第六章　统计量及其抽样分布

李德山

四川师范大学商学院

2022 年 3 月 25 日

Contents

问题的提出

- 统计量为什么不含任何未知参数？
- 什么是自由度？
- 正态分布、t 分布和 F 分布有什么区别和联系？

1. 统计量
2. 由正态分布导出的几个重要分布
3. 样本均值的分布与中心极限定理

统计量

- 设 $X_1,X_2,\cdots,X_n$ 是从总体 X 中抽取的容量为 n 的一个样本，如果由此样本构造一个函数 $T(X_1,X_2,...,X_n)$，不依赖于任何未知参数，则称函数 $T(X_1,X_2,...,X_n)$ 是一个统计量
- 样本均值、样本比例、样本方差等都是统计量
- 统计量是样本的一个函数
- 统计量是统计推断的基础

抽样分布

- 抽样分布是区间估计的理论基础
- 样本统计量的概率分布，是一种理论分布。在重复选取容量为 n 的样本时，由该统计量的所有可能取值形成的相对频数分布
- 随机变量是样本统计量。样本均值, 样本比例，样本方差等
- 结果来自容量相同的所有可能样本
- 提供了样本统计量长远而稳定的信息，是进行推断的理论基础

抽样分布

- 抽样分布是统计量的分布而不是总体或样本的分布
- 在统计推断中总体的分布一般是未知的，不可观测的（常常被假设为正态分布）
- 样本数据的统计分布是可以直接观测的，最直观的方式是直方图，可以用来对总体分布进行检验
- 抽样分布一般利用概率统计的理论推导得出，在应用中也是不能直接观测的。其形状和参数可能完全不同于总体或样本数据的分布

抽样分布

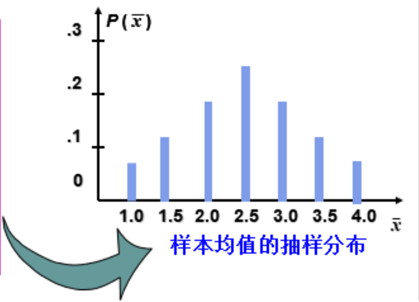
- 从总体中抽取 n = 2 的简单随机样本，在重复抽样条件下，共有 16 个样本。

所有可能的$n=2$ 的样本（共16个）				
第一个观察值	第二个观察值			
	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

抽样分布

- 各样本的均值如下表，并给出样本均值的抽样分布

16个样本的均值 ($\bar{x}$)				
第一个观察值	第二个观察值			
	1	2	3	4
1	1.0	1.5	2.0	2.5
2	1.5	2.0	2.5	3.0
3	2.0	2.5	3.0	3.5
4	2.5	3.0	3.5	4.0



抽样分布

- 所有样本均值的均值及方差

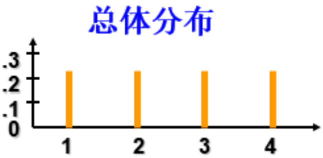
$$\mu_{\bar{x}} = \frac{\sum\limits_{i=1}^{n} \bar{x}_i}{M} = \frac{1.0 + 1.5 + \cdots + 4.0}{16} = 2.5 = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sum\limits_{i=1}^{n} (\bar{x}_i - \mu_{\bar{x}})^2}{M} = \frac{(1.0 - 2.5)^2 + \cdots + (4.0 - 2.5)^2}{16} = 0.625 = \frac{\sigma^2}{n}$$

- 样本均值的均值（数学期望）等于总体均值
- 样本均值的方差等于总体方差的 1/n

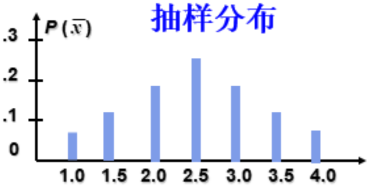
抽样分布

- 样本均值的抽样分布与总体分布的比较



$\mu = 2.5$

$\sigma^2 = 1.25$



样本均值的抽样分布

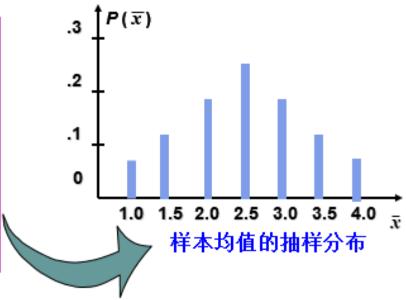
$\mu_{\bar{x}} = 2.5$

$\sigma_{\bar{x}}^2 = 0.625$

抽样分布

- 各样本的均值如下表，并给出样本均值的抽样分布

16个样本的均值（$\bar{x}$）				
第一个观察值	第二个观察值			
	1	2	3	4
1	1.0	1.5	2.0	2.5
2	1.5	2.0	2.5	3.0
3	2.0	2.5	3.0	3.5
4	2.5	3.0	3.5	4.0



χ^2 分布

- 由阿贝 (Abbe) 于 1863 年首先给出，后来由海尔墨特 (Hermert) 和皮尔逊 (Pearson) 分别于 1875 年和 1900 年推导出来
- 设 $X \sim N(\mu, \sigma^2)$ ，则 $z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- 令 $Y = z^2$，则 Y 服从自由度为 1 的 χ^2 分布，即

$$Y \sim \chi^2(1)$$

- 当总体 $X(\mu, \sigma^2)$，从中抽取容量为 n 的样本，则

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

χ^2 分布

- 分布的变量值始终为正
- 分布的形状取决于其自由度 n 的大小，通常为不对称的正偏分布，但随着自由度的增大逐渐趋于对称
- 期望为:$E(\chi^2)=n$。方差为：$D(\chi^2)=2n$
- 可加性：若 U 和 V 为两个独立的 χ^2 分布随机变量，$U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，则 $U+V$ 这一随机变量服从自由度为 n_1+n_2 的 χ^2 分布

χ^2 分布

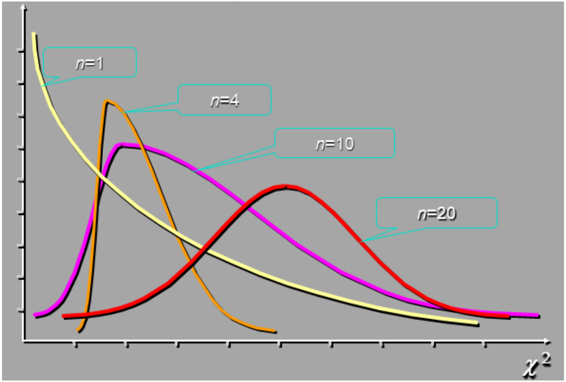
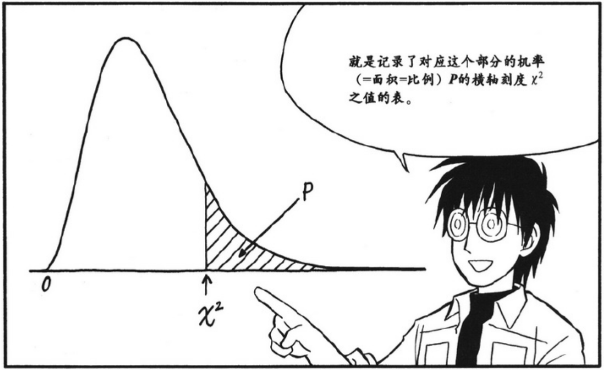


Figure: 不同容量样本的抽样分布

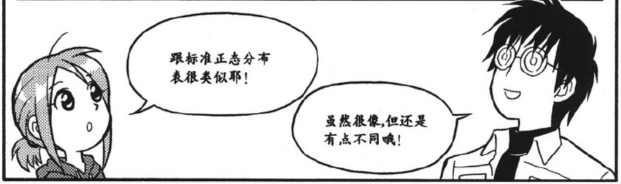
χ^2 分布



χ^2 分布

卡方分布表

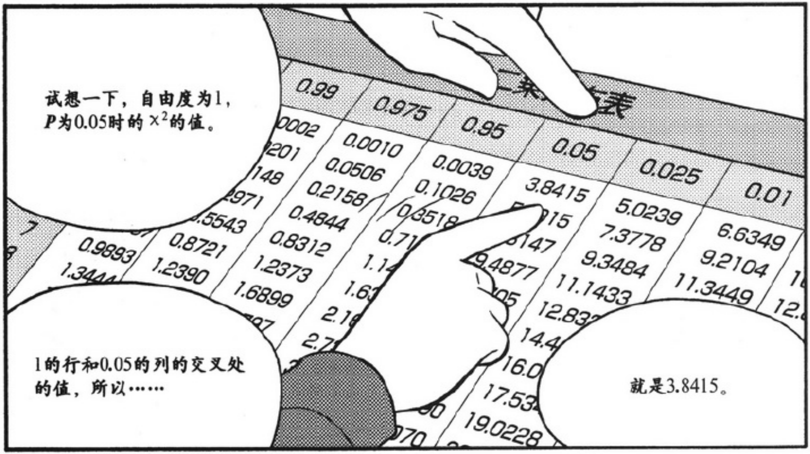
自由度 \ P	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000039	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2104	10.5965
3	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8381
4	0.2070	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767	14.8602
5	0.4118	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5475
7	0.9893	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902	21.9549
9	1.7349	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660	23.5893
10	2.1558	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093	25.1881
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



χ^2 分布


标准正态分布表为记录对应
横轴的刻度之机率的表，
欲求出的
是机率
(=面积
=比例)
卡方分布表则是记录对应
机率之横轴刻度的表。
在此，欲求出
的是这个！
χ^2
一个头变两个
大了啦！
唉呀！
先别着急嘛！

χ^2 分布



t 分布

- 高斯特 (W.S.Gosset) 于 1908 年在一篇以“Student” (学生) 为笔名的论文中首次提出
- t 分布是类似正态分布的一种对称分布，它通常要比正态分布平坦和分散
- 一个特定的分布依赖于称之为自由度的参数。随着自由度的增大，分布也逐渐趋于正态分布

t 分布

- t **分布**：假设 $Z \sim N(0,1)$,$Y \sim \chi^2(k)$，且 Z 与 Y 相互独立，则 $\frac{Z}{\sqrt{Y/k}}$ 服从自由度为 k 的 t 分布（也称 student 分布），记

$$\frac{Z}{\sqrt{Y/k}} \sim t(k)$$

k 为自由度。如果上式的分子与分母不相互独立，则一般不服从 t 分布。

- t 分布也以原点为对称。当自由度较小（20 及以下）时，t 分布尾部较厚。也就是说其与标准正态分布相比，中间的“山峰”更低（但更尖），而两侧有“厚尾”。当自由度大于等于 30 时，t 分布近似于正态分布。

t 分布

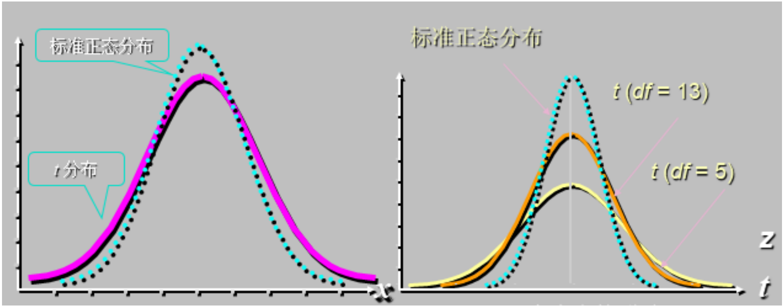


Figure: t 分布与标准正态分布的比较

F 分布

- 由统计学家费希尔 (R.A.Fisher) 提出的，以其姓氏的第一个字母来命名
- F **分布**：假设 $Y_1 \sim \chi^2(k_1)$,$Y_2 \sim \chi^2(k_2)$, 且 Y_1,Y_2 相互独立，则 $\frac{Y_1/k_1}{Y_2/k_2}$ 服从自由度为 k_1,k_2 的 F **分布**，记为

$$\frac{Y_1/k_1}{Y_2/k_2} \sim F(k_1, k_2)$$

- F 分布的取值也只能为正数，其概率密度形状与 χ^2 分布相似。t 分布的平方就是 F 分布，即如果 $X \sim t(k)$, 则 $X^2 \sim F(1, k)$（证略）。

F 分布

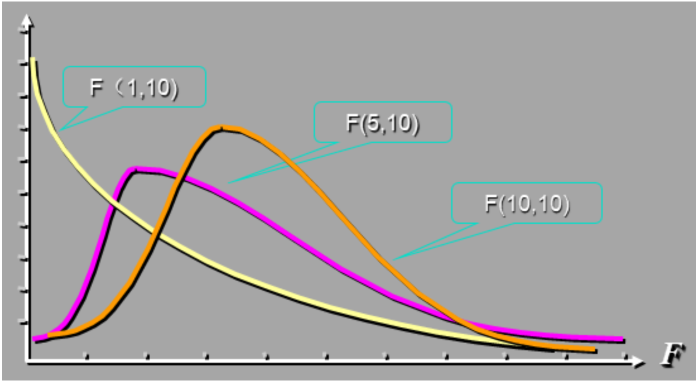


Figure: 不同自由度的 F 分布

分布	函数	函数的特征
正态分布[1]	NOPMDIST	可计算对应横轴刻度的机率
正态分布	NORMINV	可计算对应机率的横轴刻度
标准正态分布	NOPMDIST	可计算对应横轴刻度的机率
标准正态分布	NORMSINV	可计算对应机率的横轴刻度
卡方分布	CHIDIST	可计算对应横轴刻度的机率
卡方分布	CHIINV	可计算对应机率的横轴刻度
t分布	TDIST	可计算对应横轴刻度的机率
t分布	TINV	可计算对应机率的横轴刻度
F分布	FDIST	可计算对应横轴刻度的机率
F分布	FINY	可计算对应机率的横轴刻度

Figure: 不同分布对应 EXCEL 的函数

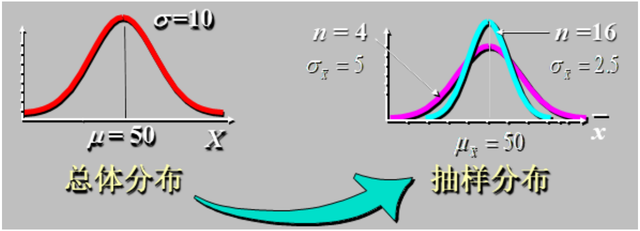
1. 统计量
2. 由正态分布导出的几个重要分布
3. 样本均值的分布与中心极限定理

样本均值的分布与中心极限定理

- 在重复选取容量为 n 的样本时，由样本均值的所有可能取值形成的相对频数分布
- 一种理论概率分布
- 推断总体均值 μ 的理论基础

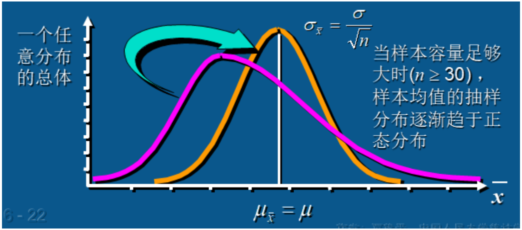
样本均值的分布与中心极限定理

- 当总体服从正态分布 $N(\mu,\sigma^2)$ 时，来自该总体的所有容量为 n 的样本的均值 $\bar{x}$ 也服从正态分布，$\bar{x}$ 的数学期望为 μ，方差为 σ^2/n。即 $\bar{x} \sim N(\mu,\sigma^2/n)$

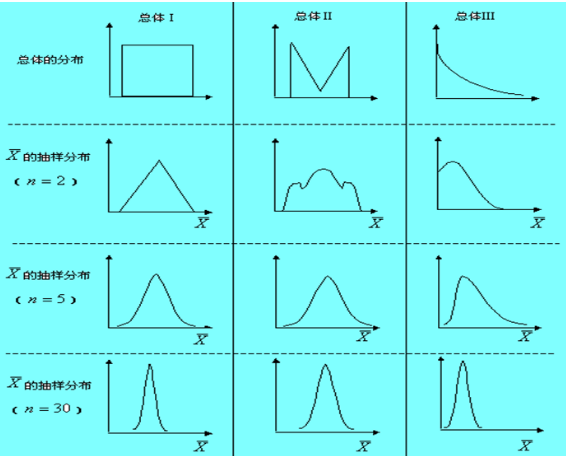


中心极限定理 central limit theorem

- 从均值为 μ，方差为 σ^2 的一个任意总体中抽取容量为 n 的样本，当 n 充分大时，样本均值的抽样分布近似服从均值为 μ、方差为 σ^2/n 的正态分布



中心极限定理



总体 I
总体 II
总体III
总体的分布
$\bar{X}$ 的抽样分布
（ n ＝2 ）
$\bar{X}$ 的抽样分布
（ n ＝5 ）
$\bar{X}$ 的抽样分布
（ n ＝30 ）
$\bar{X}$

有关样本均值抽样分布的一些结论

- 简单随机抽样、重复抽样时，样本均值抽样分布的标准差等于 $\sigma/\sqrt{n}$，这个指标在统计上称为标准误。统计软件在对变量进行描述统计时一般会输出这一结果。

有关样本均值抽样分布的一些结论

- 简单随机抽样、不重复抽样时，样本均值抽样分布的方差略小于重复抽样的方差，等于
$$\frac{\sigma^2}{n} \bullet \frac{N-n}{N-1}$$
- $\frac{N-n}{N-1}$ 这一系数称为有限总体校正系数（Finite Population Correction Factor）
- 当抽样比（n/N）<0.05 时可以忽略有限总体校正系数

补充内容

- 内容参见 Bilibili《计量经济学》–概率论与统计相关知识

参考资料

- 贾俊平.《统计学》(第八版) [M]. 北京：中国人民大学出版社，2021。
- 洪永淼. 《概率论与统计学（第二版）》[M]. 北京：中国统计出版社，2022。

Q&A

THANK YOU