布

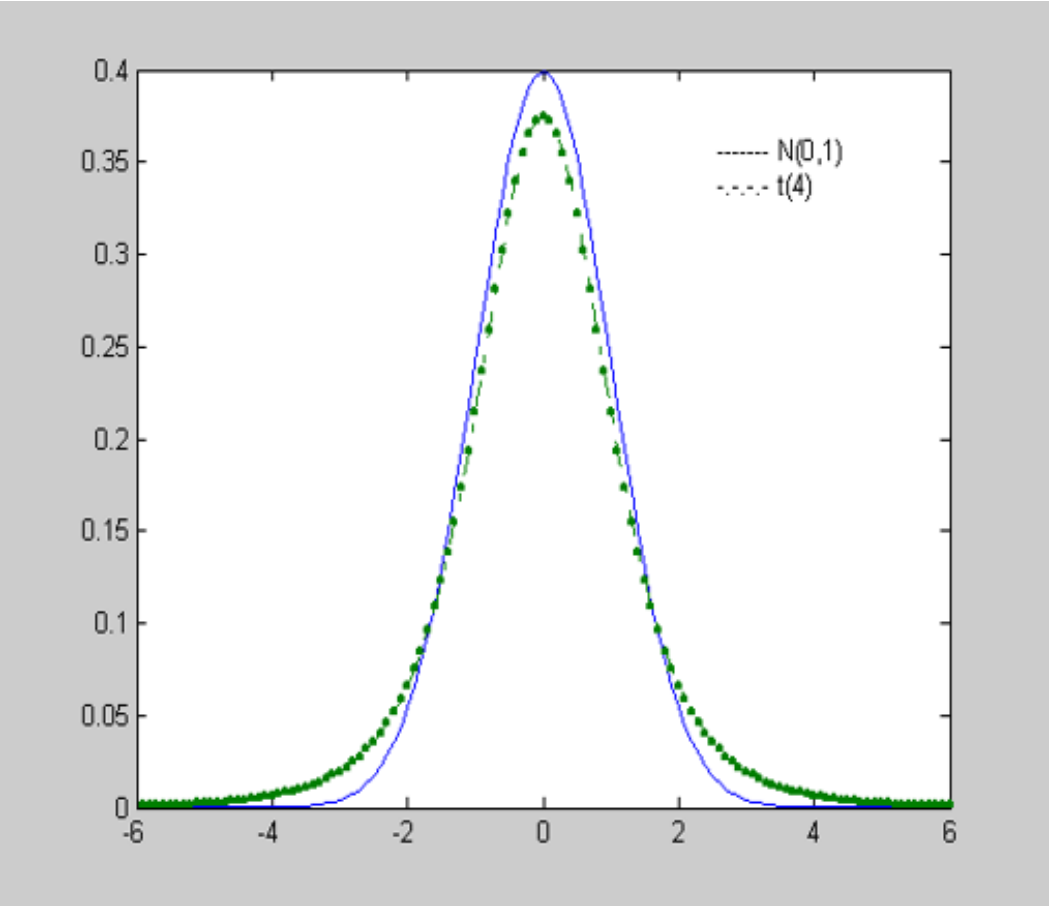
定义 5.4.3 设随机变量X_1 与X_2独立, 且$X_1 \sim N(0,1), X_2 \sim \chi^2(n)$, 则称$t = \dfrac{X_1}{\sqrt{X_2/n}}$的分布为自由度为$n$ 的t 分布, 记为$t \sim t(n)$.

- t分布的密度函数(推导见教材):

$$p(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}$$

- 若 $t \sim t(n)$, 则$E(t) = 0, \quad n > 1;$
  $Var(t) = \dfrac{n}{n-2}, \quad n > 2$

- 可应用于单总体均值检验; (最常用)可应用于回归模型的参数检验.



- t 分布的密度函数的图象是一个关于纵轴对称的分布, 与标准正态分布的密度函数形状类似, 只是峰比标准正态分布低一些尾部的概率比标准正态分布的大一些.

- 当自由度较大 (如$n \geqslant 30$) 时, t 分布可以用正态分布$N(0,1)$近似.
- 自由度为1的t分布就是标准柯西分布, 它的均值不存在.

- 当随机变量$t \sim t(n)$时, 称满足$P(t \leqslant t_{1-\alpha}(n)) = 1-\alpha$的$t_{1-\alpha}(n)$ 是自由度为 n 的t 分布的$1-\alpha$分位数.
- 分位数 $t_{1-\alpha}(n)$ 可以从附表4中查到.
- 譬如 $n = 10, \alpha = 0.05$, 那么从附表4上查得$t_{1-0.05}(10) = t_{0.95}(10) = 1.812$.

- 由于t 分布的密度函数关于0对称, 故其分位数间有如下关系$t_{\alpha}(n) = -t_{1-\alpha}(n)$.

用STATA计算$t_{1-\alpha}(n)$ 分位数：

```
help function
help invt // the inverse cumulative Student's t distribution:
// if t(df,t) = p, then invt(df,p) = t
// 计算自由度为11的t分布的0.95分位数
di invt(11, 0.95)

1.7958848
```

5.4.4 一些重要结论

定理5.4.1 设 $x_1, x_2, \ldots, x_n$是来自$N(\mu, \sigma^2)$ 的样本, 其样本均值和样本方差分别为 $\bar{x} = \sum_{i=1}^{n} x_i/n$ 和$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, 则

- (1) $\bar{x}$和s^2相互独立 (用于推导t检验)
- (2) $\bar{x} \sim N(\mu, \sigma^2/n)$ (5.3中已证明)
- (3) $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$ (用于置信区间推导)

推论5.4.1 设 $x_1, x_2, \ldots, x_m$ 是来自$N(\mu_1, \sigma_1^2)$ 的样本, $y_1, y_2, \ldots, y_n$ 是来自$N(\mu_2, \sigma_2^2)$ 的样本, 且此两样本相互独立, 则有

$$F = \frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} \sim F(m-1, n-1)$$

特别地, 若$\sigma_1^2 = \sigma_2^2$, 则$F = s_x^2/s_y^2 \sim F(m-1, n-1)$

推论5.4.2 设 $x_1, x_2, \ldots, x_n$ 是来自$N(\mu, \sigma^2)$ 的样本, 其样本均值和样本方差分别为 $x = \sum_{i=1}^{n} x_i/n$ 和$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, 则有

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{s} \sim t(n-1)$$

- 小样本, σ未知时, 统计量分布.

推论5.4.3 在**推论5.4.1**的记号下, 设$\sigma_1^2 = \sigma_2^2 = \sigma^2$, 并记

$$s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} = \frac{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{m+n-2}$$

(s_w^2可认为是样本方差的加权平均)

则

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_w \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

- 用于构造两总体期望之差的置信区间

§5.5 充分统计量(complete statistics) (略)

5.5.1 充分性的概念

例5.5.1 为研究某个运动员的打靶命中率, 我们对该运动员进行测试, 观测其10次, 发现除第三, 六次未命中外, 其余8次都命中. 这样的观测结果包含了两种信息:

- (1) 打靶10次命中8次;
- (2) 2次不命中分别出现在第3次和第6次打靶上.

第二种信息对了解该运动员的命中率是没有什么帮助的. 一般地, 假设对该运动员进行n次观测, 得到 $x_1, x_2, \ldots, x_n$, 每个x_j 取值非0即1.

令 $T = x_1 + \ldots + x_n$, T为观测到的命中次数. 在这种场合仅仅记录T 不会丢失任何与命中率θ有关的信息, 统计上将这种“**样本加工不损失信息**”称为“**充分性**”.

定义5.5.1 设 $x_1, x_2, \ldots, x_n$ 是来自某个总体的样本, 总体分布函数为$F(x;\theta)$, 如果在给定T的取值后, $x_1, x_2, \ldots, x_n$ 的条件分布与θ 无关, 则统计量 $T = T(x_1, x_2, \ldots, x_n)$ 称为θ的**充分统计量**.

小结

1. 了解数理统计的基本内容, 基本概念(总体, 样本, 抽样, 简单随机样本, 参数, 统计量等).

2. 熟悉掌握常用的统计量的定义, 计算, 性质及其抽样分布

3. 理解并熟悉掌握三大抽样分布(χ^2分布, t分布, F分布)的定义, 性质, 分布形状及其之间的关系

4. 理解并熟悉掌握正态总体样本均值和样本方差的抽样分布

5. 初步了解*STATA*软件, 能够运用*STATA*软件实现数据整理, 实现正态分布, χ^2分布, t分布, F分布的分位数计算, 常用统计量的计算.

作业1

5.1课后习题： 1-3, 5

5.3课后习题： 1, 4-5, 9-10, 15-18, 23, 25

5.4课后习题： 1-3, 8-11, 19

上机实验1

1. 载入stata内置数据auto.dta, 74个车型的价格及基本配置(sysuse auto.dta). 1) 分别有放回和无放回抽取10个车型组成一个随机样本, 输出10个车型名称. 2) 分别有放回和无放回抽取10%的车型组成一个随机样本, 同时输出车型名称和其价格.

2. 以上例中汽车价格数据为例, 输出常用的统计量 (样本均值, 方差, 标准差, 偏度, 峰度, 最小和最大次序统计量).

3. 利用STATA软件验证总体分别为正态分布, 均匀分布, 指数分布时样本均值的抽样分布(示例代码见课程中心).

4. 绘制不同自由度下的χ^2分布, t分布, F分布密度函数.

5. 给定不同的α,计算不同自由度下的χ^2分布, t分布, F分布密度函数$1-\alpha$分位数.