区 间 估 计

目录

CONTENTS

3 引　　言

我们无法期望点估计量能给出总体参数的精确值，所以经常在点估计上加减一个被称为边际误差的值来计算区间估计。区间估计的一般形式如下：

点估计 ± 边际误差

区间估计的目的在于，提供基于样本得出的点估计与总体参数值的接近程度方面的信息。

在计算区间估计时，点估计量的抽样分布起着非常重要的作用！

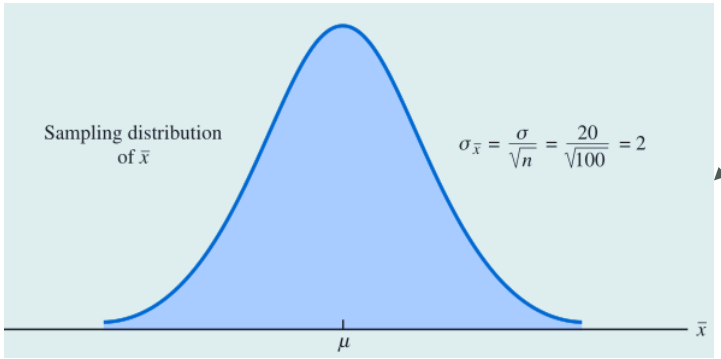
总体均值的区间估计:σ已知

5 总体标准差已知

案例

Lloyd百货公司每周选取100名顾客组成一个简单随机样本，目的在于了解他们每次购物的消费额，令 x 表示每次购买的消费额，样本均值 $\bar{x}$ 是Lloyd全体顾客每次购物消费额的总体均值μ的点估计。Lloyd公司的这项周度调查已经进行了很多年。根据历史数据，假定总体标准差已知，为σ =20美元，并且历史数据还显示总体服从正态分布。

最近抽样的100名顾客，得到的样本均值$\bar{x}$=82美元

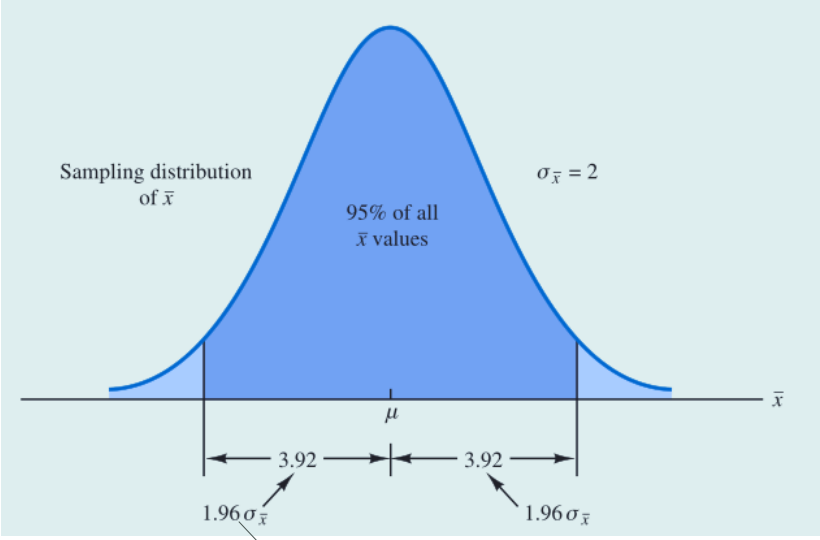


分布解释

点估计章节中，我们发现可以利用$\bar{x}$的抽样分布来计算$\bar{x}$在μ附近一定范围内的概率。

图中展示了$\bar{x}$的抽样分布。因为抽样分布说明了$\bar{x}$的值如何分布在总体均值μ 附近，所以$\bar{x}$的抽样分布提供了关于$\bar{x}$的和μ 之间可能存在的差别的信息。

6 边 际 误 差



任何正太分布随机变量都有95%的值在均值±1.96个标准差以内

边际误差

查标准正态概率分布表，任何正态分布随机变量都有95%的值在均值附近±1.96个标准差以内。因此当$\bar{x}$的抽样分布是正态分布时，一定有95%的$\bar{x}$的值在均值$\mu \pm 1.96\sigma_{\bar{x}}$以内。

案　　例

在Lloyd公司的例子中，我们已知$\bar{x}$的抽样分布是正态分布并且标注误差$\sigma_{\bar{x}}$=2。因为$1.96\sigma_{\bar{x}}$=1.96x2=3.92，所以在n=100的样本容量下，$\bar{x}$的所有值中有95%落在总体均值μ附近±3.92以内

7 置 信 区 间

总体均值的区间估计:σ已知

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$1-\alpha$为置信系数；$z_{\alpha/2}$表示正态概率分布上侧面积为$\alpha/2$时的z值

```
from scipy.stats import norm
import numpy as np
def confidence_interval(sample_mean,population_std,
                                sample_size,confidence):
    alpha = 1 - confidence
    # 根据概率求解z值，或者是求解阴影面积，它的逆操作是norm.cdf(z_score)
    z = abs(norm.ppf(alpha/2))
    upper_limit = sample_mean + z * population_std/np.sqrt(sample_size)
    lower_limit = sample_mean - z * population_std/np.sqrt(sample_size)
    return [lower_limit,upper_limit]
```

计　　算

对于95%的置信区间，置信系数$1-\alpha$=0.95，于是α=0.05. 查正态分布表可知，上侧面积为0.025时对应的$z_{0.25}$ =1.96。Lloyd公司的样本均值为82，总体标准差为20，样本容量为100，于是得到

$$82 \pm 1.96\frac{20}{\sqrt{100}} \quad 82 \pm 3.92$$

于是，当边际误差为3.92时，95%的置信区间为[78.08,85.92]

常用置信水平

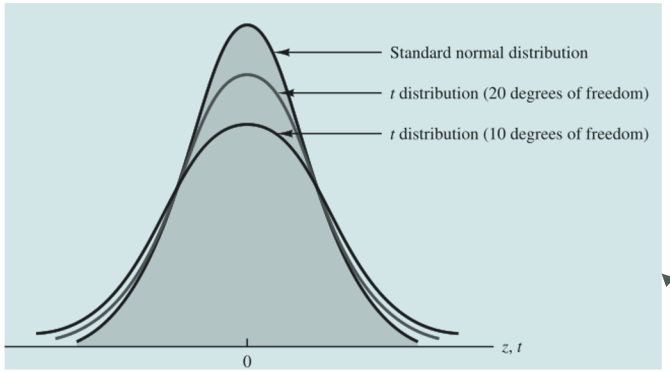
Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.05	1.645
95%	.05	.025	1.960
99%	.01	.005	2.576

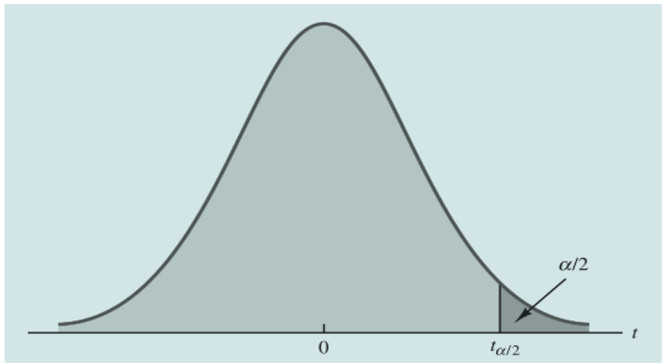
想要达到较高的置信水平，必须加大边际误差，既加大置信区间的宽度

02

总体均值的区间估计:σ未知

9 t 分 布





给t加上下标表明其在t分布上侧的面积

σ未 知

在建立总体均值的区间估计时，我们通常并没有关于总体标准差一个号的估计。在这种情形下，我们必须利用同一样本估计μ和σ两个参数。

t 分 布

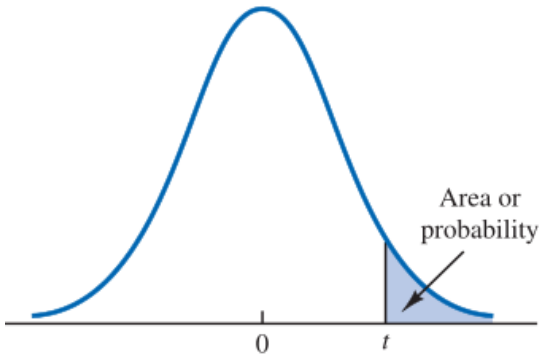
定义：在概率论和统计学中，t-分布（t-distribution）用于根据小样本来估计呈正态分布且方差未知的总体的均值。如果总体方差已知（例如在样本数量足够多时），则应该用正态分布来估计总体均值。

当利用s估计σ时，边际误差和总体均值的区间估计都以t分布的概率分布为依据进行的。虽然t分布的数学推导是以假设总体服从正态分布为依据的，但是许多研究表明在总体分布偏态的情形下，t分布效果也相当不错。

t 分布依赖于自由度的参数。当自由度为1,2,3,...时，有且仅有唯一的t分布与之相对应。随着自由度的增大，t分布与标准正态分布之间的差别变得越来越来越小

t分布的均值为0

10 t 分 布 表



Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	.848	1.296	1.671	2.000	2.390	2.660
61	.848	1.296	1.670	2.000	2.389	2.659
62	.847	1.295	1.670	1.999	2.388	2.657
63	.847	1.295	1.669	1.998	2.387	2.656
64	.847	1.295	1.669	1.998	2.386	2.655
65	.847	1.295	1.669	1.997	2.385	2.654
66	.847	1.295	1.668	1.997	2.384	2.652
67	.847	1.294	1.668	1.996	2.383	2.651
68	.847	1.294	1.668	1.995	2.382	2.650
69	.847	1.294	1.667	1.995	2.382	2.649
⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	.846	1.291	1.662	1.987	2.368	2.632
91	.846	1.291	1.662	1.986	2.368	2.631
92	.846	1.291	1.662	1.986	2.368	2.630
93	.846	1.291	1.661	1.986	2.367	2.630
94	.845	1.291	1.661	1.986	2.367	2.629
95	.845	1.291	1.661	1.985	2.366	2.629
96	.845	1.290	1.661	1.985	2.366	2.628
97	.845	1.290	1.661	1.985	2.365	2.627
98	.845	1.290	1.661	1.984	2.365	2.627
99	.845	1.290	1.660	1.984	2.364	2.626
100	.845	1.290	1.660	1.984	2.364	2.626
∞	.842	1.282	1.645	1.960	2.326	2.576

11 置 信 区 间

总体均值的区间估计:σ未知

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

s为样本标准差,$1-\alpha$为置信系数，自由度为n-1的分布中，$t_{\alpha/2}$上侧的面积恰好等于$\alpha/2$

```
from scipy.stats import t
import numpy as np
def confidence_interval(sample_mean,sample_std,
                                sample_size,confidence):
    alpha = 1 - confidence
    # 根据概率求解t值，或者是求解阴影面积，它的逆操作是t.cdf(t_score)
    t = abs(t.ppf(alpha/2,sample_size -1))
    upper_limit = sample_mean + t * sample_std/np.sqrt(sample_size)
    lower_limit = sample_mean - t * sample_std/np.sqrt(sample_size)
    return [lower_limit,upper_limit]
```

自 由 度 解 释

自由度通常就是样本数量减去一。我们知道样本标准差：

$$s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$

自由度是计算$\sum(x_i-\bar{x})^2$时所用到的信息中独立信息的个数，在计算该项时，我们用到了n条信息： $x_1-\bar{x}$ ，$x_2-\bar{x}$，..., $x_n-\bar{x}$。可以证明对于任何数据集$\sum(x_i-\bar{x})=0$,因此$x_i-\bar{x}$中只有n-1项是独立的，既我们知道了n-1个值，则由所有$x_i-\bar{x}$值之和为0，可以确定余下的值。

12 置信区间案例

样本中70个家庭的信用卡余额数据

9430	14661	7159	9071	9691	11032
7535	12195	8137	3603	11448	6525
4078	10544	9467	16804	8279	5239
5604	13659	12595	13479	5649	6195
5179	7061	7917	14044	11298	12584
4416	6245	11346	6817	4353	15415
10676	13021	12806	6845	3467	15917
1627	9719	4972	10493	6191	12591
10112	2200	11356	615	12851	9743
6567	10746	7117	13627	5337	10324
13627	12744	9465	12557	8372	
18719	5742	19263	6232	7445	

计　　算

对于95%的置信区间，置信系数$1-\alpha$=0.95，于是α=0.05.。计算得出样本均值为9312美元，样本标准差为4007美元，自由度为70-1=69，那么查表计算$t_{0.025}$=1.995。于是得到置信区间估计：

$$9312 \pm 1.995\frac{4007}{\sqrt{70}} \quad 9213 \pm 955$$

总 体 容 量 的 确 定

14 样本容量的确定

总体原则：样本容量越大，估计的置信区间就越准确。尤其是总体偏态严重时，样本容量需要足够大。
确定的目标：说明如何确定足够的样本容量以达到所希望的边际误差。

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$就是边际误差，令E代表希望达到的边际误差

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

解出$\sqrt{n}$：

$$\sqrt{n} = z_{\alpha/2}\frac{\sigma}{E}$$

总体均值区间估计的样本容量

$$n = \frac{{z_{\alpha/2}}^2\sigma^2}{E^2}$$

15 样本容量的确定

总体均值区间估计的样本容量

$$n = \frac{{z_{\alpha/2}}^2 \sigma^2}{E^2}$$

美国汽车租赁成本的已有调查发现，租赁一辆中型汽车的平均费用大约为每天55美元，租赁费用的样本标准差为9.65美元。假设现在需要对美国一辆中型汽车的租赁费用的总体均值进行估计，设定置信水平为95%，边际误差为2美元。

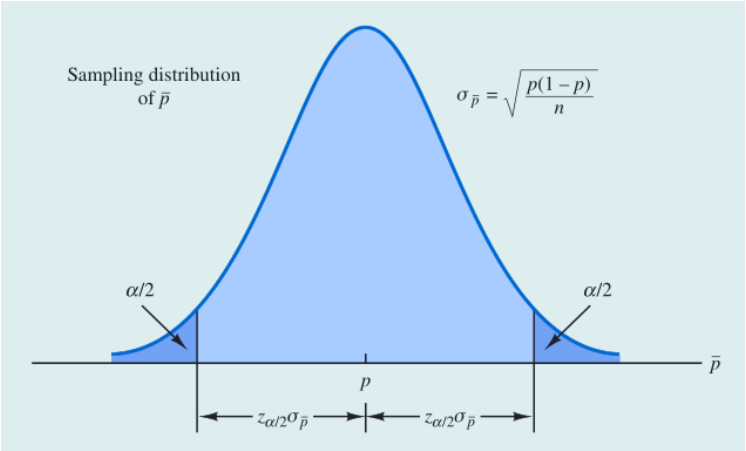
$$n = \frac{{z_{\alpha/2}}^2 \sigma^2}{E^2} = \frac{1.96^2 \times 9.65^2}{2^2} = 89.43$$

那么样本中至少应该选取89.43笔中型汽车租赁业务才能满足估计的边际误差为2美元。

04

总 体 比 率

17 总 体 比 率



总体比率的区间估计

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

区间估计

$\bar{p}$ ±边际误差

在计算区间估计的边际误差时，$\bar{p}$的抽样分布至关重要

抽样分布

当$np \geq 5$ 和$n(1-p) \geq 5$时，$\bar{p}$的抽样分布近似服从正态分布。 $\bar{p}$的抽样分布的均值是总体比率p， $\bar{p}$的标准差为：$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$

但是由于总体比率是未知的，我们必须使用样本比率来估算总体比率

边际误差为：

$$z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$