Sampling Distribution for a Mean

Topics

1. μ **vs.** $\bar{x}$
2. Sampling distribution for a mean
3. Central limit theorem

我们来看一下sample mean的分布到底是什么样子的

μ vs. $\bar{x}$

μ = **population** mean

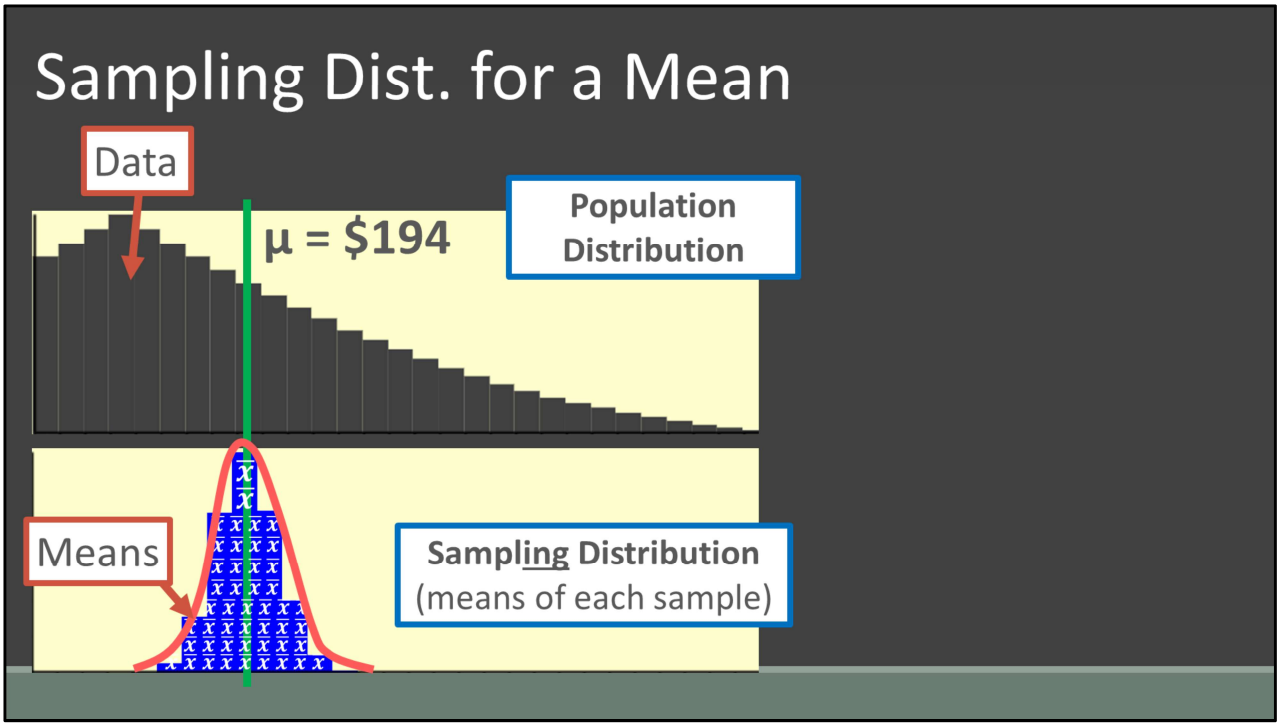
- Parameter

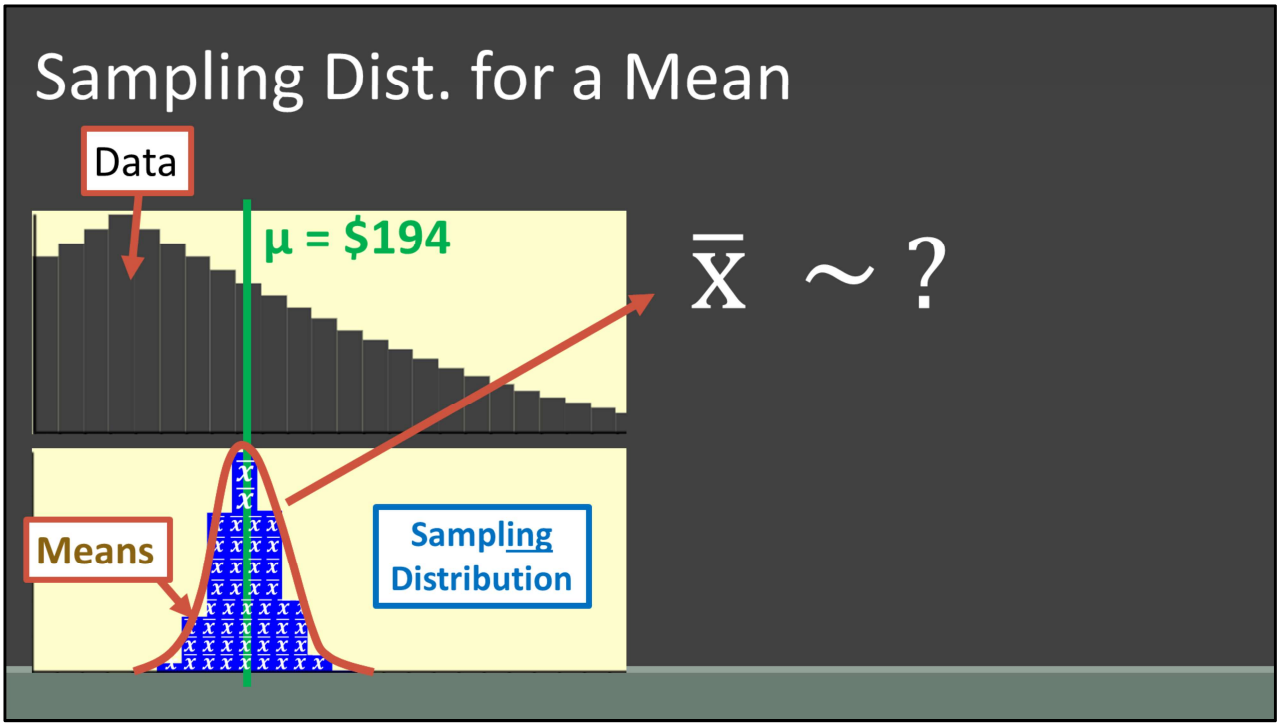
$\bar{x}$ = **sample** mean

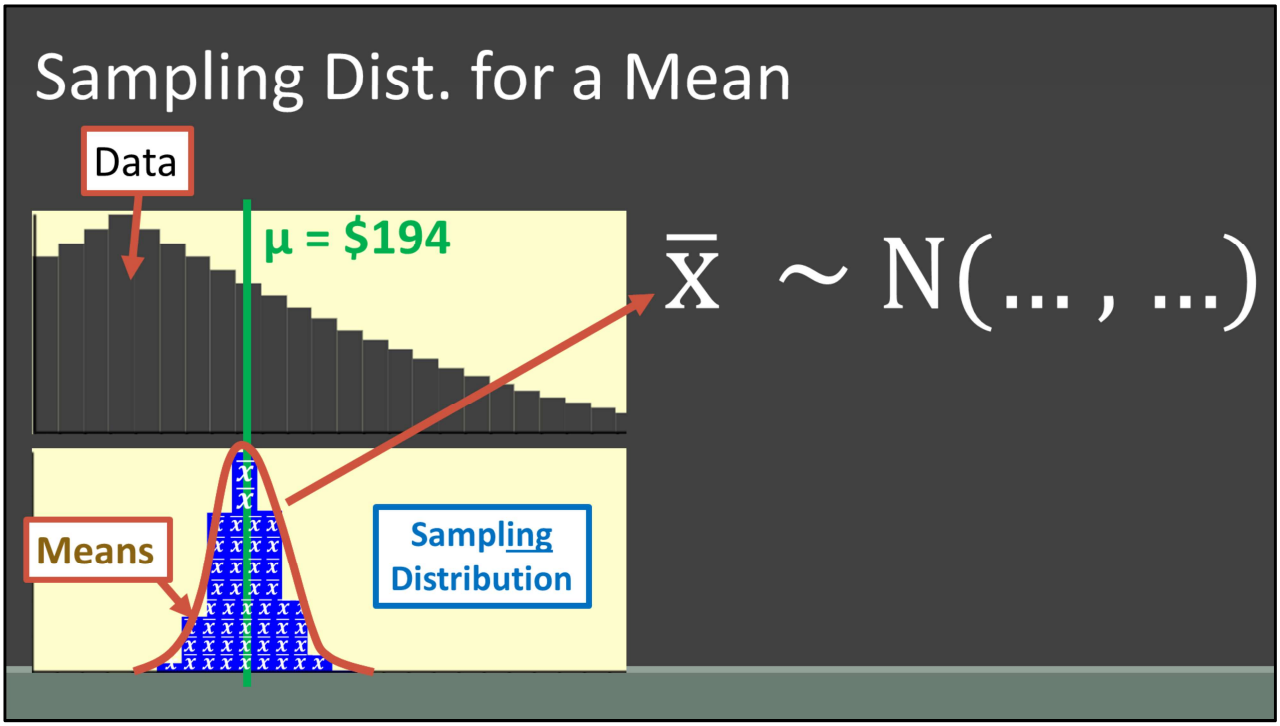
- Statistic
- Estimator of μ

Topics

1. μ vs. $\bar{x}$
2. **Sampling distribution for a mean**
3. Central limit theorem


Sampling Dist. for a Mean
Data
μ = $194
Population Distribution
Means
Sampling Distribution
(means of each sample)


Sampling Dist. for a Mean
Data
μ = $194
$\bar{x} \sim ?$
Means
Sampling Distribution



我们发现都比较像Normal，事实也是如此，满足一定条件的情况下，sample mean的分布就趋近于Normal Distribution
那这里也是需要我们来确认一下两个系数： xbar的期望和方差
先来看一下它的定义

Population Mean： $\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$

Estimator (Statistic)： $\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

这里面，对于xbar来说我们回想一下得到均值的过程
我们是不是从population里面随机抽取一个个体，记录它的值，比如就是这里的x1，x1是不是从population中随机抽取得到的？
那是不是x1的取值有很多种可能性？
那它是什么？RV!
那同理，x2到xn其实都是随机变量，如果我们每次抽取都是with replacement，那每个随机变量还都是独立的

Population Mean: $\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$

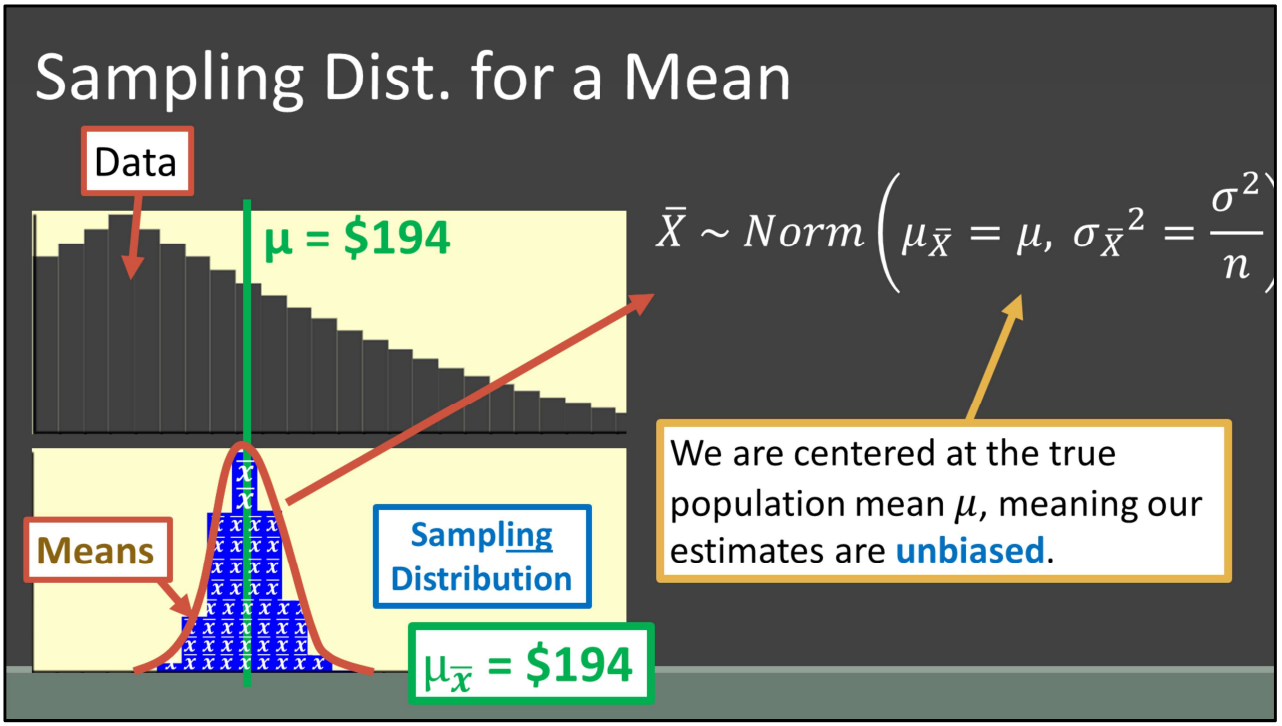
Estimator (Statistic): $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$

Independent Random Variables:

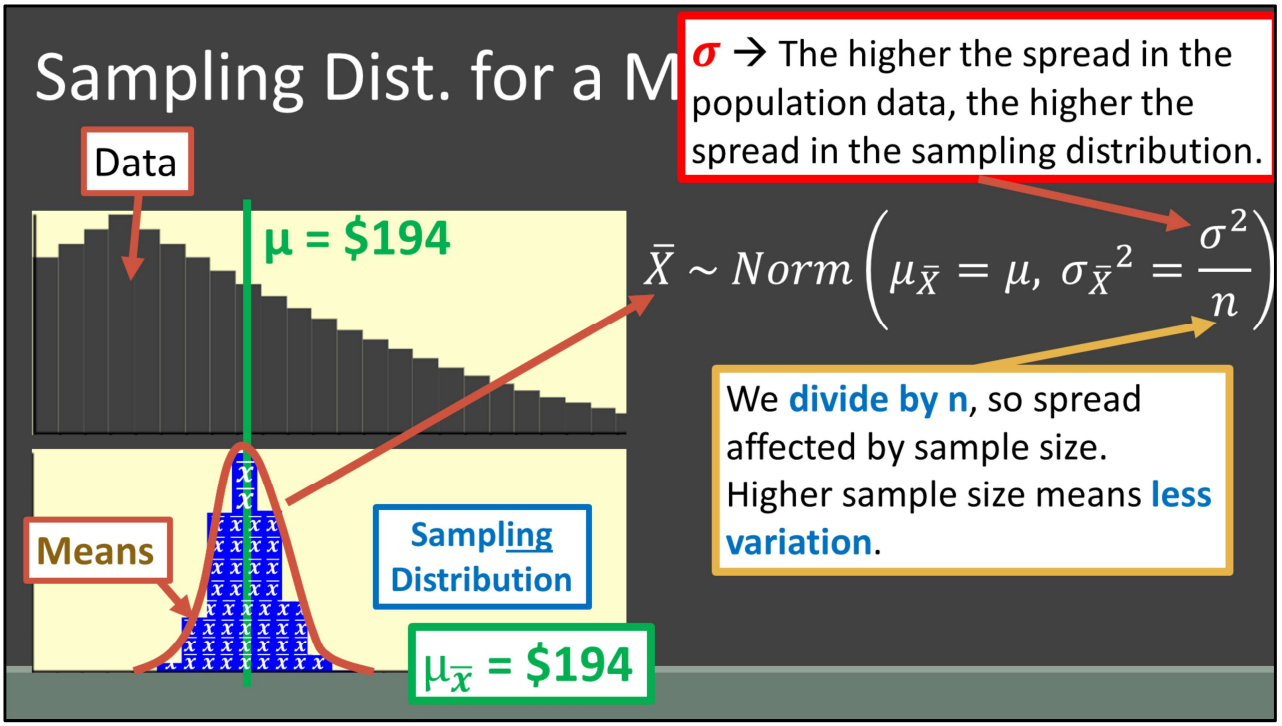
$$X_1, X_2, \ldots, X_n \sim Distribution\ of\ X\ with\ (\mu, \sigma^2)$$

所以我们可以把xbar改写成：$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$
其中，Xi是独立且拥有共同分布的随机变量，他们都服从于population Distribution，也就是他们的期望都是mu

那xbar的均值是？ Mu
方差是？ Sigma^2/n



所以我们得到了xbar的分布


Sampling Dist. for a M
σ → The higher the spread in the population data, the higher the spread in the sampling distribution.
Data
μ = $194
$\bar{X} \sim Norm\left(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}{}^2 = \frac{\sigma^2}{n}\right)$
We divide by n, so spread affected by sample size. Higher sample size means less variation.
Means
Sampling Distribution
$\mu_{\bar{x}}$ = $194

Topics

1. μ vs. $\bar{x}$
2. Sampling distribution for a mean
3. **Central limit theorem**

Central Limit Theorem (CLT)

Central limit theorem (CLT):
If you have a population with mean μ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

- Useful because we **never know** the population's shape!
- Now if there is a large enough sample, we can model the sampling distribution for the mean.

If you have a population with mean μ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

也就是说如果我们的样本满足两个条件
一个是 sample size足够大
另一个是，抽样的方式是有放回的

那么xbar的sampling Distribution就是可以近似看作Normal Distribution

这个定理非常有用，因为大多数情况我们都不知道population的分布是什么样的
那在这种情况下，我依旧可以知道sample mean的分布大概什么样子，只要满足了刚刚提到的两个条件，我们就可以把mean的sampling Distribution当做Normal来处理。这个定理也是我们后面做假设检验的一个最基础的理论依据。非常有意义

Condition 1: Random

You must ensure the sample was **randomly selected** from the **target population** (in the case of an experiment, you must ensure there was a random assignment to treatment)

Why must this condition be satisfied?

If this condition is not satisfied, your estimator is **biased**.
If it is satisfied, your sampling distribution is **centered** at the true population mean value → $\mu_{\bar{x}} = \mu$

我们昨天得出了xbar的分布，但是有一些条件，我们今天来具体的看一看
第一个就是 random sampling method
这个条件满足的话，我们就会得到一个unbiased sample，那也就得到了unbiased estimator
所以就可以保证E(Xbar）=mu

Condition 2: 10% Condition

a) What is the c

The **sample size (** **n**

size (N): n < 0.10

If sampling without replacement, the standard deviation of the sample mean is smaller than what is given by the formula above. If the sample size is less than 10% of the population size, the difference is negligible.

b) Why must this condition be satisfied?

It ensures this is true: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

For example if sample = population, $\sigma_{\bar{x}}$ **= ?**

第二个条件是用来保证variance的，上节课我们说到，如果抽样的时候是with replacement的话，xbar的方差是可以求出来的，但是如果不放回抽样的话，就比较复杂，可能很难算出方差是多少，
但是现实中抽样基本都是without replacement，但如果满足这个10% condition的话，也就是population size 比sample size大很多的话，就可以近似成with replacement。

我们也可以举个反例，如果sample很大，大到趋近于population的话，xbar就变成了mu，那是不是就失去了波动性？

Condition 2: 10% Condition

a) What is the condition?

The **sample size** (n) must be less than 10% of the **population size** (N): n < 0.10(N)

b) Why must this condition be satisfied?

It ensures this is true: $\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$

For example if sample = population, $\boldsymbol{\sigma_{\bar{x}} = 0}$

Condition 3: Normal/Large Sample

a) What is the condition?

i) The **population** distribution must be normal **OR**

ii) The sample size is **30 or more** ($n \geq 30$)

b) Why must this condition be satisfied?

This provides evidence that the sampling distribution is **approximately normal** in shape.

For example if sample size = 1, sampling distribution = ?

第三个也就是最后一个条件，是来保证xbar的分布是正态分布的
这里面有两个条件都是可以的
第一个，如果population Distribution是normal的话，那无论sample size是多少，我们得到的sample mean都是正态分布的（可以详细解释一下）

如果我们不知道population Distribution或者它不是正态分布的话，就需要满足sample size larger than or equal to 30这个条件，也就是用到了central limit theory

我们可以举个反例说明一下
假设现在的sample size=1，那 sampling Distribution是啥？

Condition 3: Normal/Large Sample

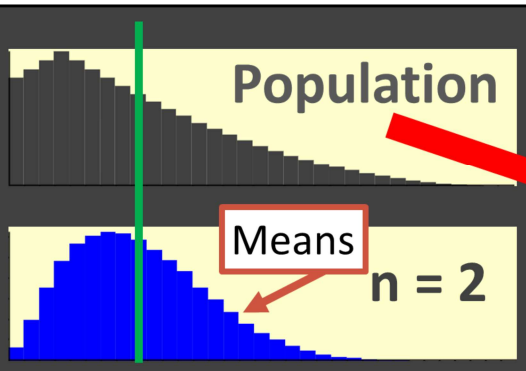
a) What is the condition?

i) The **population** distribution must be normal **OR**

ii) The sample size is **30 or more** ($n \geq 30$)

b) Why must this condition be satisfied?

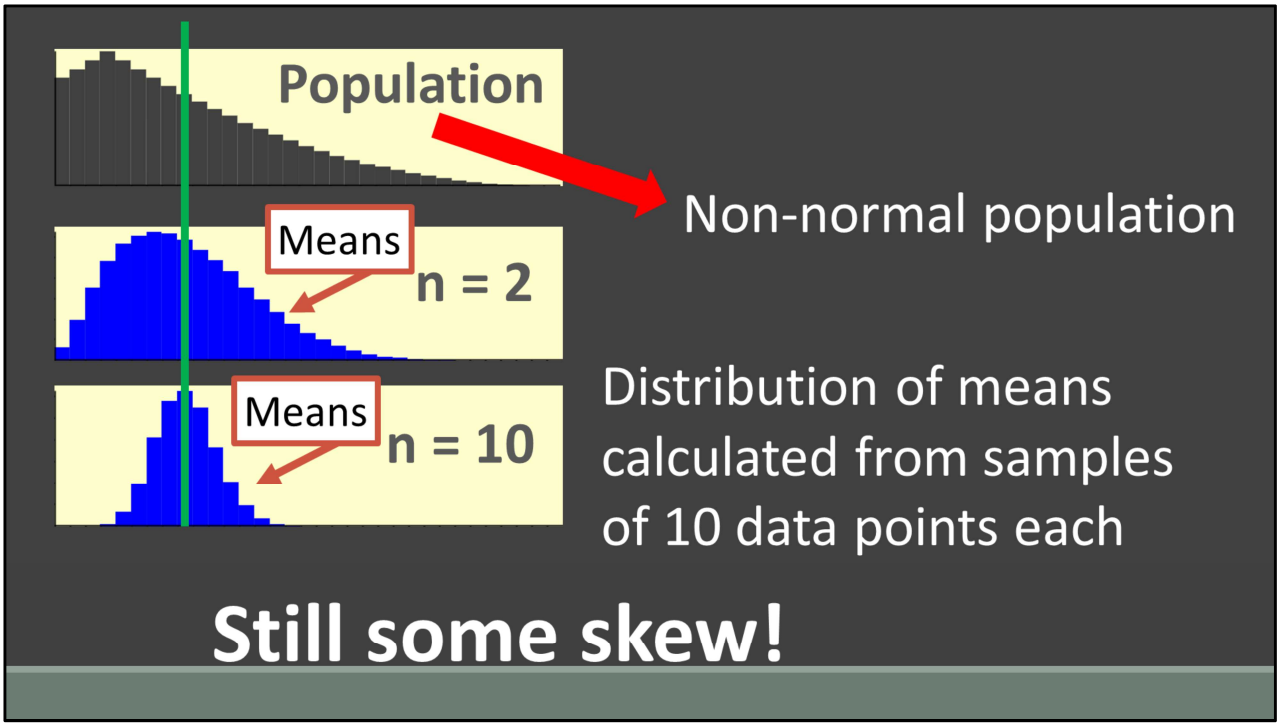
For example if sample size = 1, sampling distribution = **population distribution.**

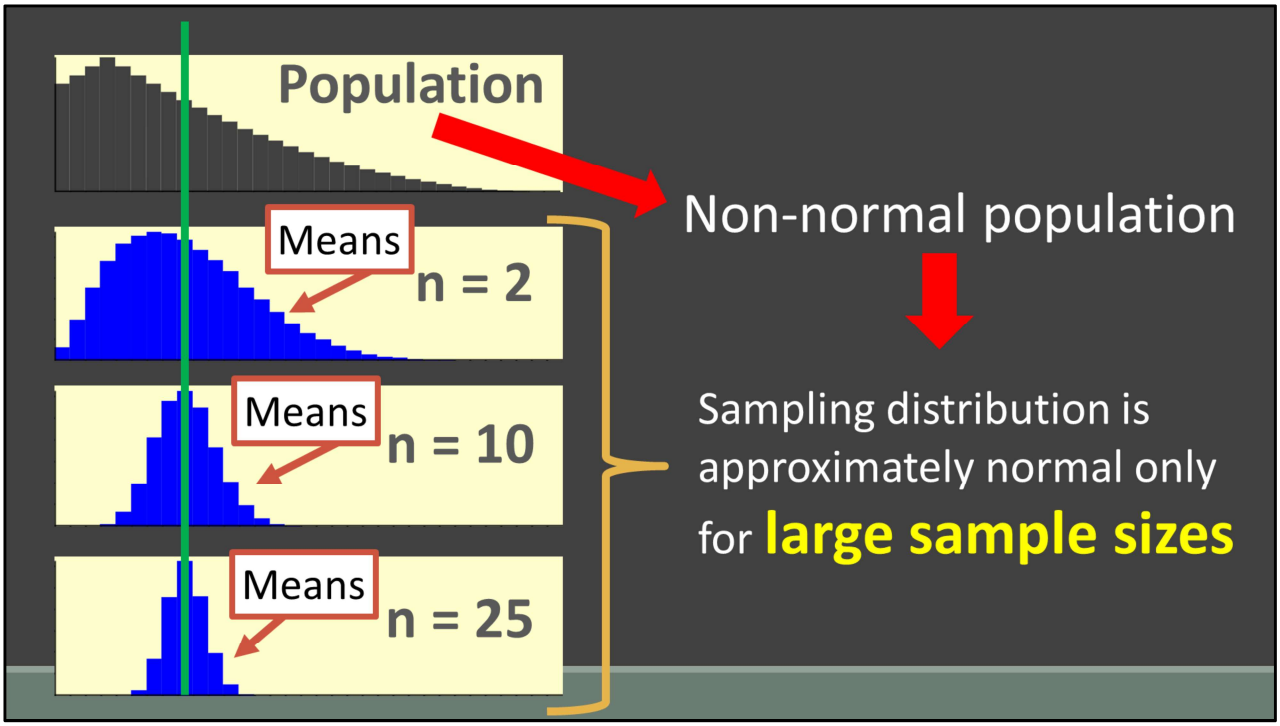


Non-normal population

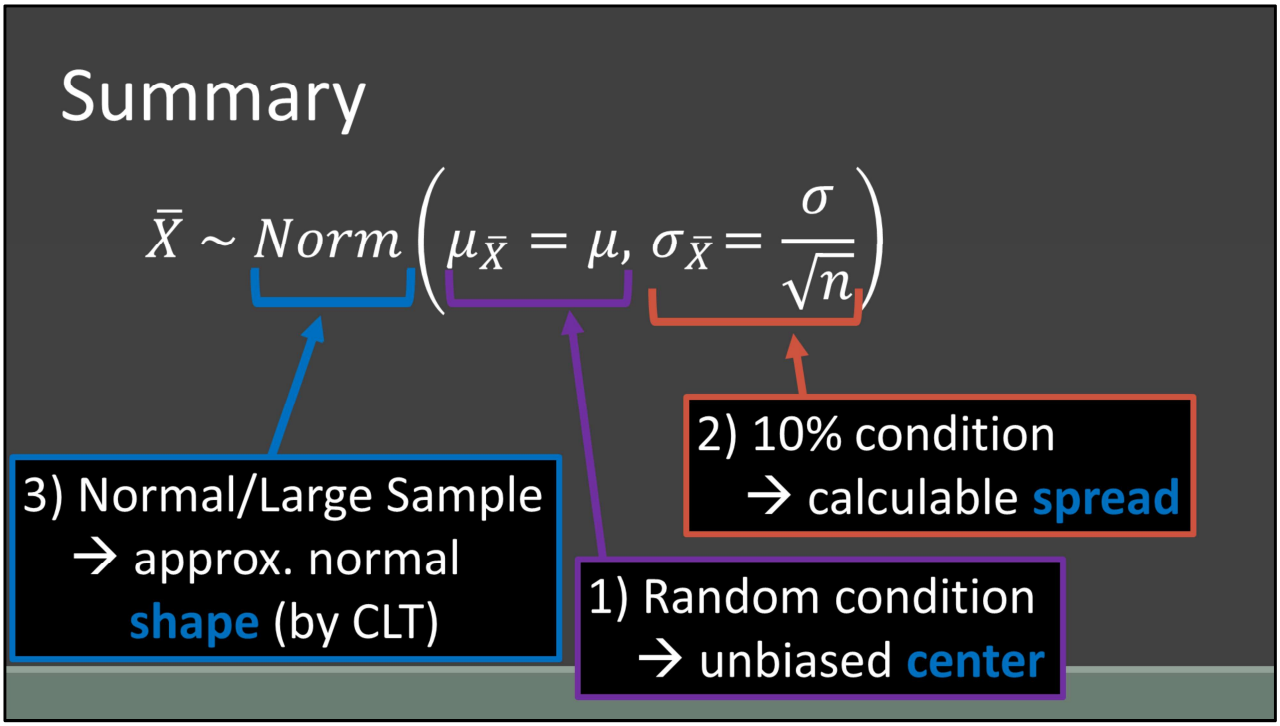
Sampling Distribution of means calculated from samples of 2 data points each

Sampling distribution has skew!


Population
Non-normal population
Means
n = 2
Means
n = 10
Distribution of means calculated from samples of 10 data points each
Still some skew!



所以Sampling Distribution是一个随着sample size的变大，逐渐趋近于Normal



我们来总结一下
对于xbar的分布，我们需要满足几个条件
第一个，如果保证它的期望就是mu的话，需要满足random sample
如果要保证它的方差可以算出来的话，需要满足with replacement或者10% condition，一般来说都是无法满足without replacement，但是这里考点比较灵活，大家也最好知道一下原理，原则上如果with replacement的话，我们的方差是可以保证能算出来的，10% condition只是一个近似。
第三个条件是针对normal的，如果population 就是 normal，那xbar就是完美的normal
如果不是，但是满足了n>=30,那xbar就可以趋近于normal

这三个条件分别对应了三部分，期望，方差和分布类型
这三个条件是相互独立的，比如，如果第一个random sample的条件满足了，其他的都没满足，我还是可以得到xbar的期望是mu这个结论的。

Sampling distributions for differences in sample means

For a numerical variable, when randomly sampling with replacement from two independent populations with population mean $\boldsymbol{\mu_1}$ **and** $\boldsymbol{\mu_2}$ and population standard deviation $\boldsymbol{\sigma_1}$ **and** $\boldsymbol{\sigma_2}$, the sampling distribution of the difference in sample means $\overline{X_1} - \overline{X_2}$ has mean $\boldsymbol{\mu_{\overline{X_1}-\overline{X_2}} = \mu_1 - \mu_2}$ and standard deviation

$$\sigma_{\left(\overline{x_1}-\overline{x_2}\right)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The sampling distribution of the difference in sample means $\overline{X_1} - \overline{X_2}$ can be modeled with a normal distribution **IF**

the two sampling distributions can be modeled with a normal distribution.

If $X \sim N(\mu_X, {\sigma_X}^2)$, $Y \sim N(\mu_Y, {\sigma_Y}^2)$,

X and Y are independent random variables,

then X+Y follows the normal distribution.

The sampling distribution of the difference in sample means $\overline{X_1} - \overline{X_2}$ can be modeled with a normal distribution **IF**

the two sampling distributions can be modeled with a normal distribution:

1. if the population distributions can be modeled with a normal distribution.
2. if cannot, sample sizes have to be greater than or equal to 30.