Bootstrap Estimation of a Sampling Distribution

Background

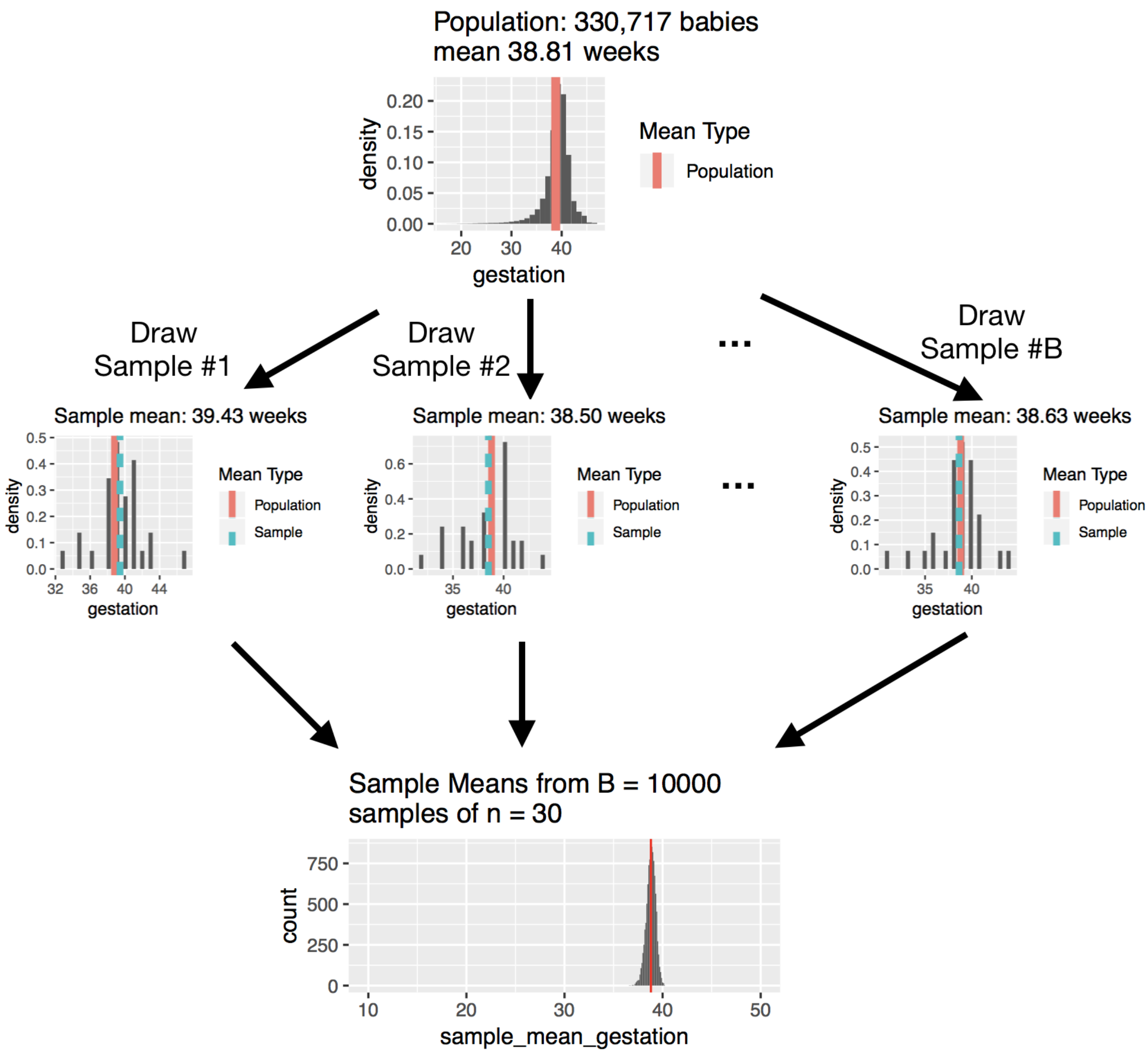
- Confidence intervals are derived from the sampling distribution of a pivotal quantity T
  - Often involves $\hat{\Theta}_{MLE}$ and θ.
- Approaches so far:
  - Get exact sampling distribution (not always possible, depends on correct model specification):
    * If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu, \sigma^2)$ then $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$
    * If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu, \sigma^2)$ then $T = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
  - If n is large, parameter is not on boundary of parameter space, everything is differentiable, …, then approximately $T = \frac{\hat{\Theta}^{MLE} - \theta}{\sqrt{\frac{1}{I(\hat{\theta}^{MLE})}}} \sim \text{Normal}(0, 1)$
- New approach: **simulation-based approximation** to the sampling distribution.
  - In general, this may be used to approximate the sampling distribution of either:
    * the original estimator $\hat{\Theta}$; or
    * a pivotal quantity T based on the estimator, like $T = \frac{\hat{\Theta} - \mu}{SE(\hat{\Theta})}$

Simulation-based approximation to sampling distribution of random variable $\hat{\Theta}$:

Observation: The sampling distribution of $\hat{\Theta}$ is the distribution of values $\hat{\theta}$ obtained from all possible samples of size n.

1. For $b = 1, \dots, B$:
   a. Draw a sample of size n from the population/data model
   b. Calculate the value of the estimate $\hat{\theta}_b$ based on that sample (a number)
2. The distribution of $\{\hat{\theta}_1, \dots, \hat{\theta}_B\}$ from different simulated samples approximates the sampling distribution of the estimator $\hat{\Theta}$.

Example: We have data that contains a record of the gestation time (how many weeks pregnant the mother was when she gave birth) for the population of every baby born in December 1998 in the United States.



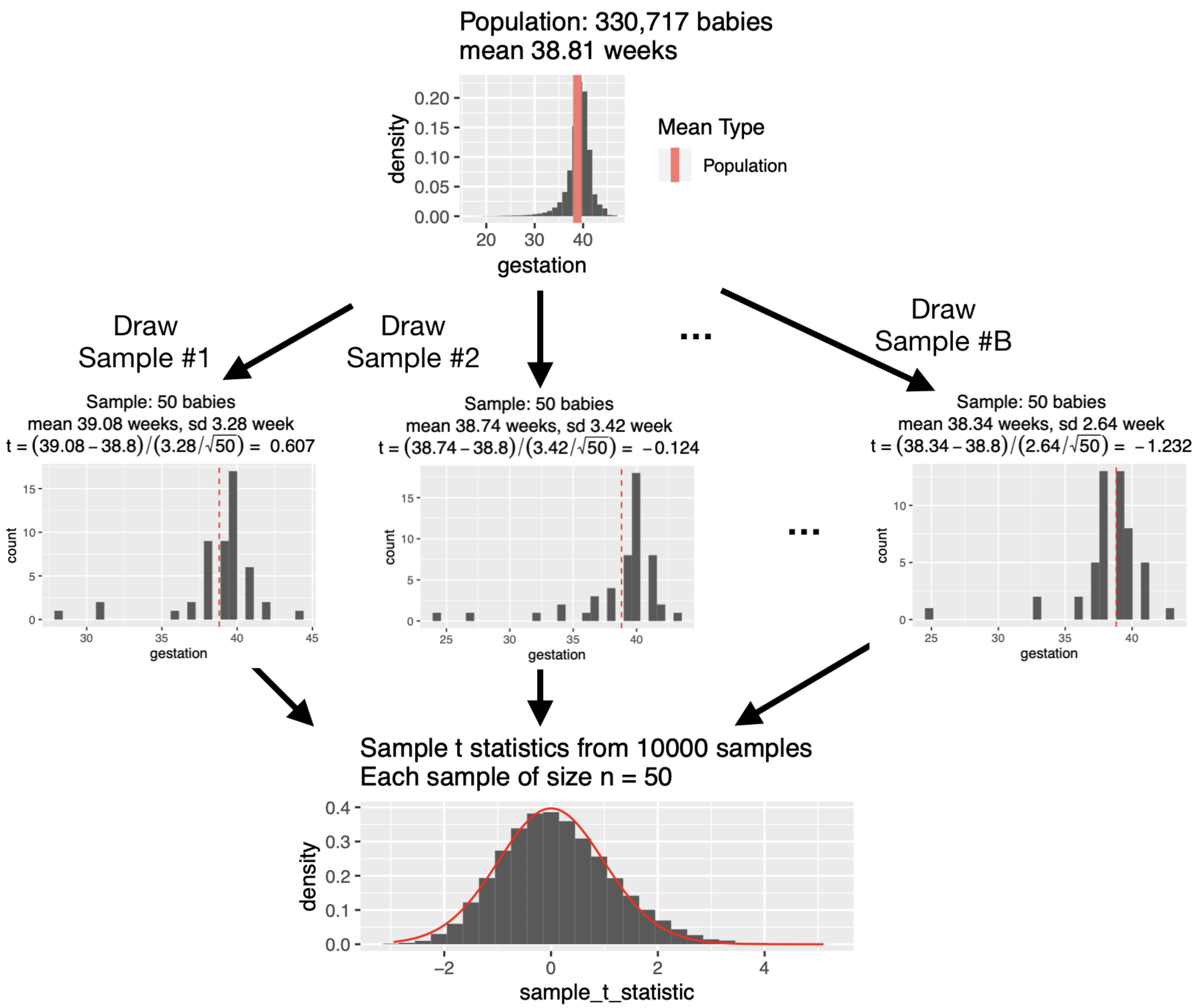
- As $B \to \infty$, we get a better approximation to the distribution of $\hat{\Theta}$
- **Challenge**: If we don't know the population distribution, we can't simulate samples from the population

Simulation-based approximation to sampling distribution of random variable T:

Observation: The sampling distribution of T is the distribution of values t obtained from all possible samples of size n.

1. For $b = 1, \dots, B$:
   a. Draw a sample of size n from the population/data model
   b. Calculate the value of t_b based on that sample (a number)
2. The distribution of $\{t_1, \dots, t_B\}$ from different simulated samples approximates the sampling distribution of the pivotal quantity T.

Example: We have data that contains a record of the gestation time (how many weeks pregnant the mother was when she gave birth) for the population of every baby born in December 1998 in the United States.



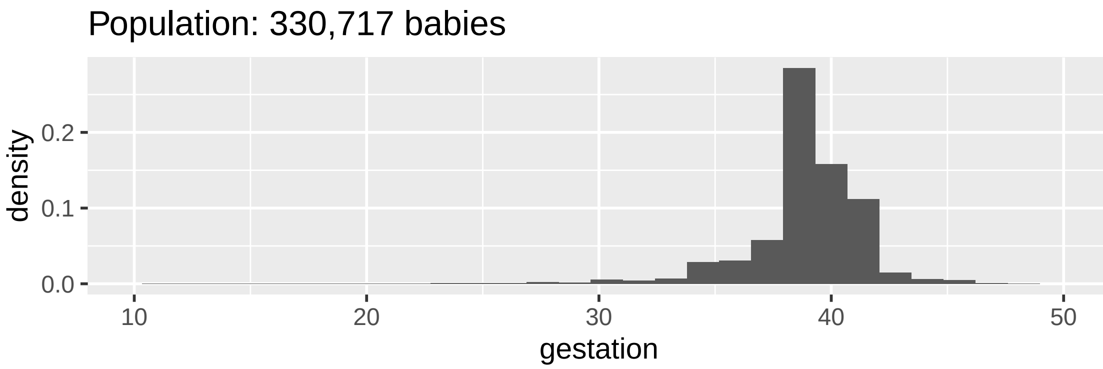
- As $B \to \infty$, we get a better approximation to the distribution of T
- **Challenge**: If we don't know the population distribution, we can't simulate samples from the population

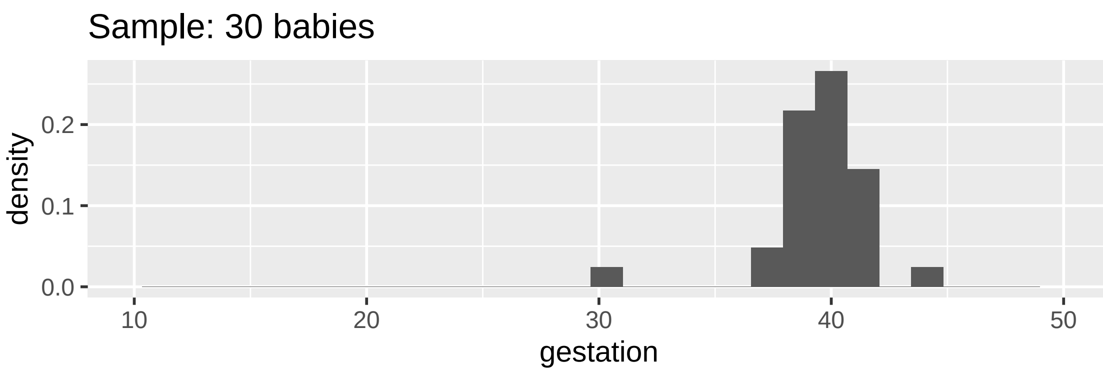
Idea:

- Treat the distribution of the data in our sample as an estimate of the population distribution

Suppose we have a sample of 30 babies. How does its distribution compare to the population distribution?

View 1: In terms of histograms (think pdfs):

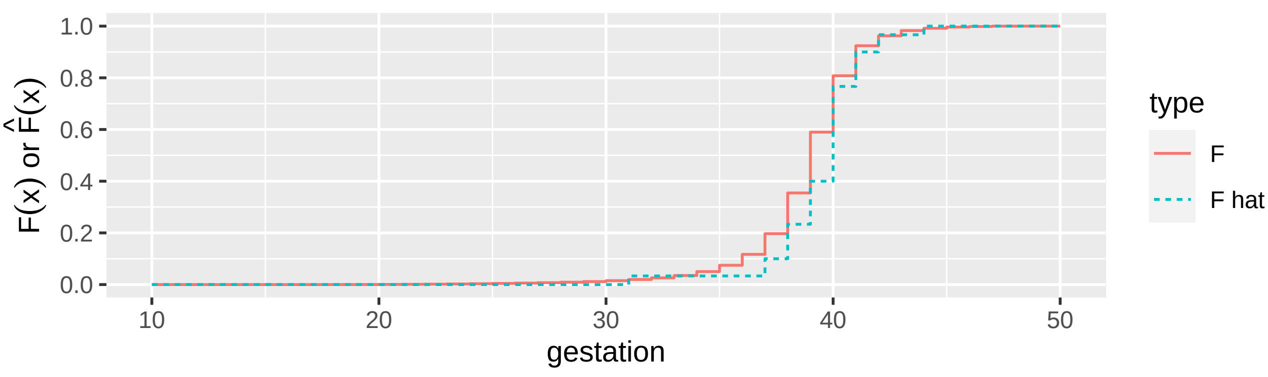




View 2: In terms of cdfs

Recall that $F_X(x) = P(X \leq x)$

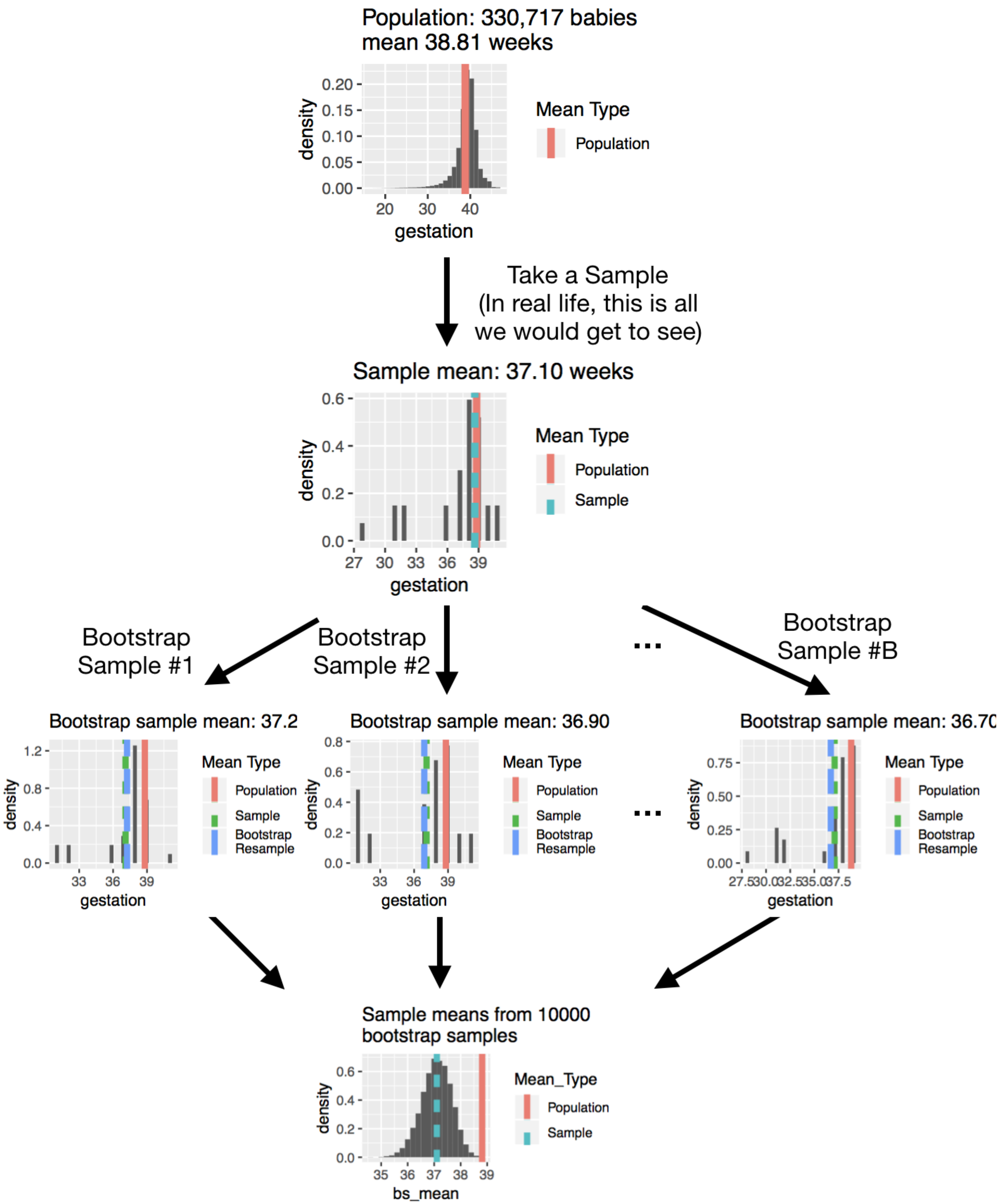
Based on an observed sample $x_1, \ldots, x_n$ each drawn independently from the distribution with pdf $f_X(x_i)$ and cdf $F_X(x_i)$, we estimate $F_X(x)$ by the *empirical cdf*: $\hat{F}_X(x) = \frac{\text{\# in sample} \leq \text{x}}{n}$



If $\hat{F}_X(x)$ (or $\hat{f}_X(x)$) is a good estimate of $F_X(x)$ (or $f_X(x)$), then a sample drawn from the distribution specified by $\hat{F}_X(x)$ will look similar to a sample drawn from $F_X(x)$.

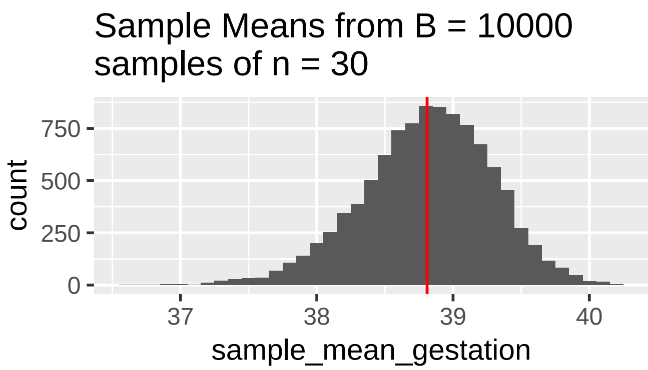
- Instead of repeatedly drawing samples from $F_X(x)$ to approximate the sampling distribution of $\hat{\Theta}$ or T, repeatedly draw samples from $\hat{F}_X(x)$.
- In practice, this means (repeatedly) draw a sample of size n **with replacement** from our observed data.

1. For $b = 1, \ldots, B$:
   a. Draw a bootstrap sample of size n **with replacement** from the observed data
   b. Calculate the estimate $\hat{\theta}_b$ based on that bootstrap sample (a number)
2. The distribution of estimates $\{\hat{\theta}_1, \ldots, \hat{\theta}_B\}$ from different simulated samples approximates the sampling distribution of the estimator $\hat{\theta}$ (the random variable).

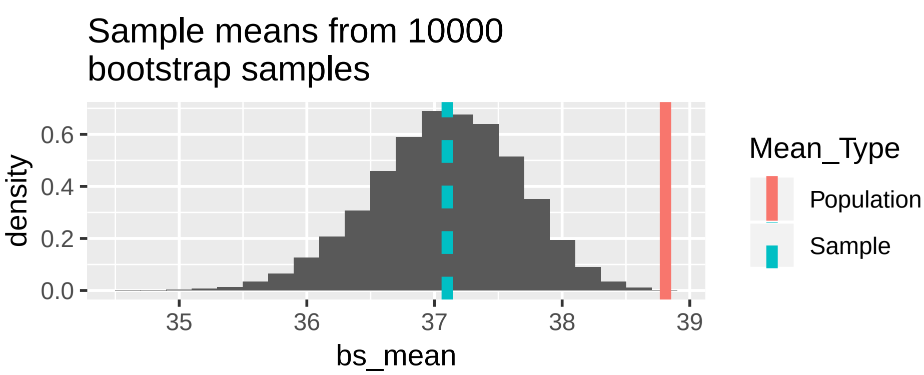


Compare the approximations from sampling directly from the population and from bootstrap resampling:

Many means, based on samples from the population:



Many means, based on bootstrap resamples with replacement from the sample:



- Properties:
  - Bootstrap distribution **reproduces shape, variance, and bias** of actual sampling distribution
  - Bootstrap distribution **does not reproduce mean** of actual sampling distribution
    - E.g., centered at sample mean instead of population mean