Bootstrap Estimation of a Sampling Distribution

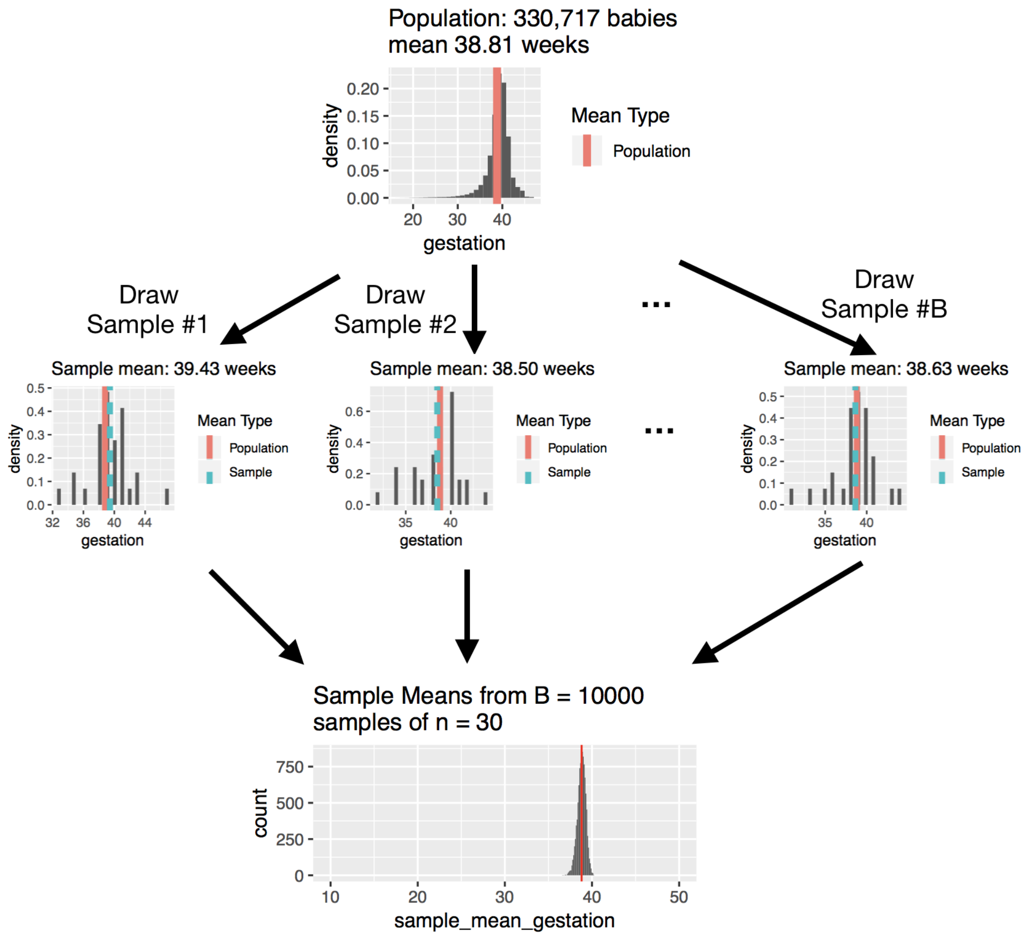
Background

- Confidence intervals are derived from the sampling distribution of an estimator like $\hat{\theta}_{MLE}$.
- The sampling distribution is the distribution of estimates $\hat{\theta}_{MLE}$ obtained from all possible samples of size n.
- Approaches so far:
  - Get exact sampling distribution (not always possible, depends on correct model specification):
    - If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu, \sigma^2)$ then $t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$
    - If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu, \sigma^2)$ then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
  - If n is large, parameter is not on boundary of parameter space, everything is differentiable, $\dots$, then $\hat{\theta}^{MLE} \sim \text{Normal}(\theta, \frac{1}{I(\hat{\theta}^{MLE})})$
- New approach: **simulation-based approximation** to the sampling distribution

Simulation-based approximation to sampling distribution, if population distribution is known:

1. For $b = 1, \dots, B$:
   a. Draw a sample of size n from the population/data model
   b. Calculate the estimate $\hat{\theta}_b$ based on that 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 $\hat{\theta}$ (the random variable).

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.



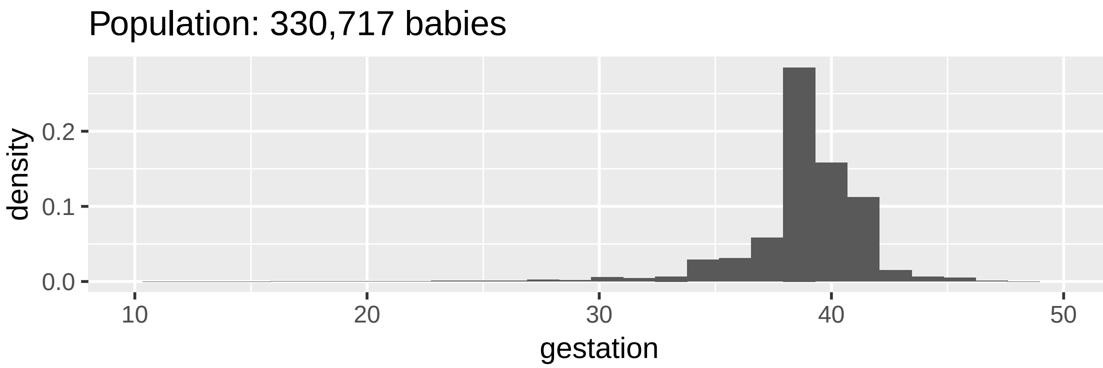
- As $B \to \infty$, we get a better approximation to the distribution of $\hat{\theta}$
- **Challenge**: If we don't know the population distribution, we can't simulate samples from the population

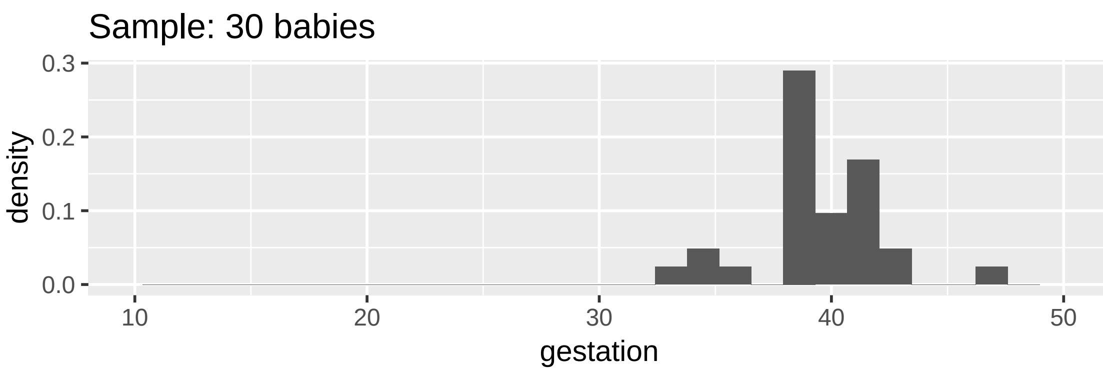
Idea:

- 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):



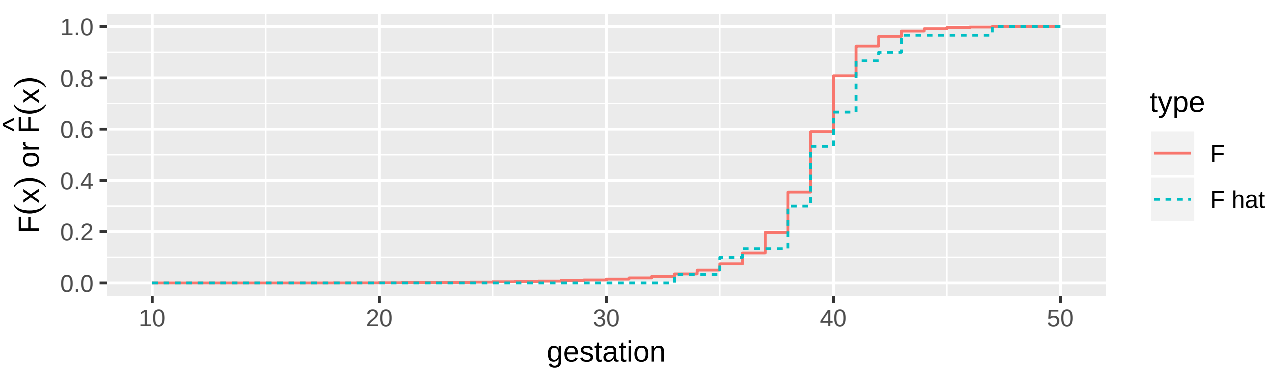


View 2: In terms of cdfs

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

Based on a sample, this is estimated by the *empirical cdf*: $\hat{F}_X(x) = \frac{\text{\# in sample} \leq \text{x}}{n}$

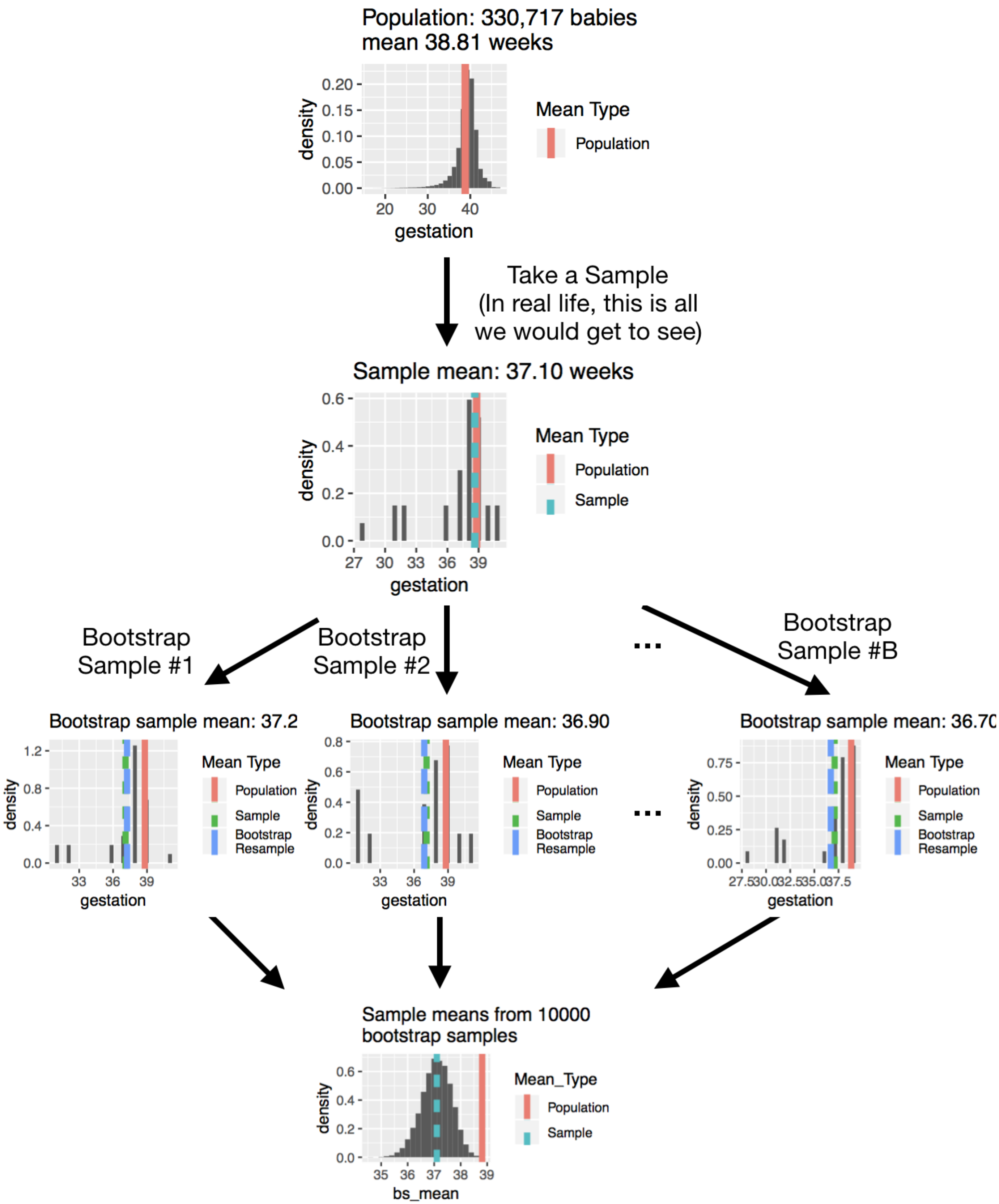
If n is large, $\hat{F}_X(x)$ will be a good approximation to $F_X(x)$.



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 $F_X(x)$.

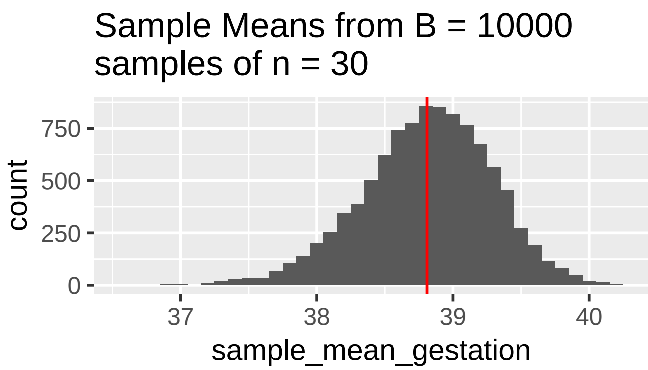
- Instead of repeatedly drawing samples from $F_X(x)$ to approximate the sampling distribution of $\hat{\theta}$, repeatedly draw samples from $\hat{F}_X(x)$.
- In practice, this means (repeatedly) draw a sample of size n **with replacement** from our observed data.

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 $\hat{\theta}$ (the random variable).

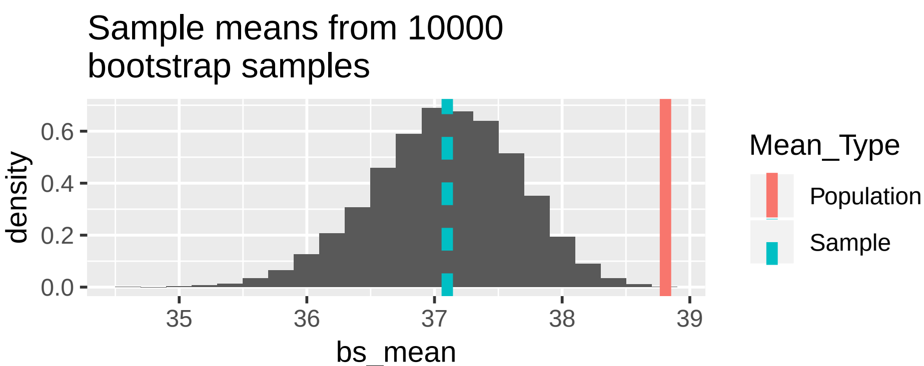


Compare the approximations from sampling directly from the population and from bootstrap resampling:

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
- Sketch of more formal justification:
  - Suppose $X_1, \ldots, X_n$ are i.i.d. with pdf $f_X(x)$
  - Our estimator is a function of $X_1, \ldots, X_n$: $\hat{\theta} = g(X_1, \ldots, X_n)$
  - The sampling distribution of $\hat{\theta}$ is determined by its cdf

$$
\begin{aligned}
F_{\hat{\theta}}(\theta^*) &= P(\hat{\theta} \leq \theta^*) \\
&= P(g(X_1, \ldots, X_n) \leq \theta*) \\
&= \int \cdots \int_{\{x_1, \ldots, x_n : g(x_1, \ldots x_n) \leq \theta^*\}} f_X(x_1) \cdots f_X(x_n) dx_1 \cdots dx_n \\
&= \int \cdots \int_{\{x_1, \ldots, x_n\}} \mathbb{I}_{(-\infty, \theta^*]}\{g(x_1, \ldots x_n)\} f_X(x_1) \cdots f_X(x_n) dx_1 \cdots dx_n \\
&\approx \int \cdots \int_{x_1, \ldots, x_n} \mathbb{I}_{(-\infty, \theta^*]}\{g(x_1, \ldots x_n)\} \hat{f}_X(x_1) \cdots \hat{f}_X(x_n) dx_1 \cdots dx_n && \text{if } n \text{ is large, } \hat{f}_X(x) \approx f_X(x) \\
&\approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}_{(-\infty, \theta^*]}\{g(x_1^{(b)}, \ldots x_n^{(b)})\} && \text{Law of Large Numbers, if } x_1^{(b)}, \ldots, x_n^{(b)} \overset{\text{i.i.d.}}{\sim} \hat{f}_X(x)
\end{aligned}
$$

- The last two lines above involve approximations.
- **Note 1:** It's sometimes claimed that the bootstrap can help with small sample sizes; this is FALSE. Second-to-last equation above is a large-n approximation of $f_X(x)$ with $\hat{f}_X(x)$. In practice, this is useful for moderate sample sizes.
- **Note 2:** As long as $B \approx 1000$ or so, the approximation in the last equation is typically good enough

Example with Poisson data (one last time!)

Recall our Poisson data about asbestos fiber counts:

31, 29, 19, 18, 31, 28, 34, 27, 34, 30, 16, 18, 26, 27, 27, 18, 24, 22, 28, 24, 21, 17, 24

Sample mean: $\bar{x} = 24.9$

Model: $X_i \sim \text{Poisson}(\lambda)$

The maximum likelihood estimate is $\hat{\lambda}_{MLE} = \bar{X} = 24.9$

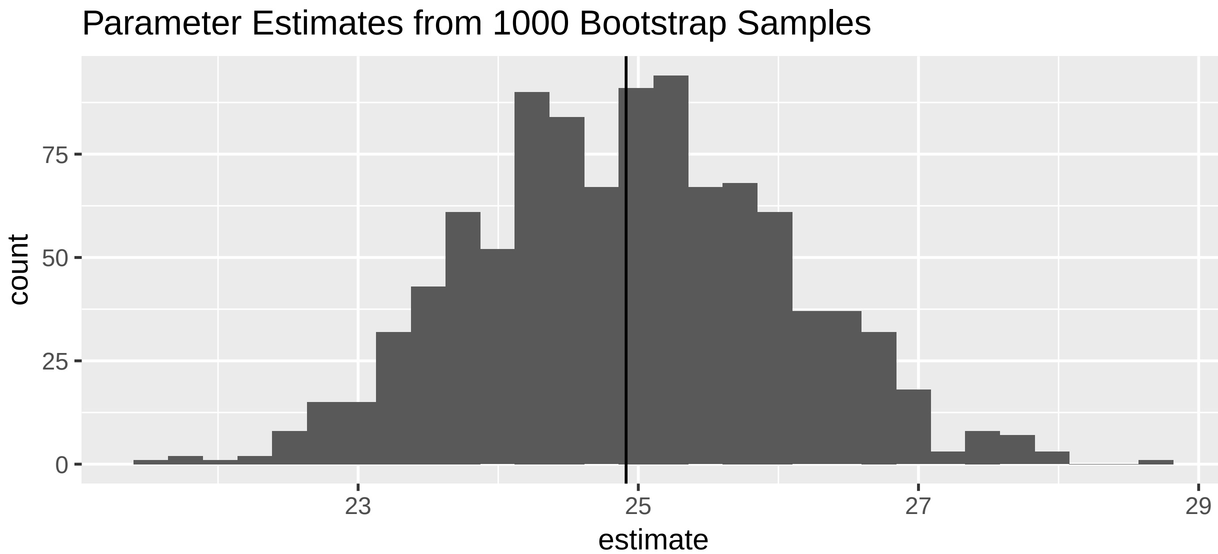
A bootstrap-based estimate of the sampling distribution of $\hat{\lambda}_{MLE}$:

```
# the dplyr package contains the sample_n function,
# which we use below to draw the bootstrap samples
library(dplyr)

# observed data: 23 counts of asbestos fibers
sample_obs <- data.frame(
  fiber_count = c(31, 29, 19, 18, 31, 28, 34, 27, 34, 30, 16, 18, 26, 27, 27, 18, 24,
                  22, 28, 24, 21, 17, 24)
)
# number of observations in sample_obs
n <- 23

# how many bootstrap samples to take, and storage space for the results
num_bootstrap_samples <- 10^3
bootstrap_estimates <- data.frame(
  estimate = rep(NA, num_bootstrap_samples)
)

# draw many samples from the observed data and calculate mean of each simulated sample
for(i in seq_len(num_bootstrap_samples)) {
  ## Draw a bootstrap sample of size n with replacement from the observed data
  bootstrap_resampled_obs <- sample_obs %>%
    sample_n(size = n, replace = TRUE)

  ## Calculate mean of bootstrap sample
  bootstrap_estimates$estimate[i] <- mean(bootstrap_resampled_obs$fiber_count)
}
```

Plot of bootstrap estimate of sampling distribution

- Note that this is centered at $\hat{\lambda}_{MLE}$ based on our sample, not at the true λ – but it should otherwise look similar to the actual sampling distribution (if we think $n = 23$ is large enough).

```
library(ggplot2)
ggplot(data = bootstrap_estimates, mapping = aes(x = estimate)) +
  geom_histogram(bins = 30) +
  geom_vline(
    mapping = aes(xintercept = mean(sample_obs$fiber_count))) +
  ggtitle("Parameter Estimates from 1000 Bootstrap Samples")
```



Bootstrap Estimate of Bias:

Actual bias is $E(\hat{\lambda}_{MLE}) - \lambda$, which we have shown to be 0

Estimate bias by (Average of bootstrap estimates) − (Estimate from our actual sample) $= \frac{1}{B}\sum_{i=1}^{n}\hat{\lambda}^{(b)} - \hat{\lambda}_{MLE}$

```
mean(bootstrap_estimates$estimate) - mean(sample_obs$fiber_count)
```

```
## [1] 0.01821739
```