

政治经济学前沿方法论与量化分析

第三讲 推断性统计的初步

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



清华大学
Tsinghua University

目录

CONTENTS



描述性统计在Stata中的实现



抽样与初步的推断性统计



初步推断性统计在Stata中的实现



参考文献

1

描述性统计在Stata中的实现

统计工作不是把数字随便填到几个格格里去，而应当是用数字来说明所研究的现象在实际生活中已经充分呈现出来或正在呈现出来的各种社会类型。

——列宁

※本节介绍如何在stata中实现数量型变量的描述性统计，质量型变量的描述性统计在逻辑回归中介绍。

※上节课通达信的数据文件，时间是以字符形式储存的，需要转化为数值形式（stata处理时间序列需要先指定时间变量，字符形式的变量不能被指定为时间变量），需要使用以下命令（无法通过菜单操作时间）或者先在excel中转变好格式（数据清洗最好在导入之前就完成）：

```
gen date=date(日期, "YMD" /* YMD表示原变量“日期”是年月日格式 */
format date %td /* 用format将数值型变量date定义为月日年格式，更详细的格式使用help datetime 查看帮助文件*/
```

※注释：/* 注释*/ 在代码中加入注释，以免时间过长搞忘是什么意思

※时间序列定义时间变量：

```
statistics>time series >
```

```
setup and utilities>
```

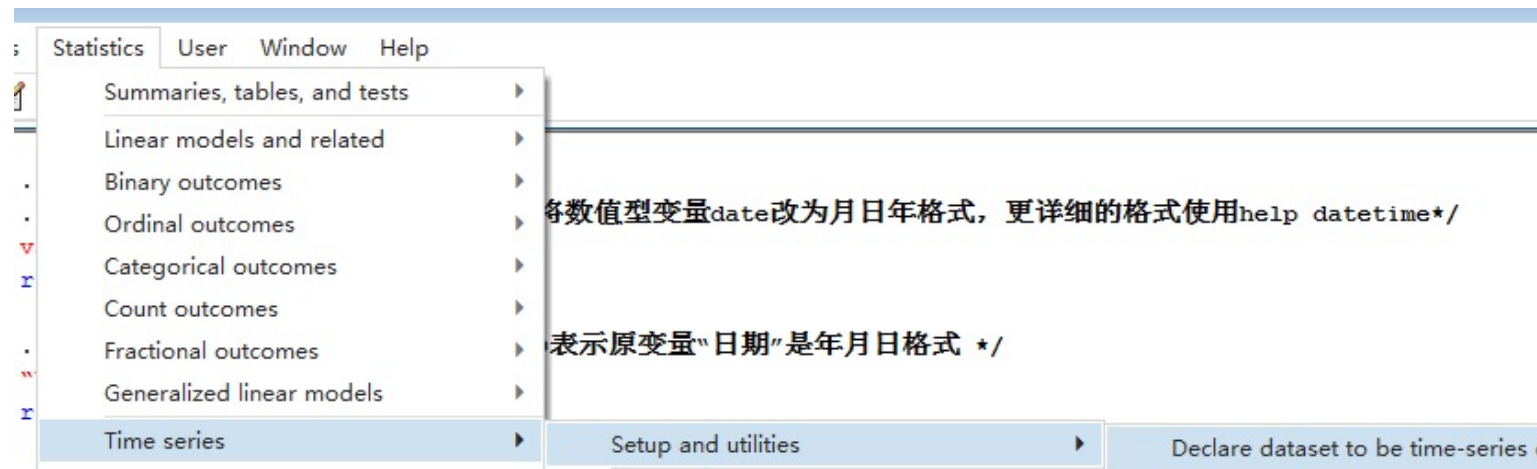
```
Declare dataset to be time-series data
```

```
Time variable 下拉选date
```

```
或者使用代码：tsset date
```

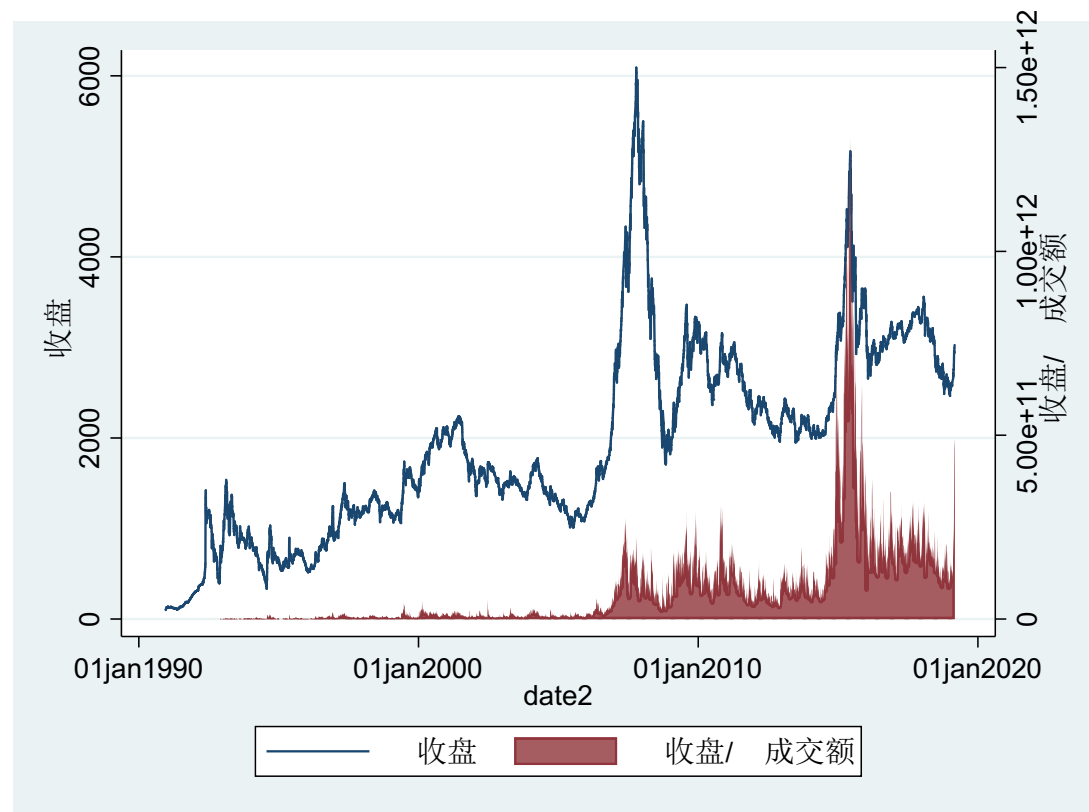
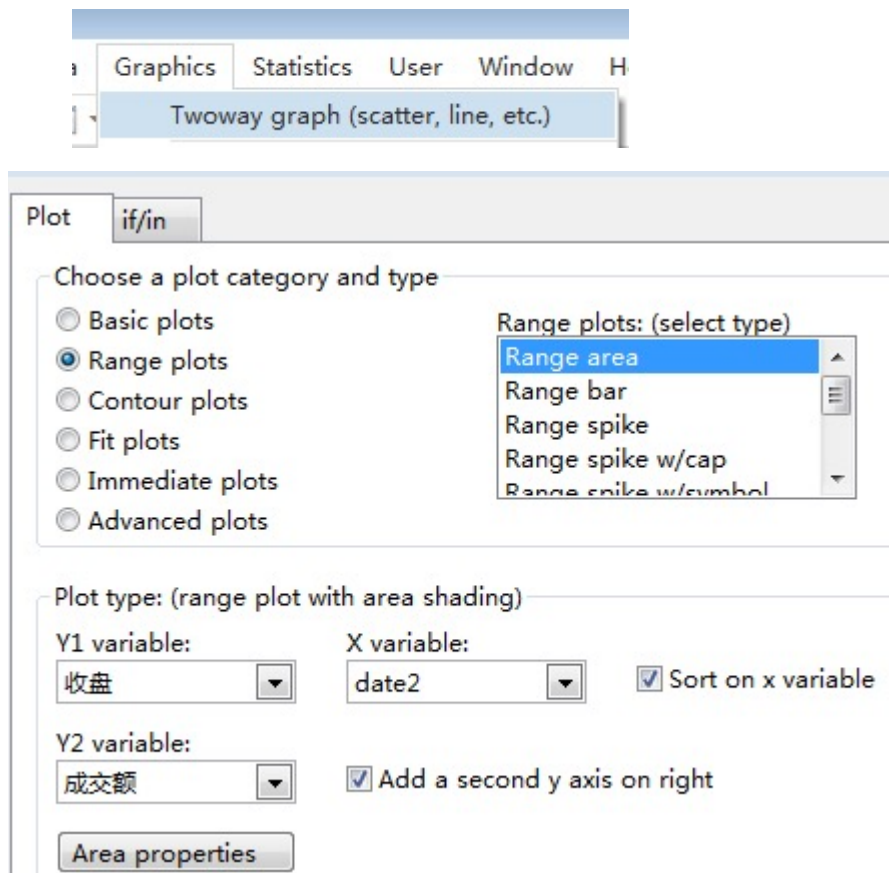
Variables	
Name	日期
Label	日期
Type	str22
Format	%22s
Value label	
Notes	

Variables	
Name	date
Label	
Type	float
Format	%9.0g
Value label	
Notes	



※描述性统计（Descriptive statistics）的意义：通过观察大量重复试验的结果，发现可能的统计规律

※画图是发现统计规律的第一步和最好的办法：直观
Graphics>two-way graph



通过观察图像，我们发现如下规律：成交额和价格走势高度相关

※我们想知道，股市是否存在规律？每天的涨跌幅度由什么决定？

假设1：股市的涨跌幅没有规律，由无穷多相互独立的事件决定，因此涨幅服从正态分布

假设2：股市的涨跌幅有规律，股市是经济的晴雨表，由经济基本面、流动性和风险偏好等决定，因此涨幅不是正态分布

※首先定义涨幅=当日收盘价/上日收盘价-1 (增长率) 或者涨幅= $\Delta \log(\text{收盘价})$

generate 涨幅 = 收盘/收盘[_n-1]-1 /*或者分两步*/

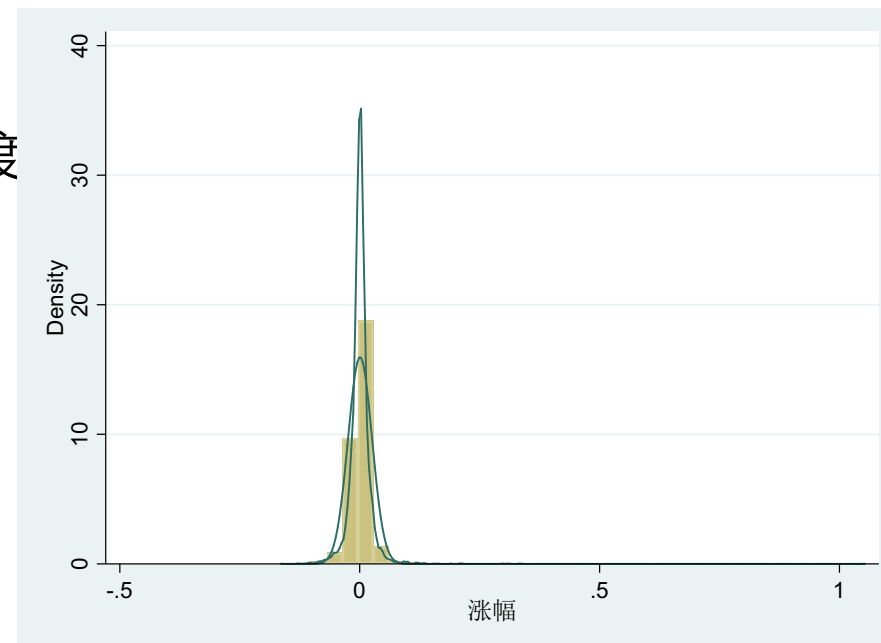
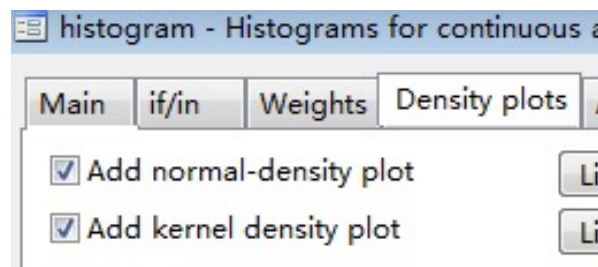
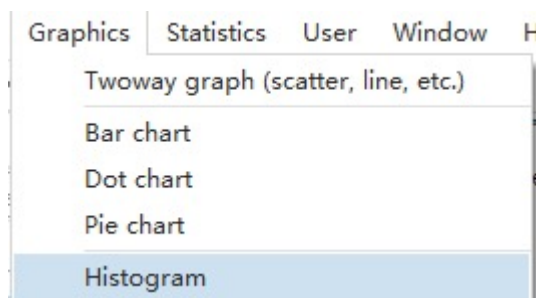
gen l收盘 = log(收盘)

gen g涨幅2 = 100*D.l收盘 (注意：谨慎使用L. D. F. 因为节假日周末不开市，会造成缺失值)

※绘制涨幅的直方图

Graphics>Histogram 在选项Density plots中勾选正态密度和核密度

结论：离正态分布很远，股市的涨跌可能存在某种规律



※我们想知涨幅的大概轮廓是怎样的呢？

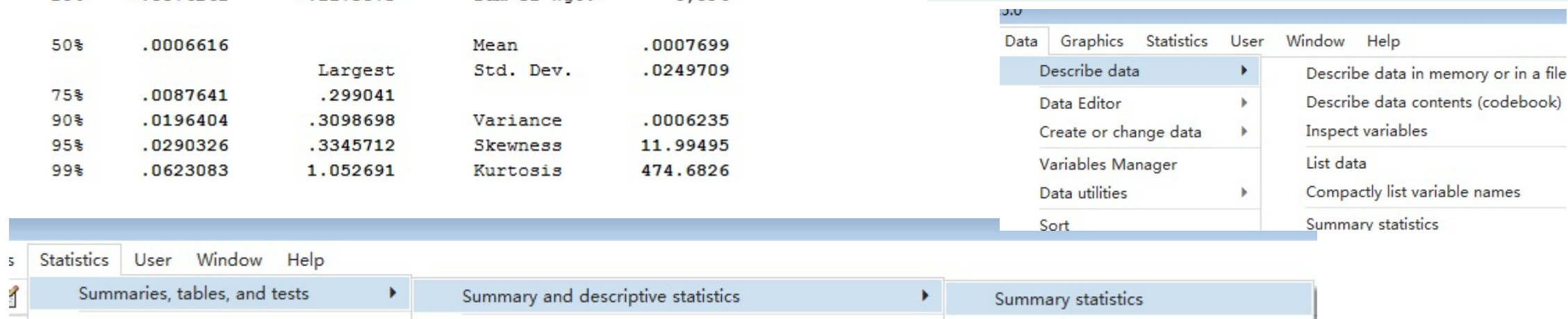
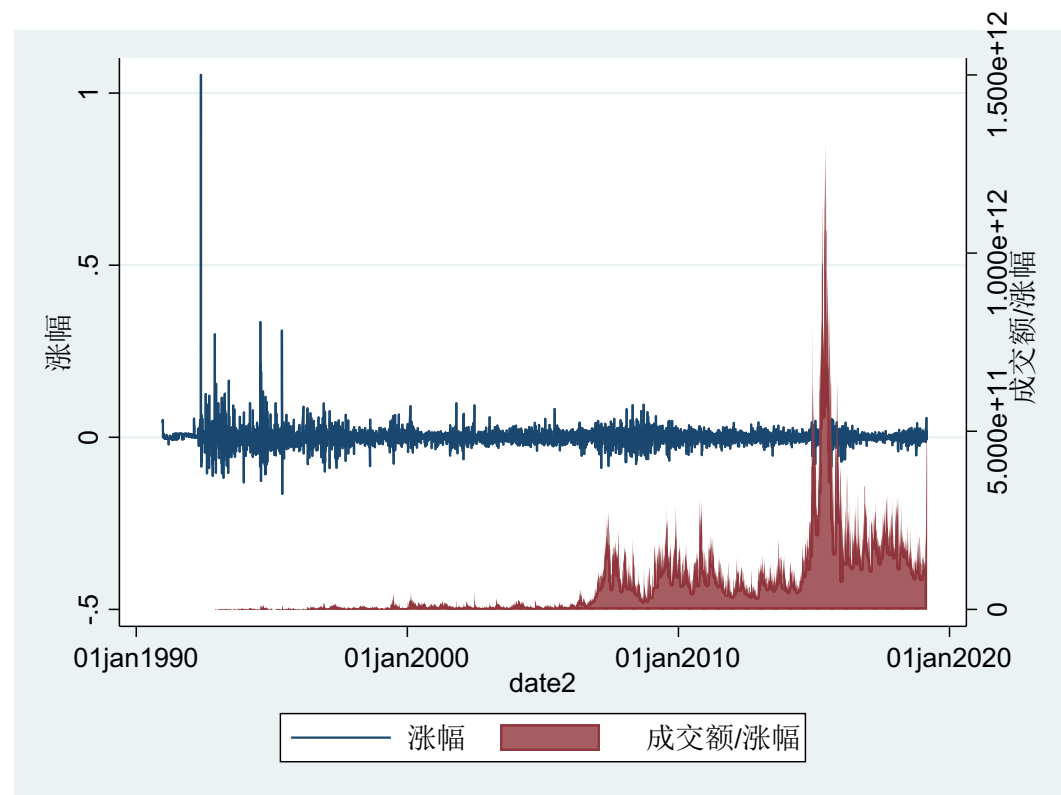
图像显示：

涨幅在0附近波动（数学期望接近于0），
波动较为剧烈（方差很大），
波动随时间变化（峰度较高）

※在stata中很容易实现准确的数字特征

```
. summarize 涨幅, detail
```

涨幅				
Percentiles		Smallest		
1%	-.0622251	-.1639366	Obs	6,894
5%	-.0289717	-.1307641		
10%	-.0190777	-.1267491		
25%	-.0074241	-.1175073		
50%	.0006616		Mean	.0007699
75%	.0087641	.299041	Std. Dev.	.0249709
90%	.0196404	.3098698	Variance	.0006235
95%	.0290326	.3345712	Skewness	11.99495
99%	.0623083	1.052691	Kurtosis	474.6826
		Largest	Sum of Wgt.	6,894



※我们想知涨幅的大概轮廓是怎样的呢？

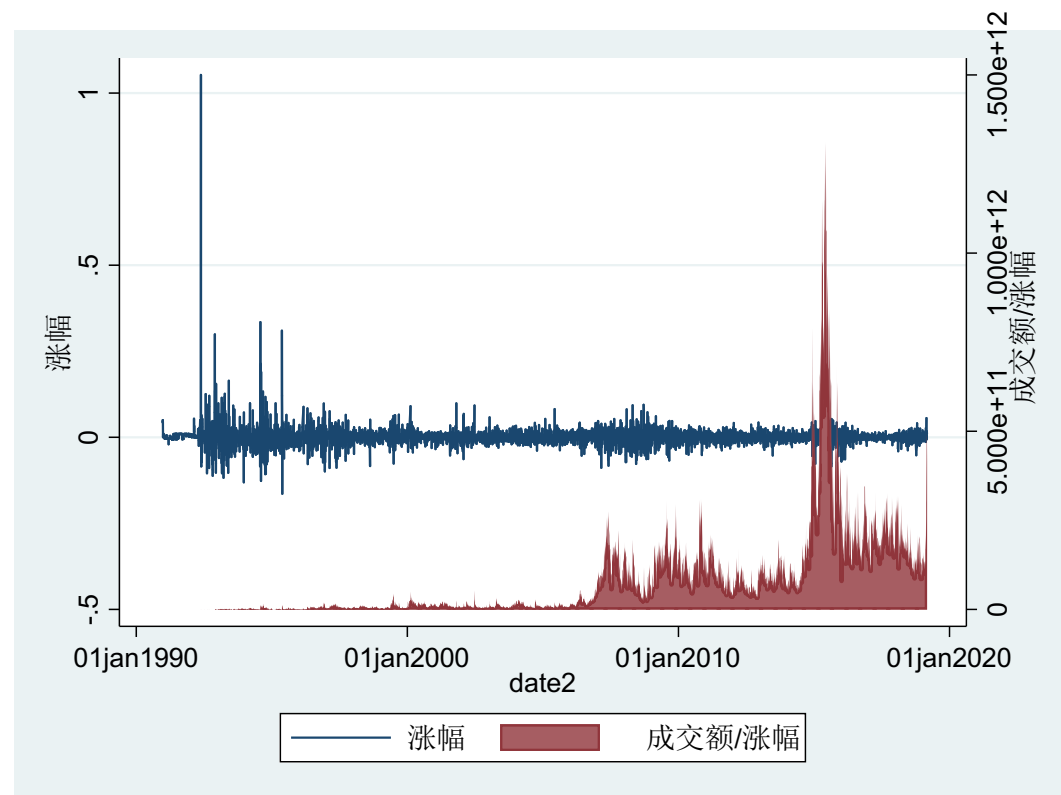
图像显示：

涨幅在0附近波动（数学期望接近于0），
波动较为剧烈（方差很大），
波动随时间变化（峰度较高）

※在stata中很容易实现准确的数字特征

```
. summarize 涨幅, detail
```

涨幅				
	Percentiles	Smallest		
1%	-.0622251	-.1639366		
5%	-.0289717	-.1307641		
10%	-.0190777	-.1267491	Obs	6,894
25%	-.0074241	-.1175073	Sum of Wgt.	6,894
50%	.0006616		Mean	.0007699
		Largest	Std. Dev.	.0249709
75%	.0087641	.299041		
90%	.0196404	.3098698	Variance	.0006235
95%	.0290326	.3345712	Skewness	11.99495
99%	.0623083	1.052691	Kurtosis	474.6826



期望：略微为正，长期慢涨。方差：很小，分布集中（涨跌停限制）偏度：为正，右偏态，集中分布在左侧，牛短熊长。
峰度：极高，离群值很多，小概率事件频发。

A股大概轮廓：长期慢涨，波动集中，牛短熊长，频繁爆雷

※我们通过对涨幅的描述性统计分析，大概了解了A股市场特点。但是主要指标之间的关系是怎样的呢？会不会存在某种相关性？

※计算多个变量之间的协方差矩阵（多元线性回归的基础）

```
. correlate 开盘 最高 最低 收盘 成交量 成交额  
(obs=6,895)
```

	开盘	最高	最低	收盘	成交量	成交额
开盘	1.0000					
最高	0.9998	1.0000				
最低	0.9996	0.9995	1.0000			
收盘	0.9994	0.9997	0.9997	1.0000		
成交量	0.6933	0.6959	0.6923	0.6949	1.0000	
成交额	0.6864	0.6888	0.6848	0.6876	0.9751	1.0000

我们发现：

1) 不同价格（成交量与成交额）之间相关系数过高，几乎接近于1，表明几乎是共线的（多重共线性，下一讲解释）因此不适合相互解释。

2) 成交量和价格相关性较高，可能可以相互解释

（内生性问题，下一讲线性回归）。

可能的理论基础：供给和需求（Paul Samuelson：You can make even a parrot into a learned economist; all it must learn are the two words, "supply" and "demand".）

Summary statistics

Means

Proportions

Ratios

Totals

Pairwise comparisons of means

Confidence intervals

Normal mean CI calculator

Poisson mean CI calculator

Proportion CI calculator

Variance CI calculator

Standard deviation CI calculator

Correlations and covariances

correlate - Display correlation matrix or covariance ...

Main by/if/in Weights Options

☒ Display means, std. dev., min, and max with matrix

☐ Ignore display format associated with variables

☒ Display covariances

☐ Allow wide matrices to wrap

2

抽样与初步的推断性统计

在终极的分析中，
一切知识都是历史；
在抽象的意义下，
一切科学都是数学；
在理性的世界里，
所有的判断都是统
计学。

——C.R.Rao 《统计
与真理》

※描述性统计的局限性：

1) 描述性统计对随机变量的描绘仍然是较为模糊的，如无法回答随机变量是正态分布的可靠性具体是多少（例如通过计算一个随机变量的样本观测值，我们得到偏度为0.2，峰度为2.9，那随机变量究竟服不服从正态分布呢？如果声称它服从正态分布，我们有太多风险犯错误？）

2) 描述性统计对某些总体的描述是存在偏差的，如无限总体或总体个数很大(不可能将所有观测值都列出来)或试验不可能重复，我们需要知道通过样本为总体描述的“肖像”，在多大风险上是错误的？

※从描述性统计到推断性统计（statistical inference），对于无法一一观测的总体（例如，不可能将50岁以上的广东人的肺都打开来看一看），我们只能通过样本来推断总体的特征，并作出概率形式的推断（我们的假设在多大风险上是错误的）

钟南山：数据显示50岁以上广州人肺脏呈黑色

<http://www.sina.com.cn> 2008年06月13日00:48 金羊网-新快报



中国工程院院士钟南山指出，因吸入污染物过多，广州人一旦超过50岁，肺部就变成了黑色

钟南山院士（2008）声称：无论是有病还是没病，50岁以上的广州人肺都是黑色的”。

钟南山院士错在哪里？

※抽样：样本容量越大，抽样次数越多，样本的分布就越接近总体（理论基础：大数定理和中心极限定理）

※简单随机抽样：每个样本单位被抽中的概率相等，样本单位相互独立

※幸存者偏差：每个样本单位被抽中的概率不相等，样本不能反映总体的真实情况。

Sampling bias：以偏概全

经济学例子：库兹涅茨曲线

※解决办法：

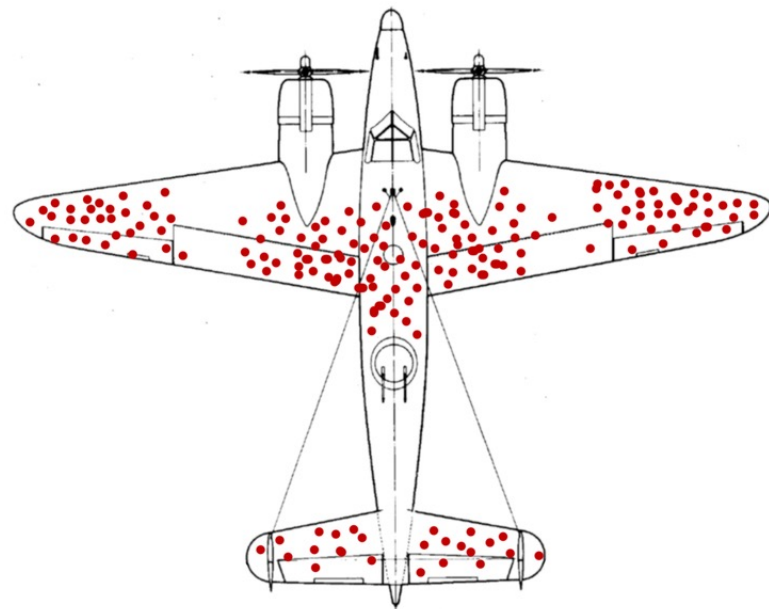
1) 正确的理解总体的性质和可能的样本偏差

例：去年龙老师马原课上有400名同学，其中男生约70%，女生约30%，某同学进行了大学生抖音使用时间的问卷调查，得到100人样本，男女各50人，女生沉迷于抖音的时间是否显著地比男生高？

2) 找到合适的解决办法，设计抽样方法

加权权重：男生权重70/50，女生权重30/50

分层抽样：博一的同学和博三是不一样的，先分层，后独立抽样



2018年全国卷2的作文：“二战”期间，为了加强对战机的防护，英美军方调查了作战后幸存飞机上弹痕的分布，决定哪里弹痕多就加强哪里。然而统计学家沃德力排众议，指出更应该注意弹痕少的部位，因为这些部位受到重创的战机，很难有机会返航，而这部分数据被忽略了。

※假设我们已经得到一个无抽样误差的样本，如何通过样本了解总体的性质（如期望、方差、概率密度等。）呢？质量如何？可靠性又是多少呢？

※常用的估计方法（不展开讲）：

方法	思想	备注
最小二乘法OLS	误差最小	GMM特殊形式
最大似然估计MLE	概率最大	GMM特殊形式
广义矩估计GMM	随机变量的矩决定了它的分布，样本矩依概率收敛于总体矩	

※估计的质量：

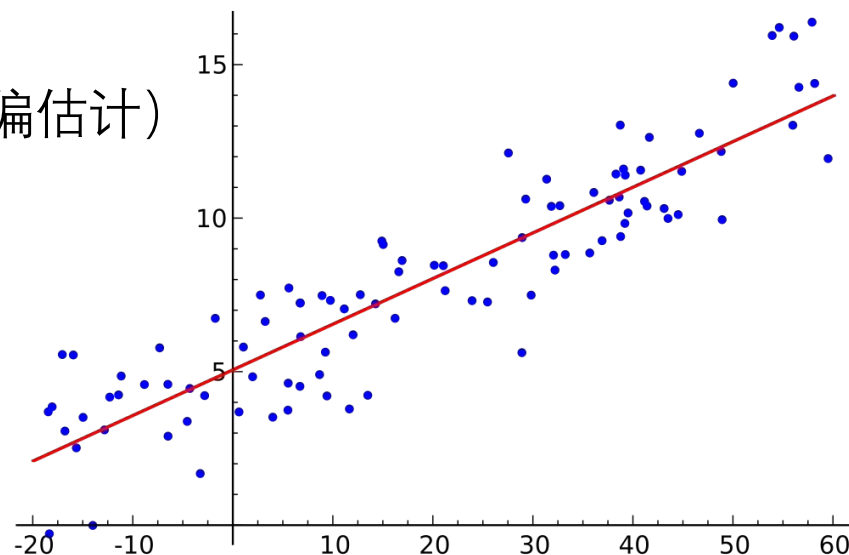
无偏性（unbiased estimator）无系统性偏差（注意样本方差的无偏估计）

有效性（efficient estimator）方差更小

一致性（相合性consistent estimator）依概率收敛

※高斯—马尔可夫定理（Gauss-Markov Theorem）：

在理想情况下，最小二乘法估计量是最佳线性无偏估计
(BLUE estimator, Best Linear unbiased estimator)



※估计的可靠性：即使得到了一个无偏的一致估计量（方差最小通常较为困难），这也是远远不够的，我们还需要知道它究竟在多大程度上是可靠的。

※ $1-\alpha$ 置信区间：参数的估计量有 $1-\alpha$ 的概率在某个区间内。

一般来说，估计量落入95%的置信区间表明估计比较可靠。

※单侧置信区间：参数的估计量有 $1-\alpha$ 的概率大于（或小于）某个值

※正态分布的“ 3σ ”原则：

$\mu \pm \sigma$: 68.27%

$\mu \pm 2\sigma$: 95.45%

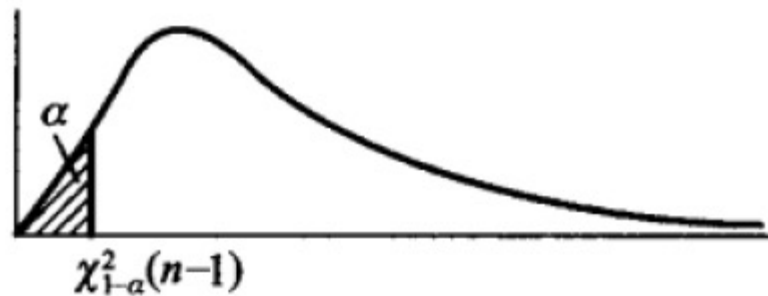
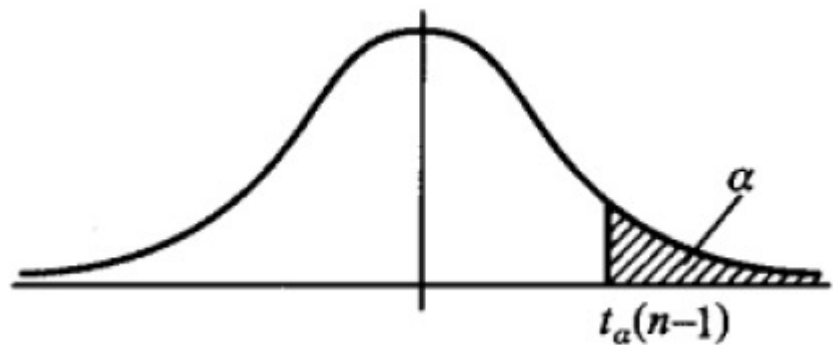
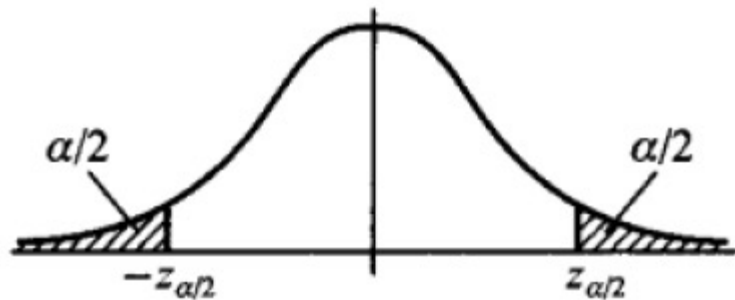
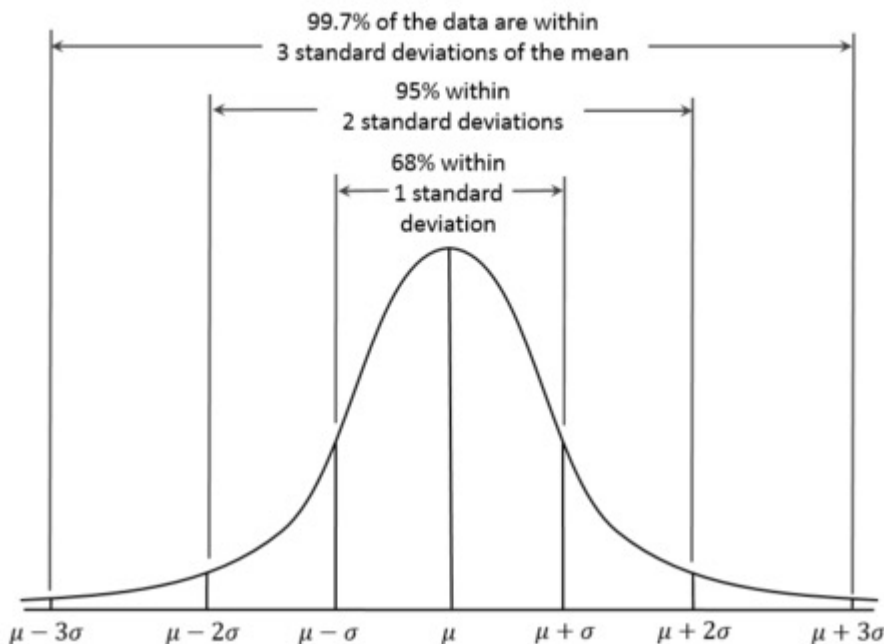
$\mu \pm 3\sigma$: 99.73%

※重要的 α 分位点：

90%置信区间：1.64

95%置信区间：1.96

更多的见标准正态分布表



参数估计——标准正态分布表

※标准正态分布表：

阴影部分表示随机变量取值小于x的概率。

第一列+第一行的值是X的取值
表中数据对应该x取值的阴影部分面积

※使用：

90%置信区间，两侧空白各5%，需要找使得阴影部分面积为95%的x
表中1.5这一行都小于0.95，而1.7这一行都大于0.95，因此在1.6这一行。第一个最接近0.95的是0.04这一列。因此90%置信区间对应的是1.64.

同理95%置信区间对应1.96

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

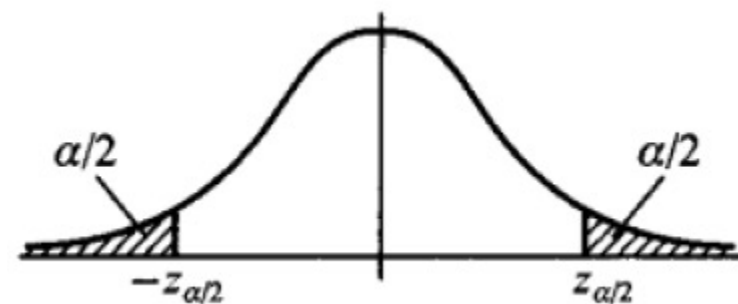


x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

※ 总体的参数或者分布未知，我们做出了一定的合理假设（如假定价格是一个随机游走，那么价格的变动，即涨跌额度是一个白噪音），我们基于对样本的观察，做出拒绝或者接受假设的决策，叫做假设检验。

※ 第一类错误和第二类错误：但是样本具有随机性，我们可能做出错误的判断。

假设检验的两类错误		
真实情况 (未知)	所作决策	
	接受 H_0	拒绝 H_0
H_0 为真	正确	犯第 I 类错误
H_0 不真	犯第 II 类错误	正确



※ 原则：谨慎性原则-错误后果严重的作为零假设（某种药品是否为真、某个经济政策变量是否有效）
显著性检验：样本容量无法增加的时候，控制犯第一类错误的概率（即 α 很小），使

$$P\{\text{当 } H_0 \text{ 为真拒绝 } H_0\} \leq \alpha$$

称为显著性检验。

零假设处于拒绝域或者P值较大时，不能拒绝零假设。

※什么是p值: 简单地讲, 就是拒绝零假设犯错误的概率。

P值越小, 表明拒绝原假设所要冒的风险越小; p值越大, 表明不能拒绝原假设。

注意: 很多统计教材说, p值大, 原假设成立, 这是错误的! 准确的说法是“我们不能拒绝H0”

※ p值的滥用与经济学的帝国主义扩张: 一篇没有p值的经济学论文是几乎无法被发表的, 甚至一篇没有p值政治学、历史学、社会学论文也几乎无法发表。

争论的关键: 非常重要的经济、社会政策是否应该仅仅建立在p值之上?

例: 分数每提高1分, 学生评教分数显著上升2%。

李克强总理: GDP增长1个百分点能拉动150万人就业

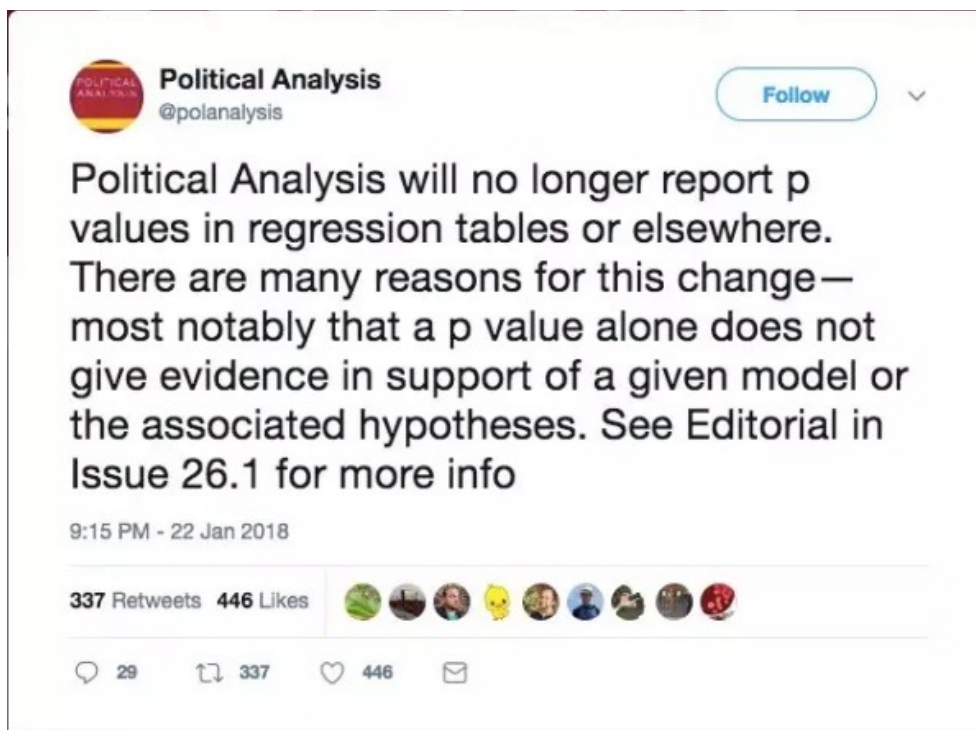
菲利普曲线

美国政治学顶级学术期刊《政治分析》在他们的官方twitter上宣布从2018年的开始的第26辑起禁用p值。

美国统计学会的回应

Ronald L. Wasserstein & Nicole A. Lazar (2016):

- 1) 对p值的误解和滥用
- 2) 建议取代p值的其他方法



※这6条原则包括：

1. P-values can indicate how incompatible the data are with a specified statistical model. P值可以表示数据与一个特定的统计模型是否相容。
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. P值不能代表研究假设为真的概率，也不代表数据完全是由随机因素造成的概率。
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. 科研结论、商业决定和政策制定不能完全凭P是否小于一个特定的值来决定。
4. Proper inference requires full reporting and transparency. 正确的推理需要全面的报告和透明度。
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result. 一个P值，或者显著性，不能表示一个效应的大小，或者一个结果的重要性。
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. P值本身不能作为判断一个模型或假设的良好量度。单独的P值只能提供有限信息。

※建议取代p值的其他方法：

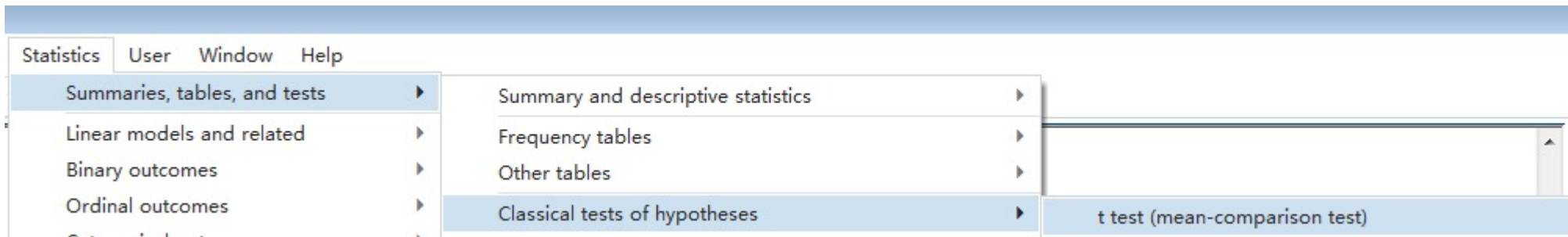
- Confidence, credibility, or prediction intervals
- Bayesian methods
- Alternative measures of evidence, such as likelihood ratios or Bayes Factors
- Other approaches such as decision-theoretic modeling and false discovery rates

3

初步推断性统计在Stata中的实现

※ 检验A股涨幅长期是否为0：A股十年前的指数和现在一样，似乎长期来看没有投资价值？

均值的检验服从学生分布（证明从略），操作方法：statistics>classical tests of hypotheses>t test (mean)



Statistics User Window Help

Summaries, tables, and tests ▶ Summary and descriptive statistics ▶

Linear models and related ▶ Frequency tables ▶

Binary outcomes ▶ Other tables ▶

Ordinal outcomes ▶ Classical tests of hypotheses ▶ t test (mean-comparison test)

Main by/if/in

t tests

☒ One-sample

☐ Two-sample using groups

☐ Two-sample using variables

☐ Paired

One-sample mean-comparison test

Variable name: 涨幅 Hypothesized mean: 0

95 Confidence level

```
. ttest 涨幅 = 0
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
涨幅	6,894	.0007699	.0003007	.0249709	.0001804	.0013595

mean = mean(涨幅) t = 2.5600

Ho: mean = 0 degrees of freedom = 6893

Ha: mean < 0 Ha: mean != 0 Ha: mean > 0

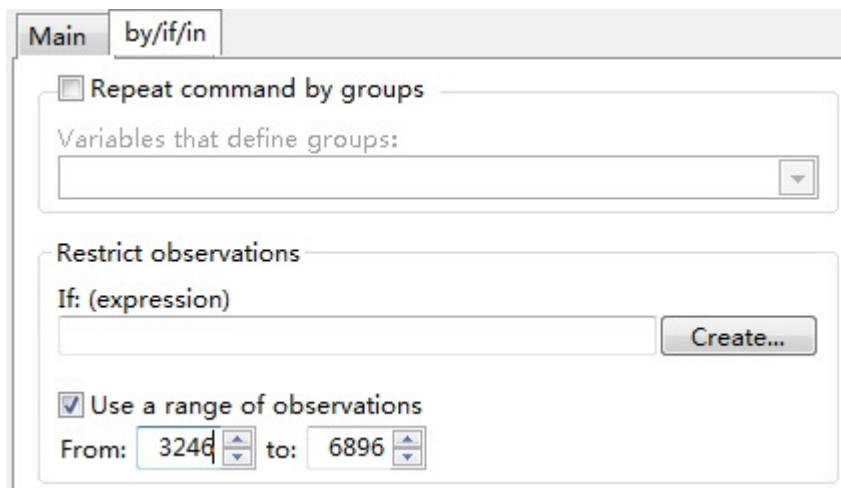
Pr(T < t) = 0.9948 Pr(|T| > |t|) = 0.0105 Pr(T > t) = 0.0052

※ 三种判断方法：0不在95%置信区间，我们拒绝H0。

t检验值等于2.56大于1.96，我们拒绝H0。（学生分布自由度很大的时候，接近正态分）

P-value=0.0105<0.05,我们拒绝H0

※前面检验表明，从三十年的尺度来看，A股是长期缓慢上涨的，那么近十年呢？
在by/if/in 选项中，勾选 “use a range of observations” ,输入 From 3246 to 6896



The image shows the 'Main' tab of the 'by/if/in' dialog box in Stata. The 'Repeat command by groups' section is collapsed. The 'Restrict observations' section is expanded, showing the 'If: (expression)' field is empty. The 'Use a range of observations' checkbox is checked. The 'From' field is set to 3246 and the 'to' field is set to 6896. A 'Create...' button is visible next to the 'If: (expression)' field.

```
. ttest 涨幅 == 0 in 3246/6896
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
涨幅	3,650	.0002921	.0002687	.0162333	-.0002347	.0008189

```
mean = mean(涨幅)                                t = 1.0872
Ho: mean = 0                                     degrees of freedom = 3649

Ha: mean < 0                                     Ha: mean != 0                                     Ha: mean > 0
Pr(T < t) = 0.8615                             Pr(|T| > |t|) = 0.2770                             Pr(T > t) = 0.1385
```

※三种判断方法：0在95%置信区间，我们不能拒绝H0。
t检验值等于1.0872大于1.96， 我们不能拒绝H0。
P-value=0.2770,我们不能拒绝H0

※注意：std.err 和std.dev是两个不同的量，后者是变量“涨幅”的标准差，前者是统计检验量t（一个新的变量）的标准误差。也就是说假设检验的本质是，构造一个新的变量，即统计检验量，通过这个统计检验量的分布判断是否拒绝原假设。

※结论：近三十年A股市场显著上涨，近十年A股上涨是非显著的（原因：近十年包括了2015年股灾）。

※前面的例子都是双侧检验，这里举一个单侧检验的例子

※A股市场有10%的涨跌幅度限制，因此波动性-即方差被限制在一定的范围内。我们有99%的把握说，涨幅在区间 $[-0.1, 0.1]$ （A股市场价格最小单位为分，四舍五入后涨幅有可能是10.5%之类），近十年均值又接近于0，那么根据 3σ 法则，标准差应该小于0.03.

※ $H_0: \sigma > 0.03$ $H_1: \sigma < 0.03$

方差检验服从卡方分布（ χ^2 , 证明从略）操作方法：statistics>classical tests of hypotheses>variance-comparison test

※判断：

0.03大于95%置信区间上限，我们拒绝 H_0
通过查表得知，统计检验值 χ^2 远大于临界值，我们拒绝 H_0 .

Stata给出的p-value对应的零假设是等于，这里不能用

```
. sdtest 涨幅 == 0.03
```

One-sample test of variance

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
涨幅	6,894	.0007699	.0003007	.0249709	.0001804	.0013595

```
sd = sd(涨幅)
```

```
Ho: sd = 0.03
```

```
c = chi2 = 4.8e+03
```

```
degrees of freedom = 6893
```

```
Ha: sd < 0.03
```

```
Pr(C < c) = 0.0000
```

```
Ha: sd != 0.03
```

```
2*Pr(C < c) = 0.0000
```

```
Ha: sd > 0.03
```

```
Pr(C > c) = 1.0000
```

※结论：A股市场波动因涨跌幅限制的存在，涨幅的方差较小。

※通过观察第一节描述性统计里的图像和表格，我们得知，A股市场早年有过一段时间没有涨跌幅限制，大盘曾经一天暴涨105%，一天暴跌16%。在限定了涨跌幅之后，样本标准差约为 $0.025 < 0.03$ 。

※我们检验，在存在涨跌幅限制的情况下，若标准差已知（用样本标准差替代），A股近十年是否长期不涨？

方差已知，均值的z检验使用正太分布表（证明从略），操作方法：statistics>classical tests of hypotheses> test (mean)

※结论：限定方差后，0在95%区间内，我们不能拒绝原假设

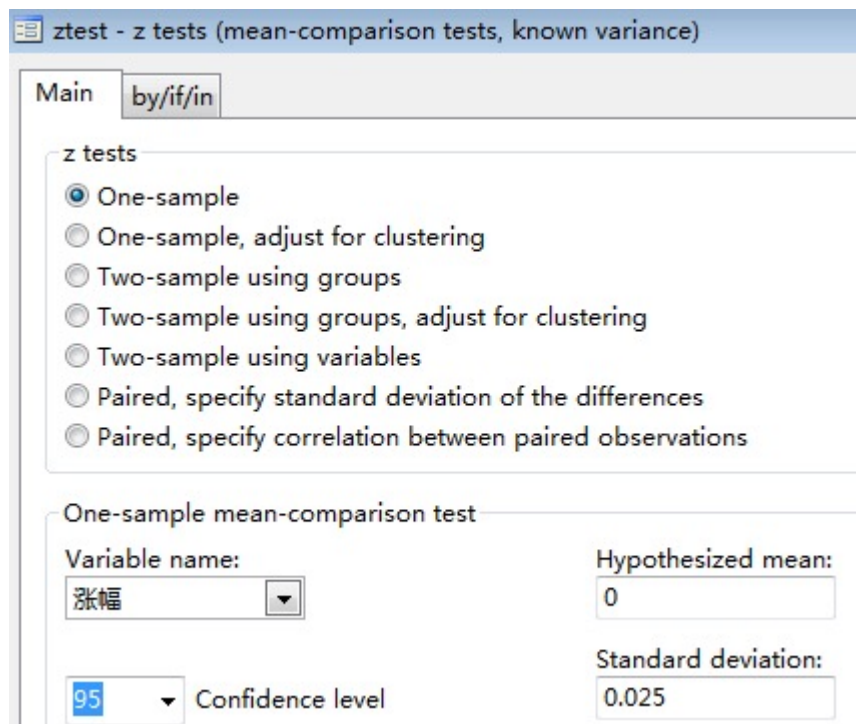
One-sample z test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
涨幅	3,650	.0002921	.0004138	.025	-.0005189	.0011032

mean = mean(涨幅) z = 0.7059

Ho: mean = 0

Ha: mean < 0	Ha: mean != 0	Ha: mean > 0
Pr(Z < z) = 0.7599	Pr(Z > z) = 0.4802	Pr(Z > z) = 0.2401



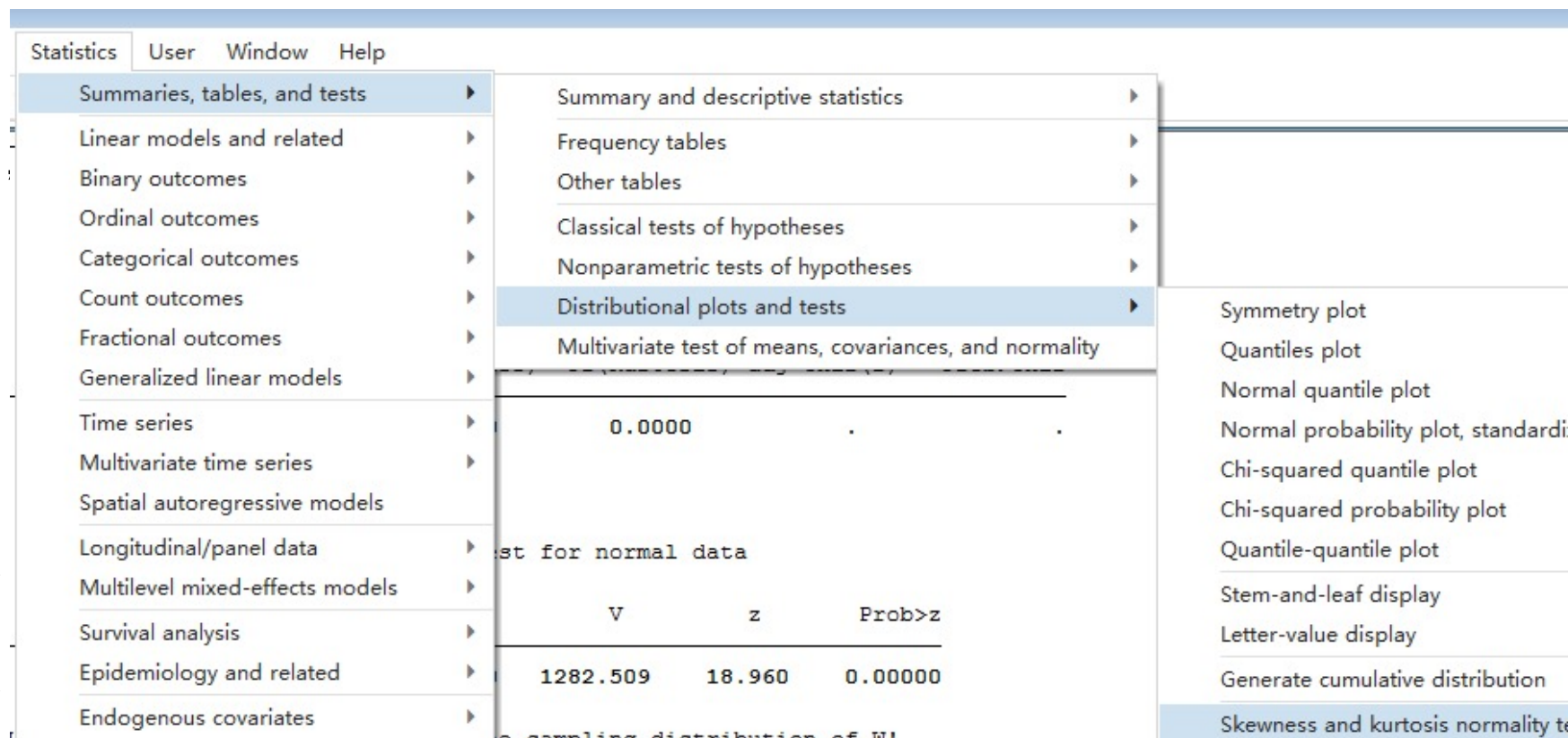
※ Z检验的其他用途：同一总体（方差已知）的两个样本均值是否有显著差异

※通过涨幅是不是没有规律？由无穷多相互独立的因素组成，因此服从正态分布呢？多种方法检验随机变量的分布是否服从正态分布。

※Jarque-Bera test：正态分布的偏度为0，峰度为3，如果样本的偏度和峰度都接近这个数字，我们有理由怀疑变量服从正态分布。

※不同统计量的构造和分布函数是不同，甚至非常复杂的。直接使用P-value进行判断是最为简单有效的。这也是相比于拒绝域和置信区间判定法，P-value越来越流行的原因。

※结论：涨幅不服从正态分布。价格具有一定的惯性（新凯恩斯主义：价格刚性），因此可以用过去的记忆来描述，如果价格被自己过去所解释，剩下的部分可能服从正态分布）



V	z	Prob>z
1282.509	18.960	0.00000

※ Shapiro–Wilk test：我国国标《GB/T4882-2001数据的统计处理 and 解释：正态性检验》中推荐了 Shapiro–Wilk 检验和 Epps-Pulley 检验，并明确 Shapiro–Wilk 检验适用于 $8 \leq n \leq 50$ 的样本数据，Epps-Pulley 检验适用于 $n \geq 8$ 的样本。

※ 操作方法：Jarque–Bera test 下面那一行即是。

※ 问题：我们的样本较大，stata 的临界值适用于 4–2000 的样本，推荐使用 JB test。

※ Epps-Pulley test 在实际中使用较少，stata 也未封装该命令，谨慎使用外部命令，不做介绍。

Epps, T.W. and Pulley, L.B. (1983), A test of normality based on empirical characteristic function, *Biometrika*, 70(3), 723–726.

※ Stata 中还提供了 Shapiro-Francia test，使用方法类似于 Shapiro–Wilk test

```
. swilk 涨幅
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
涨幅	6,894	0.64380	1282.509	18.960	0.00000

Note: The normal approximation to the sampling distribution of W' is valid for $4 \leq n \leq 2000$.

※Kolmogorov–Smirnov test 是世界上最重要的统计检验，因为它可以检验样本分布是否服从于任意的分布（当然也可以检验是否服从正态分布），是一种非参数方法。

操作方法：statistics>nonparametric tests of hypotheses> one sample Kolmogorov–Smirnov test

※原理：比较样本的累积分布函数和假定分布函数，若二者之间的差异很小，那么可以认为样本服从特定的分布。

※在expression中写入分布函数的表达式

正态分布可以用这样的命令：

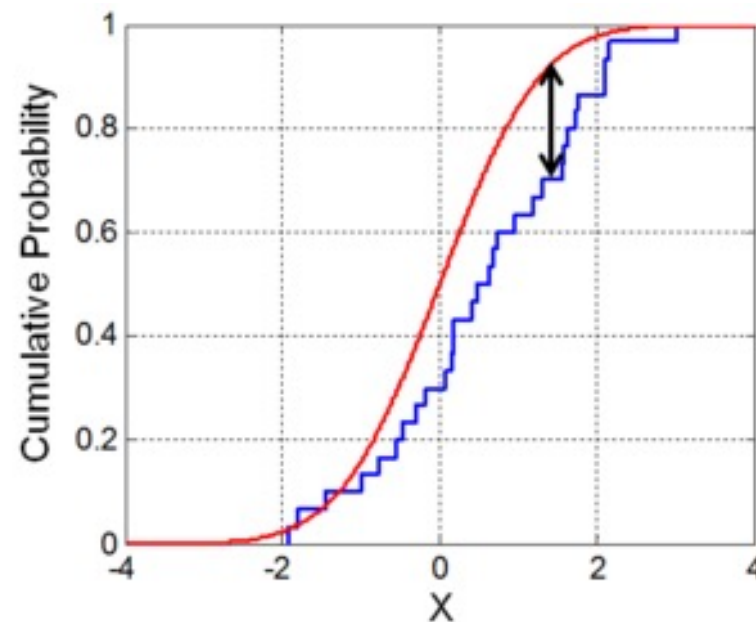
summarize 涨幅

ksmirnov 涨幅 = normal((涨幅-r(mean))/r(sd))

```
One-sample Kolmogorov-Smirnov test against theoretical distribution  
normal((涨幅-r(mean))/r(sd))
```

Smaller group	D	P-value
涨幅:	0.1486	0.000
Cumulative:	-0.1369	0.000
Combined K-S:	0.1486	0.000

```
Note: Ties exist in dataset;  
there are 6892 unique values out of 6894 observations.
```



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

见网络学堂附件

下节课见

马克思主义学院

龙治铭



清华大学
Tsinghua University