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.

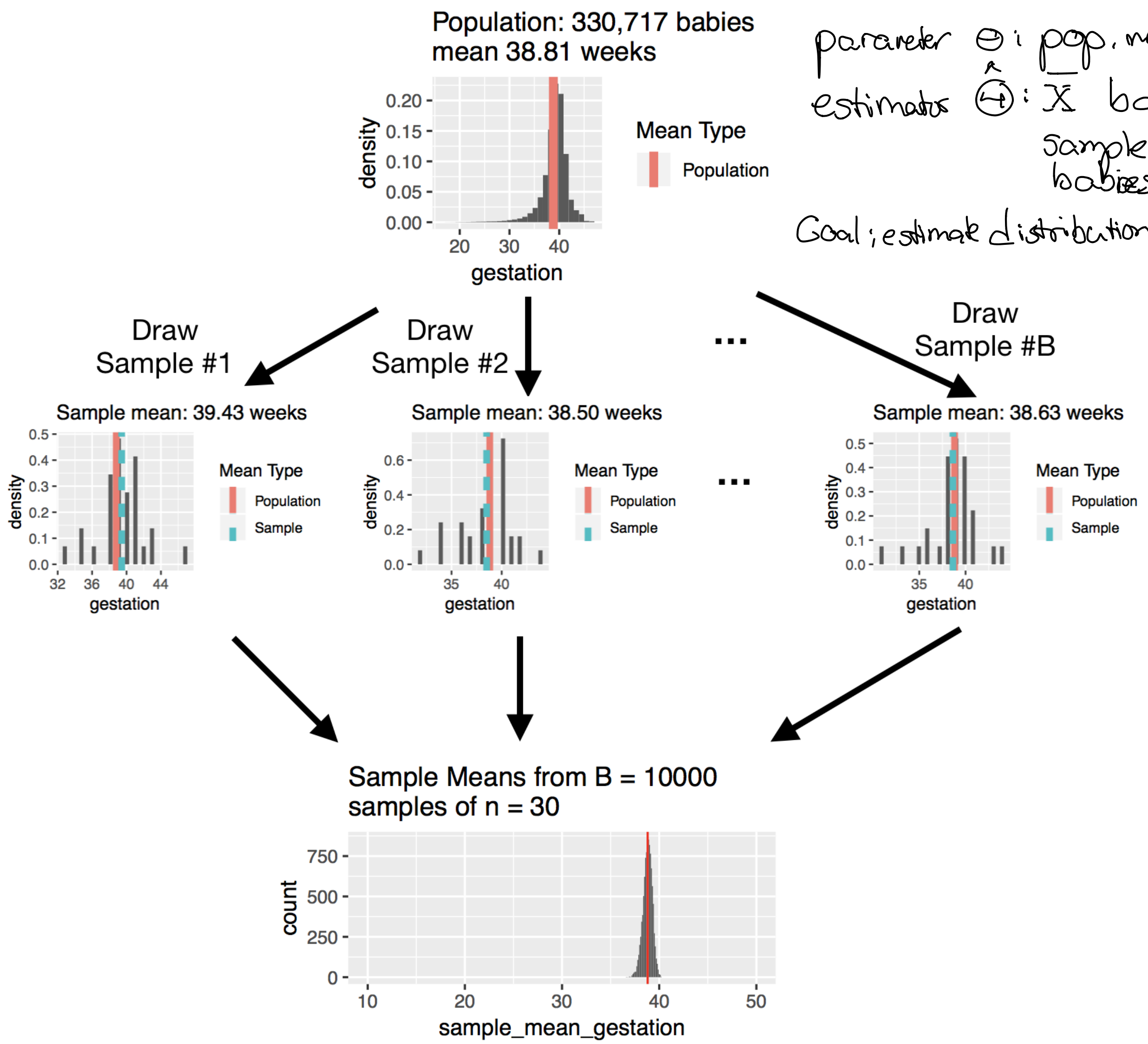
1. For $b = 1, \dots, B$:
   a. Draw a sample of size n from the population/data model
   b. Calculate the value of the estimate $\hat{\theta}_b$ based on that sample (a number)
2. The distribution of $\{\hat{\theta}_1, \dots, \hat{\theta}_B\}$ from different simulated samples approximates the sampling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

parameter Θ: pop. mean, 38.81

estimator $\hat{\Theta}$: $\bar{X}$ based on a sample of $n=30$ babies.

Goal: estimate distribution of $\hat{\Theta}$.



- 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

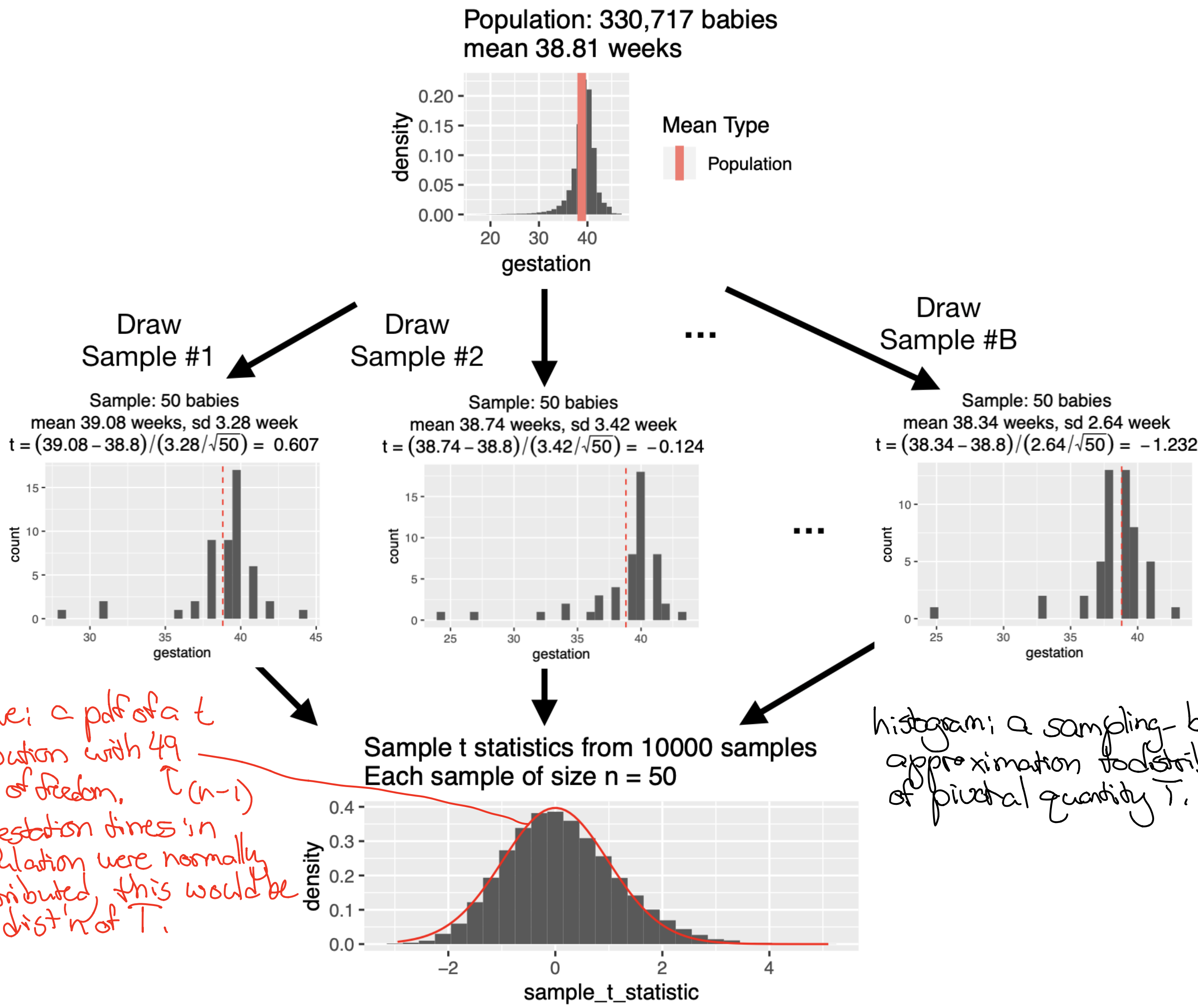
$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

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.



red curve: a pdf of a t distribution with 49 degrees of freedom. $t_{(n-1)}$

- If gestation times in population were normally distributed, this would be the dist'n of T.

histogram: a sampling-based approximation to distribution of pivotal quantity T.

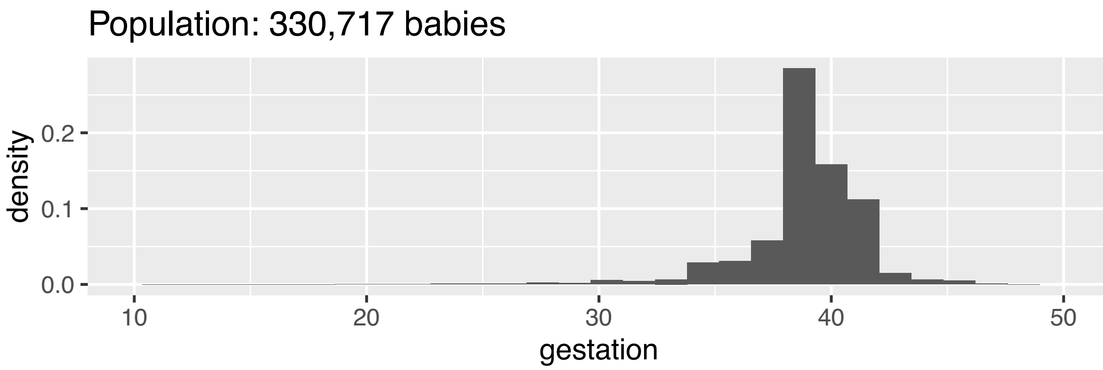
- As $B \to \infty$, we get a better approximation to the distribution of T
- **Challenge**: If we don't know the population distribution, we can't simulate samples from the population

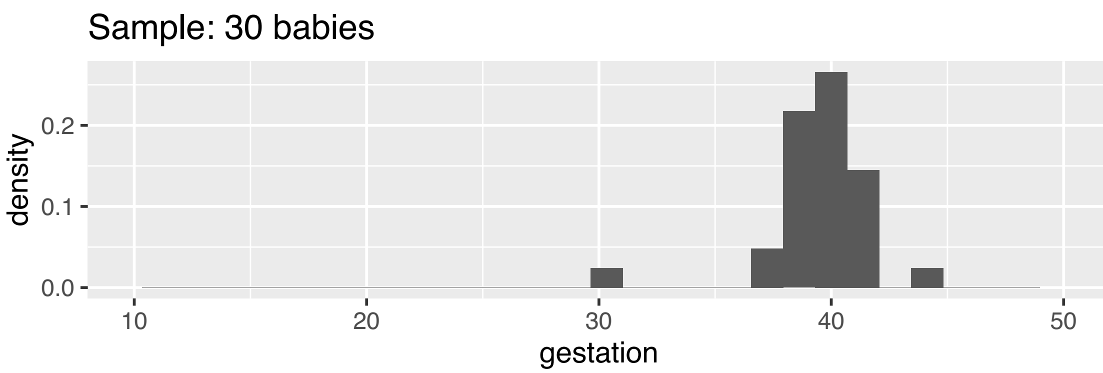
Idea:

- 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):



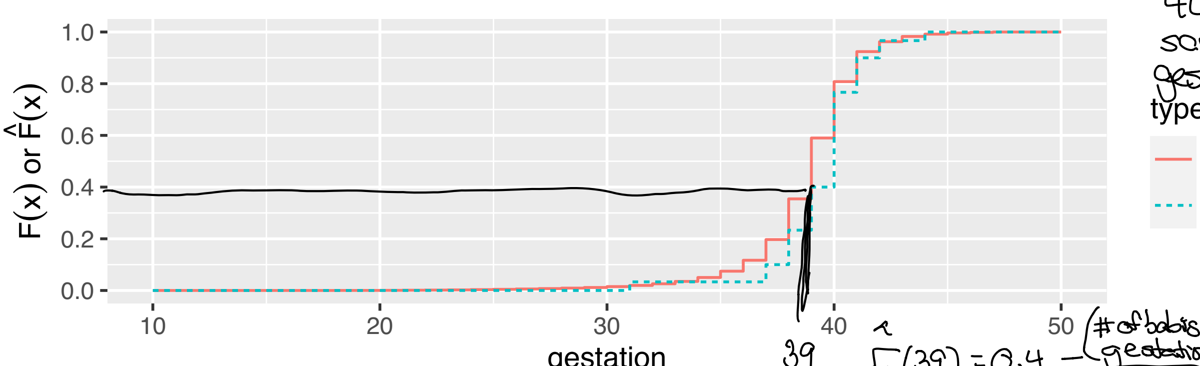


View 2: In terms of cdfs

Recall that $F_X(x) = P(X \leq x)$ ← describing distribution in population.

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*: $\hat{F}_X(x) = \frac{\text{\# in sample} \leq \text{x}}{n}$

40% of babies in sample had a gestation time ≤ 39 weeks



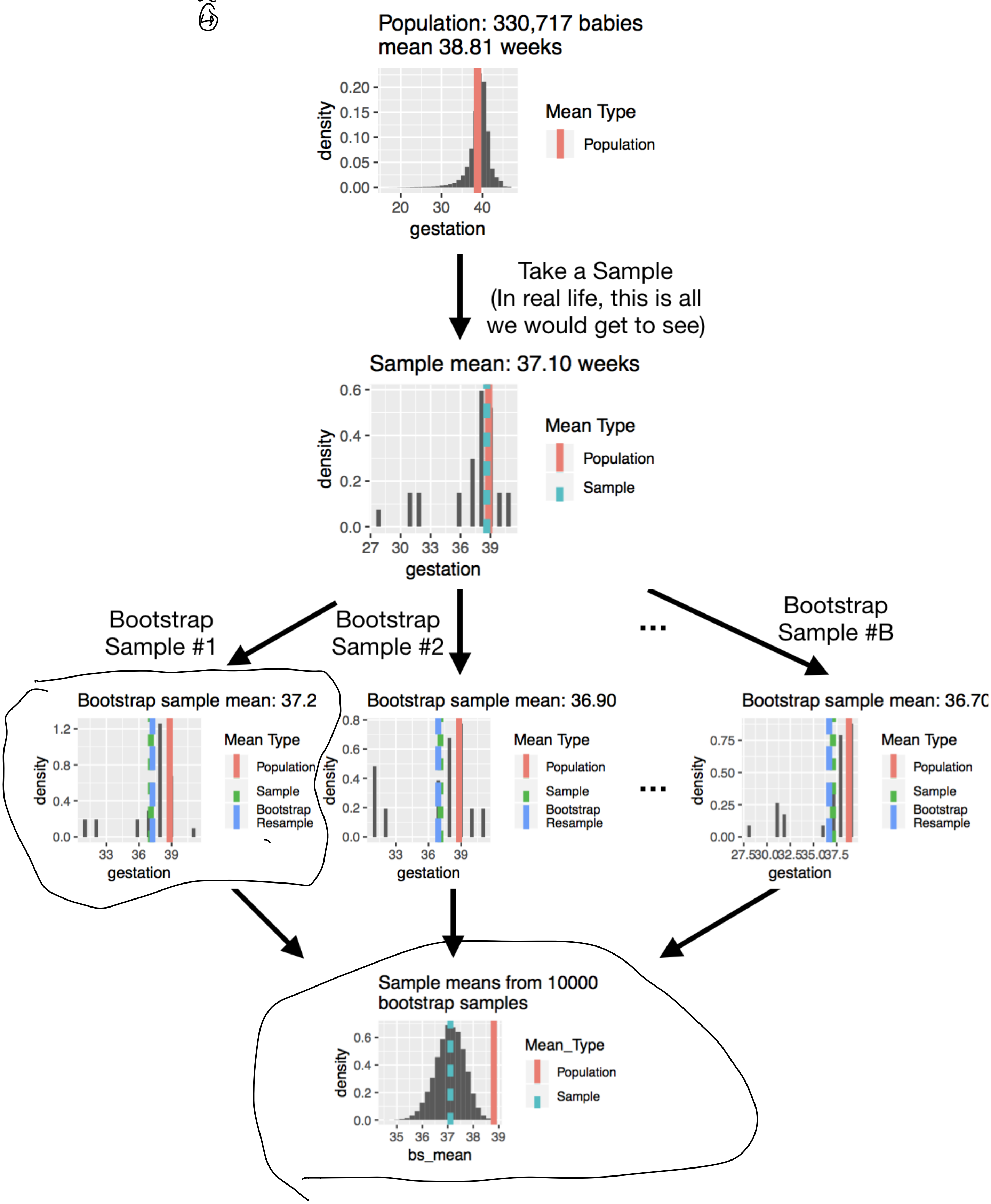
39

$F(39) = 0.4 = \frac{(\text{\# of babies in sample with gestation time} \leq 39)}{30}$

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 $F_X(x)$.

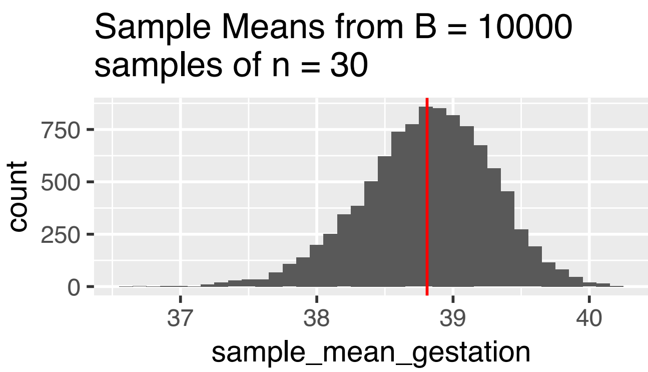
- Instead of repeatedly drawing samples from $F_X(x)$ to approximate the sampling distribution of $\hat{\Theta}$ or T, repeatedly draw samples from $\hat{F}_X(x)$.
- In practice, this means (repeatedly) draw a sample of size n **with replacement** from our observed data.

1. For $b = 1, \dots, B$:
   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 $\{\hat{\theta}_1, \dots, \hat{\theta}_B\}$ from different simulated samples approximates the sampling distribution of the estimator $\cancel{\hat{\theta}}$ $\hat{\Theta}$ (the random variable).

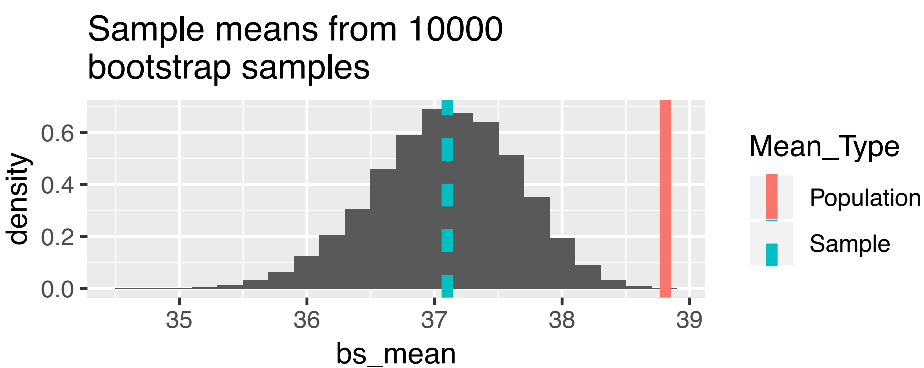


Compare the approximations from sampling directly from the population and from bootstrap resampling:

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

Justification part 1:

Show $\hat{F}_{\underline{X}}(x) \to F_{\underline{X}}(x)$ as $n \to \infty$

$$\lim_{n\to\infty} \hat{F}_{\underline{X}}(x) = \lim_{n\to\infty} \frac{\#\text{ of obs in sample} \leq x}{n}$$

$$\mathbb{I}_{(-\infty,x]}(x_i) = \begin{cases} 1 \text{ if } x_i \leq x \\ 0 \text{ if } x_i > x \end{cases}$$

$$= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(x_i)}{n}$$

$$= E\left[\mathbb{I}_{(-\infty,x]}(\underline{X}_i)\right] \qquad \text{(Law of Large Numbers)}$$

$$= \int_{-\infty}^{\infty} \mathbb{I}_{(-\infty,x]}(x_i) \cdot f_{\underline{X}_i}(x_i)\, dx_i$$

$$= \int_{-\infty}^{x} \cancelto{1}{\mathbb{I}_{(-\infty,x]}}(x_i) \cdot f_{\underline{X}_i}(x_i)\, dx_i + \int_{x}^{\infty} \cancelto{0}{\mathbb{I}_{(-\infty,x]}}(x_i)\, f_{\underline{X}_i}(x_i)\, dx_i$$

$$= \int_{-\infty}^{x} f_{\underline{X}_i}(x_i)\, dx_i$$

$$= P(\underline{X}_i \leq x)$$

$$= F_{\underline{X}_i}(x)$$

$$= F_{\underline{X}}(x)$$ (since observations all from population with cdf $F_{\underline{X}}(x)$)

Suppose $X_1, \ldots, X_n \overset{iid}{\sim}$ with pdf $f_{X|\Theta}(x|\theta)$

Our estimator is a function of $X_1, \ldots, X_n$:

$$\hat{\Theta} = g(X_1, \ldots, X_n)$$

For example if $\hat{\Theta} = \bar{X}$, $g(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$,

The distribution of $\hat{\Theta}$ is determined by its cdf:

$$F_{\hat{\Theta}}(\theta^*) = P(\hat{\Theta} \leq \theta^*)$$

$$= P(\underline{g(X_1, \ldots, X_n) \leq \theta^*})$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{I}_{(-\infty, \theta^*]}(g(x_1, \ldots, x_n)) \cdot f_X(x_1) \cdots f_X(x_n)\, dx_1 \cdots dx_n$$

$$\approx \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{I}_{(-\infty, \theta^*]}(g(x_1, \ldots, x_n))\, \hat{f}_X(x_1) \cdots \hat{f}_X(x_n)\, dx_1, \ldots, dx_n$$

(for large n, by LLN from previous page)

$$\approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}_{(-\infty, \theta^*]}(g(x_1^{(b)}, \ldots, x_n^{(b)}))$$

where $x_1^{(b)}, \ldots, x_n^{(b)} \overset{iid}{\sim} \hat{f}_X(x)$ for $b = 1, \ldots, B$

(by LLN)

Notes:

1) Bootstrap is not theoretically justified if n is small. Proof relied on law of large numbers.

   → in practice, OK for a "moderately large" sample size.

   → better than a large-sample normal approx.

2) We get to pick B (# of bootstrap samples). The larger the better.

   If feasible, $B \approx 10,000$.