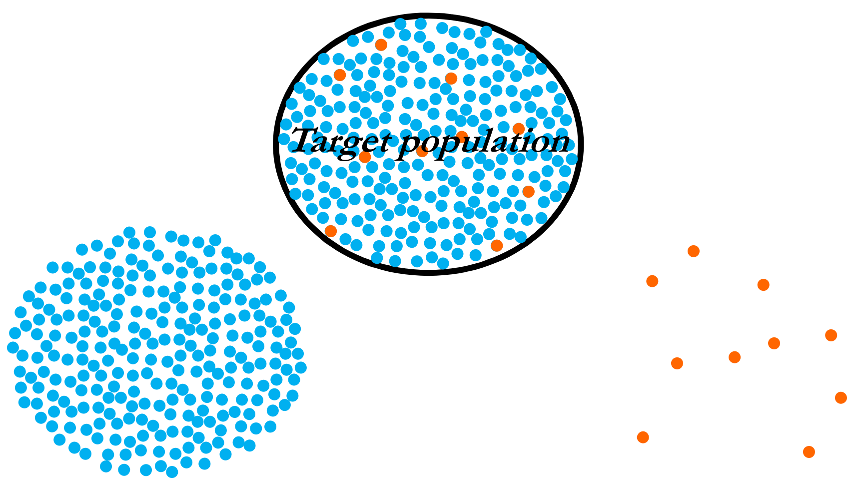
Sampling distributions

PSLS chapter 13 (sections 13.1 through 13.4 only)

Part II of flipped lesson

Two ways to collect data



Population data
(data for all individuals)

→ parameter

Sample data
(data for some individuals)

→ statistic

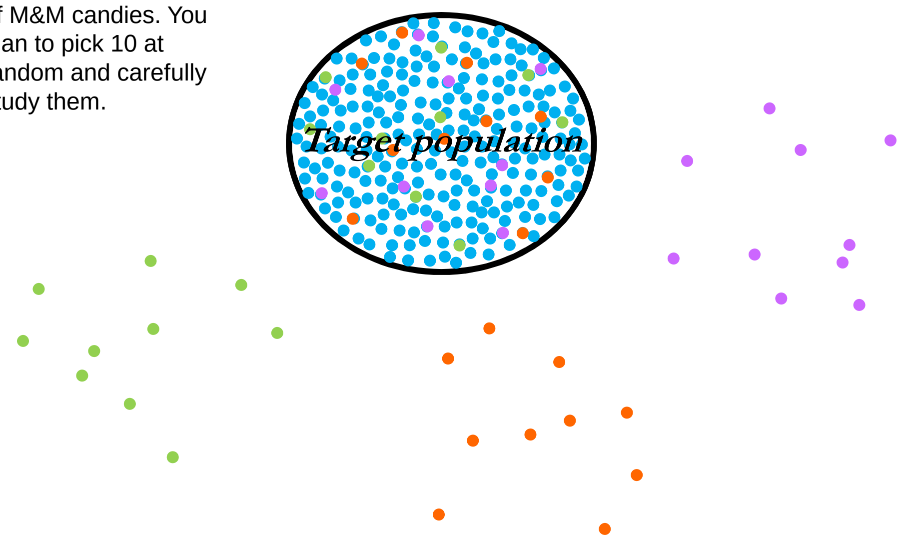
When the sample is random and representative, the sample statistic should be pretty close to the population parameter.

Seems logical, but:

Can we prove that it's true?

And how close is "pretty close"?

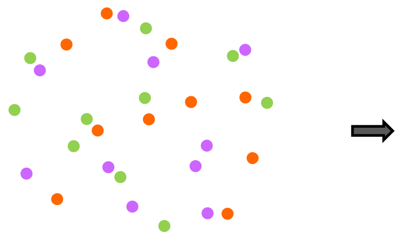
Imagine a huge batch of M&M candies. You plan to pick 10 at random and carefully study them.

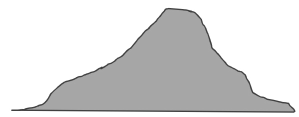


Every random sample will be different

→ Sample statistics vary

Sampling distributions





Values of statistics summarizing all random samples of *n* observations from this population (and their probability distribution)

A statistic computed from a random sample is a random variable.

The probability distribution of a statistic for all random samples of *n* observations from a given population is called the **sampling distribution of the statistic**.

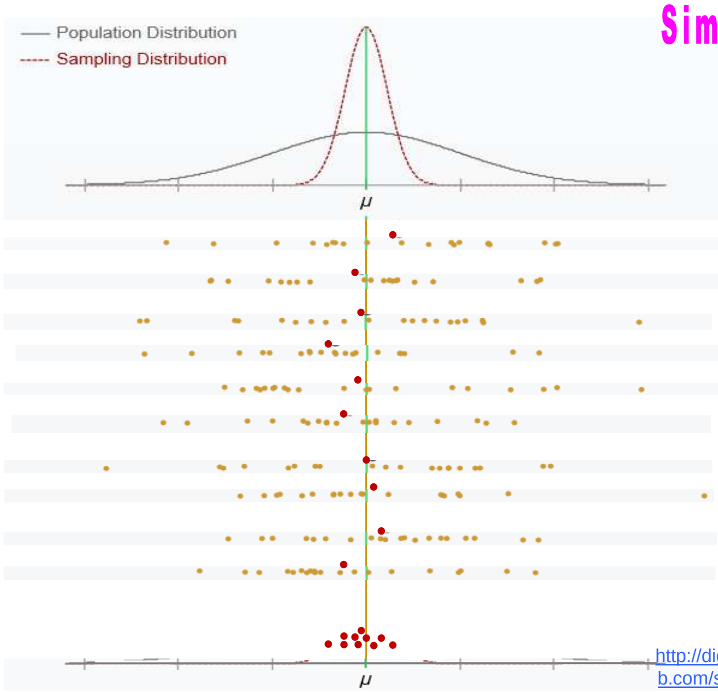
Which characteristic describes the fact that the sample mean does not tend to over- or under-estimate the population mean?

A. sample
B. unbiased
C. consistent
D. statistic
E. parameter

$$\mu_{\bar{x}} = \mu_x$$

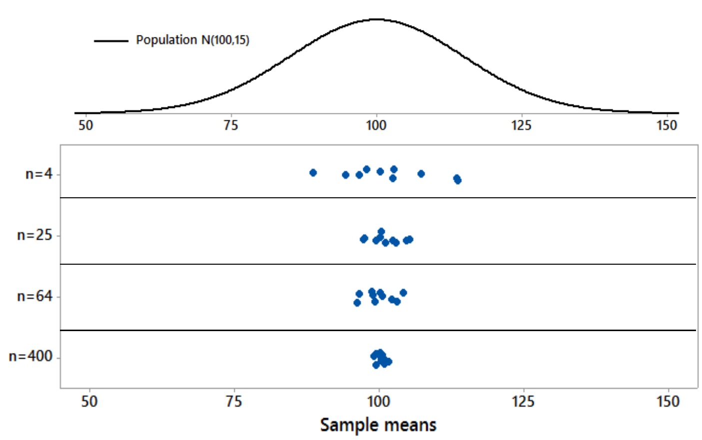
If a population (the variable x) has mean μ, the sampling distribution of sample means (the variable $\bar{x}$) has mean μ as well.

Simulation



http://digitalfirst.bfwpub.com/stats_applet/stats_applet_4_ci.html

Simulation



Each row shows the means $\bar{x}$ of ten simulated random samples for different chosen sample sizes *n*

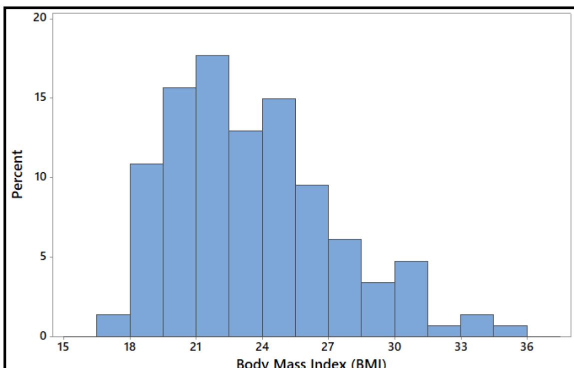
What is the variability of the statistic $\bar{x}$ described by?

The spread of its sampling distribution
The amount of bias present
The vagueness in the wording of the question used to collect the sample data
The stability of the population it describes

If a population (the variable x) has standard deviation σ, the sampling distribution of sample means (the variable $\bar{x}$) has standard deviation $\sigma/\sqrt{n}$.

➔ Larger samples tend to give closer estimates of μ.

$$\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$$



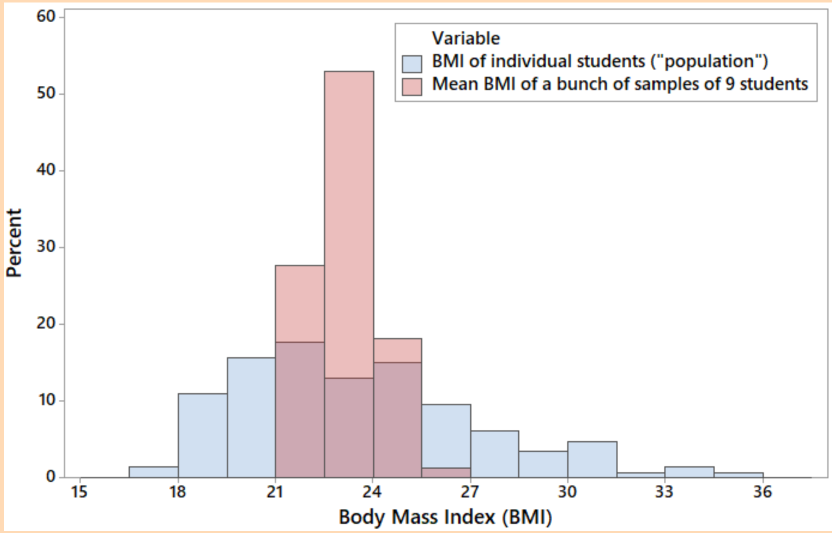
Class survey : weight (lbs) and height (in) used to compute BMI.

Describe the distribution of BMI values in this sample.

What can you assume about the shape of the distribution of BMI values for all UCI students?

Which would be approximately Normal?

A. The sampling distribution of mean BMI for random samples of 9 students
B. The sampling distribution of mean BMI for random samples of 60 students
C. The sample distribution of BMI values in a random sample of 500 students
D. A, B, and C
E. B and C but not A



Variable	N	Mean	StDev	Minimum	Median	Maximum
BMI	147	23.469	3.706	16.820	22.860	35.680
avg9	83	23.021	0.948	21.063	22.853	25.662

(Technically, there are 68.8 trillion different random samples of 9 students that could be picked from a population of 147 students, but I stopped after 83 different random samples…)

Central limit theorem: shape of sampling distribution

The central limit theorem states that, for large enough sample sizes n (depending on the shape of the population), the **sampling distribution** of $\bar{x}$ is approximately Normal.

Very rough guidelines:

- A sample size of 10+: *good enough for most symmetric data distributions*
- A sample size of 25+: *good enough for many situations (no extreme skew)*
- A sample size of 30-40+: *usually good enough even for extreme skews*

http://digitalfirst.bfwpub.com/stats_applet/stats_applet_3_cltmean.html