经济统计 – 4

刁莉男

diaoln@jlu.edu.cn

吉林大学商学院

March 28, 2012

第三章

- 集中趋势的度量
- 离散趋势的度量

第四章

- 点图、盒形图、偏度、峰度

提纲

第五章 概率

- 概率
- 客观概率、主观概率
- 计算概率的一些原则

第六章 离散型概率分布

- 概率分布
- 二项概率分布

第七章 连续型概率分布

- 均匀概率分布
- 正态分布

第八章 抽样方法及中心极限定理

- 抽样方法
- 抽样“误差”
- 中心极限定理

概率（Probability）

介于0和1之间，描述一个事件发生的相对可能性。

- 试验：引致几个可能观测结果中的一个结果而且只有一个结果发生的这样的一个过程叫做试验。
- 结果：试验的一个特定结果称为结果。
- 事件：一个试验中一个或多个结果的集合称为事件。

概率

一网络游戏公司开发了一个网络游戏，邀请了80个玩家进行测试（喜欢、不喜欢）。

- 在本例中什么是试验?
- 举例说明一个可能的结果。
- 假设65个玩家喜欢这款游戏，65是概率吗?
- 举例说明一个事件。

客观、主观概率

- 客观概率 （Objective Probability）
  - 古典概率（Classical Probability）
  - 经验概率（Empirical Probability）
- 主观概率（Subjective Probility）

客观概率-古典概率

每个试验的结果发生的可能性都是相同的。

P(A)=m/n，m表示事件A包含的基本事件数，n表示所有的基本事件数。

- 互斥：一个事件的发生意味着其它事件不可能同时发生（例如：性别）。
- 完全穷尽：一个试验的事件集合包含所有可能结果，则此事件集合是完全穷尽的。

客观概率-经验概率

又叫相对频率，基于一个事件过去发生的频率而计算出来。

P(A)=m/n，m表示事件A过去发生的频率，n表示所有观测次数。

大数定律：在一个大量的重复试验中，某一事件的经验概率将接近其真实概率。

主观概率

个体基于可获得信息（观念、理论、预感等）对某一事件发生的概率（可能性）进行估计。

主观概率有可能产生偏差（Bias）。

例子

- 从一副扑克中随机抽取一张牌，被抽到的牌是红桃A的概率是多少？属于哪类概率？
- 儿童托管中心调查539名儿童父母的情况，其中有333对结婚、182对离异、24对丧偶。随机选中的某一儿童父母离异的概率多少？属于哪类概率？
- Dow-Jones工业指数下一年超过15,000的概率是多少？属于哪类概率？

计算概率的一些原则

加法原则：

- 特殊原则：如果A和B互斥，
  P(A or B)= P(A)+P(B)
- 互补：P(A)+P(∼ A)=1
- 联合概率：测量两个或两个以上事件发生的概率。
- 一般原则：P(A or B)=P(A)+P(B)-P(A and B)

计算概率的一些原则

乘法原则：

- 独立：一个事件的发生对于另一个事件发生的概率不产生影响。
- 特殊原则：如果A和B互相独立，

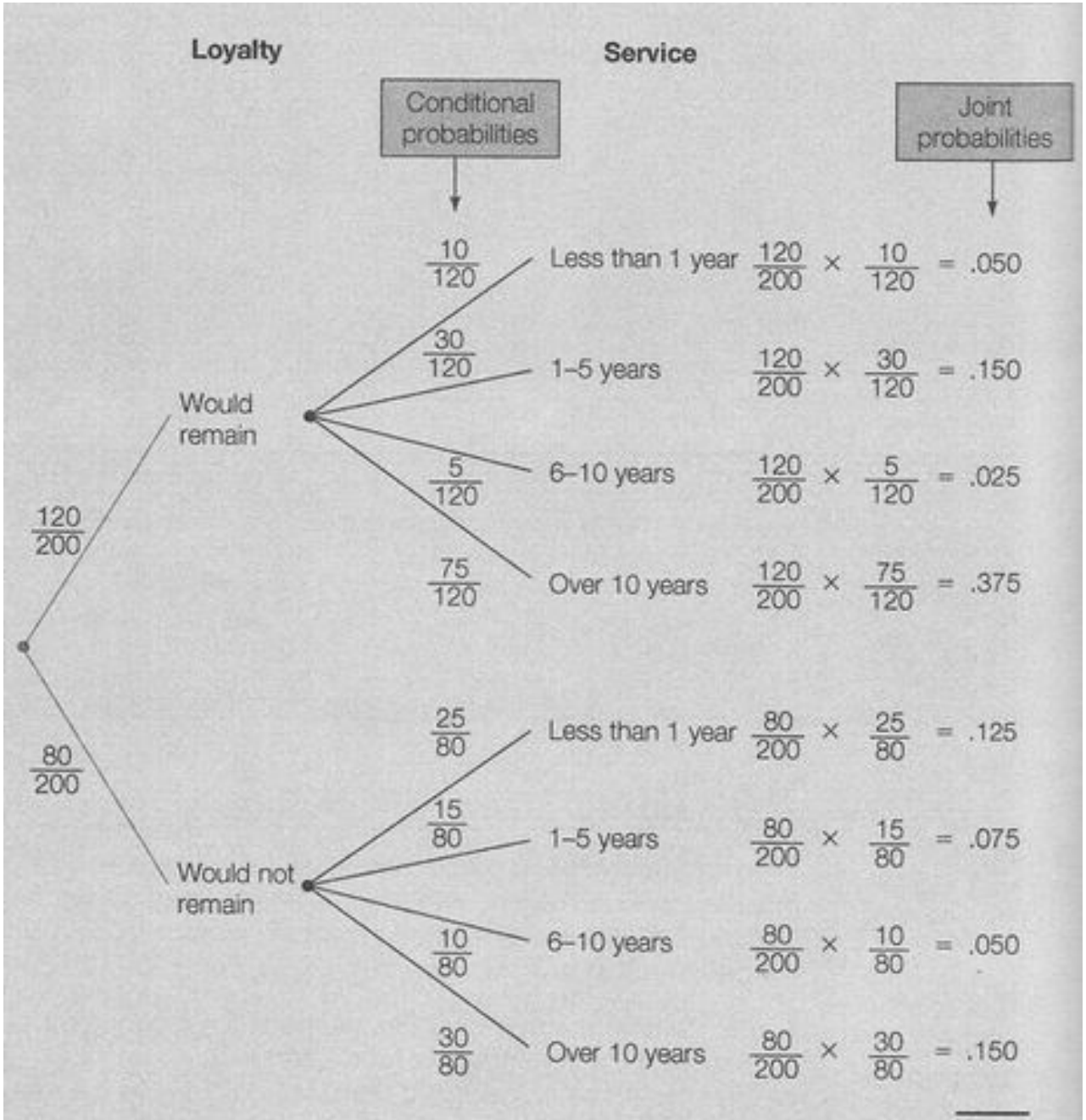
  P(A and B)= P(A)P(B)
- 条件概率：在其它事件发生的条件下，某一特殊事件发生的概率。
- 一般原则：P(A and B)=P(A)P(B|A)

条件概率的表示：列联表（Contingency Table）

将样本观测值根据两个或多个可识别特征进行交叉分类的表格。

	公司工作年限				
忠诚度	少于1年	1-5年	6-10年	10年以上	Total
	B1	B2	B3	B4	
留下 A1	10	30	5	75	120
不留 A2	25	15	10	30	80
Total	35	45	15	105	200

条件概率的表示：概率树（Tree Diagram）



一些概念

概率分布（Probability Distribution）：一个试验的所有可能结果以及每个结果的概率。

随机变量（Random Variable）：一种变量，该变量的结果依机率原则可以取不同的值。

- 离散型随机变量：只能取某些清晰可分的值的随机变量。
- 连续型随机变量：可以取某一区间内所有值的随机变量。

概率分布的均值、方差和标准差

- 均值：$\mu = \Sigma[xP(x)]$
- 方差：$\sigma^2 = \Sigma[(x-\mu)^2 P(x)]$
- 标准差：$\sigma = \sqrt{\Sigma[(x-\mu)^2 P(x)]}$

二项概率分布（Binomial Probability Distributuion）

- 一次试验只有两个可能的互斥结果，例如：成功、失败。
- 二项分布的随机变量是固定试验次数下成功的次数。
- 每次试验成功或失败的概率相同。
- 每次试验之间是相互独立的。

$$P(x) = C_n{}^x \pi^x (1-\pi)^{n-x}$$

均值：$\mu = n\pi$ 方差：$\sigma^2 = n\pi(1-\pi)$

二项概率分布，例子

从Pittsburgh到Bradford，每天有5次航班。每个航班晚点的概率都相同，为0.20。

一天中没有航班晚点的概率是多少？

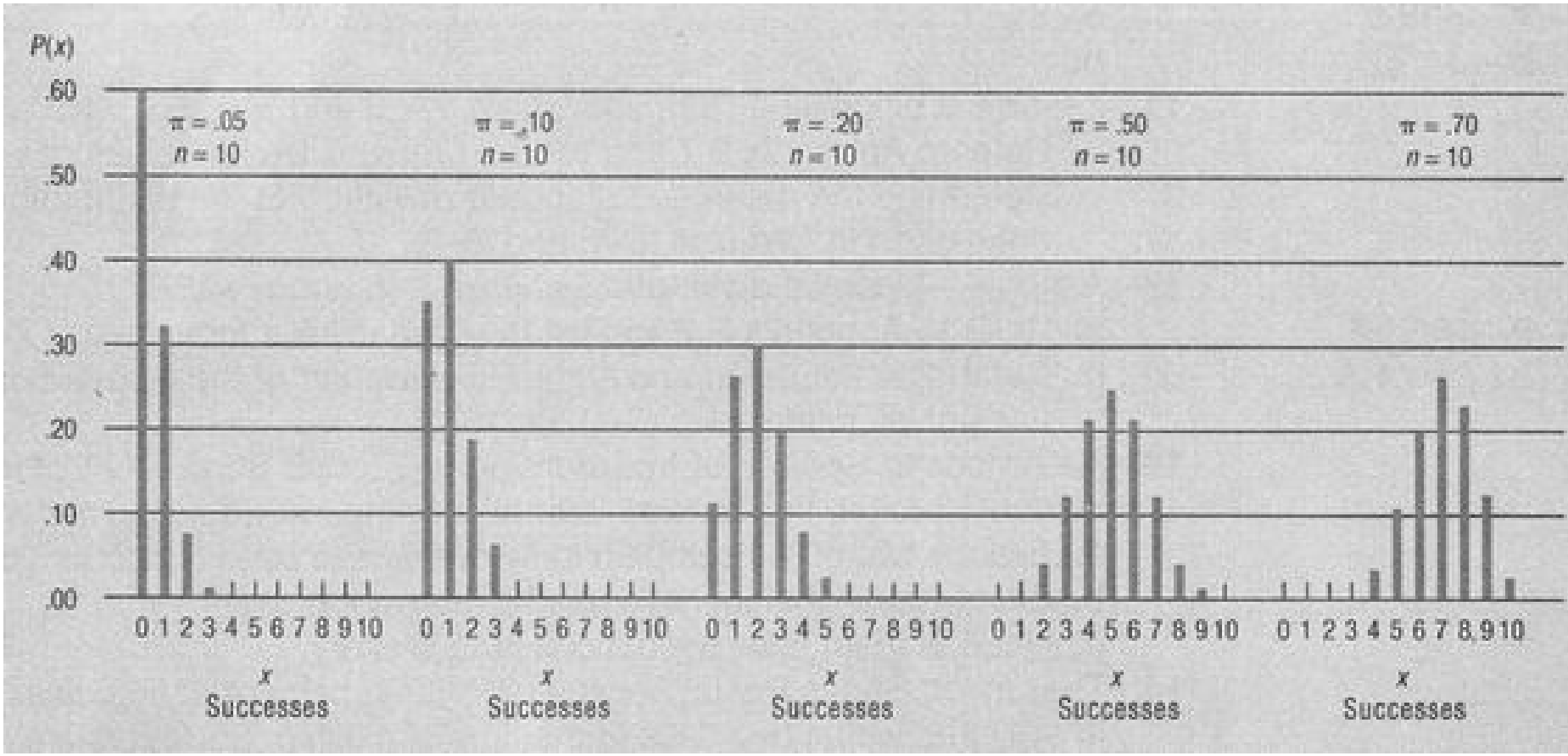
有一个航班晚点的概率是多少？

二项概率表（n=6）

x\π	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95
0	.599	.349	.107	.028	.006	.001	.000	.000	.000	.000	.000
1	.315	.387	.268	.121	.040	.010	.002	.000	.000	.000	.000
2	.075	.194	.302	.233	.121	.044	.011	.001	.000	.000	.000
3	.010	.057	.201	.267	.215	.117	.042	.009	.001	.000	.000
4	.001	.011	.088	.200	.251	.205	.111	.037	.006	.000	.000
5	.000	.001	.026	.103	.201	.246	.201	.103	.026	.001	.000
6	.000	.000	.006	.037	.111	.205	.251	.200	.088	.011	.001
7	.000	.000	.001	.009	.042	.117	.215	.267	.201	.057	.010
8	.000	.000	.000	.001	.011	.044	.121	.233	.302	.194	.075
9	.000	.000	.000	.000	.002	.010	.040	.121	.268	.387	.315
10	.000	.000	.000	.000	.000	.001	.006	.028	.107	.349	.599

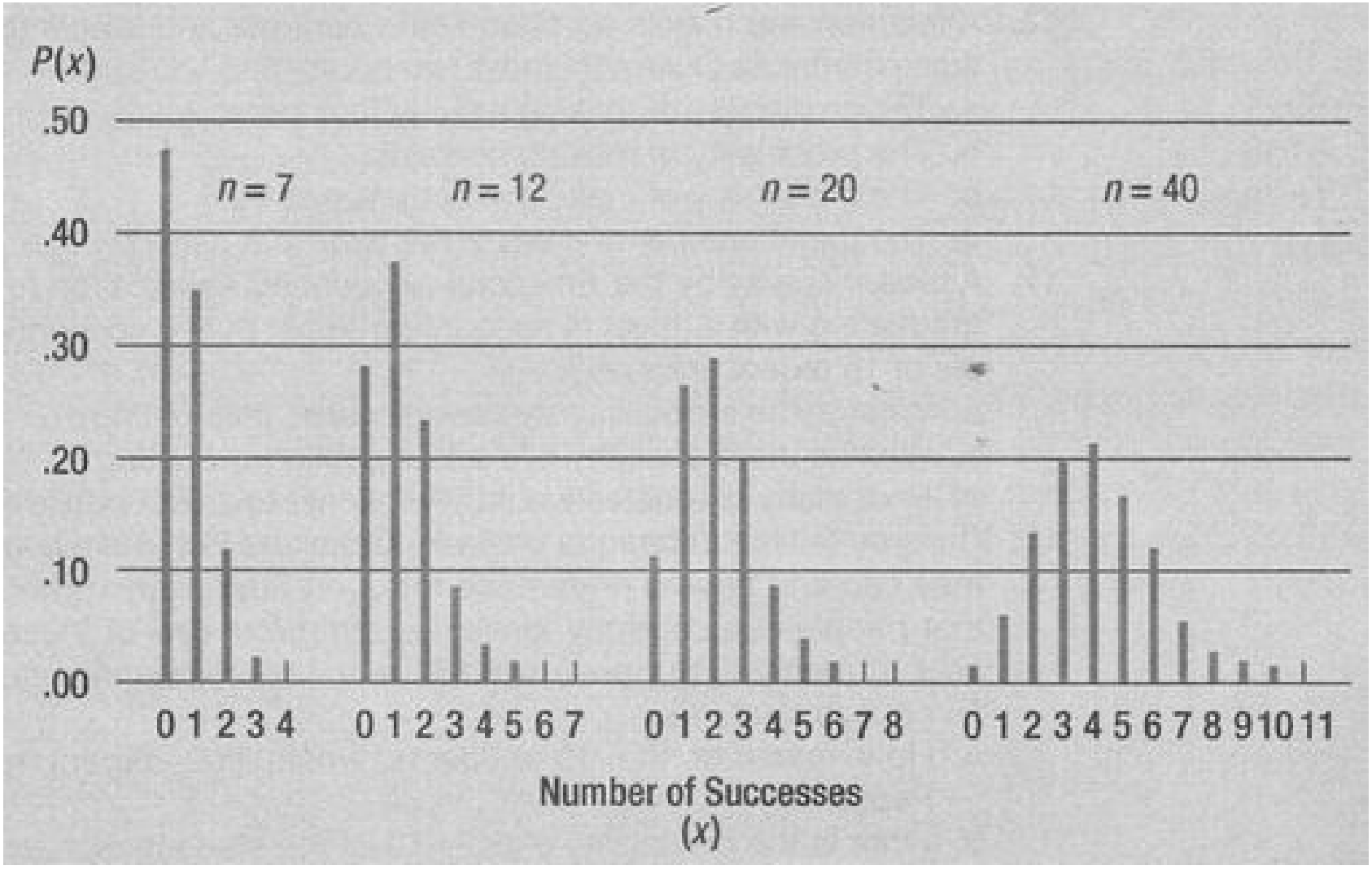
二项概率分布

固定n值，考察π值变动情况。



二项概率分布

固定π值，考察n值变动情况。



泊松概率分布（Poissonl Probability Distributuion）

- 泊松分布的随机变量是一些事件在某一特定区间发生次数的概率分布。
- 事件的概率与区间长度成正比。
- 区间相互独立并且没有交叉。

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

均值：$\mu = n\pi$ 方差：$\sigma^2 = \mu$

泊松分布概率表（n=6）

	μ								
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494
4	0.0000	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

泊松分布与二项分布

例：Costal Insurance Company 为海岸线上的一些房子承保。保险公司估计每年出现三级以及三级以上海啸的概率为0.01。一个海岸线上的房子如果贷款30年，那么30年内遭受一次以上海啸的概率是多少?

用泊松分布求解：$\mu = n\pi = 30 * (0.10) = 0.30.$

$P(x \geqslant 1) = 1 - P(x = 0) = 1 - \frac{\mu^0 e^{-0.3}}{0!} = .2592$

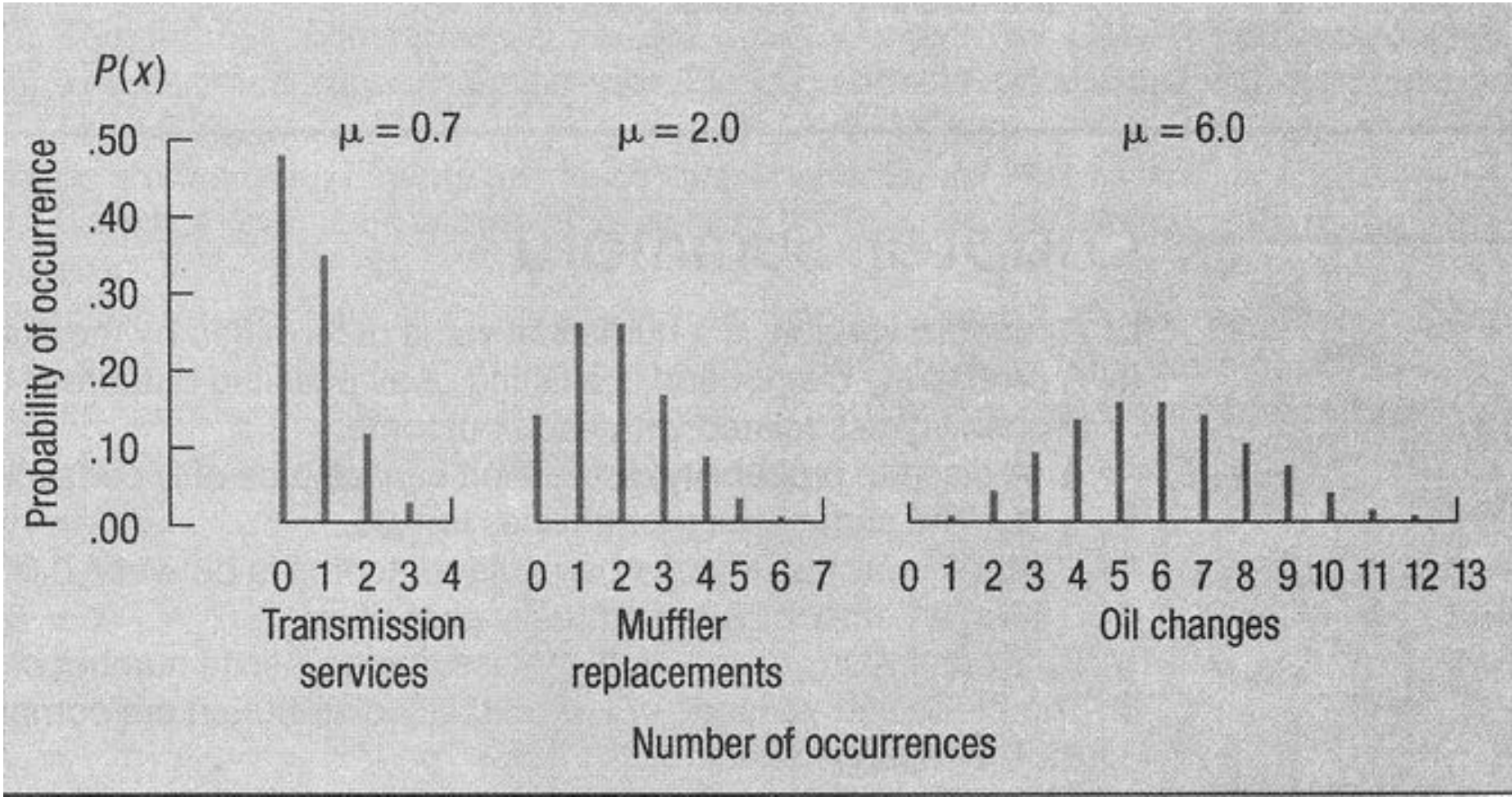
用二项分布求解：

$P(x \geqslant 1) = 1 - P(x = 0) =$

$1 - C_3 0^0 (.01)^0 (.99)^3 0 = .2603$

泊松概率分布

考察μ值变动情况。



均匀概率分布（Uniform Probability Distribution）

$$P(x) = \frac{1}{b-a}$$

均值：$\mu = \frac{a+b}{2}$

标准差：$\sigma = \sqrt{\frac{(b-a)^2}{12}}$

面积：$Area = \frac{1}{(b-a)}(b-a)$

均匀概率分布：例子

公共汽车每30分钟一班，在30分钟期间到达汽车站的人服从均匀分布。

- 画出分布图。
- 计算均匀分布的面积。
- 等待的平均时间是多少？等待时间的标准差是多少？
- 等待时间超过25分钟的概率？
- 等待10-20分钟的概率？

正态概率分布（Normal Probability Distribution）

钟形分布、对称、渐近于0、形状取决于均值和方差。

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-[\frac{(x-\mu)^2}{2\sigma^2}]}$$

标准正态分布： $z = \frac{x-\mu}{\sigma}$

$$P(z) = \frac{1}{\sqrt{2\pi}} e^{-[\frac{z^2}{2}]}$$

经验法则。

抽样方法

- 简单随机抽样（Simple Random Sampling）；
- 系统随机抽样（Systematic Random Sampling）；
- 分层随机抽样（Stratified Random Sampling）；
- 整群抽样（Cluster Sampliing）。

简单随机抽样

保证总体中的每个个体具有相同的机率被抽中，用这种方法抽选出来的样本称作简单随机样本。

即：N个总体中随机抽取n个样本，或者在总体中不放回地抽取n次。

- 抽签法；
- 随机数字表示法。

简单随机抽样：例子

一个小旅馆Foxtrot Inn，有8间房。

2007年6月每天出租的房屋数如下：

June	Rentals	June	Rentals	June	Rentals
1	0	11	3	21	3
2	2	12	4	22	2
3	3	13	4	23	3
4	2	14	4	24	6
...	...	...	...	...	...
10	7	20	2	30	3

系统随机抽样

若总体中的抽样单元都按一定顺序排列，在规定范围内随机抽取一个单元作为初始单元，然后按照一套事先定好的规则确定其他单元。

等距抽样：首先从总体中随机地选择第一个被抽选的样本单位，然后每隔k个单位从总体中抽选一个样本单位。

分层随机抽样

当总体能够根据某特性分成若干组时，我们可以用分层随机抽样。首先将总体分成若干类或若干层，然后从每一类或每一层中再随机抽选样本。

例如：以美国352家最大公司为总体，研究股本回报率（Return on Equity, ROE）与投资到广告的费用是否成正比。

分层随机抽样

Table: 1. 分层随机抽样所选取的样本个数

分层	ROE	企业个数	相对频率	抽样个数
1	30%以上	8	0.02	1
2	20%-30%	35	0.10	5
3	10%-20%	189	0.54	27
4	0-10%	115	0.33	16
5	赤字	5	0.01	1
Total	-	352	1.00	50

整群抽样

运用自然存在的、地理的或其他的某种特征或边界，将总体划分成不同的整群，然后从整群中随机地选择若干整群，最后再从已被选中的各个整群中随机地选择样本单位。

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 BB。

样本均值的抽样分布

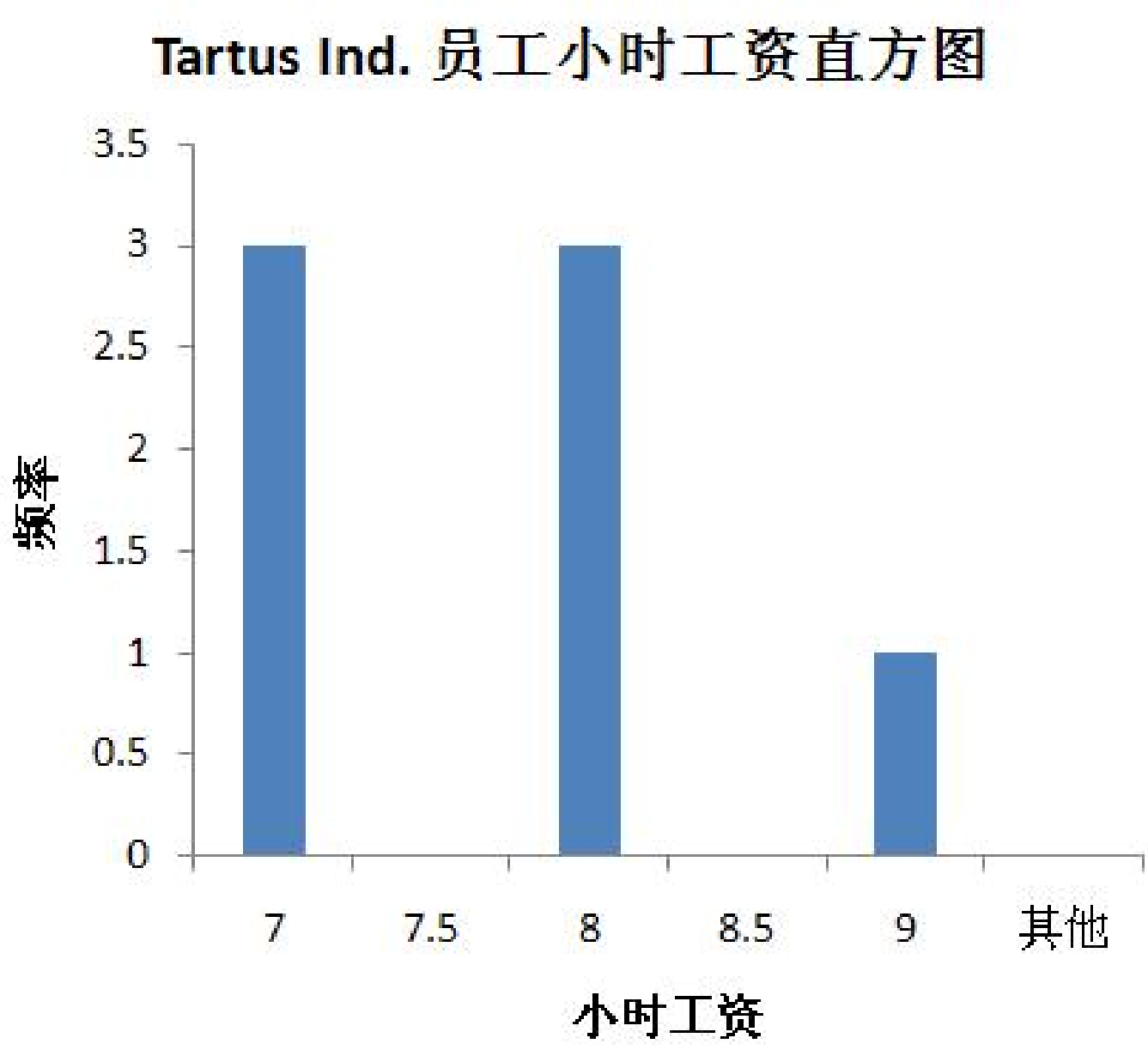
给定样本容量条件下所有可能的样本均值的概率分布，称作样本均值的概率分布。

例：Tartus Industries 7名员工每小时工资如下：

Table: 2.

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		

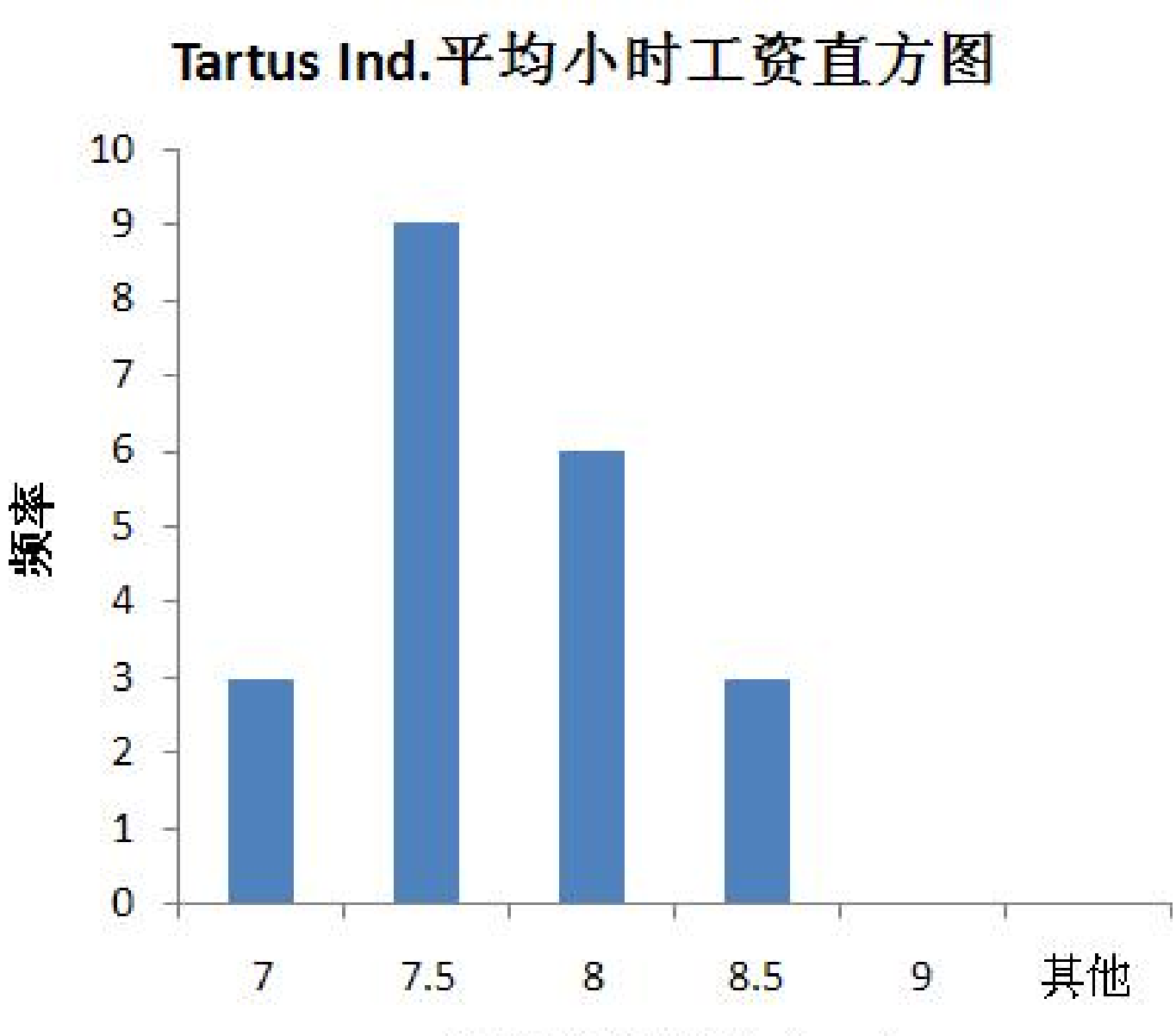
Tartus Ind. 员工小时工资直方图



Tartus Ind. 员工小时工资n=2时24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



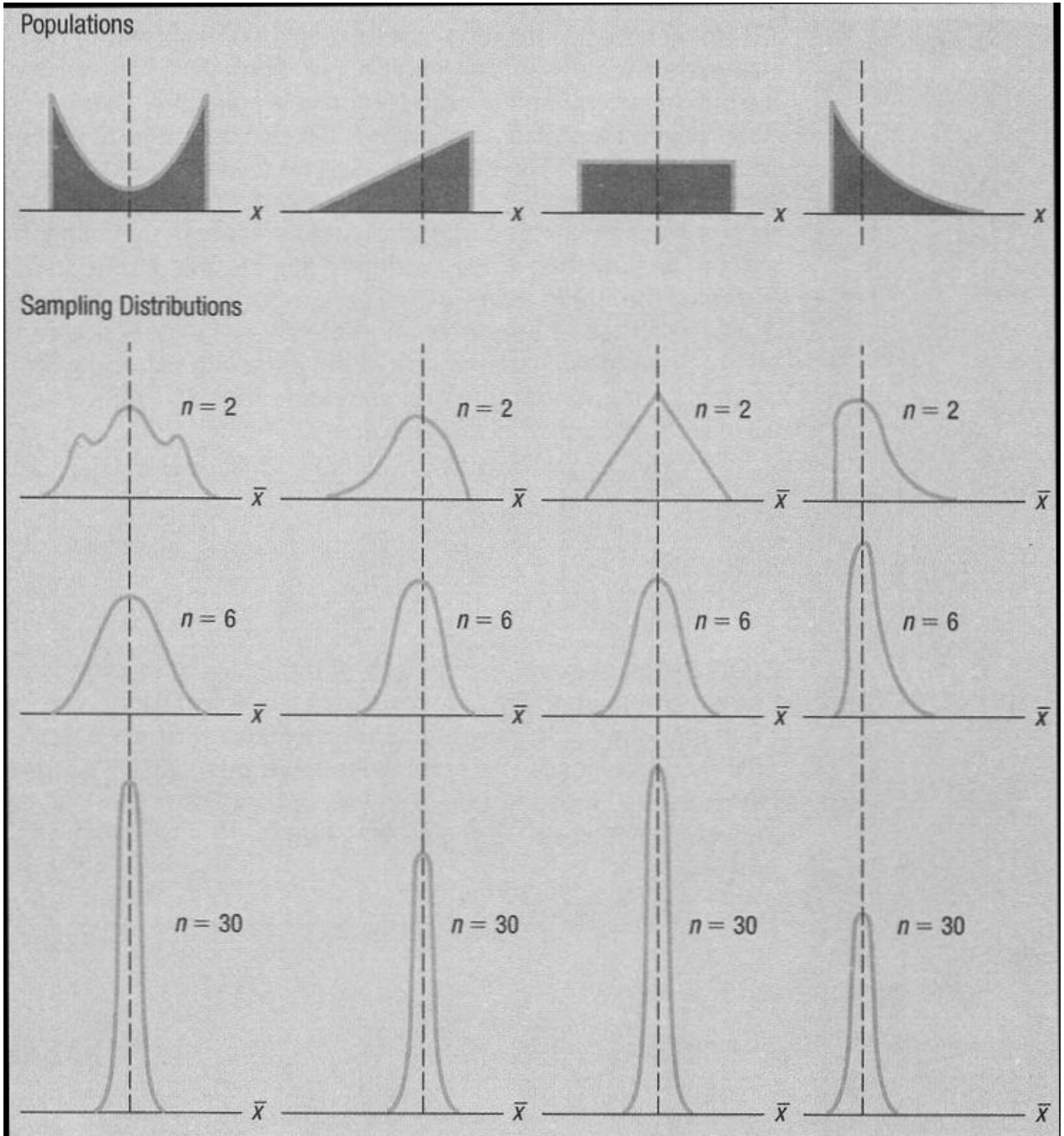
样本均值的抽样分布

- 样本均值的均值与总体均值完全相等；
- 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。

- 总体为正态分布，则$\bar{X}$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。

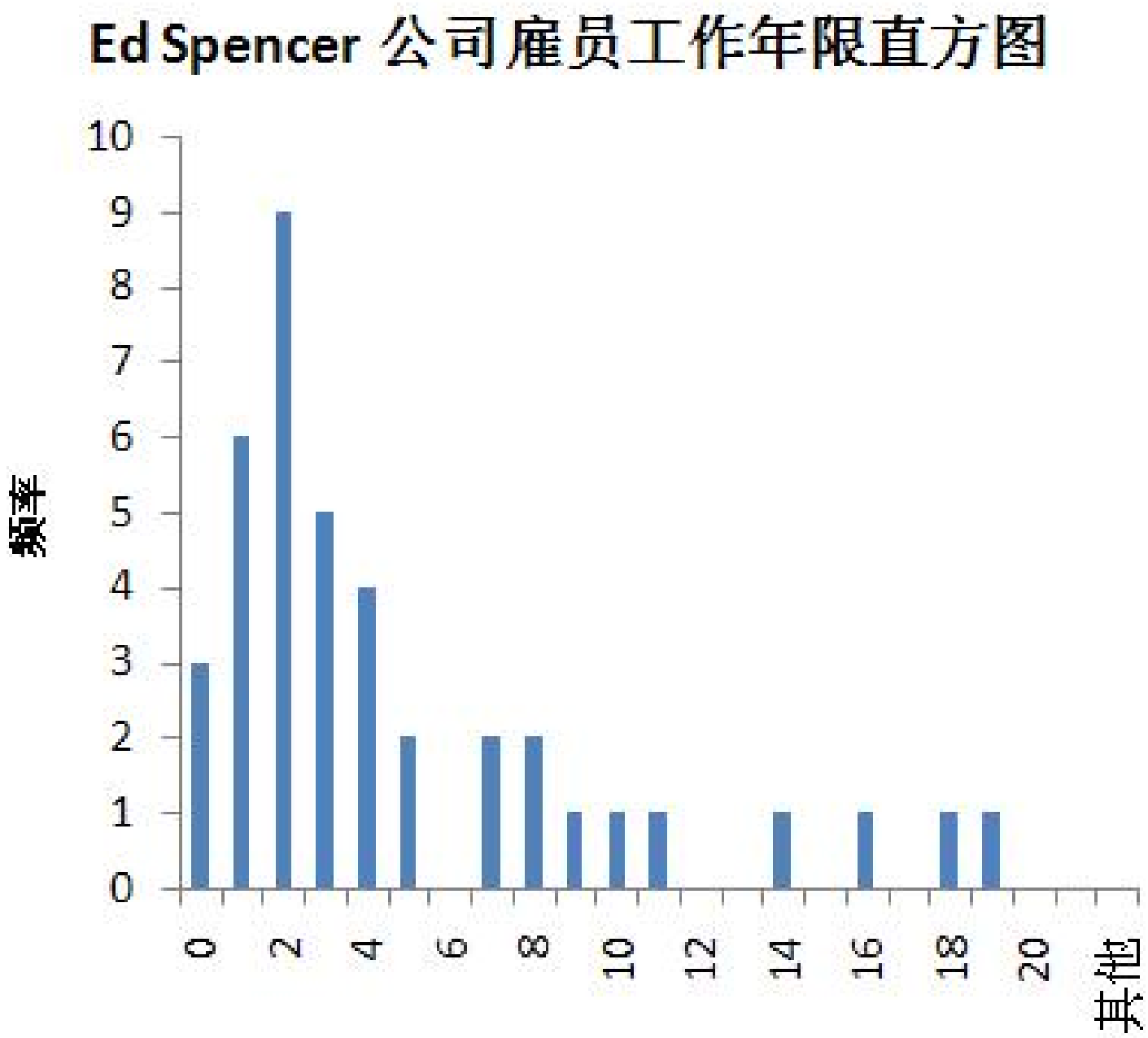

Populations
x
x
x
x
Sampling Distributions
n = 2
n = 2
n = 2
n = 2
n = 6
n = 6
n = 6
n = 6
n = 30
n = 30
n = 30
n = 30
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄
x̄

例子：Ed Spence公司

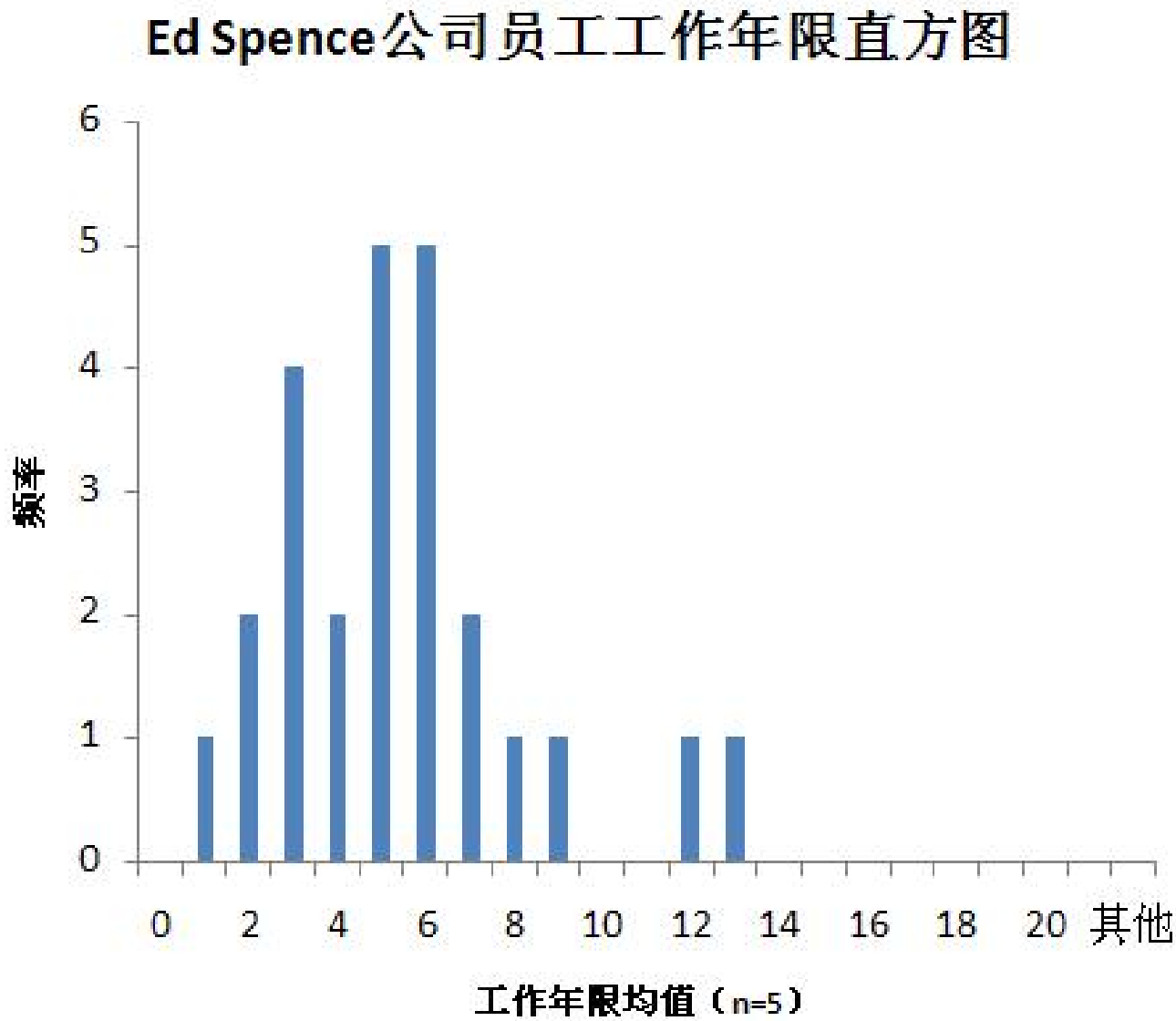
Table: 3.Ed Spence公司员工工作年限：

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

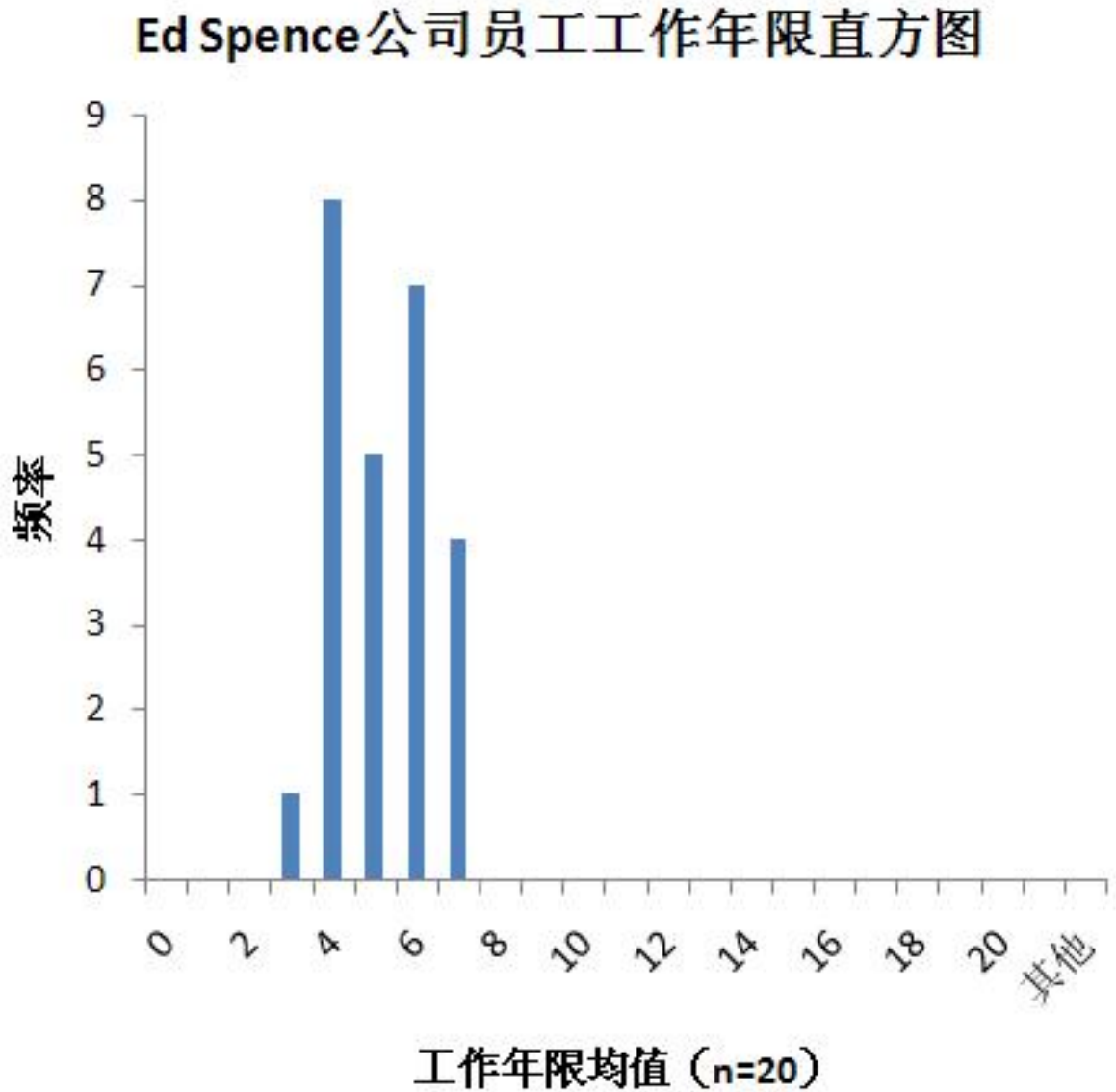
例子：Ed Spence公司员工工作年限分布



例子：Ed Spence公司员工工作年限样本均值（n=5）抽样分布



例子：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）

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