# Example: Bootstrap Estimation of a Sampling Distribution

## Example with Poisson data

The National Institute of Standards and Technology conducted a study to evaluate a method for measuring concentration of asbestos (a cancer-causing substance). Their procedure involves counting the number of asbestos fibers captured in several regions of a filter. Here are the resulting asbestos fiber counts from the study:

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

The sample mean is $\bar{x} = 24.9$

We adopt a Poisson model for these data: $X_i \sim \text{Poisson}(\lambda)$

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

Here is code to obtain a bootstrap-based estimate of the sampling distribution of $\hat{\lambda}_{MLE}$:

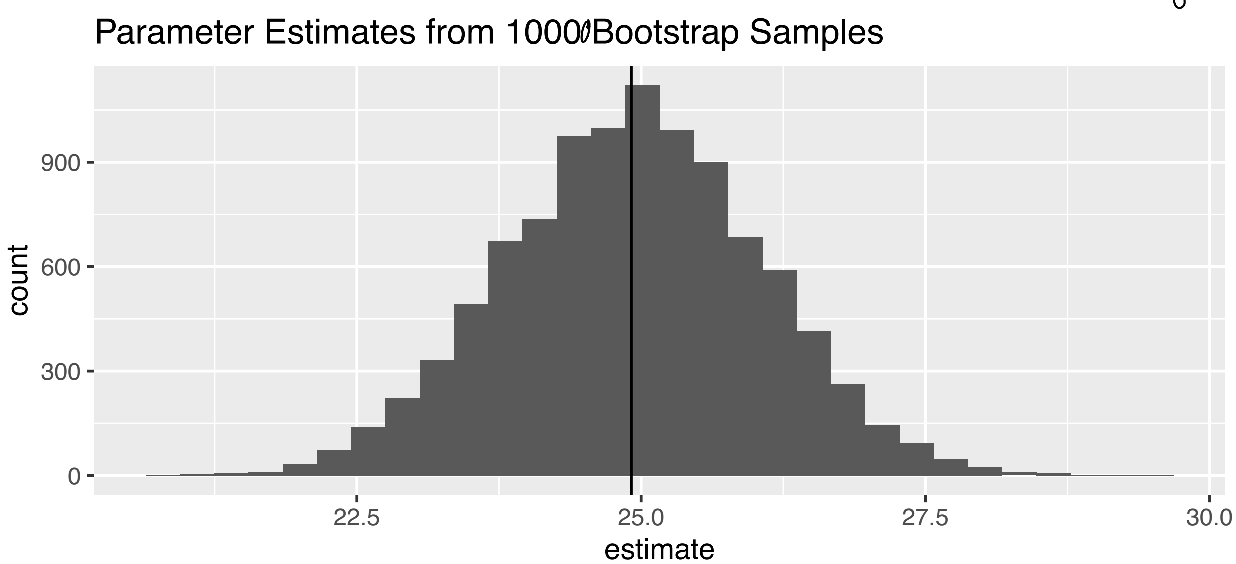
```
# the dplyr package contains the sample_n function,
# which we use below to draw the bootstrap samples
library(dplyr)

# 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 <- 10^4
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 $\lambda$ – 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")
```

→ 24.9 (mean from original sample)




**Bootstrap Estimate of Bias:**

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

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

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

```
## [1] 0.01425217
```

**Bootstrap Standard Error:**

```
sd(bootstrap_estimates$estimate)
```

```
## [1] 1.12311
```

Translating bias from
population ⟷ estimator based on sample
to
sample ⟷ estimator based on a bootstrap sample.

Bias: $E(\hat{\lambda}^{MLE}) - \lambda$

↳ $E[\hat{\lambda}^{(b)}]$ − (mean from original sample)

$\frac{1}{B}\sum_{i=1}^{B} \hat{\lambda}^{(b)} - \hat{\lambda}^{MLE}$