上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

上海交通大学生物医学工程学院
SJTU School of Biomedical Engineering

生物医学统计概论(BI148-2)

Fundamentals of Biomedical Statistics

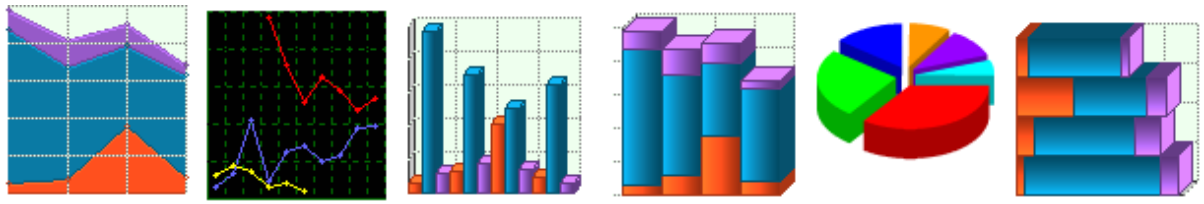
2022 春季

授课老师：林关宁

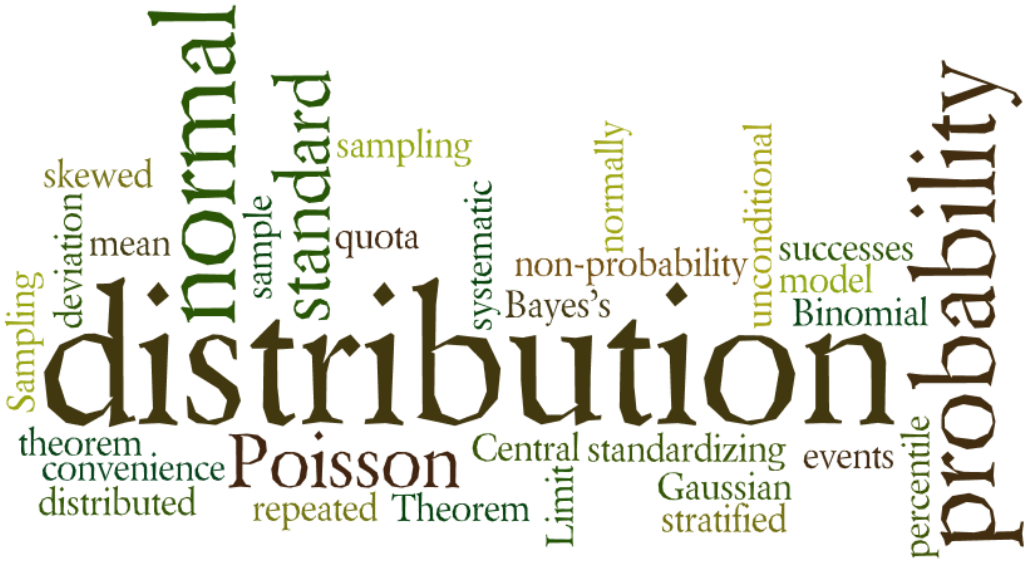
课程内容安排

上课日期	章节	教学内容	教学要点	作业	随堂测	学时
2.16	1	数据可视化，描述性统计	1. 课程介绍 & 数据类型	作业1（8%）	测试1（8%）	2
2.23			2. 描述性统计Descriptive Statistics & 数据常用可视化			2
3.2			3. 常用概率分布			2
3.9			4. 大数定理 & 中心极限定理			2
3.16	2	推断性统计，均值差异检验	5. 统计推断基础-1：置信区间 Confidence Interval *	作业2（12%）	测试2（12%）	2
3.23			6. 统计推断基础-2:假设检验 Hypothesis Test			2
3.30			7. 数值数据的均值比较-1： 单样本t-检验			2
4.6			8. 数值数据的均值比较-2： 独立双样本t-检验，配对样本t-检验			2
4.13			9. 数值数据的均值比较-3：One-Way ANOVA			2
4.20			10. 数值数据的均值比较-4：Two-way ANOVA			2
4.27	3	比例差异检验	11. 类别数据的比例比较-1： 单样本比例推断 *	作业3（4%）	测试3（4%）	2
5.7（调）			12. 类别数据的比例比较-2： 联立表的卡方检验			2
5.11	4	协方差，相关分析，回归分析	13. 相关分析（Pearson r, Spearman rho, Kendal’s tau) *	作业4 （6%）	测试4 （6%）	2
5.18			14. 简单回归分析			2
5.25			15. 多元回归Multiple Regression			2
6.1	5	Course Summary	16. 课程总结 *			2
			Total	30%	30%	32

*** 随堂测试**



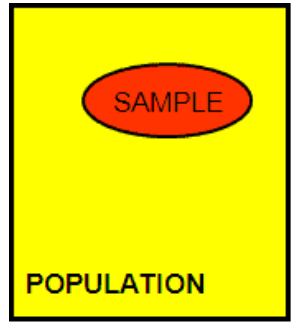
Week 4

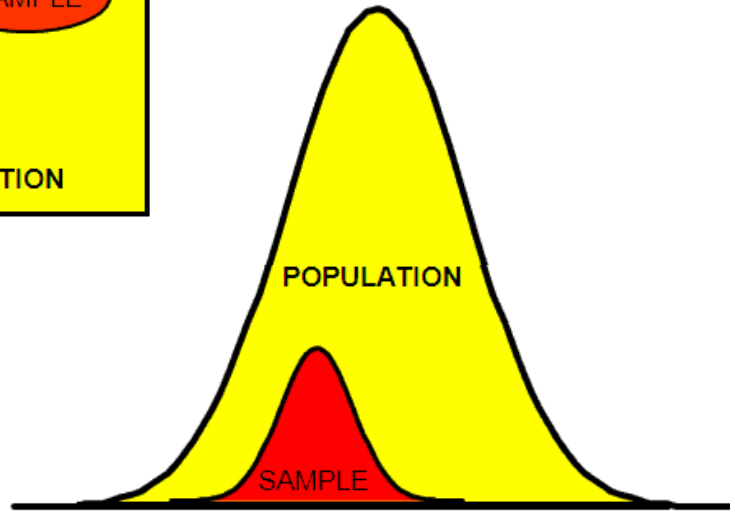

normal
distribution
probability
standard
sampling
skewed
deviation
mean
sample
quota
systematic
normally
non-probability
successes
unconditional
model
Bayes's
Binomial
Sampling
theorem
convenience
distributed
Poisson
repeated
Central
Theorem
Limit
standardizing
Gaussian
stratified
events
percentile

常用生物医学统计术语

(1) 总体与样本

- 研究对象的全体称为**总体**（population)
- 总体中的一个研究单位称为**个体**（individual)
- 总体的一部分称为**样本**（sample)

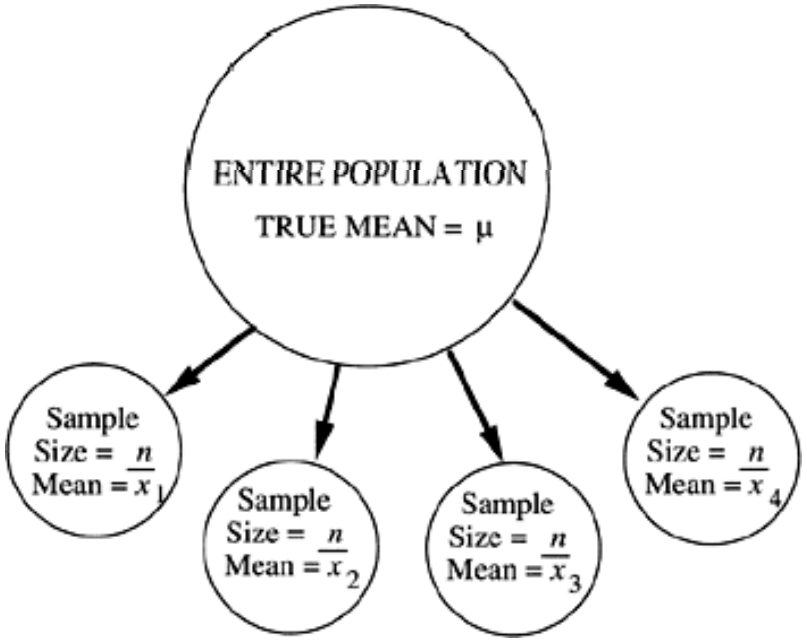


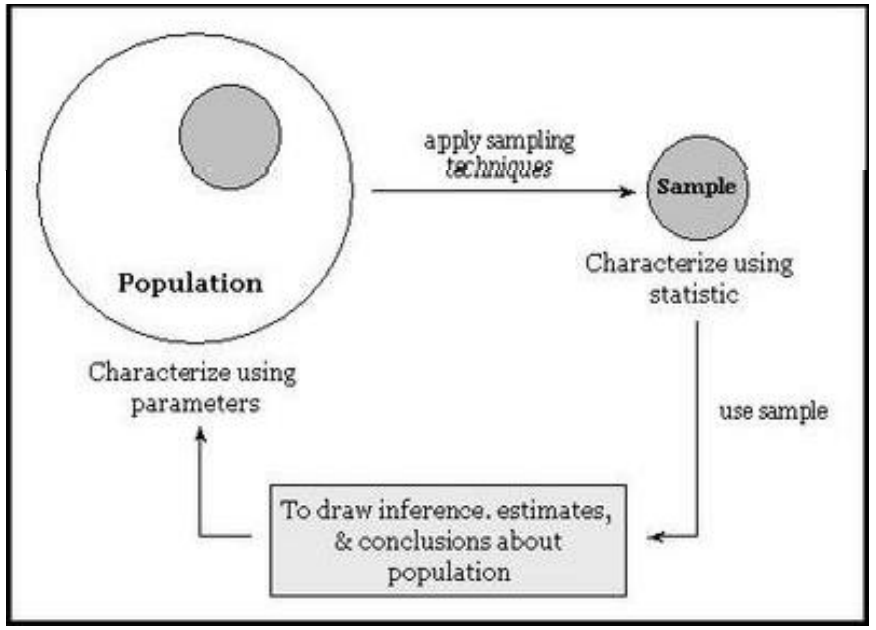


- 个体数有限的总体称为有限总体
- 个体数无限的总体叫无限总体
- 假想总体：把所进行的实验看成是假想总体的一个样本

常用生物医学统计术语

- 样本中所包含的个体数目叫样本容量或**样本大小** (sample size)，记为n。通常把 $n \leq 30$ 的样本叫小样本，$n > 30$ 的样本叫大样本
- 科学研究目的是了解总体，而能观测到的却是样本，通过样本推断总体是统计分析的基本特点

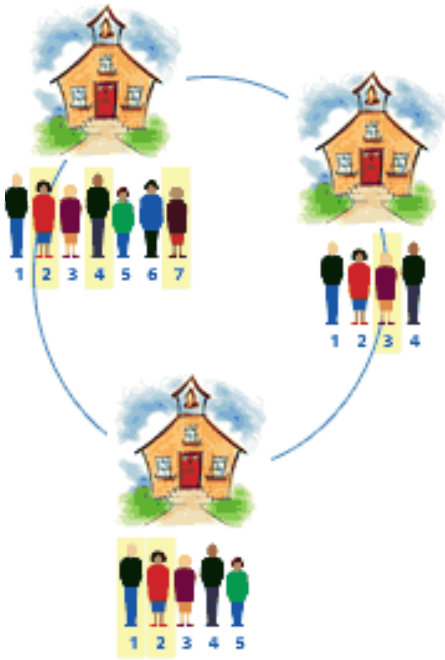




常用生物医学统计术语

可靠地从样本推总体，要求样本具有一定的含量和代表性。从总体随机抽取的样本才具有代表性。

随机抽样（random sampling）：总体的每一个体都有同等被抽取的机会。



Total Sample = Those Randomly Selected from Each Schoool

通过样本推断总体不可能是百分之百正确，有一定的错误率，这是统计分析的又一特点。

常用生物医学统计术语

从有限总体抽样分为放回式抽样和非放回式抽样。

从总体中抽出一个个体，记下特征后放回总体再做第二次抽 样，这种方式抽取的样本为随机样本。抽样的概率不变。

从总体中抽出个体后不再放回，抽取的样本为非随机样本， 每次抽样的概率发生了变化。

随机抽样的方法

- 抽签、拈阄（Draw lots)、抛钱币（Tossing）
- 使用随机数字表。随机数字表是采用完全随机化方法制成，在随机抽样、随机化实验设计中应用很广。(课后阅读)

常用生物医学统计术语

（2）参数与统计量

为了表示总体和样本的数量特征，需计算几个特征数。

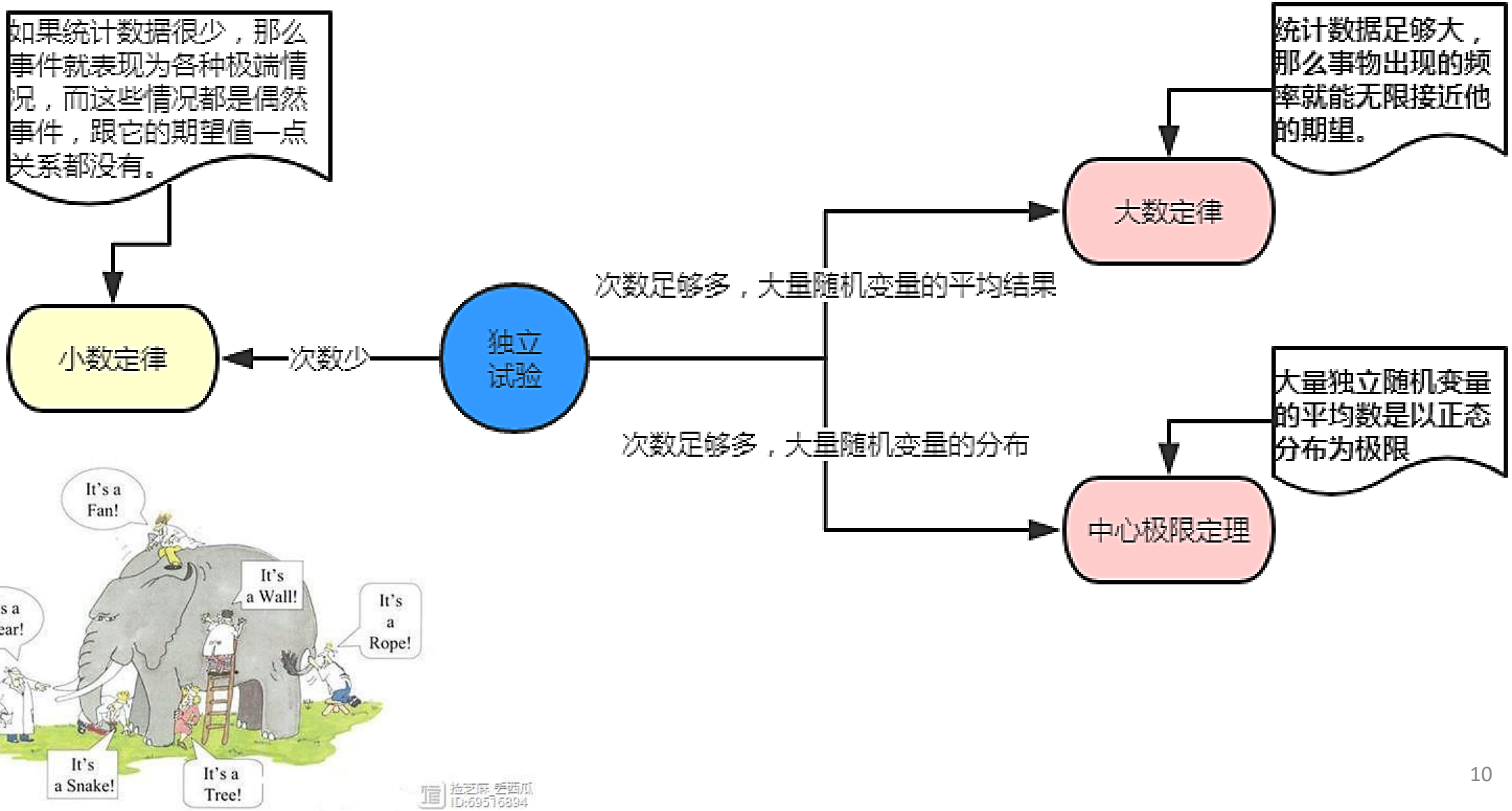
总体的特征数叫参数(parameter)，用希腊字母表示，例如用 μ 表示总体平均数，用 σ （sigma）表示总体标准差，σ^2 总体方差。

样本的特征数叫统计量（statistics)。用拉丁字母表示，例如用 $\bar{X}$ 表示样本平均数，用S表示样本标准差，S^2样本方差。个体的某一性状、特性的测定数值叫做观察值 (Observation) ,以 x_1，x_2，…… x_i …… x_n 表示。

同一总体内每一个体的观察值有变异。

极限定理

- 极限定理是关于随机变量序列极限特性的一簇定理的总称
- 有**大数定律（law of large number）**和**中心极限定理（central limit theorem CLT）**两大最基本的类型
- 前者用于**描述平均结果和频率**的稳定性
  - 阐明**大量重复实验的平均结果具有稳定性**的一系列定律都称为大数定律
- 后者用于**描述分布**的稳定性
  - 论证随机变量（实验结果）**之和渐进服从某一分布**的定理称为极限定理


如果统计数据很少，那么事件就表现为各种极端情况，而这些情况都是偶然事件，跟它的期望值一点关系都没有。
小数定律
次数少
独立试验
次数足够多，大量随机变量的平均结果
大数定律
统计数据足够大，那么事物出现的频率就能无限接近他的期望。
次数足够多，大量随机变量的分布
中心极限定理
大量独立随机变量的平均数是以正态分布为极限
It's a Fan!
It's a Wall!
It's a Spear!
It's a Rope!
It's a Snake!
It's a Tree!

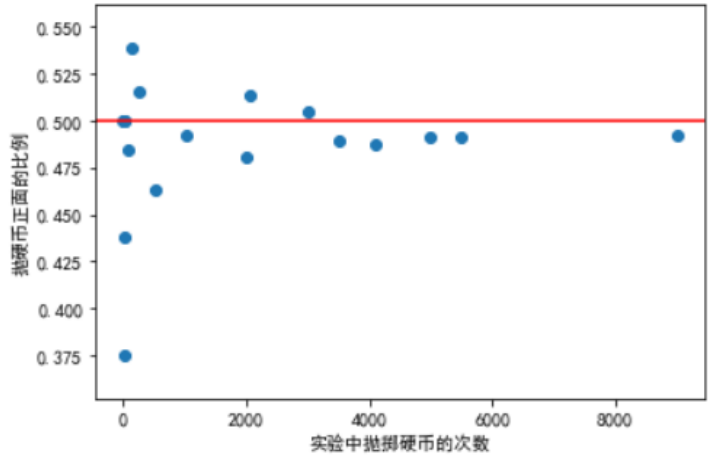
1. The Law of Large Numbers（大数定律）

对于理论，维基百科有这样的介绍：大数定律又称大数法则、大数律，是描述相当多次数重复实验的结果的定律。根据这个定律知道，样本数量越多，则其算术平均值就有越高的概率接近期望值。在此我们利用python直观演示：

:
```
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt
import matplotlib as mpl
import seaborn as sns
mpl.rcParams['font.sans-serif'] = ['SimHei']
mpl.rcParams['font.family']='sans-serif'
plt.rcParams['axes.unicode_minus']=False

# 每次实验翻转硬币次数
ns = np.array([2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2000,2048, 3000,3500,4096,5000,5500,9000])
# 使用二项式函数来模拟一枚均匀的硬币,计算每次实验出现为正面的次数
heads_count = [np.random.binomial(n, 0.5) for n in ns]
# 计算硬币翻转为正面的占比
proportion_heads = heads_count/ns

fig, ax = plt.subplots()
plt.xlabel('实验中抛掷硬币的次数')
plt.ylabel('抛硬币正面的比例')
plt.axhline(0.5, color='r')
_ = ax.scatter(ns, proportion_heads)
```



从上图不难发现，当掷硬币的次数越多，翻转得到为正面的数量，越接近掷的总次数的一半，也就是越接近一枚均匀硬币的期望值0.5

案例1：
某个人乘飞机遇难，概率不可预料，对于他个人来说，飞机失事具有随机性。但是对每年100万人次所有乘机者而言，这里的100万人可以理解这100万次的重复试验，其中，总有10人死于飞行事故。**那么根据大数定律，乘飞机出事故的概率大约为十万分之一。**
这就为保险公司收取保险费提供了理论依据。对个人来说，出险是不确定的，对保险公司来说，众多的保单出险的概率是确定的。根据大数定律的定律，承保危险的单位越多，损失概率的偏差越小，反之，承保危险的单位越少，损失概率的偏差越大。
因此，保险公司运用大数定律就可以比较精确地预测危险，合理保险费率。

案例2：
有时候，在各种不知情的情况下人们的个人信息被莫名其妙地泄露。于是铺天盖地的骚扰、诈骗短信、盗号等不停地骚扰着我们的生活，只要你有手机、电话、电脑，就注定无处可逃。
有时候，人们不禁会质疑，这种像无头苍蝇一样碰运气的骗子，用这种愚蠢、低级的骗术，真的会有人上当吗？
但骗子的行为却符合“大数定律”：只要发出的诈骗短信量足够大，上当受骗的概率就会稳定在某个值附近做极小波动。
有人曾做过这样一个有趣的统计：每发出一万条这样的诈骗短信，受骗的人就有七八个，非常稳定。

中心极限定理

有时候统计概率就像魔术一样，能够从少量数据中得出不可思议的强大结论。

我们只需要对1000个美国人进行电话调查，就能去预测美国总统大选的得票数。

通过对为肯德基提供鸡肉的加工厂生产的100块鸡肉进行病毒（沙门氏菌）检测，就能得出这家工厂的所有肉类产品是否安全的结论。

这些“一概而论”的强大能力，到底是从哪里来的?

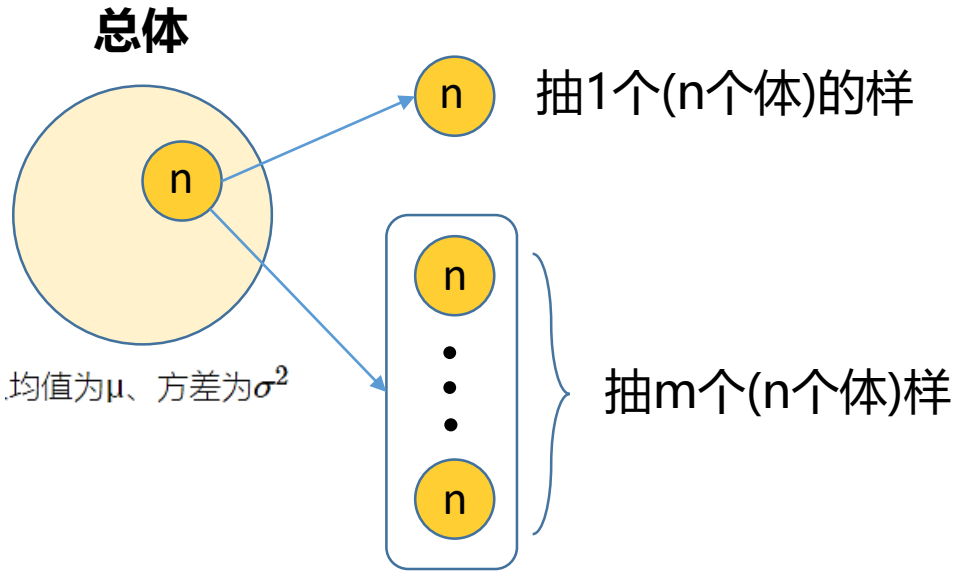
背后的秘密武器就是统计概率的第2大护法：中心极限定理
（第1大护法就是：大数定律）

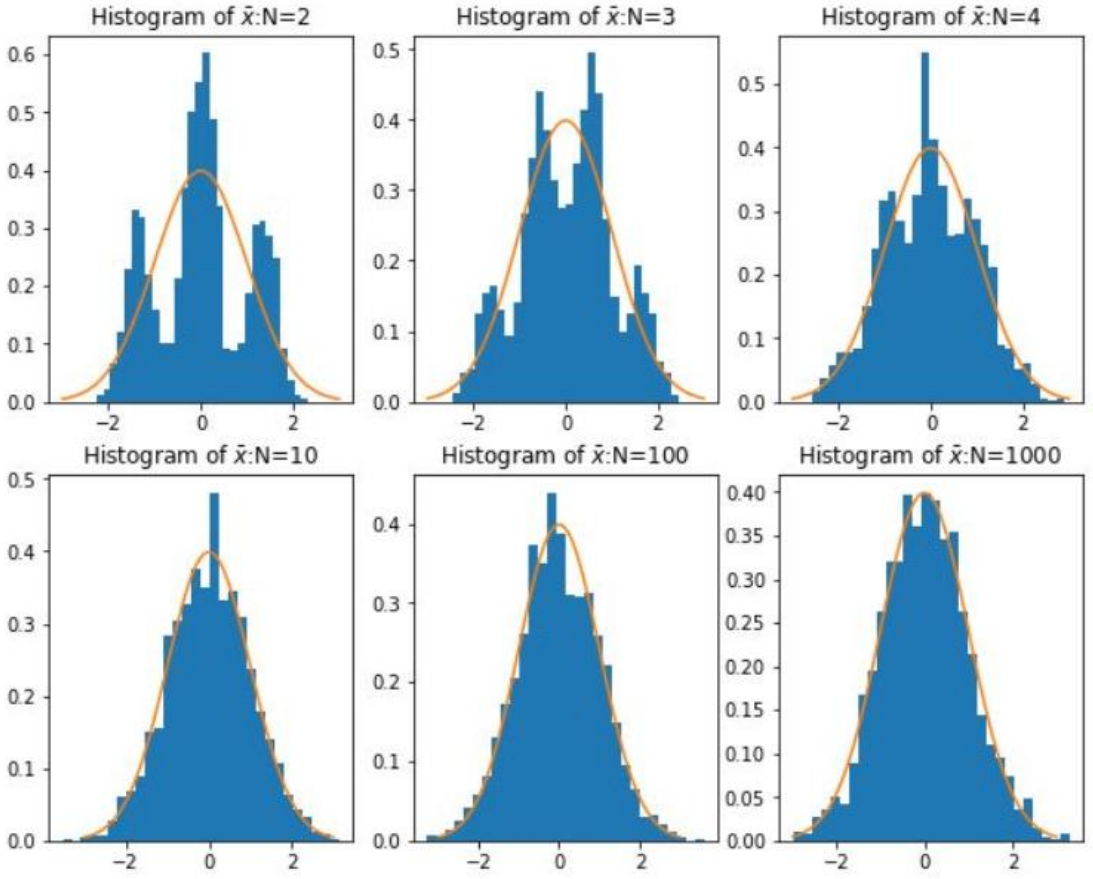
中心极限定理是许多统计活动的“动力源泉”，这些活动存在着一个共同的特点，那就是**使用样本对总体进行估计**，例如我们常看到的民意调查就是这方面的经典案例。

中心极限定理

在适当的条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于正态分布。每次从这些总体中随机抽取 n 个抽样，一共抽 m 次。 然后把这 m 组抽样分别求出平均值， 这些平均值的分布接近正态分布。设从均值为μ、方差为σ^2（有限）的任意一个总体中抽取样本量为n的样本，当n充分大时，样本均值$\overline{X}$的抽样分布近似服从均值为μ、方差为$\frac{\sigma^2}{n}$的正态分布。

中心极限定理告诉我们，当样本量足够大时，样本均值的分布慢慢变成正态分布，就像下图：



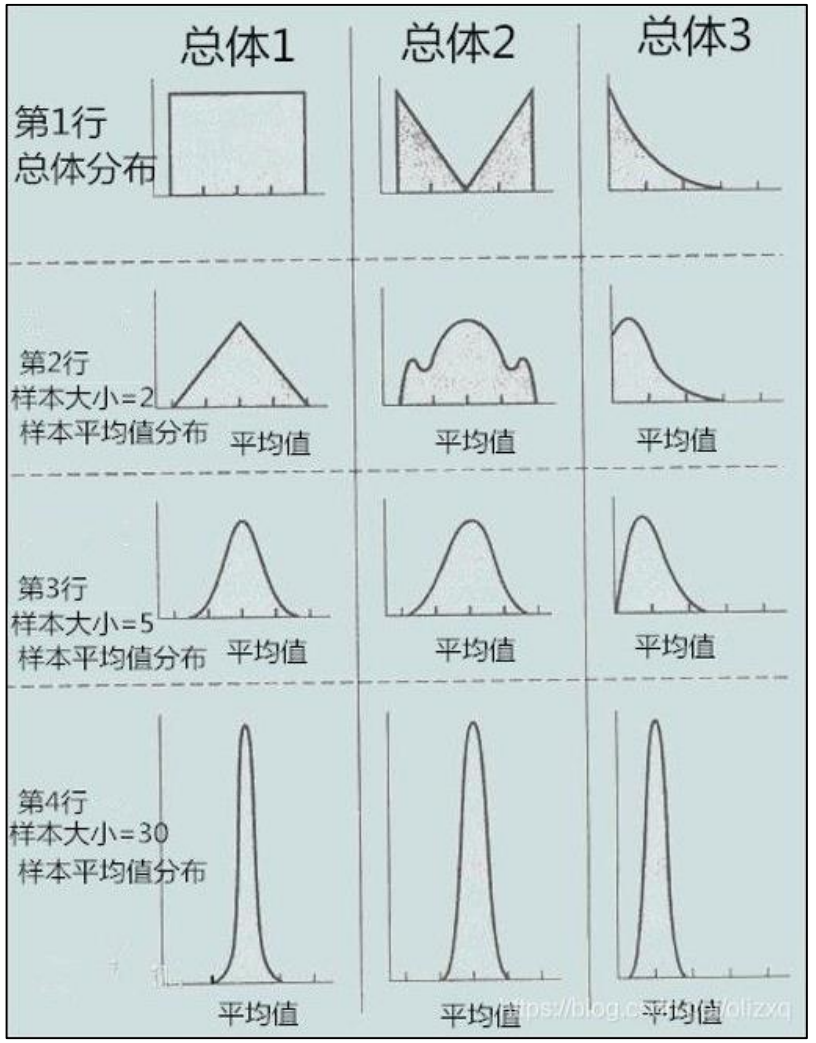


中心极限定理是说：

- **样本的均值约等于总体的平均值。**
- **不管总体是什么分布，任意一个总体的样本平均值都会围绕在总体的整体平均值周围，并且随着抽取样本容量n的增加呈正态分布。**

如右图：

- 这里第1行是3种不同分布类型的总体，用于比较不同类型下的样本均值的分布;
- 第2行每个样本大小是2，然后对每个样本求平均值，横轴表示每个样本的平均值，纵轴表示该平均值出现了多少次，最后平均值分布很不规则;
- 第3行每个样本大小是5，然后对每个样本求平均值，最后平均值分布有点接近于正态分布，但是总体3对应的第3行却不是正态分布;
- 第4行每个样本大小是30，然后对每个样本求平均值，最后平均值分布是正态分布。

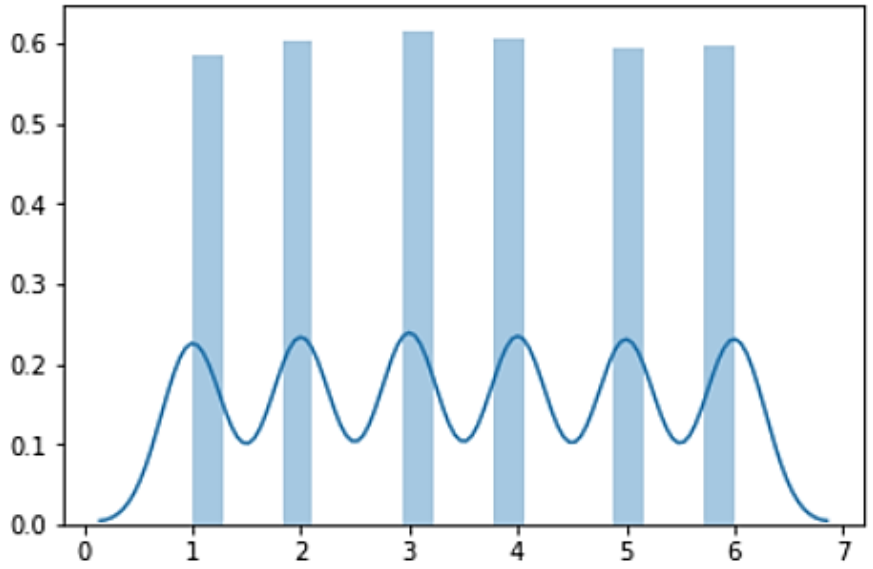


下面我们来通过实例来看看一个掷骰子的平均分布，如何变成一个正态分布

2.1 生成均匀分布的掷骰子随机数

生成在半开半闭区间[low,high)上离散均匀分布的整数值，即1-6的均匀分布

```
1.  import numpy as np
2.  import seaborn as sns
3.  data = np.random.randint(1, 7, 10000)
4.  sns.distplot(data)
```



可以看到1-6的点数是比较均匀的分布的【注意，每一次运行的图都不一样 】

下面我们来通过实例来看看一个掷骰子的平均分布，如何变成一个正态分布

2.2 抽取一组数据

通过以下程序来从data中随机抽取一组数

```
sample1 = []
for i in range(0, 10):
    rnd = int(np.random.random() * len(data))#0-9999的随机数生产
    sample1.append(data[rnd])
print(sample1)
```

返回结果：[2, 3, 2, 1, 1, 4, 2, 1, 5, 3]

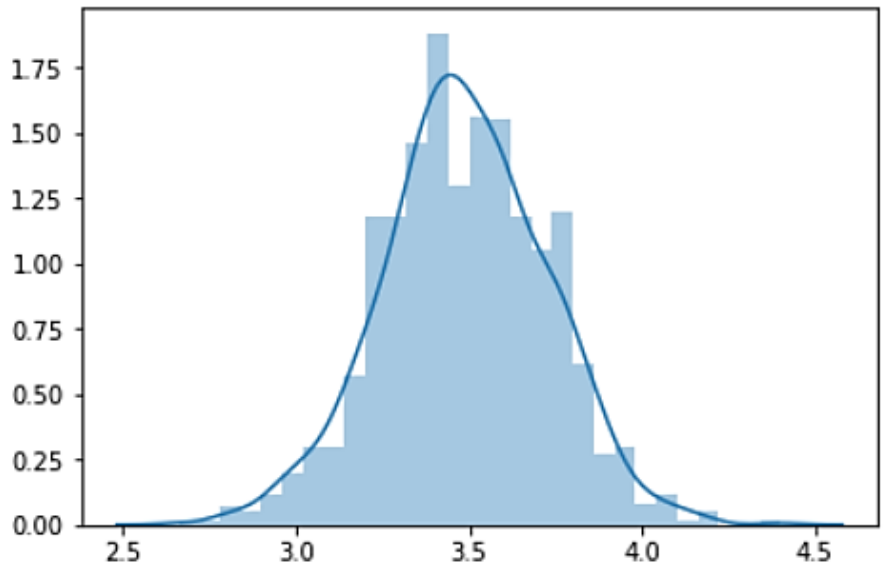
2.3 抽取1000组数据

我们在生产的随机数中，一次抽取50个作为一组并计算它的平均值，共抽取1000次，得到1000个平均值，然后通过seaborn的distplot看着1000个数值的分布。

```
samples = []
s_mean = []

for i in range(0, 1000):
    sample = []
    for j in range(0, 50):
        rnd = int(np.random.random() * len(data))#0-9999的随机数生产
        sample.append(data[rnd])#循环50次的结果存放在samples
    s_mean.append(np.mean(sample))#得到的50次sample的平均数放到s_mean后，重新循环第二次的50次循环，samples清空

sns.distplot(s_mean)
```



注意：

中心极限定理是许多统计活动的“动力源泉”，这些活动存在着一个共同的特点，那就是**使用样本对总体进行估计**，例如我们常看到的民意调查就是这方面的经典案例。

当采样的数量接近无穷大时，我们的抽样分布就会近似于正态分布。这个统计学基础理论意味着我们能根据个体样本推断所有样本。结合正态分布的其他知识，我们可以轻松计算出给定平均值的值的概率。**在理论上保证了我们可以用只抽样一部分的方法，达到推测研究对象统计参数的目的。**

其中要注意的几点：

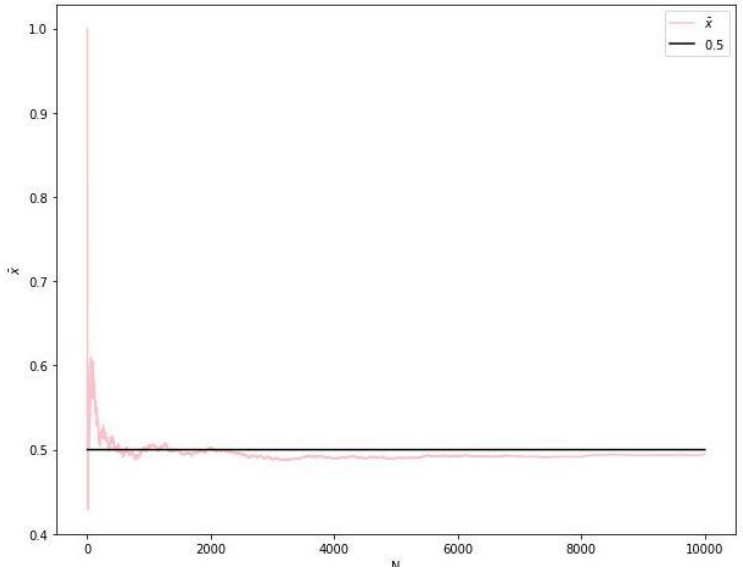
1.总体本身的分布不要求正态分布：掷一个骰子是平均分布，最后每组的平均值也会组成一个正态分布。

2.样本每组要足够大，但也不需要太大：取样本的时候，一般认为每组大于等于**30**个，即可让中心极限定理发挥作用。

大数定律及中心极限定理

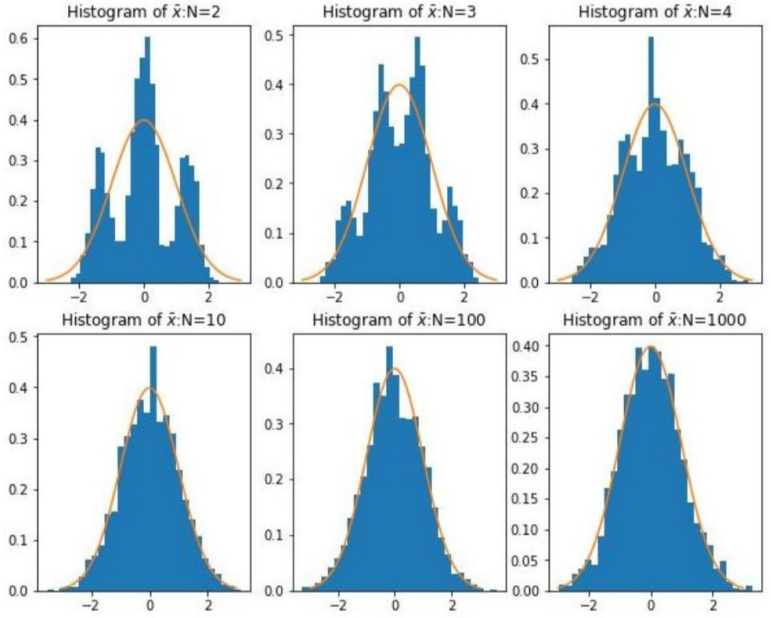
- 在统计活动中，人们发现，在相同条件下大量重复进行一种随机实验时，一件事情发生的次数与实验次数的比值，即该事件发生的频率值会趋近于**某一数值**。重复次数多了，这个结论越来越明显。这个就是**大数定律**。一般大数定律讨论的是n个随机变量平均值的稳定性。
- 而**中心极限定理**则是证明了在很一般的条件下，n个随机变量的和当n趋近于正无穷时的极限分布是**正态分布**。

大数定律讲的是样本均值收敛到总体均值，就是**期望**



$$\frac{1}{n}S_n - E(X) \xrightarrow{P} 0$$

而**中心极限定理**告诉我们，当样本足够大时，**样本均值的分布会慢慢变成正态分布**



$$\sqrt{n}(\frac{S_n}{n} - E(X)) \xrightarrow{D} N(0, \Sigma)$$

黄色的是标准正态分布的密度函数

思考题

- 两家医院，大医院每天新生儿45个，小医院新生儿15个，问一年内哪家医院男新生儿比例超过60%的天数多的可能性大?

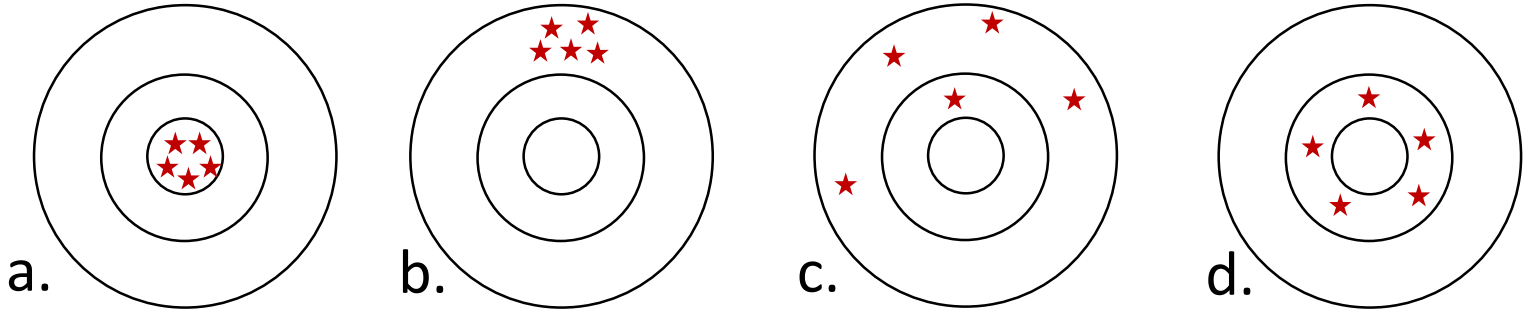
常用生物医学统计术语

精确性（precision），实验指标或性状的重复观测值彼此接近的程度。通过控制误差提高精确性。

准确性（accuracy），实验指标与真值的接近程度。

准确性与精确性合称为正确性。由于真值 μ 常常未知，所以准确性不易度量，但利用统计方法可度量精确性。

练习题：4个人在打靶场打靶，以下是他们的成果，请问他们的精确度和准确度如何？



以下为4个选项：

1. 精确度高
2. 精确度低
3. 准确度高
4. 准确度低

单元1总结

- 统计科学的主要过程（收集数据，整理数据，分析数据，解释数据）
- 描述性统计，推论性统计各是什么，区别在哪?
- 数据的类别有哪些，不同类别数据用哪些图示方式？（比如气泡图是用在什么数据）

- 描述统计里：集中趋势和离散趋势的指标 各有哪些?
- 偏度和峰度各描述的是什么?

- 离散变量，连续变量
- 4个不同的分布（伯努利，二项，泊松，正态）的区别和关联
- 辨别不同分布，及它们的使用条件和场景

- 大数定律和中心极限定理，概念，区别及应用

课程内容安排

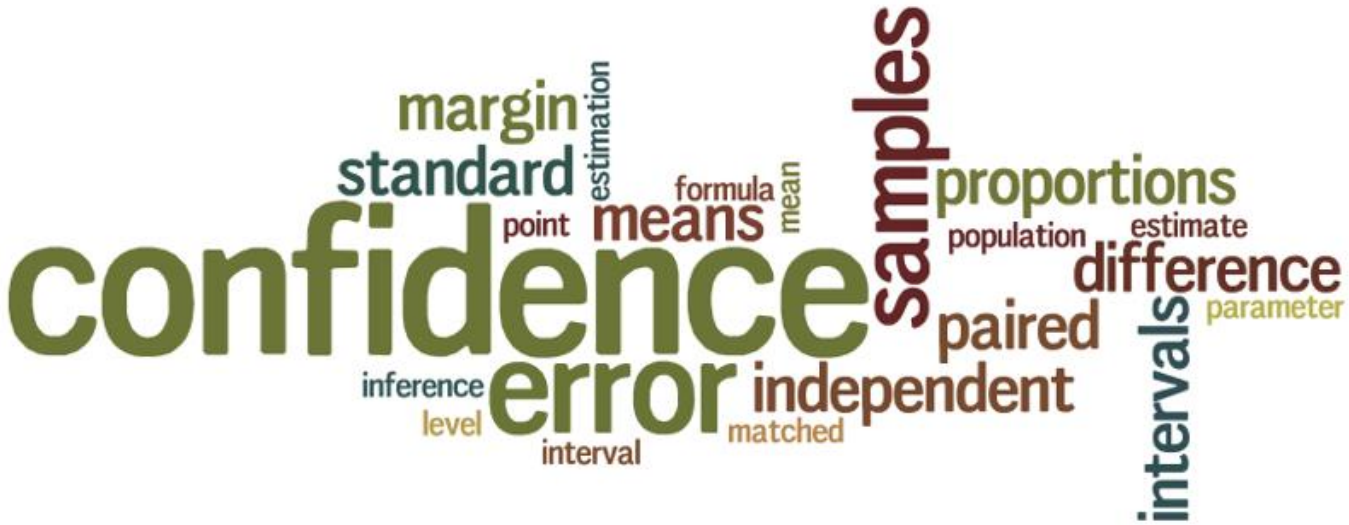
上课日期	章节	教学内容	教学要点	作业	随堂测	学时
2.16	1	数据可视化，描述性统计	1. 课程介绍 & 数据类型	作业1（8%）	测试1（8%）	2
2.23			2. 描述性统计Descriptive Statistics & 数据常用可视化			2
3.2			3. 常用概率分布			2
3.9			4. 大数定理 & 中心极限定理			2
3.16	2	推断性统计，均值差异检验	5. 统计推断基础-1：置信区间 Confidence Interval *	作业2（12%）	测试2（12%）	2
3.23			6. 统计推断基础-2:假设检验 Hypothesis Test			2
3.30			7. 数值数据的均值比较-1： 单样本t-检验			2
4.6			8. 数值数据的均值比较-2： 独立双样本t-检验，配对样本t-检验			2
4.13			9. 数值数据的均值比较-3：One-Way ANOVA			2
4.20			10. 数值数据的均值比较-4：Two-way ANOVA			2
4.27	3	比例差异检验	11. 类别数据的比例比较-1： 单样本比例推断 *	作业3（4%）	测试3（4%）	2
5.7（调）			12. 类别数据的比例比较-2： 联立表的卡方检验			2
5.11	4	协方差，相关分析，回归分析	13. 相关分析（Pearson r, Spearman rho, Kendal’ s tau) *	作业4 （6%）	测试4 （6%）	2
5.18			14. 简单回归分析			2
5.25			15. 多元回归Multiple Regression			2
6.1	5	Course Summary	16. 课程总结 *			2
			Total	30%	30%	32

* 随堂测试

单元2内容（week 4-9）

(对应参考书第4章内容)

- **总体、样本**
  - 参数、统计量
- **置信区间**
  - 区间估计、总体方差
- **均值差异**
  - One-sample 均值估计
  - Two-sample 均值差异



- **零假设检验**
  - H_0 vs H_a，Type I & II errors，P-value，One-sample vs Two sample test
- **方差分析**
  - ANOVA test（one-way ANOVA, two-way ANOVA）

什么是统计学?

- Statistics is the science of collection, analysis, interpretation, and presentation of data.

  统计学是一门收集、分析、解释和呈现数据的科学

- Descriptive statistics are numerical estimates that organize, sum up or present the data.

  描述性统计：是组织、总结或呈现一组数据的重要特征（表格，图形）

  (均值，中位数，众数，标准差)

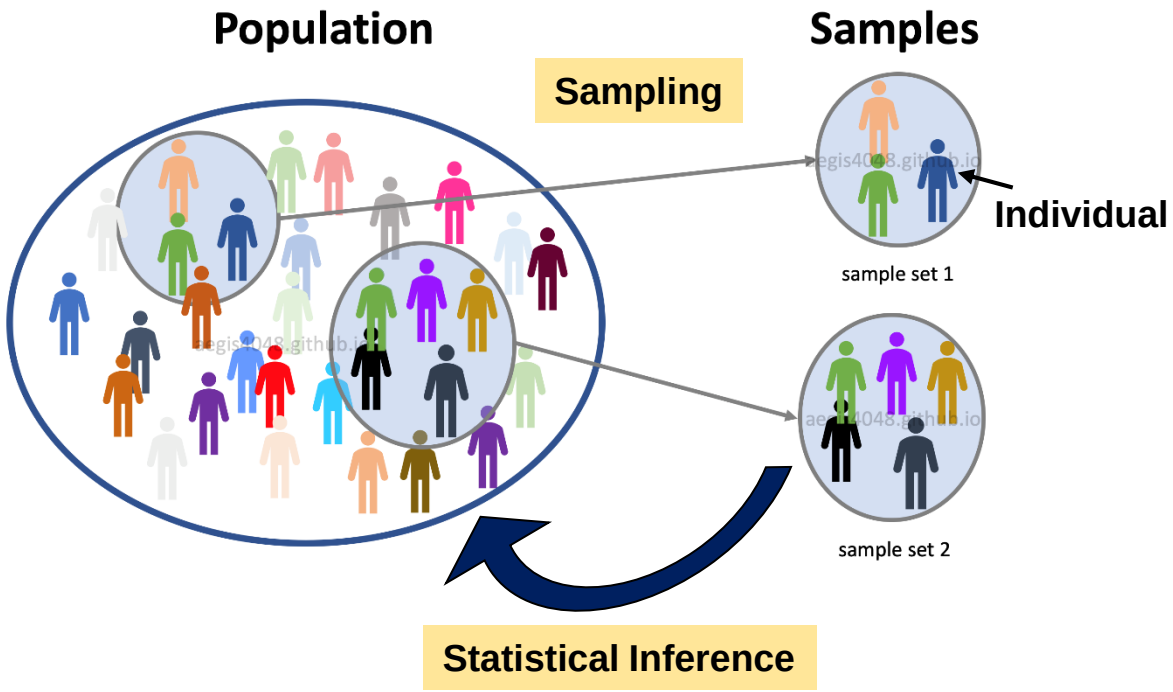
- Inferential statistics is the process of inferring from a sample to the population.

  推论统计： 利用样本信息对总体进行估计，预测或推断的一个过程（假设检验）

  (抽样，统计显著性)

总体和样本

总体：包含指定组的所有成员的数据集。例如：所有居住在中国的人。
样本：包含人口一部分或一部分的数据集。例如：居住在中国的某些人。

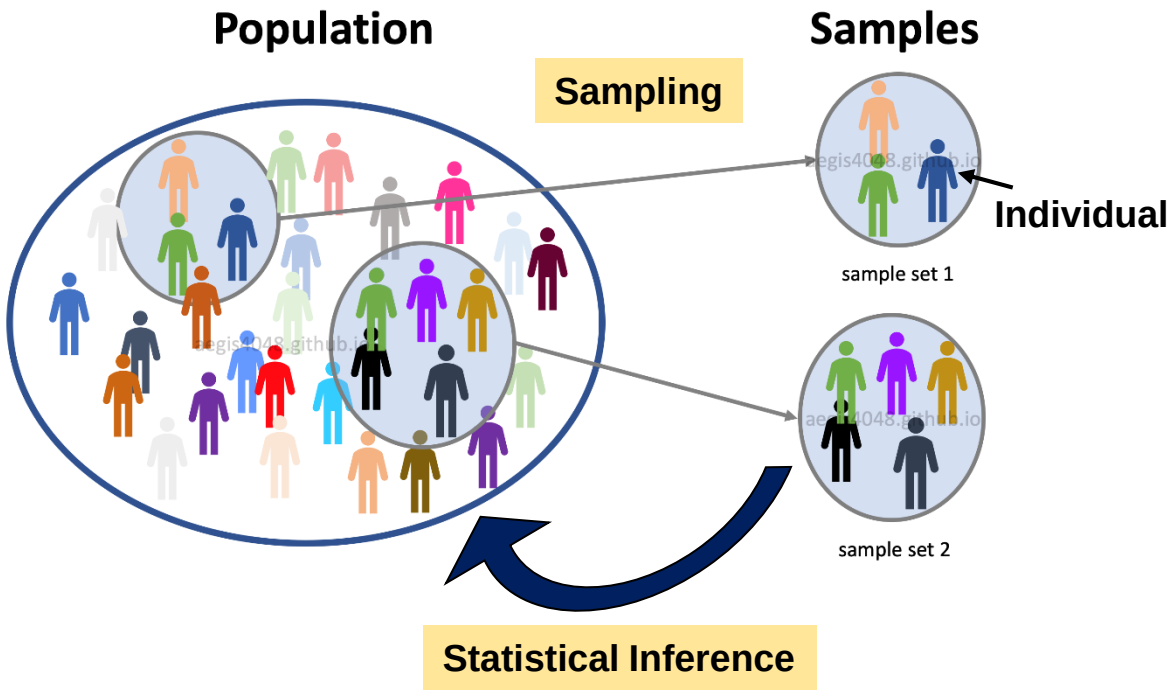


- **推断**是我们用来从数据得出结论的过程
- **推论**是使用我们从样本中计算出的统计量来做出和关于总体的有根据的猜测

我们使用样本统计量进行统计推断的方法称为估算/估计（estimation）

总体和样本

总体：包含指定组的所有成员的数据集。例如：所有居住在中国的人。
样本：包含人口一部分或一部分的数据集。例如：居住在中国的某些人。



我们通过 使用 样本统计量 (statistic) 估计 总体参数 (parameter) 来进行 统计推断

Population versus Sample

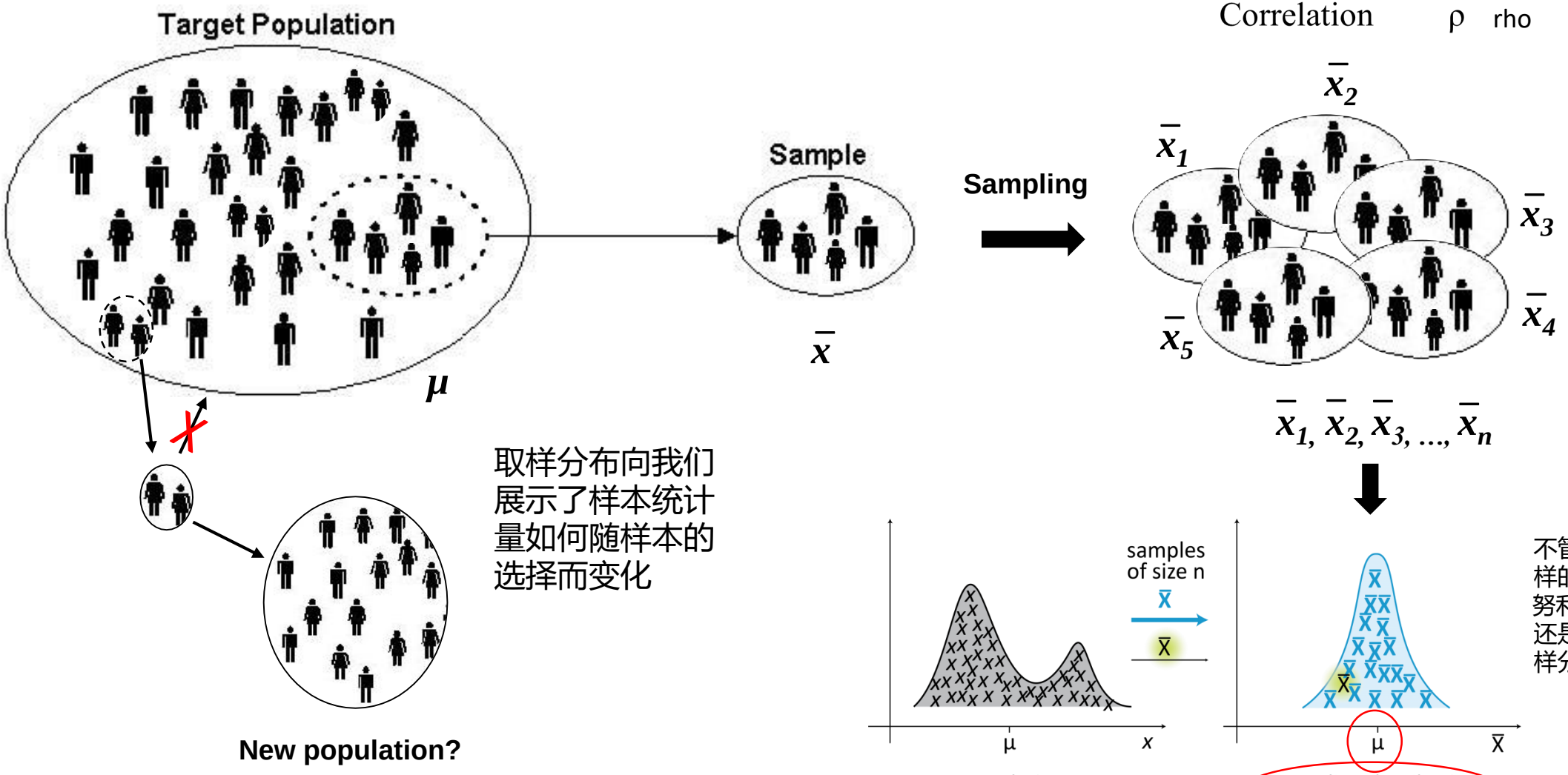
	Parameter		Statistic	
Mean	μ	mu	$\bar{x}$	x-bar
Proportion	p		$\hat{p}$	p-hat
Std. Dev.	σ	sigma	s	
Correlation	ρ	rho	r	

- **推断**是我们用来从数据得出结论的过程
- **推论**是使用我们从样本中计算出的统计量来做出和关于总体的有根据的猜测

我们使用样本统计量进行统计推断的方法称为估算/估计（estimation）

取样分布（Sampling distribution）

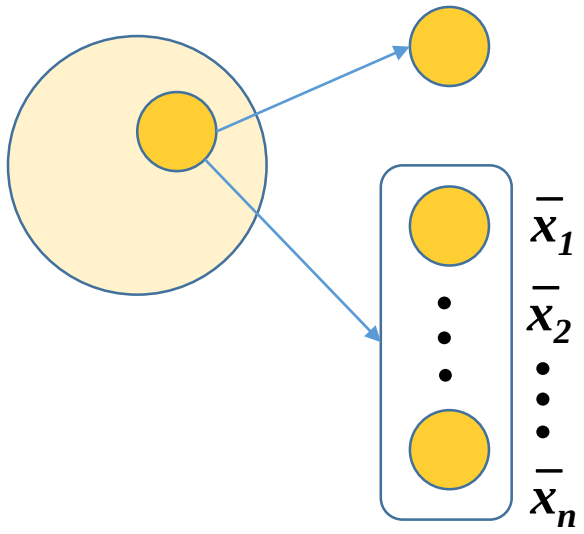
	Parameter		Statistic	
Mean	μ	mu	$\bar{x}$	x-bar
Proportion	p		$\hat{p}$	p-hat
Std. Dev.	σ	sigma	s	
Correlation	ρ	rho	r	



取样分布向我们展示了样本统计量如何随样本的选择而变化

不管总体的分布是怎样的，二项分布，伯努利分布，指数分布，还是什么。均值的取样分布将近似于一个正态分布 **(CLT)**

标准误差（standard error）



抽取一个样本，来推断总体。这时可以依据抽取的样本信息，计算出样本的均值与标准差

n次取样，n个均值 -> 计算这10个均值的标准差，此时的标准差就是**标准误差**

统计学中很常用的，来显示取样的均值离理论均值有多远

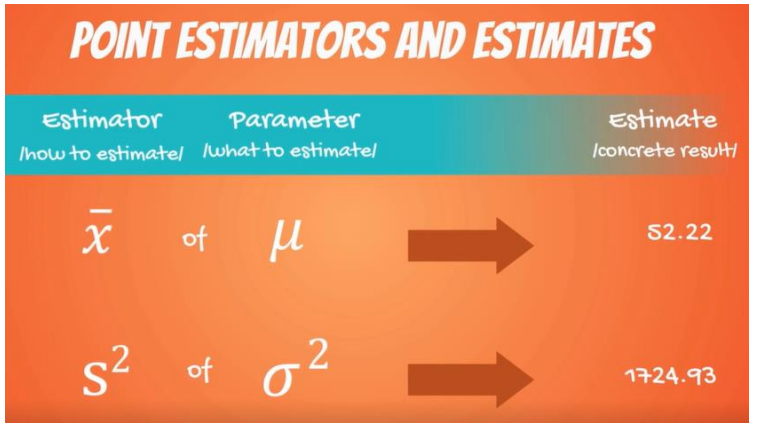
$$SE = \frac{\sigma}{\sqrt{n}}$$

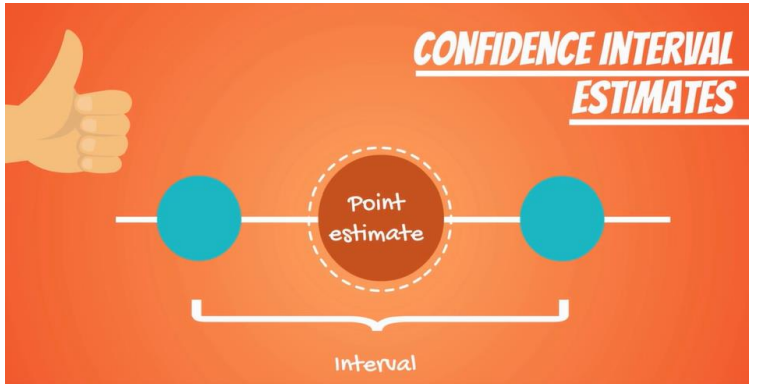
什么是“估计”

“估计”是指用抽样的数据估计全体的数据情况。
之所以这么做，是因为很多时候，想全体采集数据太难了！比如生产真空包装鸡腿的企业，要检查质量，就得把包装拆开，那鸡腿就不能再卖了。这是多大的损失呀！所以必须抽样

估计（Estimate）

- **估计量（Estimator），估计（Estimate）**
  - estimator是取决于样本的一个估计变量
  - estimate是根据某个确定的样本值而得到的一个确切地估计值（不是变量）
  - estimator是一个法则，estimate指用这个法则而算出来的具体值
- **点估计（Point estimates）**
  - 平均值：比如抽样鸡腿的平均重量为150克
- **区间估计（Confidence interval estimates）**
  - 比例：比如抽样鸡腿的卫生合格率为99.9%





通俗地讲：区间估计是在点估计的基础上，给一个合理取值范围

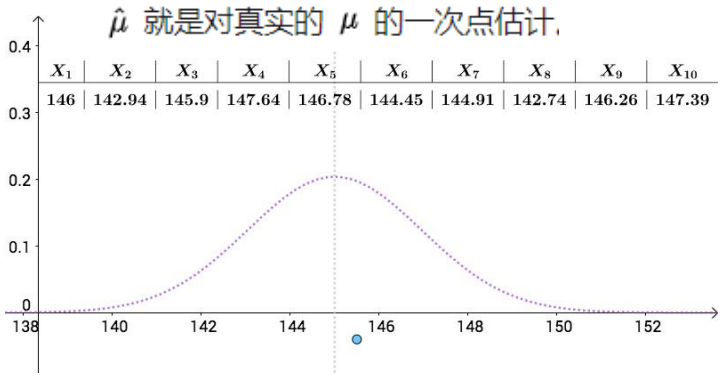
比如：抽样鸡腿的平均重量为150克，是一个点估计值。抽样鸡腿的平均重量为145克到155克之间，是一个区间估计。
其中，145到155称为置信区间。这很符合人们的常规理解：东西很难100%准确，有个范围也是可以理解的。

但这个范围有多大可信度呢？人们用置信水平来衡量，即：“我们有多大把握，真实值在置信区间内”

如何理解95%置信区间?

区间估计

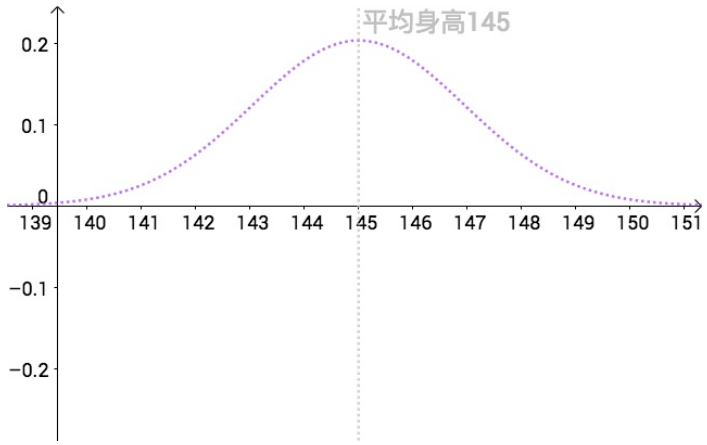
- 在一定概率保证下指出总体参数的可能范围，叫区间估计
- 所给出的可能范围叫**置信区间**（confidence interval）
- 给出的概率保证称为**置信度或置信概率**（confidence probability）



例子：人类的平均身高

假设人类的身高分布服从（$\mu = 145, \sigma = 1.4$）如下正态分布

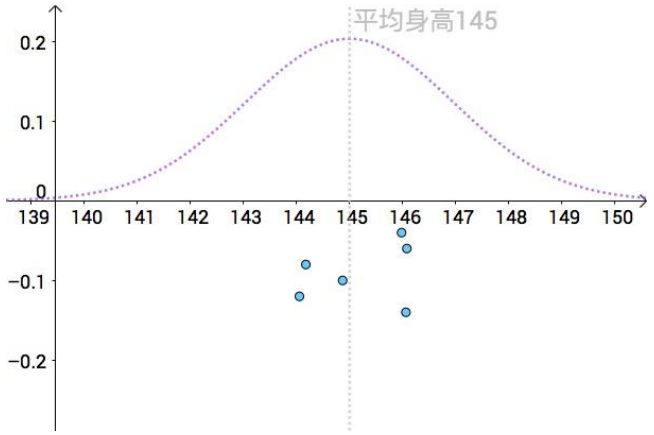
$$X \sim N(145, 1.4^2)$$

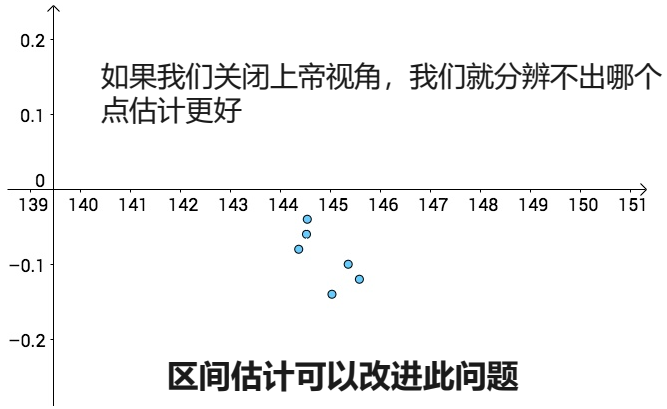


”上帝视角“

作为人类，我们只能在人群中抽样统计

通过一次次的抽样，我们可以算出不同的身高均值的点估计：





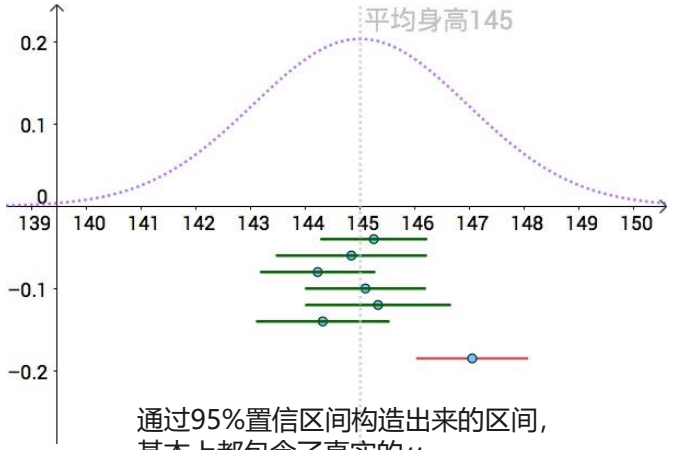
如果我们关闭上帝视角，我们就分辨不出哪个点估计更好

区间估计可以改进此问题

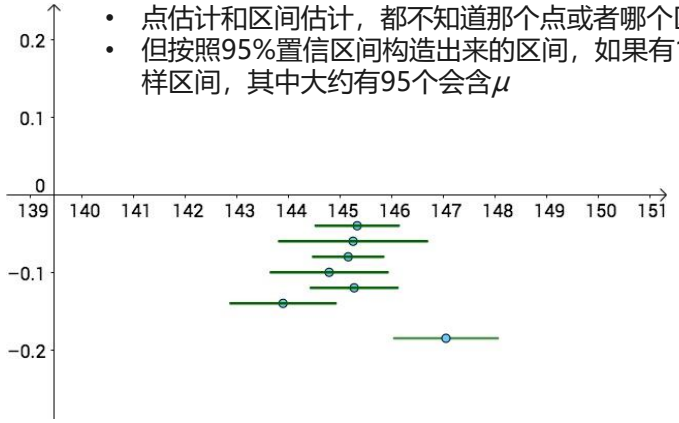
置信区间

置信区间，提供了一种区间估计的方法

下面采用95%置信区间来构造区间估计



通过95%置信区间构造出来的区间，
基本上都包含了真实的μ



- 点估计和区间估计，都不知道那个点或者哪个区间更好
- 但按照95%置信区间构造出来的区间，如果有100个这样区间，其中大约有95个会含μ

95% 置信区间

以上面的统计身高为例，假设全国人民的身高服从正态分布：

$$X \sim N(\mu, \sigma^2)$$

不断进行采样，假设样本的大小为n，则样本的均值为：

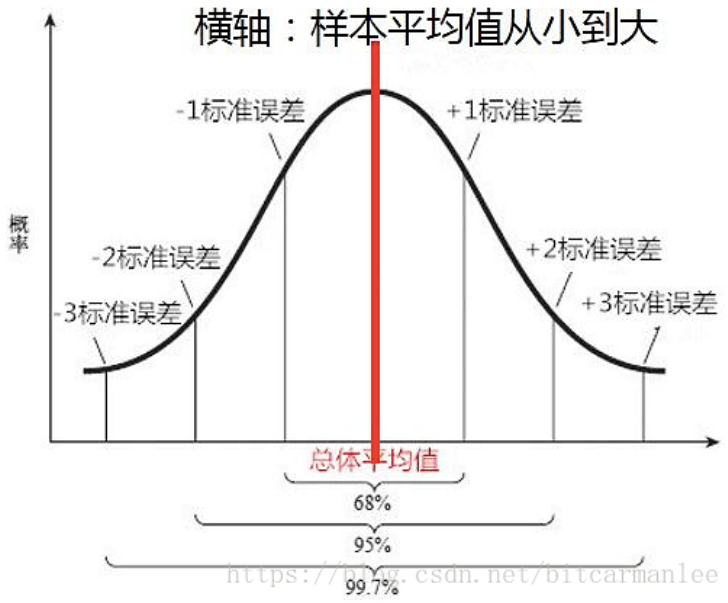
$$M = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

由大数定理与中心极限定理：

$$M \sim N(\mu, \sigma_1^2)$$

注意σ_1的计算方法为 ⟹ 标准误差!

为什么常用95%的置信水平：



用一句简单的话概括就是：
有95%的样本均值会落在2个（比较精确的值是1.96）标准误差范围内

$$P(\mu - 1.96\frac{\sigma}{\sqrt{n}} < M < \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

那么，只有一个问题了，我们不知道、并且永远都不会知道真实的 μ 是多少

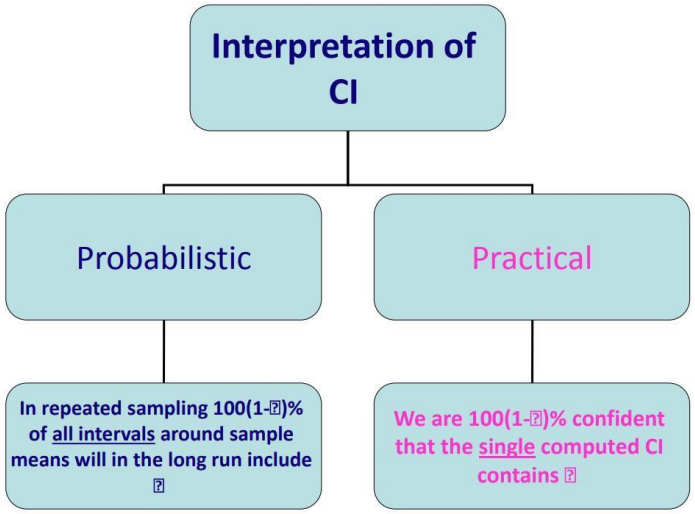
我们就只有用 $\hat{\mu}$ 来代替 μ ：

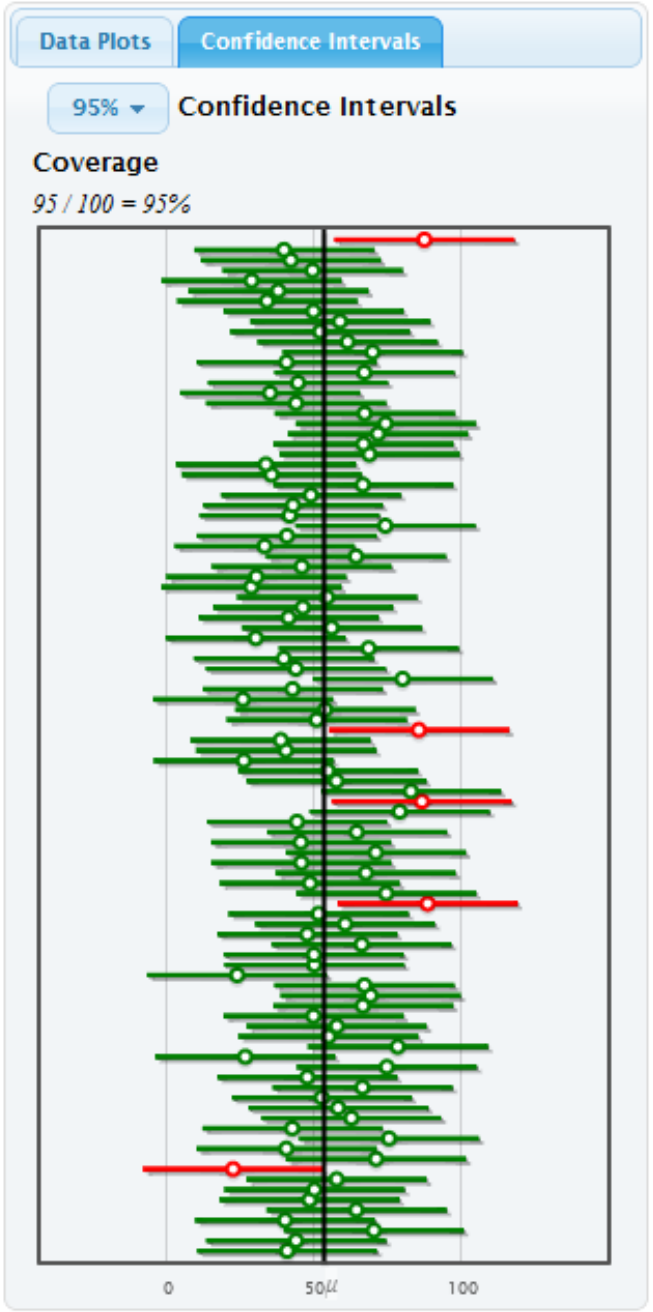
$$P(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} \le M \le \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$$

置信区间

- **It is not the probability of samples falling into this interval!!!**
- 95%的置信区间：100次抽样，有95次的置信区间包含了总体均值

这就好像用渔网捞鱼，我知道100次网下去，可能会有95次网到我想要的鱼，但是我并不知道是不是这一次的网





- 总体参数是固定的
- 样本统计量是 random（根据抽取的样本）
- 区间也是 random （根据样本统计量）
- 用中括号 [a,b] 表示样本估计总体平均值误差范围的区间。a、b的具体数值取决于你对于“该区间包含总体均值”这一结果的可信程度，因此 [a,b] 被称为置信区间

谢谢，下周见！