政治经济学前沿方法论与量化分析

第二讲 Stata的基本操作与描述性统计

上课地点： 善斋306C
上课时间：周二第六大节

龙治铭
善斋307C
zhiminglong@tsinghua.edu.cn

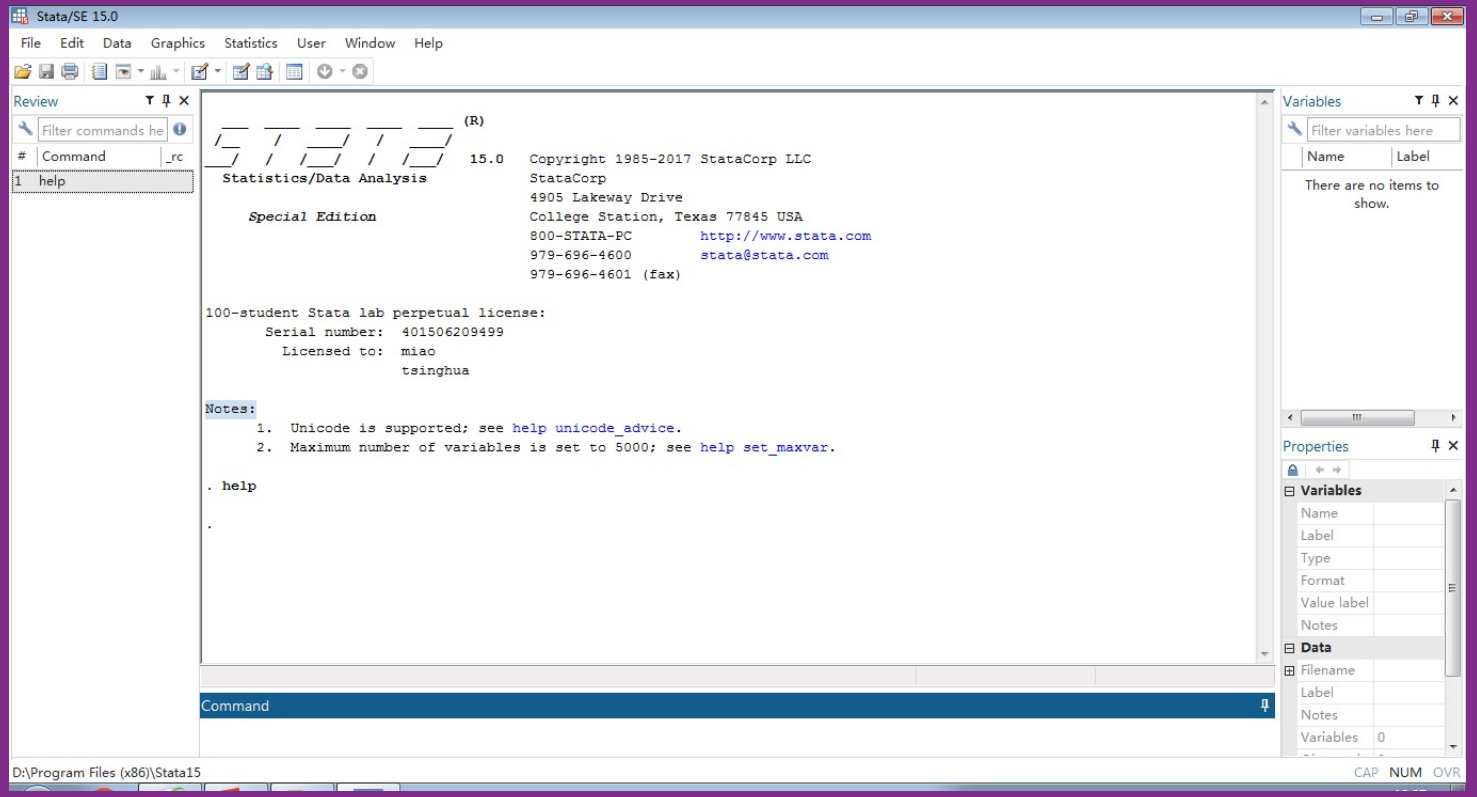
清華大學
Tsinghua University

目录

CONTENTS

1

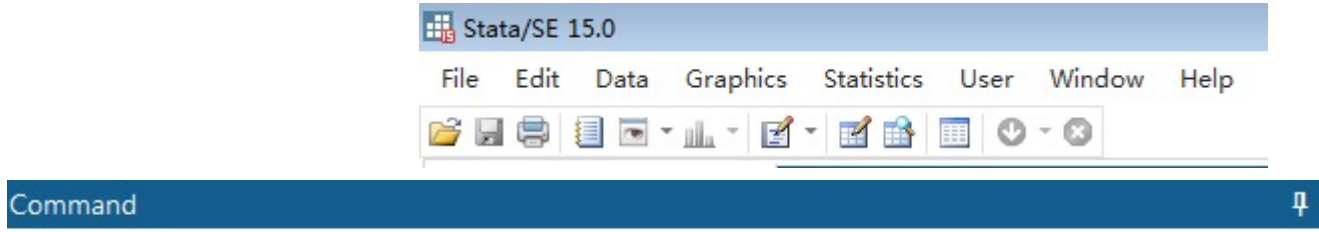
Stata的基本操作



获取&安装

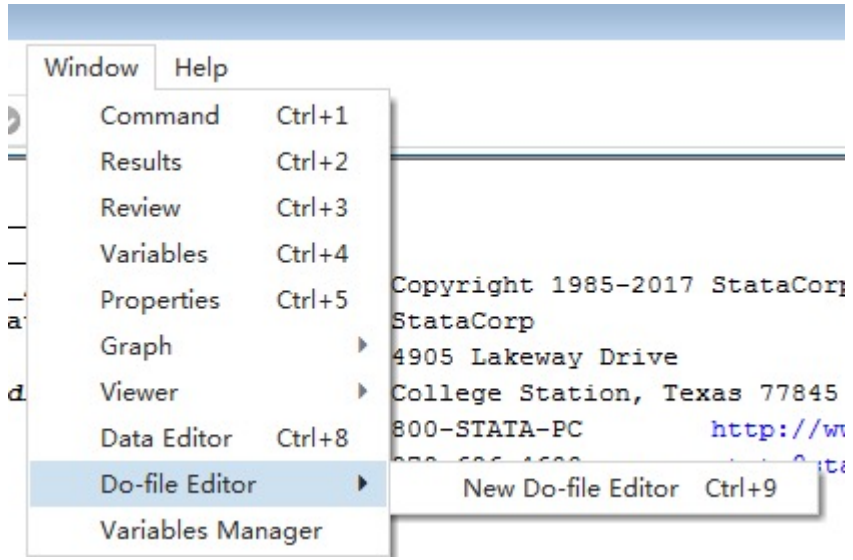
※获取和安装：见微信群聊天记录

※操作方式：菜单操作：鼠标点击
命令操作：Command栏输入命令
批量操作：do编辑器运行
只讲菜单操作

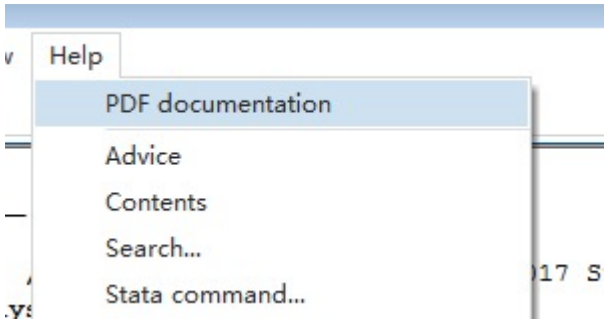


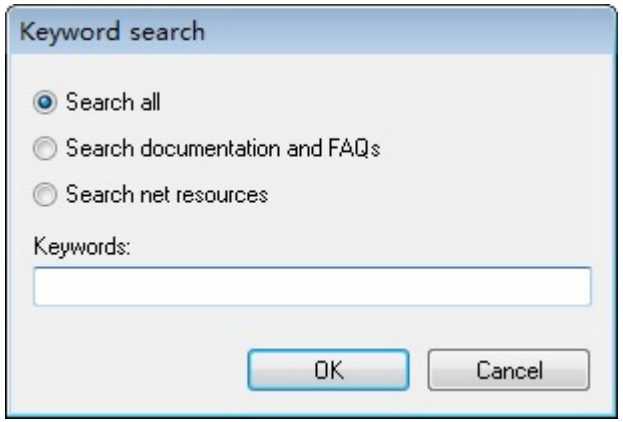
※重复实验和提供代码：菜单操作后显示结果会显示相应的命令
注意保存，以供重复、修改和审稿人审查
例子：

```
. help
```



※多使用帮助文件和用户手册

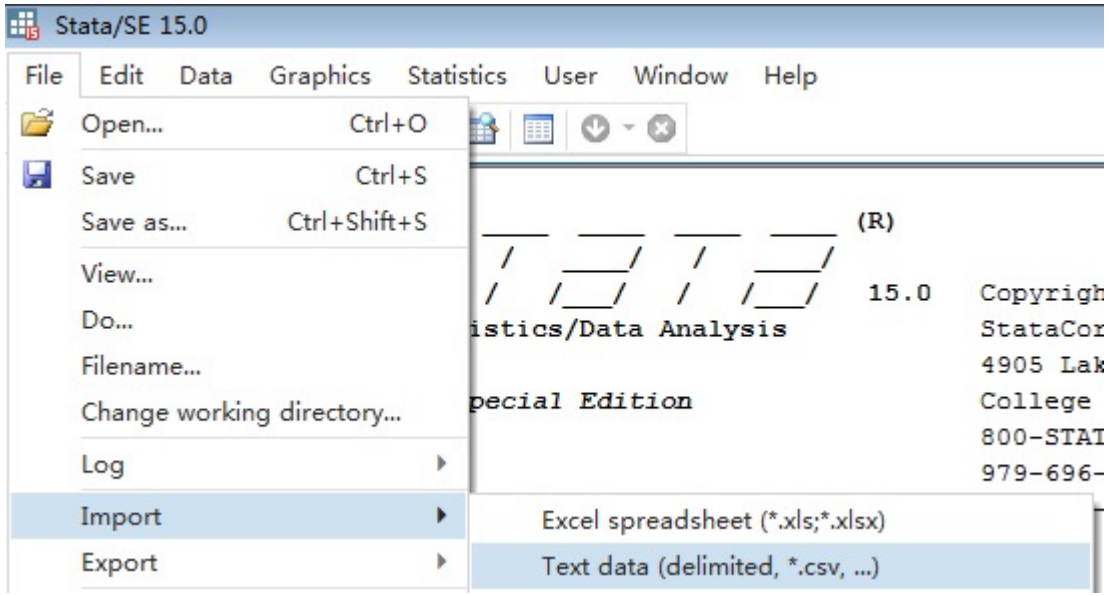


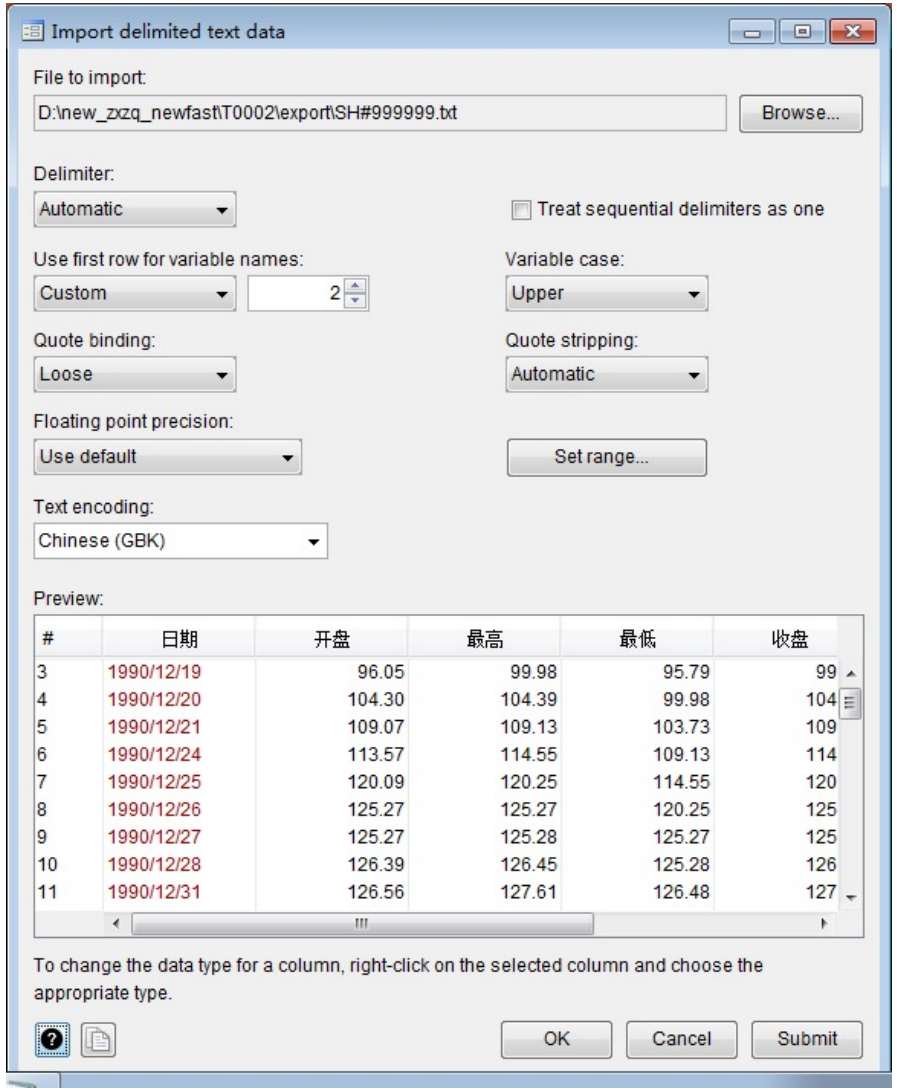


数据导入

※以导入txt文件为例 (excel等类似)
数据：上证指数20年来日线
数据来源：通达信官网

File>Import>text data





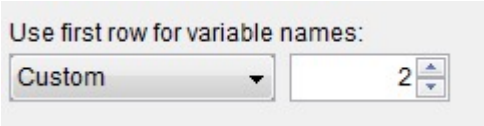
自动生成的对应代码：

```
import delimited D:\new_zxzq_newfast\T0002\export\SH#999999.txt, varnames(2) case(upper)
encoding(GBK)
```

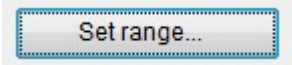
数据导入注意事项

※注意编码：含有简体中文的编码（text encoding）需要选择GBK
繁体中文BIG5 其他语言如日语、法语需要选择对应的编码

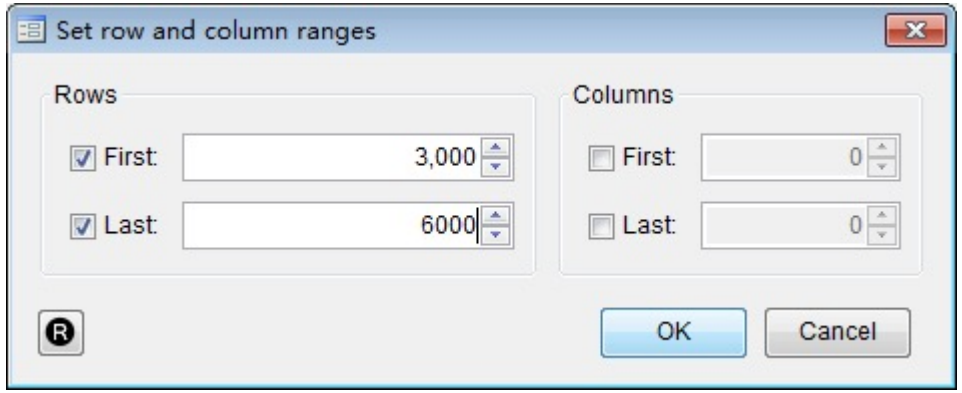
※第几行作为变量名：这个文件的第二行才是变量名称（第一行是注释 “不复权”）
选择自定义



※可以只导入特定的行/列：



例如预测明天股市可能20年前的数据没有用，只用最近的数据



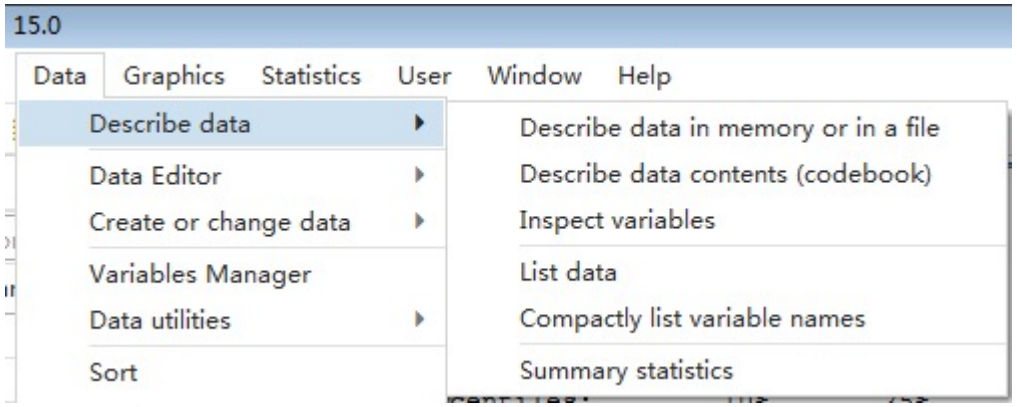
※数据清洗可以在导入前实现，这样更加简便
如：以字符形式储存还是以数字形式储存
编码的改变
清除不规则的行和列（合并的单元格）
清除注释等

※其他不常用设置见帮助手册，遇到问题时多上网搜索和查阅手册

数据导入后的查看

※导入数据后可在右侧界面查看概况

※在Data中查看具体情况

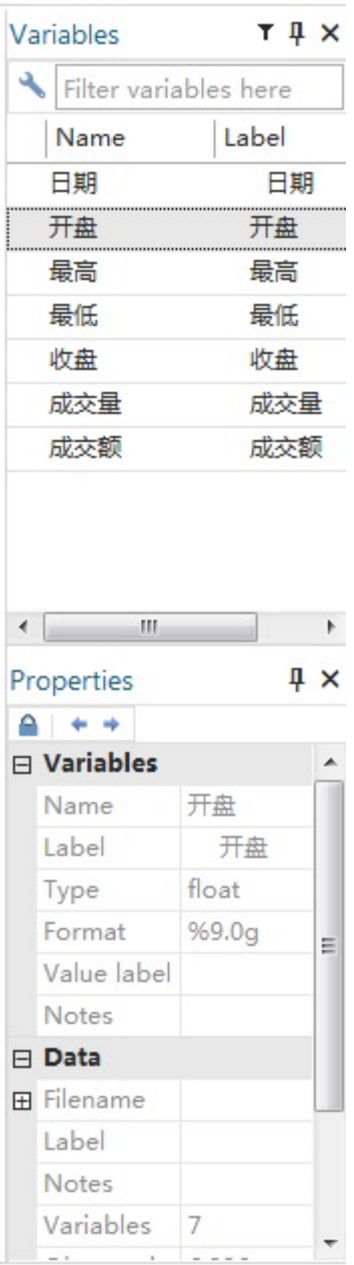


describe 日期 开盘 最高 最低 收盘 成交量 成交额
codebook 日期 开盘 最高 最低 收盘 成交量 成交额
Inspect
列出前十个观测值： list in 1/10
按字母顺序列出所有的数值型变量：ds, alpha has(type numeric)

注意：变量名较长时可以勾选 “不缩写”

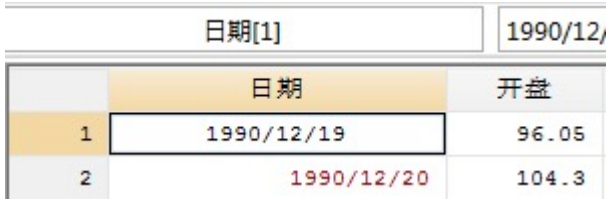
※较常用命令： summarize
会给出：
观测数Obs
均值Mean
标准差 Std.Dev （稍后讲）
最小值Min
最大值Max

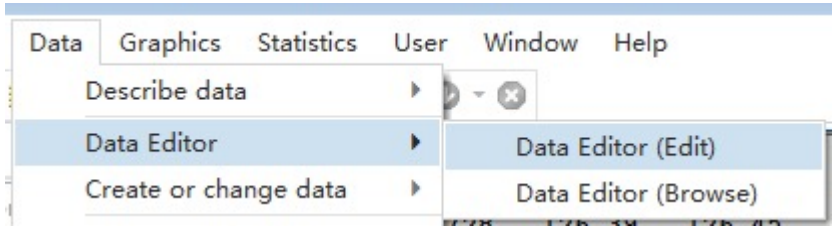
Variable	Obs	Mean	Std. Dev.	Min	Max
日期	0				
开盘	6,895	1919.441	1070.405	96.05	6057.43
最高	6,895	1938.089	1081.225	99.98	6124.04
最低	6,895	1899.295	1058.362	95.79	6040.71
收盘	6,895	1920.706	1071.772	99.98	6092.06
成交量	6,895	7.20e+07	1.03e+08	15	8.57e+08
成交额	6,895	7.58e+10	1.27e+11	6000	1.31e+12



数据的浏览与修改

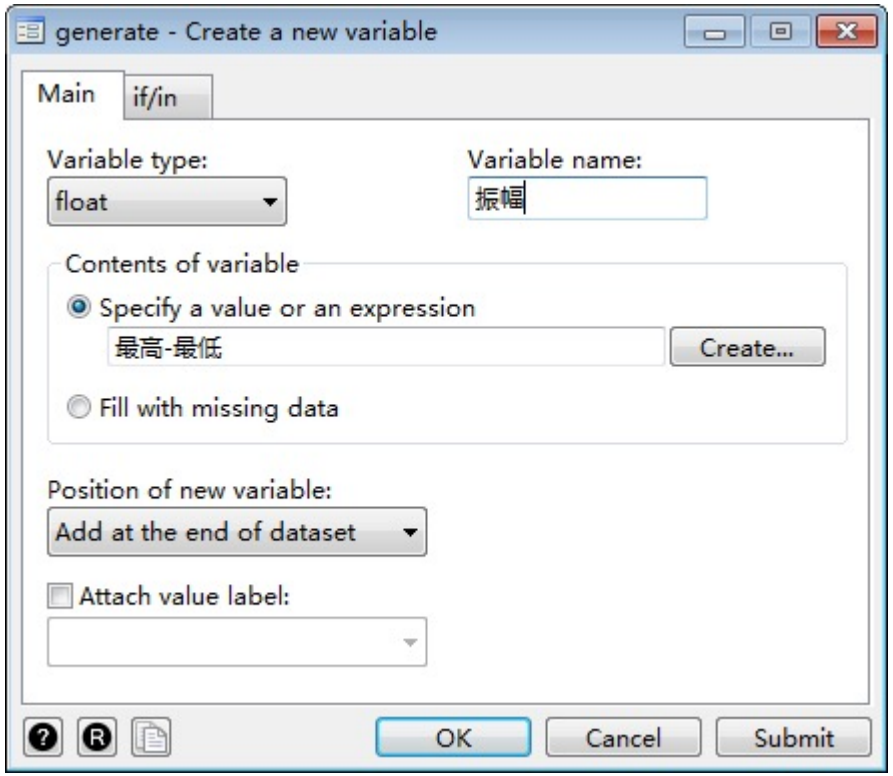
※Edit 可以编辑修改数据， Browse只能浏览





※创建新的变量， 如振幅=最高-最低

generate 振幅 = 最高-最低



※增加观测值：

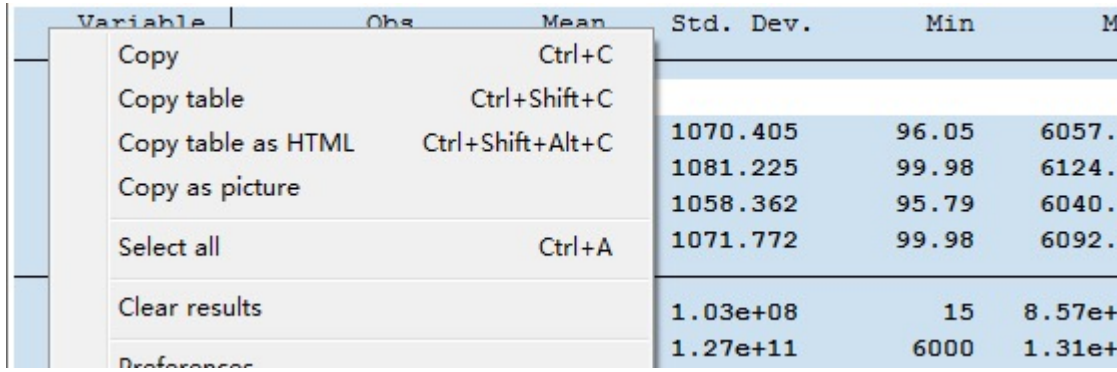
insobs 1

增加今天的信息

样本外预测

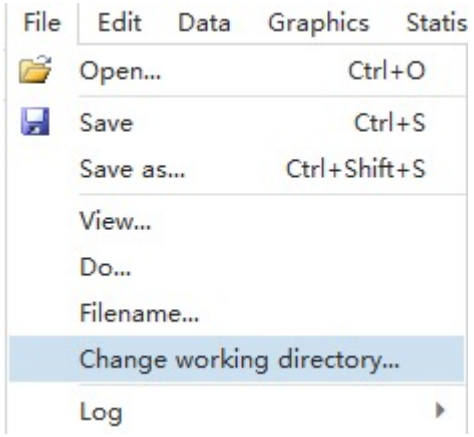
数据和结果的导出

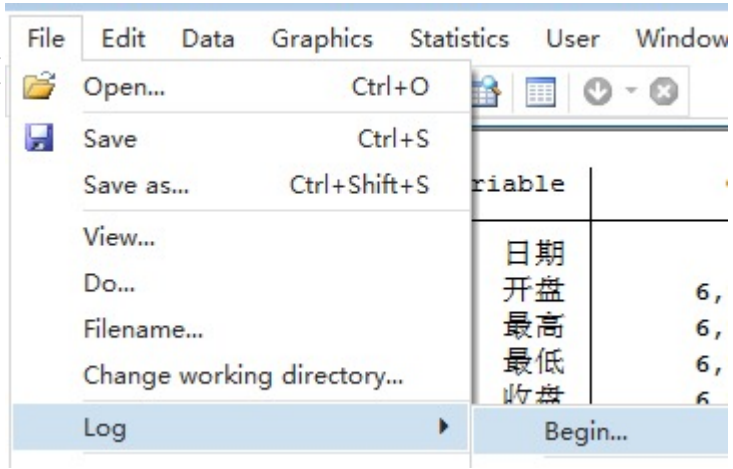
※计量结果：及时复制、粘贴保存计量结果到Word文档
注意：排列可能不是很好
可以以表格、网页以及图片的形式进行粘贴（右键）

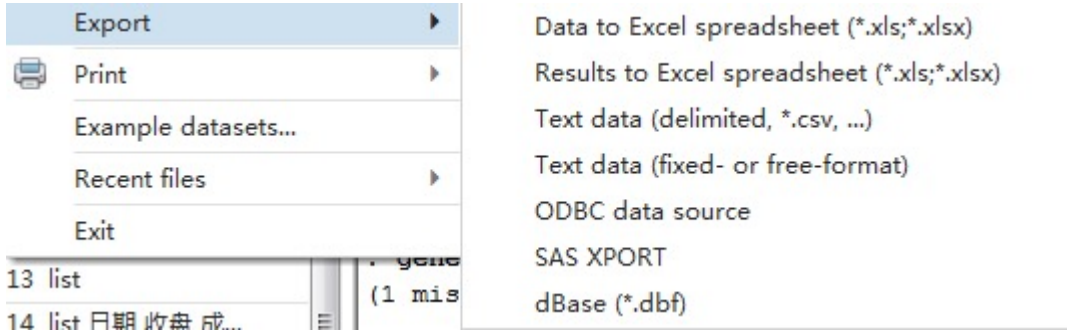


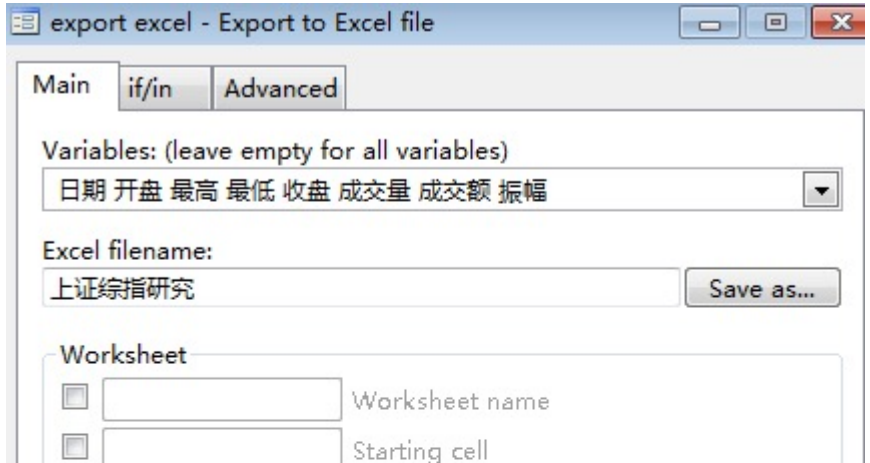
※数据的输出：
注意：结果一定要注意保存！
也可以使用log文件夹
默认smcl格式，也可以设置为文本格式（.log）

※修改默认工作文件夹：



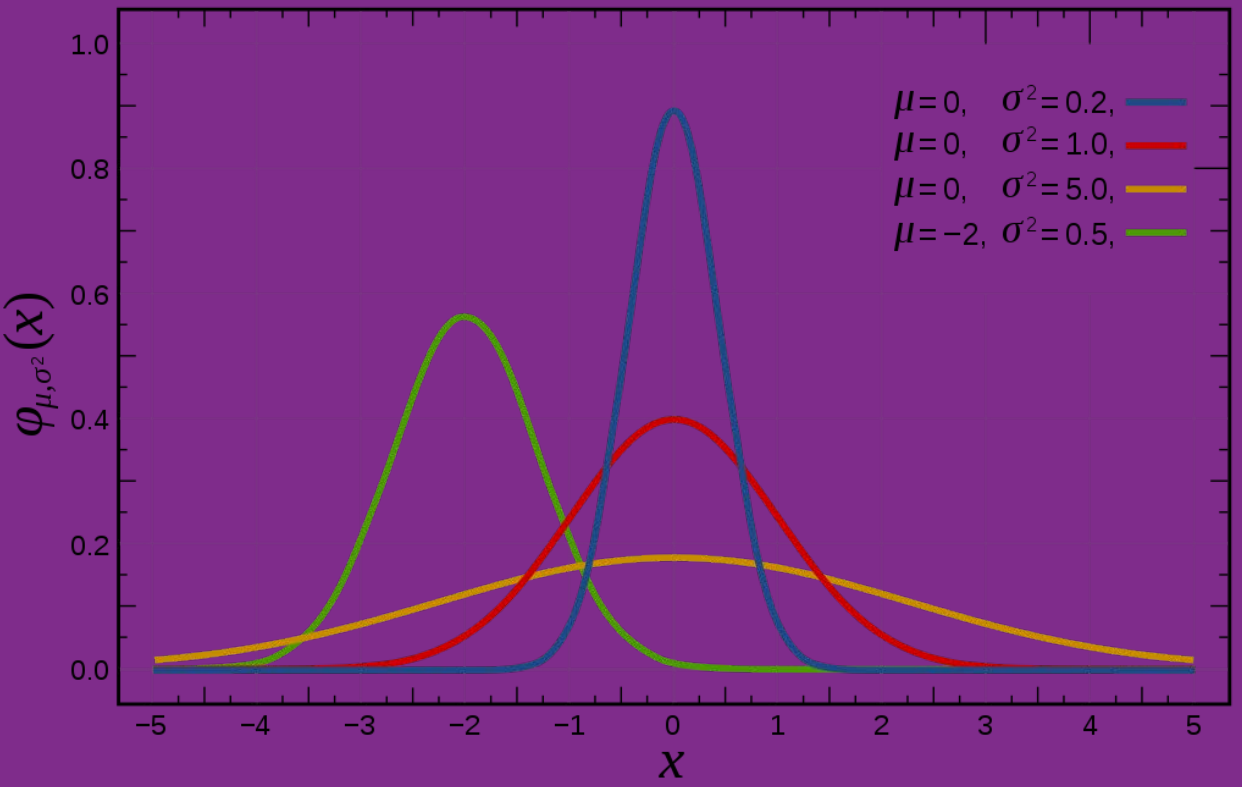






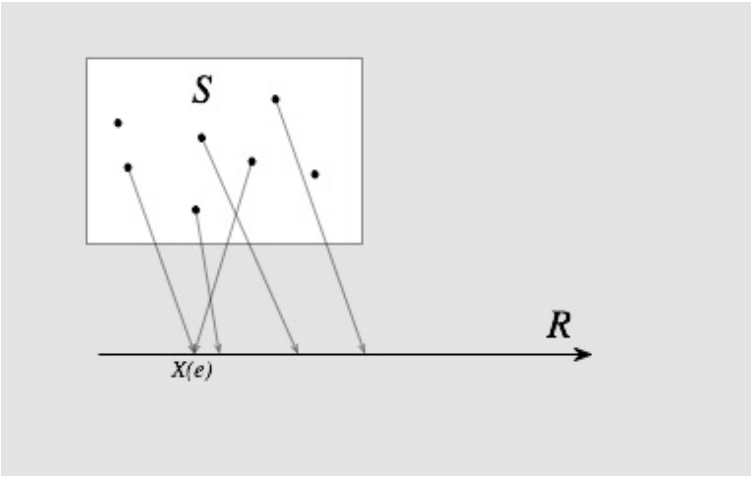
2

随机变量及其分布


1.0
0.8
0.6
0.4
0.2
0.0
$\varphi_{\mu,\sigma^2}(x)$
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
$\mu=0,\ \sigma^2=0.2,$
$\mu=0,\ \sigma^2=1.0,$
$\mu=0,\ \sigma^2=5.0,$
$\mu=-2,\ \sigma^2=0.5,$

随机变量

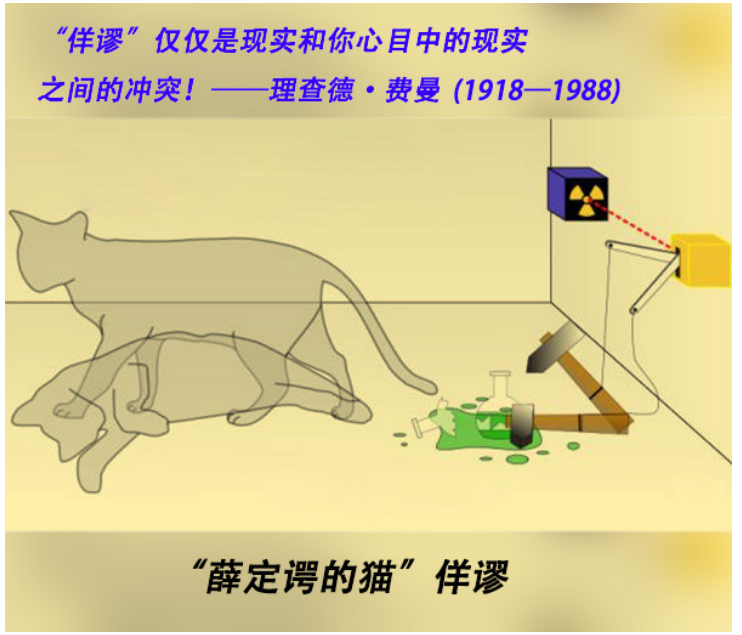
※什么叫做随机变量，跟变量有什么区别？（严格的定义需要公理集合论）
随机试验的结果e可以用数字x来表示（不用数字，难以描述和研究），这样每一个试验结果e都和一个实数x联系起来，函数X=X(e) 就是随机变量。
e是变量，X是随机变量，我们可以看到随机变量是函数



※约定：通常用大写英文字母或小写希腊字母

※数量型quantitative variable：结果e本身就是数量
百分制期末考试分数（有规律）
明天股市的收盘价（可能无规律）

※质量型qalitative variable ：结果e本身不是数量，但规定某个实数与之相联系。数值本身没有意义，因此不能用数量型变量的数字特征
等级制期末考试分数：A+、A、B+、B、B-、C+、C、C-、F
清华同学的性别：男0 女1
薛定谔猫的存活状态：生1 死0
（有人用薛定谔猫故弄玄虚证明唯心主义，错在哪儿？）



随机变量的种类

※离散型：即在一定区间内变量取值为有限个，或数值可以一一列举出来。
离散型可以是数量型也可以是质量型（注意：一一可列不一定是有限个，）
数量型：每日股市收盘价
质量型：性别

※连续型：即在一定区间内变量取值有无限个,或数值无法一一列举出来。
连续型可以是数量型也可以是质量型（注意：无限个一定无法一一列举，不讲可测度问题）
数量型：常见
质量型：光谱，陨石落入地球的坐标

与数量型/质量型的区别和联系：按照结果e的性质分和数字x的类型分
※数字特征：利用随机变量观测值（有的只可能有一个观测值，有的可以有很多观测值，想想为什么？）刻画随机变量的特征和性质，探索和总结规律。

※为什么要用数字特征和概率分布？
人文社会学的局限性：个体的特征不一定具有普遍性（数字特征进一步上升为概率和分布）。
中农办前主任陈锡文：中国农村还有5.7亿人，判断乡村情况要靠科学统计而不是返乡故事！

随机变量的分布

※概率（Probability）：随机事件发生的可能性，是一个在0到1之间的实数，一般用p表示。
随机，不代表没有规律，只是取值x随试验的结果e而定，而实验的各个结果都有一定的概率出现。
大量重复性试验或观察中所呈现出的固有规律，就是所谓的**统计性规律**。
确定性现象：国庆节会放假
随机现象：赌博（概率论起源于赌博：卡当诺、帕斯卡和费马）

※累积分布函数（Cumulative distribution function, CDF）,简称分布函数：

$$F_X(a) = \mathbb{P}(X \leq a)$$

随机变量取值小于等于a的概率，a为任意实数。
CDF刻画了不同结果出现的可能性分布情况。

※离散型随机变量的分布：
如果X的取值只有$x_1 < x_2 < \cdots < x_n$,概率分别为P(x_i)≥ 0 ,i=1,2,…n,两条性质

$$F_X(x_i) = \sum_{j=1}^{i} P(x_j)$$

$$\sum_{k=1}^{n} P(x_k) = 1$$

Girolamo Cardano1501－1576
预言自己的死亡

重要的离散型随机变量分布

※ 0 – 1 分布（Bernoulli distribution）

伯努利试验结果只可能由两个结果，若试验成功，则随机变量取值为1，若试验失败，则随机变量取值为0，分布律为（数字特征下一节讲）：

X	0	1
p_k	$1-p$	p

※二项分布（Binomial distribution）：

伯努利试验中事件 A 发生概率为 p，将伯努利试验重复 n 次，事件 A 发生的次数 X 是一个随机变量，分布律为：

$$P\{X=k\}=\binom{n}{k}p^k q^{n-k}, k=0,1,2,\cdots,n.$$

恰好是二项式定理展开的第 k 项（有放回抽样是二项式分布，无放回抽样是超几何分布）。

※泊松分布（Poisson distribution）：

小概率事件在一定时间内的发生次数通常符合泊松分布

1）事件是小概率事件

2）事件相互独立不会影响

3）事件的概率是稳定的记为 p

$$P\{X=k\}=\frac{\lambda^k e^{-\lambda}}{k!}, k=0,1,2,\cdots,$$

泊松分布可以作为二项式分布的近似

波斯公主选驸马的问题

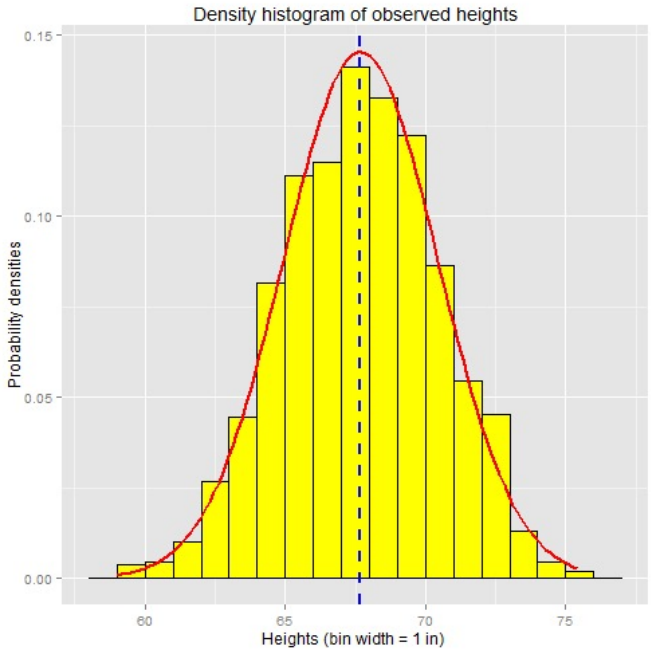
连续型随机变量分布

※概率密度：

连续型随机变量的分布律无法一一列出因此需要引入概率密度的概念。

$$f(x)=\frac{\mathrm{d}\,F(x)}{\mathrm{d}\,x}$$

密度函数f（x）是分布函数的微分，分布函数是密度函数的积分，也就是说，在密度函数下方和横轴围起来的面积才是概率



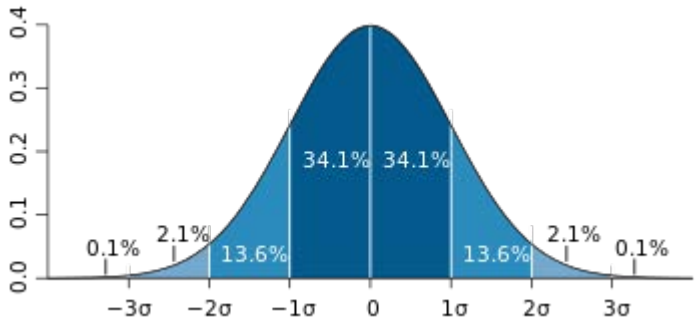
※性质及注释：

1）概率密度不是概率，面积才是概率。

2）对于连续型随机变量，取值为特定值的概率为0，但不是不可能概率为0事件≠不可能事件

3）样本的频率分布函数收敛于密度函数，稍后讲样本

4）f（x）≥ 0



最重要的分布：正态分布

※正态分布（normal distribution）又叫高斯分布（Gaussian distribution）是宇宙中最重要的分布函数

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

记为N（μ,σ^2）　（数字特征和参数意义稍后解释）

标准正态分布N（0，1）：

$$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

※正态分布的意义：

无穷多、相互独立事件影响的综合结果，

一般服从正态分布。中心极限定理：

大量统计独立的随机变量的平均值的分布趋于正态分布

比如小明同学，高三的考试几次成绩分数如下：580分,600分,680分,620分；清华的分数线大概在690分。乍一看，肯定没戏了，最高分都不都够清华的线。

但是，利用高斯分布这个方法表明，小明能考上清北的概率还是有1.5%的！

北京人大附中网红数学教师：假定学生成绩服从高斯分布，成绩不好考上清华的概率有1.5%，错在哪里？

建模不是用了数学就是对的，而是用对了数学才对的

※学生分布t分布

N 个独立同分布的正态分布的均值服从学生分布，记为t(n)，

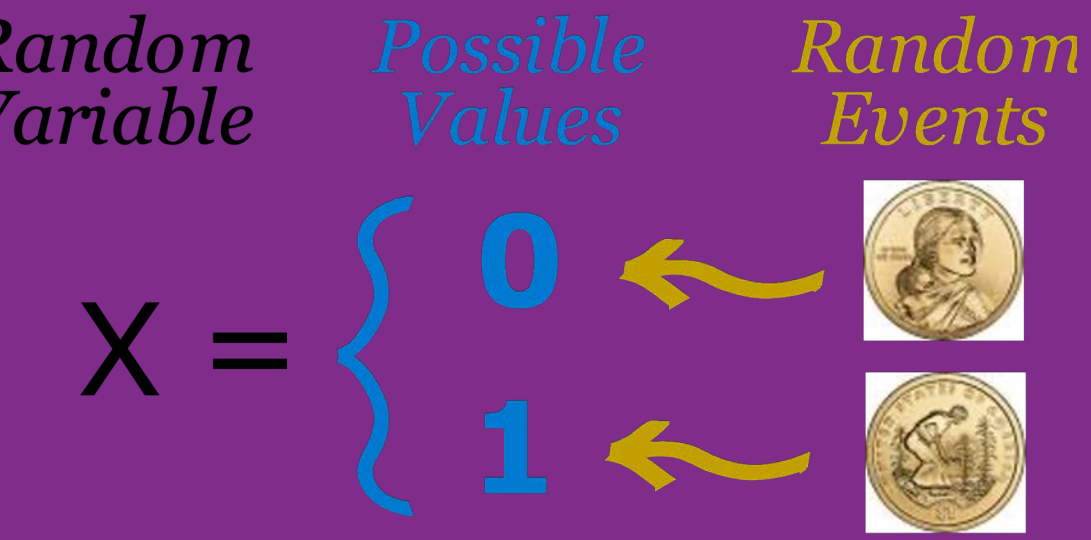
n为自由度。密度函数较复杂，从略。

学生分布是统计检验的基础，n较大时，趋向于标准正态分布。

$$\overline{X}_n=\frac{X_1+\cdots+X_n}{n}$$

3

随机变量的数字特征



平均数Mean

※算术平均Arithmetic mean (AM)；随机变量一组观测值的算术平均是样本均值（sample mean，一般记为$\bar{x}$）

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

※目的：刻画描述总体均值（population mean，一般记为μ）
样本均值不可能恰好是总体均值，抽样次数越多、样本量越大，样本均值就越接近总体均值
（原因：大数定理和中心极限定理，不讲）

※只可能有一个观测值的随机变量怎么办？
如年末GDP，只可能有国家统计局一个数据，一次观测实现。
遍历各态性假设Ergoticity：利用时间平均代替集平均（不讲，见我的一篇Working paper）
时间序列建模的基本隐含假设（通常被忽略）：时间足够长，过去发生的事情在未来还会发生
后凯恩斯主义经济学的批判：时间序列的非遍历各态性
马克思主义：事物否定之否定规律
新古典经济学建模：成立，因为经济最终趋向于稳态。大部分人既不懂也不关心。
物理学/工程学：一般直接假设成立，否则没法展开研究。
统计学家：所有的模型都是错的！
折衷办法：假定平稳的时间序列是遍历各态的（事实上二者不同）

平均数的延展（1）

※加权平均（weighted average）

$$\text{加权平均数}=\frac{x_1 f_1+x_2 f_2+\cdots+x_n f_n}{n}$$

其中：$f_1+f_2+\cdots+f_k=n$ f1，f2，…，fk叫做权重。

在一组观测值中，有的观测值可能更重要，我们赋予更高的权重。

例如：习近平总书记强调破除“唯分数”，我们在评特奖时可以将分数、社会实践、公共服务等各项得分赋予不同权重

又例：没有成交额的价格是虚假的，按照交易额作为权重调整后的指数更能反映市场的真实状况。

※几何平均（Geometric mean，GM））

$$\bar{x}=\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}=\left(x_1 x_2 \cdots x_n\right)^{\frac{1}{n}}$$

算术平均有时不能真实反映随机变量的特征。例如，一国GDP在受到金融危机冲击时增长较低，随后恢复过程又迅猛增长，此时算术平均值会夸大增长结果（因为AM>GM）。

如：新闻报道通常说中国改革开放四十年实现了平均8%以上的增长，现在增速下降到6.5%左右，是不是意味着经济衰退？早年经济基础薄弱，增长快，现在体量大，增长绝对额高，增速下降并不意味着经济衰退，这种情况几何平均优于算术平均。

又如：贷款利率（校园贷陷阱）

平均数的延展（2）

※移动平均（moving average，MA））
n 为过去期数

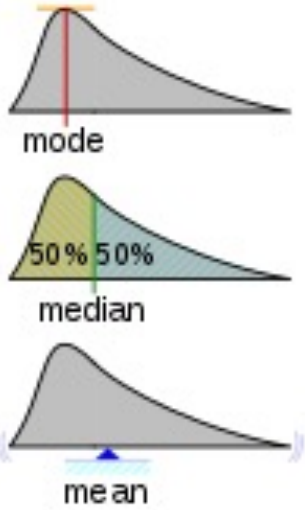
$$\bar{p}_{\mathrm{SM}} = \frac{p_M + p_{M-1} + \cdots + p_{M-(n-1)}}{n} = \frac{1}{n}\sum_{i=0}^{n-1} p_{M-i}$$

移动平均可以反映时间序列的趋势或周期，抚平波动（ARMA模型中MA部分），本质上为卷积（不讲）
扩展：中心移动平均（用于处理季节性波动，见我的硕士论文）和加权移动平均

※中位数（Median）：将数值集合划分为相等的上下两部分，中间的数即为中位数。
平均数的局限性：马云有2000亿，我有1万，平均我们有1000亿
对于收入分配而言，中位数的意义比平均数更大，平均数掩盖了贫富差异。
有限数集：从大到小排列，样本量为奇数，一个中位数；样本量为偶数，两个中位数或取平均值
无限数集：使得各自分布函数占一半（1/2）的数

※众数（Mode）：一组数据中出现次数最多的数据值。
离散变量：出现次数最多的数
连续变量：概率密度最大的数

※意义：反映了总体或随机变量有关集中趋势的重要资讯。
在对称的单峰分布（正态分布）中，均值、众数、中位数均相等。



数学期望Mathematical Expectation

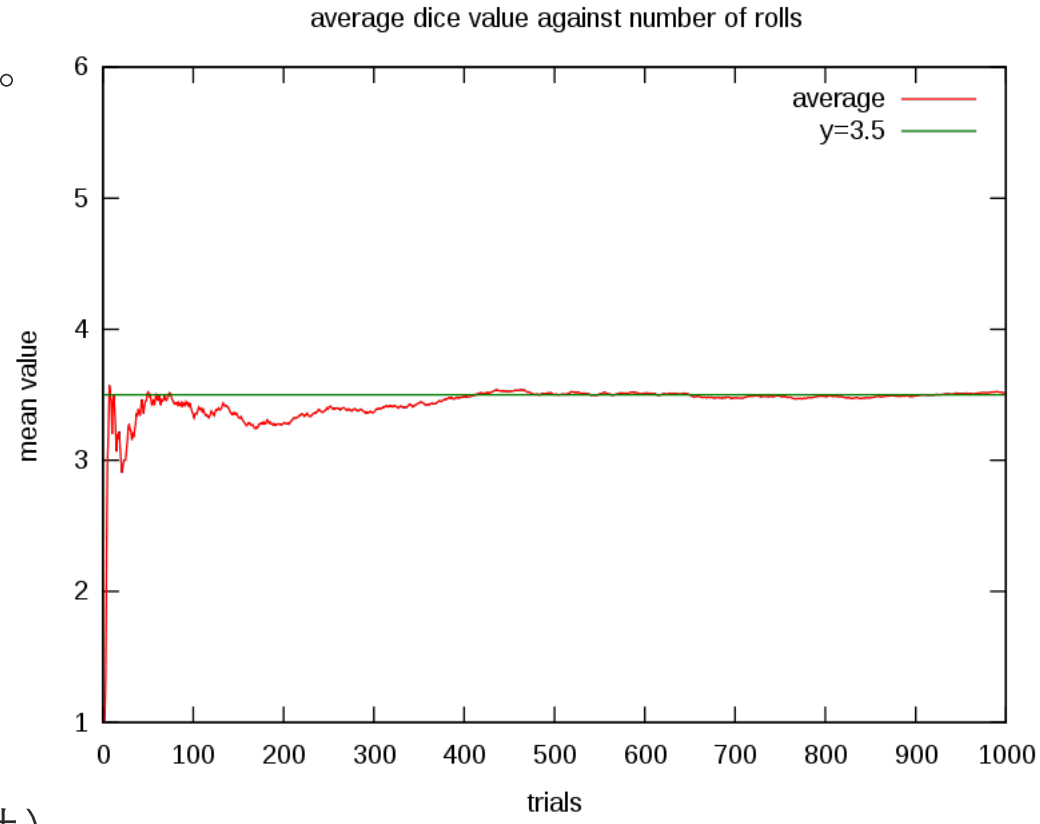
※数学期望μ与均值$\bar{x}$ ：
均值$\bar{x}$反映了样本的平均集中程度，是对数学期望μ的估计。

离散数学：
$$\mathrm{E}(X) = \sum_i p_i x_i$$

连续数学：
$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

掷骰子的数学期望：

$$\mathrm{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$
$$= \frac{1+2+3+4+5+6}{6} = 3.5$$



※各种分布的数学期望（更详细的总结见数学基础知识附件）

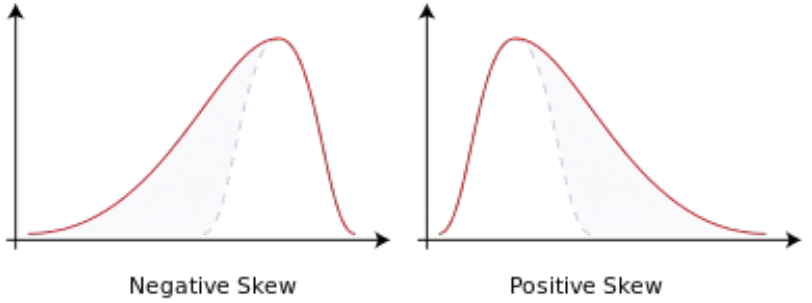
分布	正态分布	0-1分布	二项分布	泊松分布	均匀分布	指数分布
期望	μ	P(1-p)	np	λ	$\frac{a+b}{2}$	$\frac{1}{\lambda}$

方差Variance

※方差：描述的是随机变量的离散程度，也就是该变量离其期望值的距离。

$$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right]$$

正态分布
方差越大变量值的分布的越分散，方差越小，越集中在均值附近

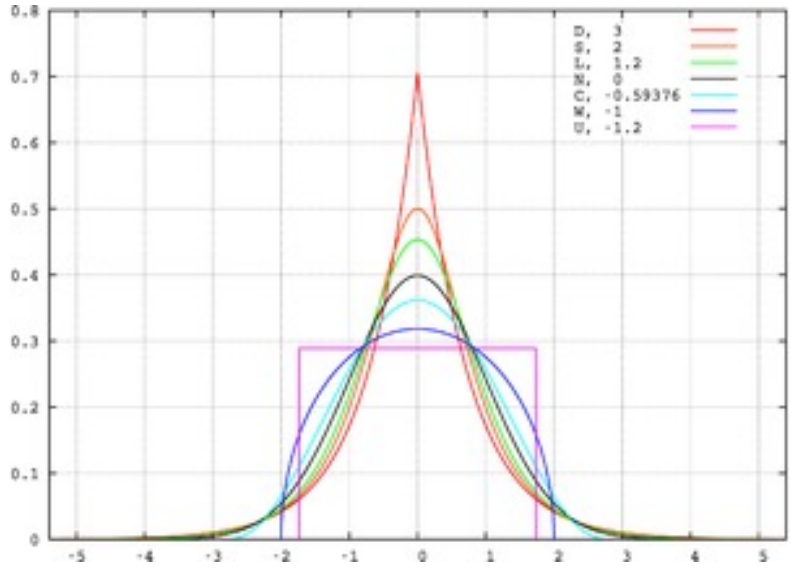


※标准差（Standard Deviation）通常记为σ
标准差是方差的平方根
为什么要有标准差? 单位一致。

※高阶矩：矩估计是一种重要的估计方法

$$\mu'_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, dx$$

N=1: 期望
N=2：方差
N=3: 偏度（Skewness）：分布不对称
负偏态或左偏态：左侧的尾部更长，分布的主体集中在右侧。
正偏态或右偏态：右侧的尾部更长，分布的主体集中在左侧。
N=4：峰度 （Kurtosis）：小概率事件或方差有变化（方差的方差）
峰度高就意味着方差增大是由低频度的大于或小于平均值的极端差值引起的。



协方差和相关系数

※协方差（Covariance）：用于衡量两个变量的总体误差。而方差是特殊的协方差

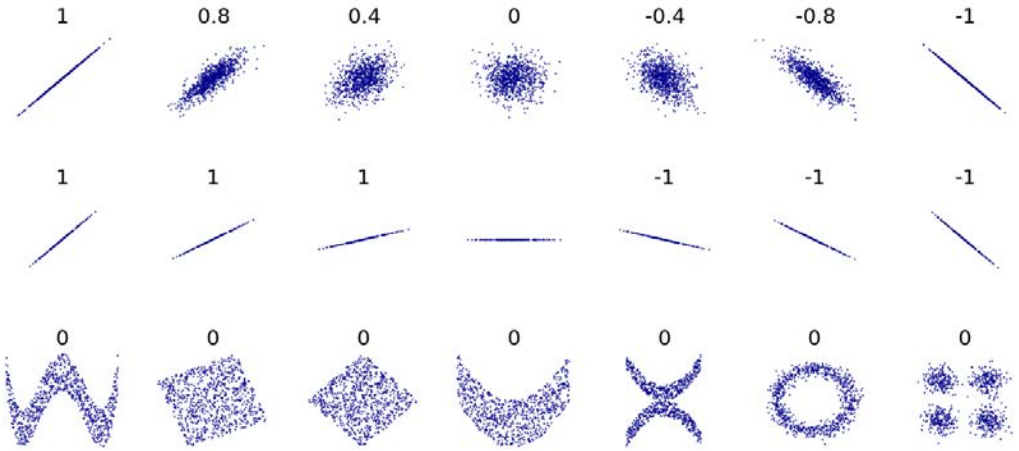
$$\operatorname{cov}(X,Y) = \operatorname{E}((X-\mu)(Y-\nu))$$

相互独立，协方差为0,；协方差为0，不一定相互独立

※线性相关系数（Linear Correlation coefficiient）：

$$\eta = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\cdot\operatorname{var}(Y)}},$$

线性回归的本质：最小二乘法估计量（OLS estimator）本质上是线性相关系数
因此线性回归的局限性也很清楚了。



※协方差矩阵：
多个变量（或两个向量随机变量）之间的协方差
用于探索多元线性回归各变量之间的关系

$$= \begin{bmatrix} \operatorname{E}[(X_1-\mu_1)(X_1-\mu_1)] & \operatorname{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \operatorname{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \operatorname{E}[(X_2-\mu_2)(X_1-\mu_1)] & \operatorname{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \operatorname{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{E}[(X_n-\mu_n)(X_1-\mu_1)] & \operatorname{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \operatorname{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}$$

4

参考文献


站在巨人的肩膀上
"If I have seen further it is by standing on the shoulders of Giants. “
by Isaac Newton in 1675

参考文献&数学基础知识&样本数据

见网络学堂附件

下节课见

马克思主义学院
龙治铭

清華大學
Tsinghua University