Sampling Distributions

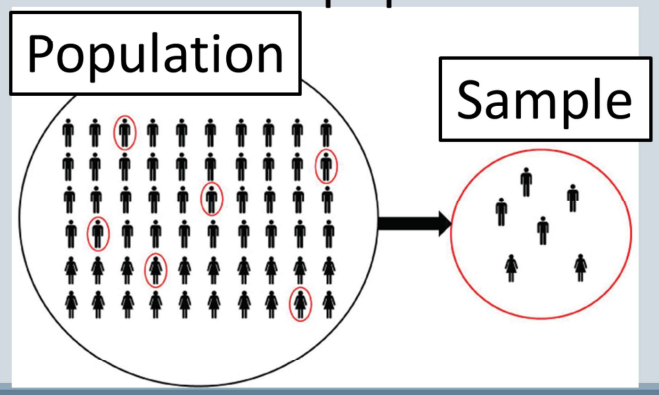
Topics

1. **Parameters vs. statistics**
2. Sampling distributions
3. Accuracy (bias) vs. precision (variation)

这节课我们开始讲Sampling Distribution，我们先来看第一个概念，inference

Inference

Inference: Estimating characteristics of a population using sample statistics.



也就是我们抽样的目的到底是什么
我们的原始问题是，想去研究整个population的一个特性，但有很多时候我们无法收集population中每一个个体的数据
那我们就会进行抽样，通过计算样本的值来估计population的值。
也就是：Estimating characteristics of a population using sample statistics.

Parameter vs. Statistic

A **parameter** is a number describing a characteristic of a whole **population**.

◆ The value of a population characteristic is fixed ! ! ! ! !

Example:

population mean (μ), population standard deviation (σ),

population variance (σ^2), etc.

这里面的population characteristic我们就叫它parameter
Parameter的值是固定的，比如population mean，这个是个常数是个固定的值
Parameter的例子还有：sigma和variance

Parameter vs. Statistic

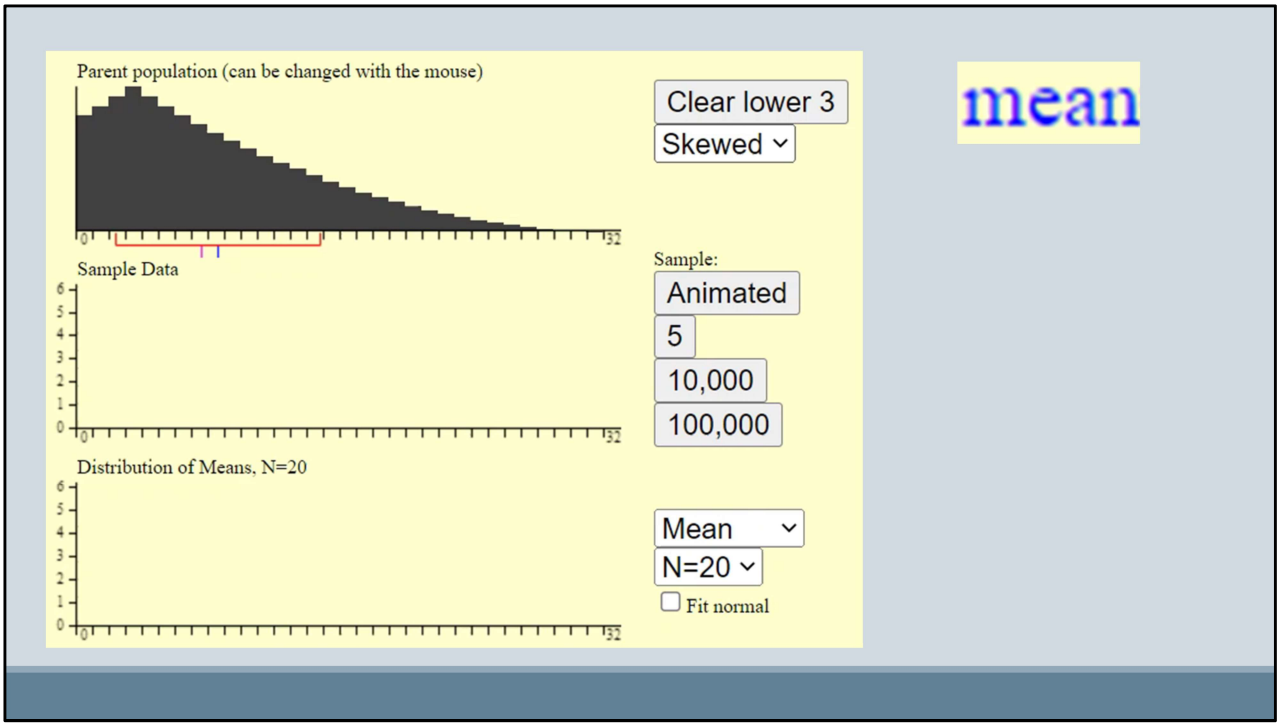
A **parameter** is a number describing a characteristic of a whole **population**.

- The value of a population characteristic is fixed ! ! ! ! !

A **statistic** is a number describing a characteristic of a **sample from the population**. $\bar{X}$, S_X, ...

- Different samples → different values of statistics

对应的样本的特征值就是sample Statistic
那对应的就是xbar，Sx，Sx^2...
对于Statistic来说，我们不同的样本有不同的统计值Statistic，所以这是一个random variable!



我们来看一下抽样过程:
最上面的是population Distribution
中间的是抽到的样本，sample size是20
抽到一个样本之后我们就可以去计算它的均值

Business Owner Problem

A business owner wants to know the average household income in her business's zip code. She randomly samples 80 households in the zip code and finds their mean income to be $46,144.

Population:

All households in zip code

Parameter:

μ - Mean income of all households in zip code

Sample:

80 surveyed households

sample size (n)

Statistic:

$\bar{x}$ - Mean income of sample ($46,144)

练习一下写出parameter和Statistics

Topics

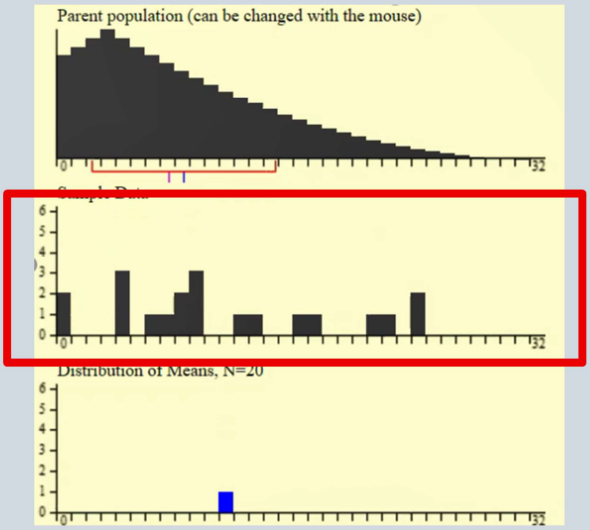
1. Parameters vs. statistics
2. **Sampling distributions**
3. Accuracy (bias) vs. precision (variation)

Sample Distribution 样本分布:

✓ Sample(s):

subset(s) of a population (the set of values actually used for estimation).

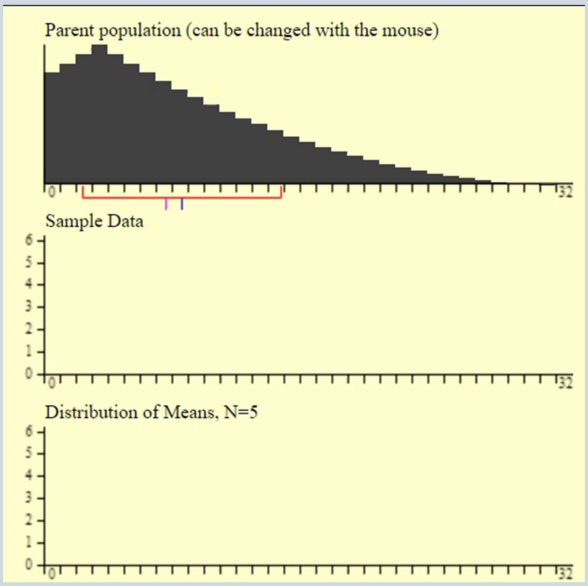
✓ Sample distribution is the distribution of the sample data.



接下来看一下sample Distribution和sampling Distribution的
最上面的叫population Distribution
中间这个是我们抽了20个个体作为样本
那这些sample data形成的分布就是sample Distribution

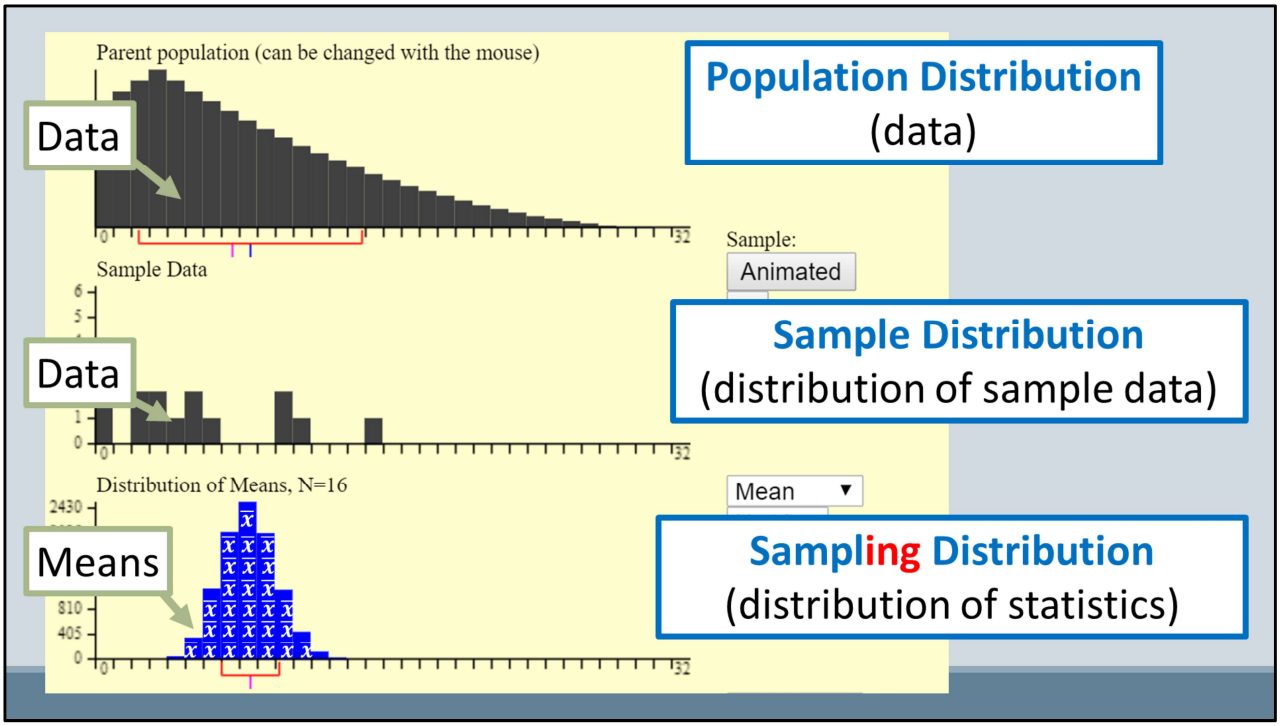
Sampling Distribution 抽样分布:

- ✓ Fix the sample size and select samples from a population
- ✓ Compute statistics from sample data
- ✓ Sampling distribution is the distribution of **statistics,** e.g. means, **not the distribution of data points**



Sampling Distribution 指的不是样本的分布了，而是样本的统计值的分布
我们最开始说统计值的时候也说了，Statistics是一个随机变量，不同的sample会有不同的统计值
我们就拿最常见的均值举例
不同的样本，算出来的均值也肯定是不一样的
那由不同的样本得到的均值的分布就叫做sampling Distribution of the mean

得到sampling Distribution的过程是这样的：
首先我们对于样本的要求是要统一的，不能说第一个样本有10个 个体，第二个有100个sample data，我们要求每次抽样的sample size都保持一致
然后每抽到一个样本，就计算出它的均值
重复抽了很多次很多个样本之后，得到的均值们的分布就叫做sampling Distribution，注意sampling Distribution是Statistics的分布，而不会sample data的。
我们来看一个仿真的过程



再次区分一下这几个概念
Sample： 是sample data的分布情况
那一个sample就对应着一个sample Distribution，就有一个sample mean
多个samples就对应着多个sample Distribution，就对应着多个sample means
这多个sample means形成的分布叫做sampling…

Population distribution

1, 2, 3

Sample distribution	Sample distribution	Sample distribution
(Sample size = 2)	(Sample size = 2)	(Sample size = 2)
1, 2	1, 3	2, 3

Sample mean : 1.5, 2, 2.5

Sampling distribution: 1.5, 2, 2.5 with prob. 1/3

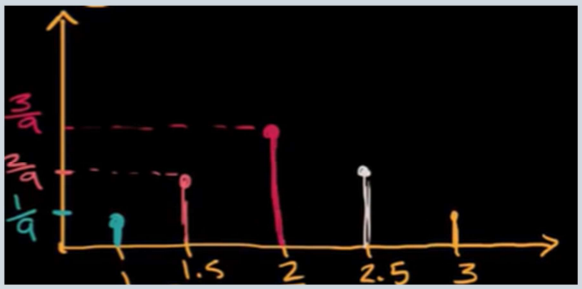
练习一下:

Identical balls in a black box with different number 1, 2 and 3. What is the distribution of sample mean with sample size of 2?

WITHOUT REPLACEMENT!

balls	Sample mean
1,1	1
1,2	1.5
1,3	2
2,1	1.5
2,2	2
2,3	2.5
3,1	2
3,2	2.5
3,3	3

Sampling distribution for the sample mean with sample size of 2



Possible sample mean

With replacement!

Topics

1. Parameters vs. statistics
2. Sampling distributions
3. **Accuracy (bias) vs. precision (variation)**

Business Owner Problem

A business owner wants to know the average household income in her business's zip code. She randomly samples 80 households in the zip code and finds their mean income to be $46,144.

Parameter:

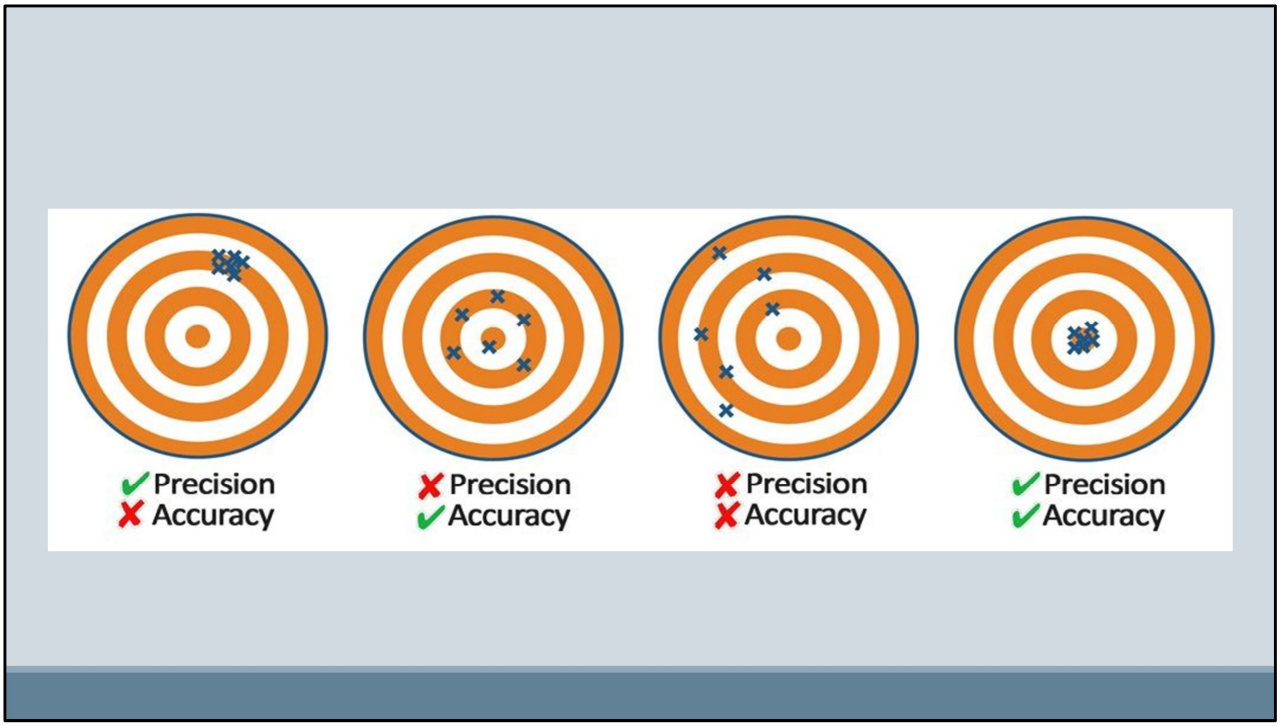
μ - Mean income of all households in zip code

Statistic:

$\bar{x}$ - Mean income of sample ($46,144)

How confident are we of our estimator?

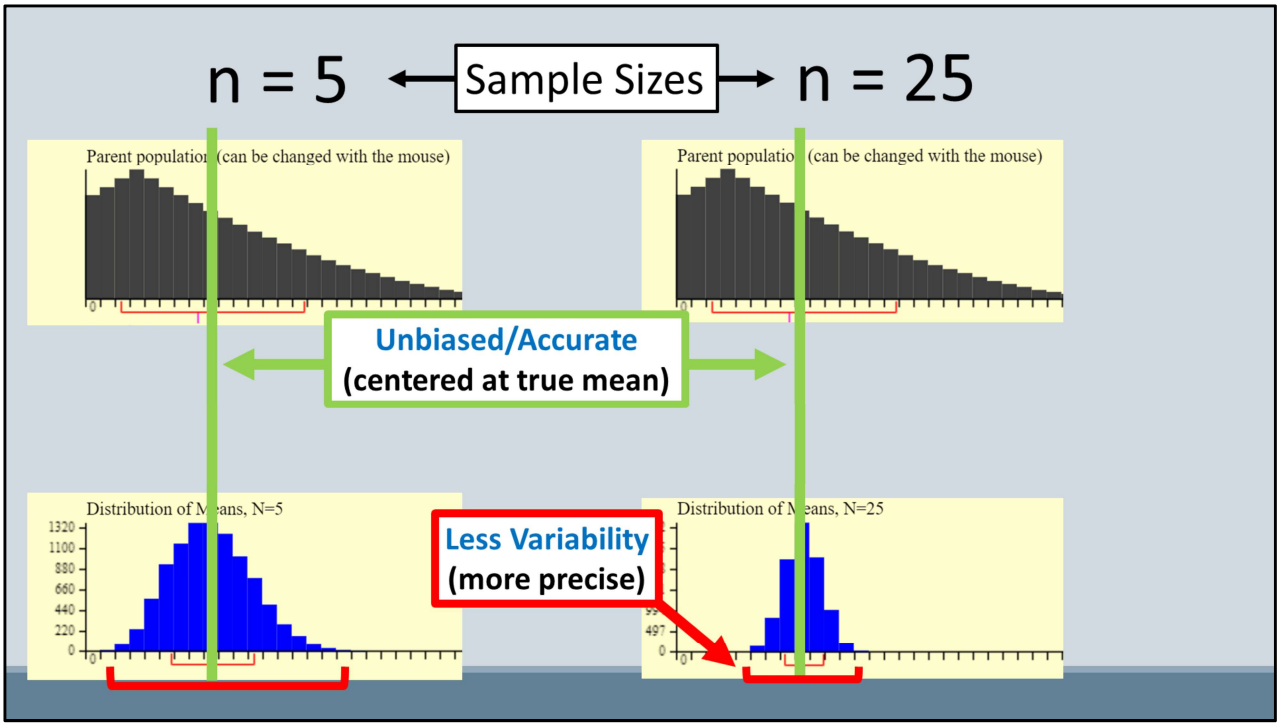
接着来看一下对于sample statistics的评判标准，我们有了sample mean，但是怎么形容它是否可信呢？主要是两个维度来形容



Accuracy and Precision
其中 accuracy指的是，估计值虽然是会波动，是个随机变量，但是它波动的中心还是真实值，也就是我们之前提到的unbiased

一般在accuracy的基础上我们才会去谈论Precision
Precision指的是，估计值的波动大小，那我们当然希望波动越小越好。

那怎么做到unbiased 和比较高的精确度呢？



对于同样的population data
我们选取不同的sample size
观察一下得到的sampling Distribution是什么样子的
这里都是随机抽样

上面的蓝色的是population mean
我们发现都是unbiased，也就是均值的均值 是真实值，我们的sample mean是围绕着mu波动的

接着我们看一下Precision
很显然左边没有右边精准，那我们就发现sample size大的时候，得到的sample mean更精准，这个也很make sense，sample size大就对应着获得的信息更多，肯定估计的就更准确了

How do you avoid bias?
(*how do you make it accurate?*)

How do you reduce variability?
(*how do you make it more precise?*)

How do you avoid bias?
(*how do you make it accurate?*)

Sample randomly

How do you reduce variability?
(*how do you make it more precise?*)

How do you avoid bias?
(*how do you make it accurate?*)

Sample randomly

How do you reduce variability?
(*how do you make it more precise?*)

Increase sample size (n)

Have a try!!!!

http://onlinestatbook.com/stat_sim/sampling_dist/index.html