经济统计 – 5

刁莉男

diaoln@jlu.edu.cn

吉林大学商学院

April 11, 2012

复习

- 5. 概率
- 6. 离散型概率分布：二项分布、泊松分布
- 7. 连续型概率分布：均匀分布、正态分布
- 8. 抽样方法、中心极限定理

提纲

8. 抽样方法、中心极限定理

抽样方法

中心极限定理

9.点估计与置信区间

均值的点估计与置信区间

抽样方法

- 简单随机抽样
- 系统随机抽样
- 分层随机抽样
- 整群抽样

抽样误差（Sample Error）

抽样误差：样本统计量和总体参数之间的差别（误差）。

例：Foxtrot Inn B&B，计算样本容量为5时，均值的抽样误差。共有多少个可能抽样?(142,506) 抽样误差之和为多少?(0)

June	Rentals	June	Rentals	June	Rentals
1	0	11	3	21	3
2	2	12	4	22	2
...	...	...	...	...	...
10	7	20	2	30	3

样本均值的抽样分布 (Sampling Distribution of the Sample Mean)

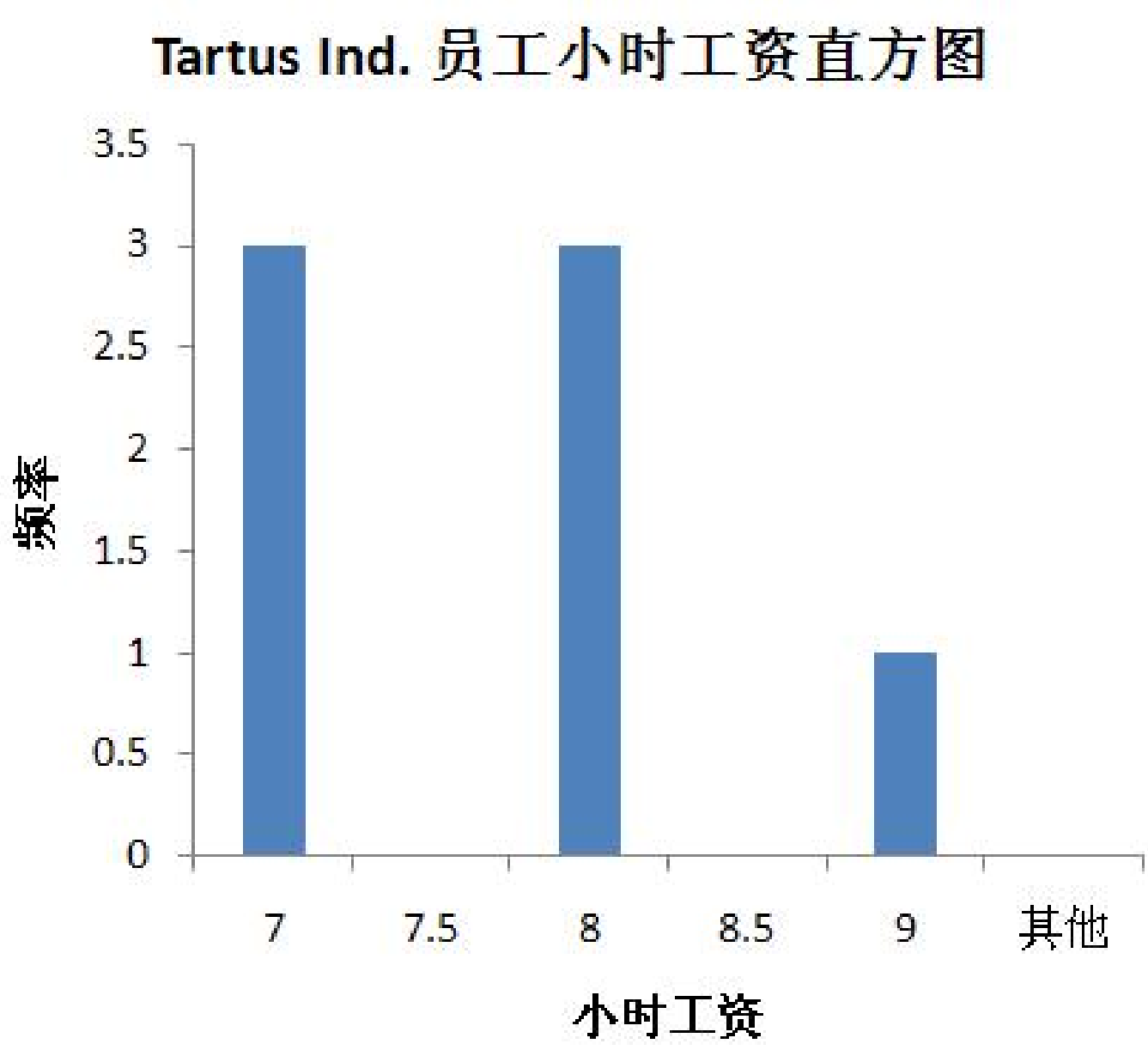
给定样本容量条件下所有可能的样本均值的概率分布。

例：Tartus Industries 7名员工每小时工资如下，构造样本均值抽样分布。

员工	小时工资	员工	小时工资
Joe	$7	Jan	$7
Sam	7	Art	8
Sue	8	Ted	9
Bob	8		

- 总体均值是多少(μ)？画出员工工资直方图。
- 样本容量为2时，构造样本均值的抽样分布？
- 样本均值抽样分布的均值是多少($\mu_{\bar{x}}$)？
- 抽样分布均值与总体均值关系是什么？

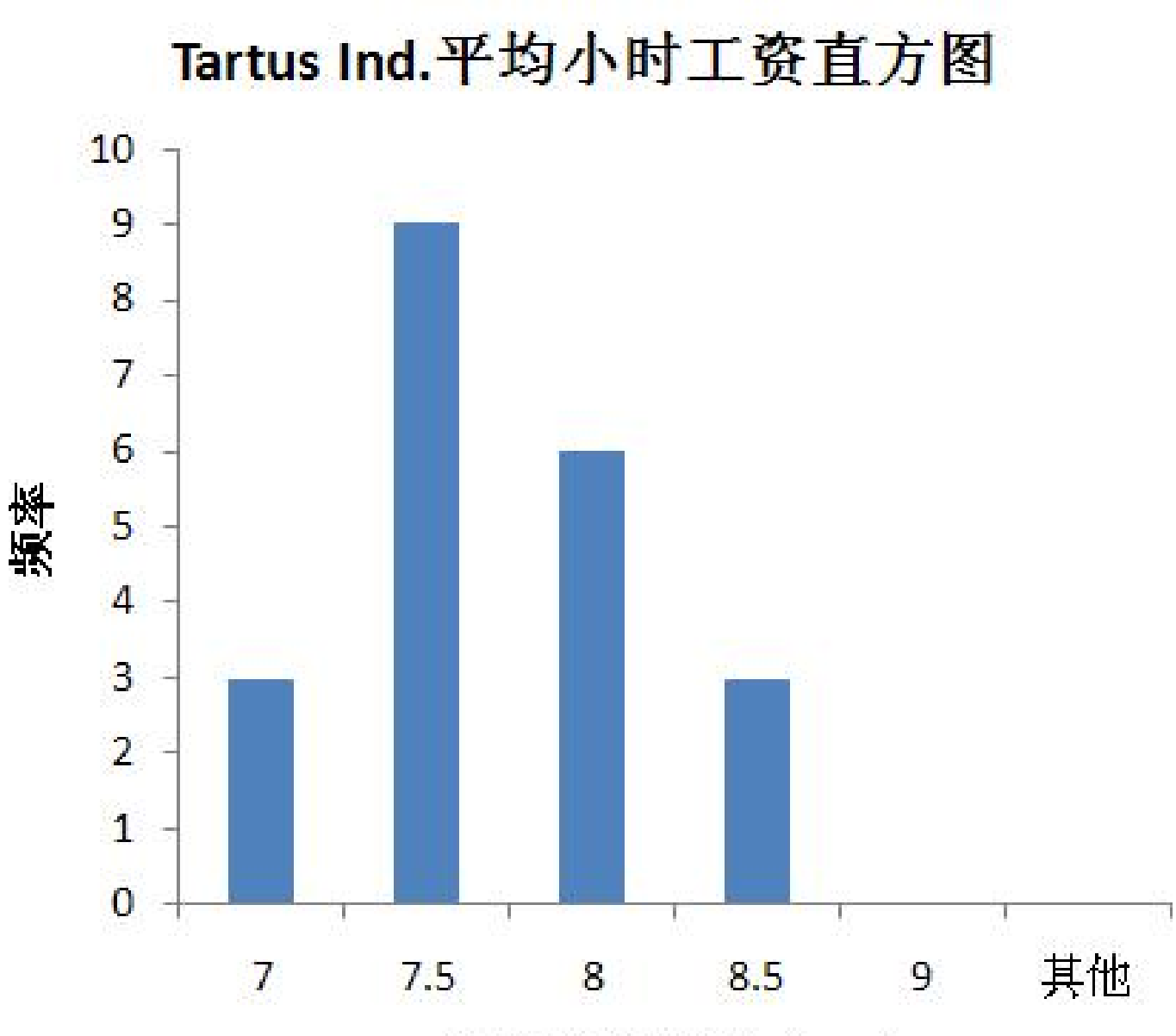
Tartus Ind. 员工小时工资直方图



Tartus Ind. 员工小时工资n=2时21个抽样样本均值

Sample	Employees	Hourly Earnings	Sum	Mean	Sample	Employees	Hourly Earnings	Sum	Mean
1	Joe, Sam	$7, $7	$14	$7.00	12	Sue, Bob	$8, $8	$16	$8.00
2	Joe, Sue	7, 8	15	7.50	13	Sue, Jan	8, 7	15	7.50
3	Joe, Bob	7, 8	15	7.50	14	Sue, Art	8, 8	16	8.00
4	Joe, Jan	7, 7	14	7.00	15	Sue, Ted	8, 9	17	8.50
5	Joe, Art	7, 8	15	7.50	16	Bob, Jan	8, 7	15	7.50
6	Joe, Ted	7, 9	16	8.00	17	Bob, Art	8, 8	16	8.00
7	Sam, Sue	7, 8	15	7.50	18	Bob, Ted	8, 9	17	8.50
8	Sam, Bob	7, 8	15	7.50	19	Jan, Art	7, 8	15	7.50
9	Sam, Jan	7, 7	14	7.00	20	Jan, Ted	7, 9	16	8.00
10	Sam, Art	7, 8	15	7.50	21	Art, Ted	8, 9	17	8.50
11	Sam, Ted	7, 9	16	8.00					

Tartus Ind. 员工小时工资样本均值（n=2）直方图



样本均值的抽样分布

- 样本均值的均值与总体均值完全相等$(\mu = \mu_{\bar{x}})$;
- 样本的抽样分布的发散程度比总体的发散程度小;
- 样本均值的抽样呈钟型分布并近似于正态。

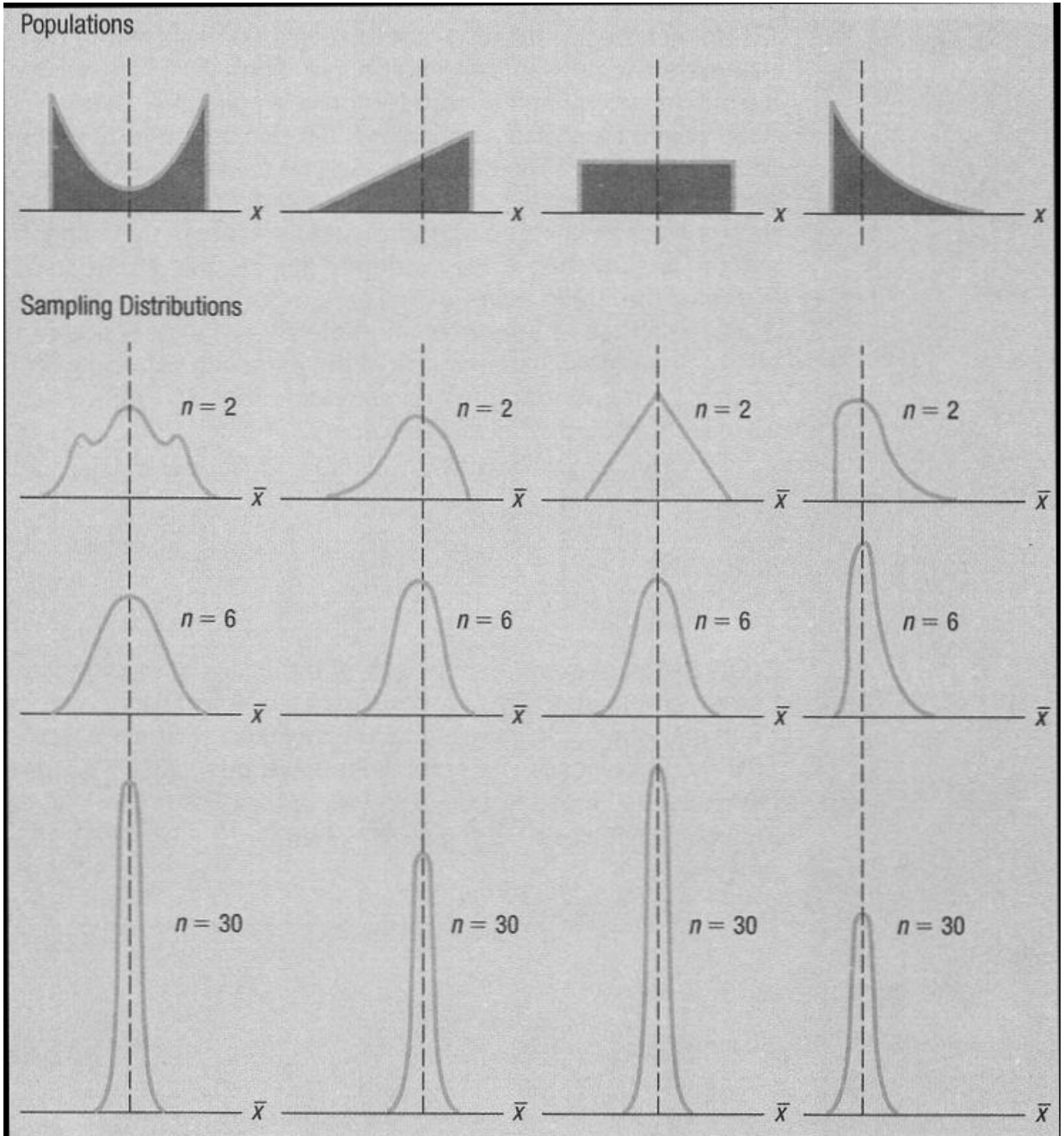
中心极限定理（Central Limit Theorem，CLT）

如果从总体中抽选出一定样本容量的所有样本，样本均值的抽样分布近似的服从正态分布，样本越大，近似的效果越好。

- 是9、10章的基础；
- 关于样本均值抽样分布$(\bar{X})$；
- 与总体分布无关；或者说，总体可以是任何分布。

中心极限定理

- 总体为正态分布，对于任何样本容量，$\bar{X}$是正态分布；
- 总体对称，但不是正态分布，则当样本容量大于10时，$\bar{X}$呈正态分布；
- 总体有偏度和峰度，则样本容量大于30时，$\bar{X}$呈正态分布。


Populations
x
x
x
x
Sampling Distributions
n = 2
n = 2
n = 2
n = 2
n = 6
n = 6
n = 6
n = 6
n = 30
n = 30
n = 30
n = 30
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄

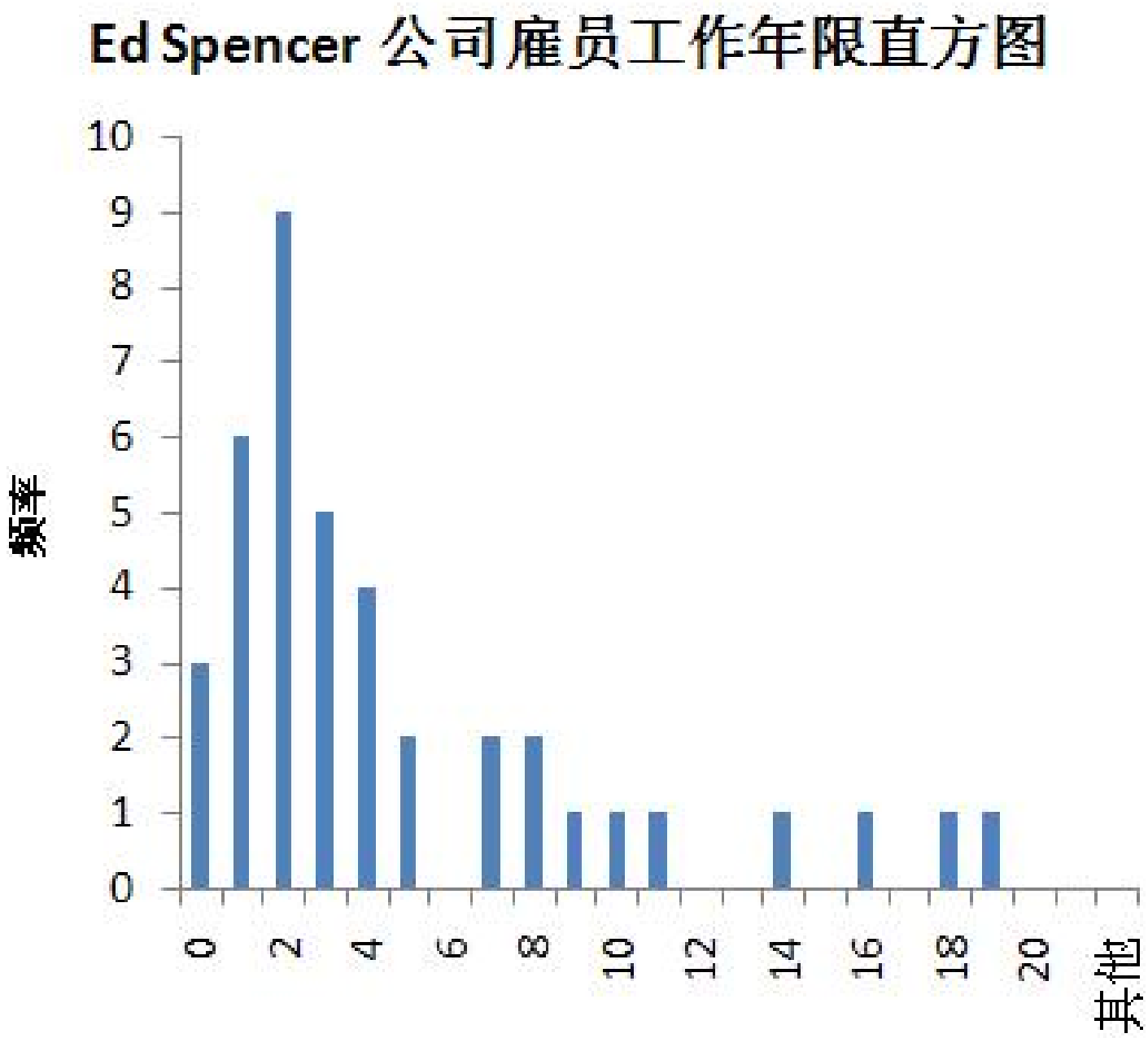
例1：Ed Spence公司

Table: Ed Spence公司员工工作年限：

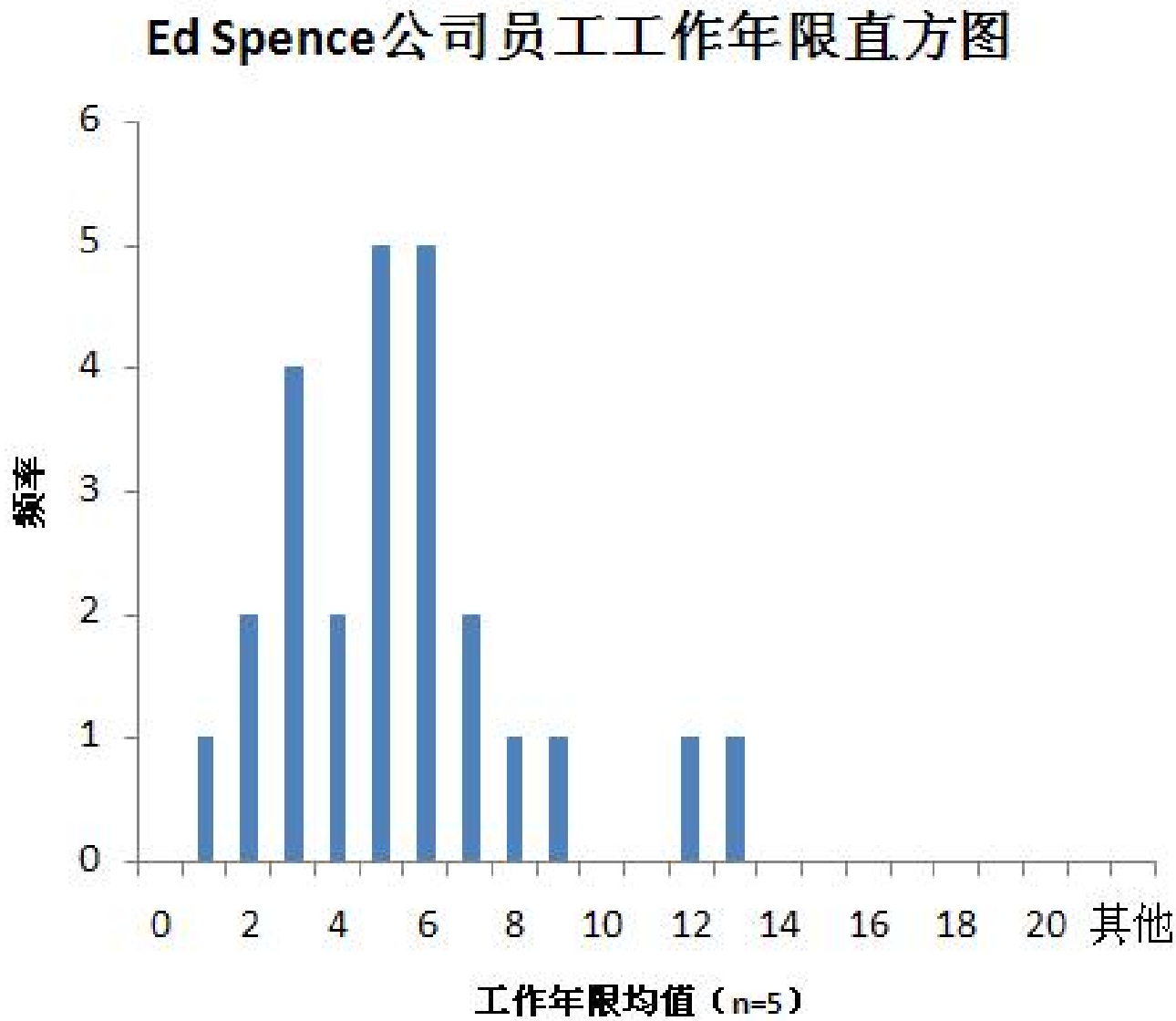
11	4	18	2	1	2	0	2	2	4
3	4	1	2	2	3	3	19	8	3
7	1	0	2	7	0	4	5	1	14
16	8	9	1	1	2	5	10	2	3

- 总体均值(μ)，总体分布（偏度、分散性）；
- n=5时，样本均值分布，与总体分布区别；
- n=25时，样本均值分布，与总体分布区别；
- CLT。

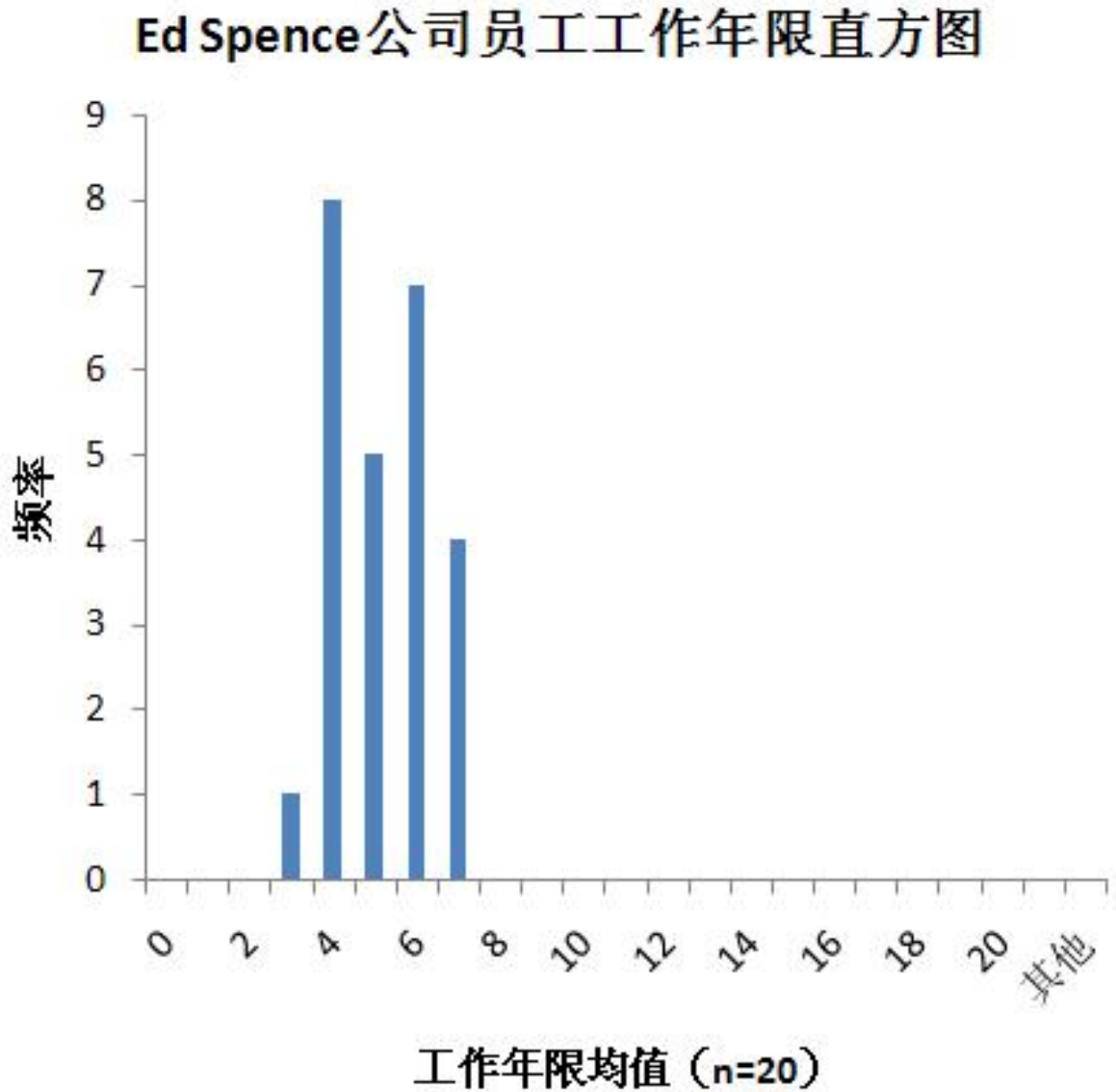
例1：Ed Spence公司员工工作年限分布



例1：Ed Spence公司员工工作年限样本均值（n=5）抽样分布



例1：Ed Spence公司员工工作年限样本均值（n=20）抽样分布



中心极限定理（Central Limit Theorem，CLT）

- 样本均值的均值与总体均值完全相等，$\mu = \mu_{\bar{x}}$；
- 样本的抽样分布的发散程度比总体的发散程度小:

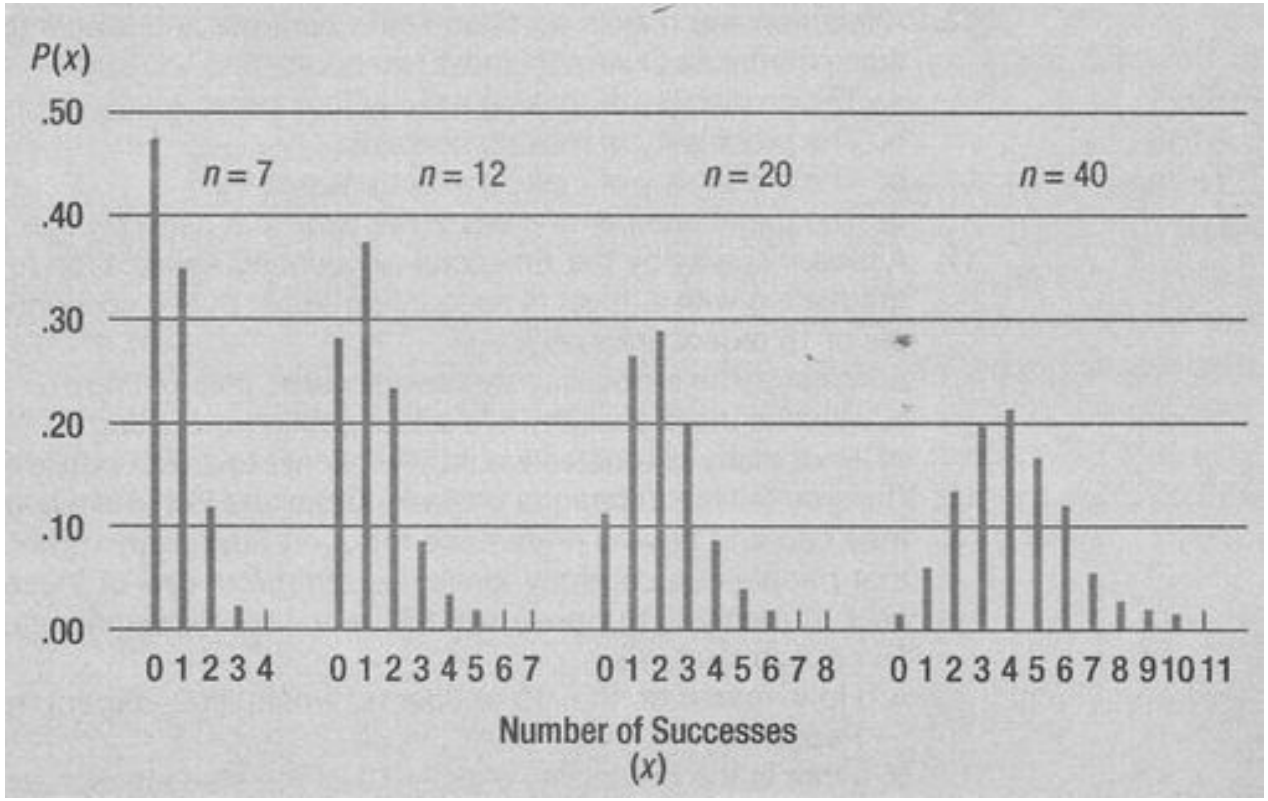
  均值的标准误差（Standard Error of the Mean, Standard Deviation of the Sampling Distribution of the Sample Mean）

  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

中心极限定理

例2：第六章二项分布，当$\pi = 0.1$时，随着n增大，趋向于服从正态分布。

？？为什么分散性没有减少（作业）？



样本均值抽样分布的应用

例1. 洗衣液公司为了确保每桶洗衣液含量都是100盎司，已知总体均值100盎司，标准差2盎司。某天早上质检员抽查了40瓶洗衣液，测量得到平均容量为99.8盎司。质检员是否可以得到洗衣液容量不足这一结论？或者说抽样误差是否在合理范围内？

样本均值抽样分布的应用

例2. 研究表明，美国成年人平均每天看电视6小时（总体均值），标准差1.5小时。我们是否可以在波士顿地区随机抽取50个样本，得到平均看电视时间为每天6.5小时？

样本均值抽样分布的应用

例3. 成年人体重的均值为160磅（总体均值），标准差为15磅。体重的分布服从正偏分布。对于一个样本容量为30的随机样本来说，平均体重为170磅的概率是多少？

样本均值抽样分布的应用

例4. 一可乐公司检查每瓶可乐容量是否标准。根据记录，可乐容量服从正态分布，均值为31.2盎司，标准差为0.4盎司。某天早上质检员抽查了16瓶包装好的可乐，可乐容量均值为31.38盎司。这是不是正常结果？是不是每瓶可乐容量过多？或者说，抽样误差（0.18）是不是正常？

样本均值抽样分布的应用

从以上4个例子可以看出，我们都需要计算样本误差是否由于抽样引起的误差。

以例4为例：

- $\mu = 31.2$，$\sigma = 0.4$;
- $\bar{X} = 31.38$，$n = 16$
- $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
- $z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- $z = \frac{31.38-31.20}{0.4/\sqrt{16}} = 1.80$，$p = 0.0359$

点估计与置信区间

- 点估计(Point Estimation)：根据样本信息计算出来的样本统计量中（单一值），被用来估计总体参数；
- 置信区间(Confidential Intervals, CI)：从样本数据构造的一个数据区间，使得总体参数以某一概率落在这一区间，概率称为置信水平；
- 确定合理的样本个数。

均值的点估计与区间估计

- 总体标准差(σ)已知；
- 总体标准差未知(σ)，用样本标准差代替(s)。

均值的点估计与区间估计，σ已知

- 点估计：$\mu = \bar{X}$;
- 区间估计：$\bar{X} \pm z\frac{\sigma}{\sqrt{n}}$;
- z由置信水平决定；
- $P(|\bar{X} - \mu| \leqslant 1.96\frac{\sigma}{\sqrt{n}}) = 95\%$；
- $P(|\bar{X} - \mu| \leqslant 2.56\frac{\sigma}{\sqrt{n}}) = 99\%$；
- 均值标准误差：$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$，大小受$\sigma$和$n$影响。

均值的点估计与区间估计，σ已知

例：一食品公司生成带果肉的果冻，每罐重量为4.01盎司（总体均值），标准差为0.02盎司，每罐重量服从正态分布。现在随机抽取16罐，得到样本均值为4.015盎司，95%置信区间是多少？

- $\mu = 4.01$，$\sigma = 0.02$
- $\bar{X} = 4.015$，
- $4.015 \pm 1.96(0.02/\sqrt{16}) = 4.015 \pm 0.0098$
- CI=[4.0052，4.0248]

均值的点估计与区间估计，σ已知

例2：美国管理协会计算零售企业中层管理者的年平均收入。随机调查了256名管理者，得到样本平均收入为$45,420。总体标准差为$2,050。现在回答已下问题：

- 总体均值是多少？用来估计总体均值的比较合理的数值是什么？
- 总体均值的合理区间是多少？
- 解释区间的含义。

均值的点估计与区间估计

例1：New York-New Jersey地区建筑工人平均年收入为$65,000，这一估计的区间为$61,000 - $69,000，总体参数落在这一区间的概率是多少？置信水平是多少？

均值的点估计与区间估计，σ未知

当总体标准差未知时，用样本标准差s代替。构造统计量

- $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$，服从$t(n-1)$分布，但是基于X服从正态分布假设；
- 置信区间：$\bar{X} \pm t\frac{s}{\sqrt{n}}$；
- t值与置信水平相关；

t分布性质

- 与正态分布相似，t分布也是连续分布；
- 对称、钟形分布；
- 均值为0，方差为$\frac{n}{n-2}$;
- 与正态分布相比更平坦，随着n增大，逐渐趋近于正态分布。

均值的点估计与区间估计，σ未知

例1：一轮胎生产厂商调查其轮胎花纹的使用寿命。随机抽取了10个行程5万英里以上的轮胎，样本均值为0.32 inch，标准差0.09 inch。构造总体均值的95%置信区间；生产厂商得到结论5万英里之后轮胎花纹的平均厚度（总体均值）为0.30 inch是否合理？

- $\bar{X} \pm t\frac{s}{\sqrt{n}} = 0.32 \pm 2.262\frac{0.09}{\sqrt{10}} = [0.256, 0.384]$；

均值的点估计与区间估计，σ未知

例2：一商场管理者想估计每位顾客的平均消费金额，下表给出了20个顾客的消费金额：

48.16	42.22	46.82	51.45	23.78	41.86	54.86
37.92	52.64	48.59	50.82	46.94	61.83	61.69
49.17	61.46	51.35	52.68	58.84	43.88	

- 总体均值的最好估计量是多少？
- 确定总体均值的95%置信区间，
- 解释结果，
- 总体均值为50（60）这一结论是否合理？