Bootstrap Confidence Intervals

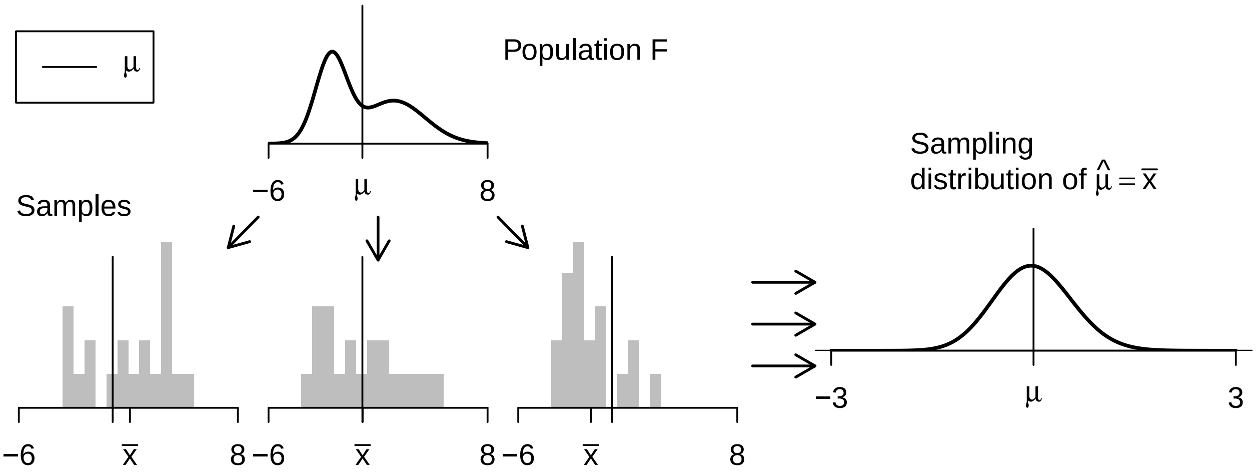
Part 1 - Bootstrap Percentile Confidence Intervals

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.

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

1) Simulate many samples from the population/data model
2) For each simulated sample, calculate the estimate $\hat{\theta}_{MLE}$
3) The distribution of estimates from different simulated samples approximates the sampling distribution of the estimator $\hat{\theta}_{MLE}$.



Challenge:

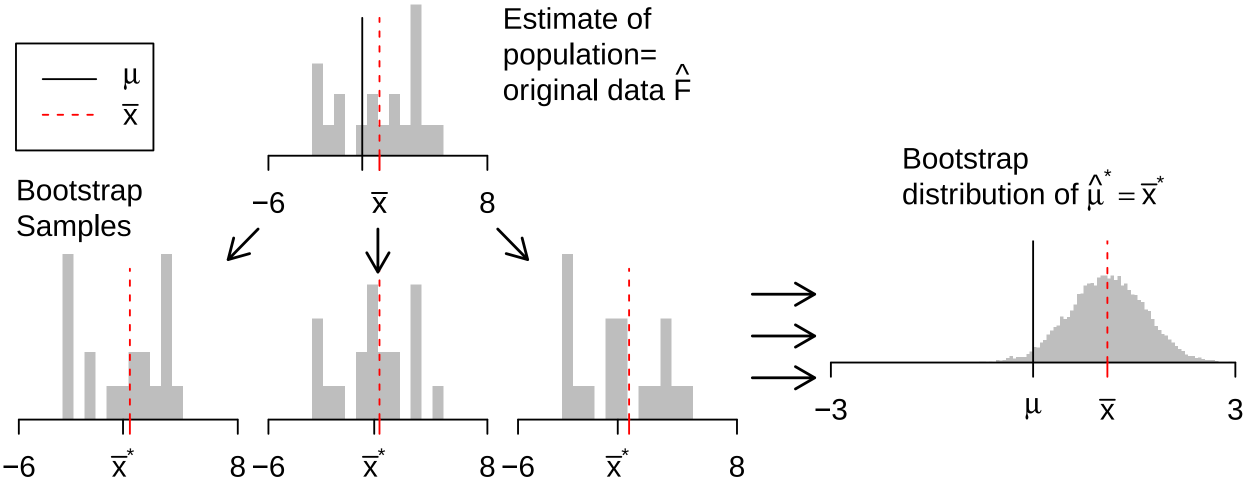
- If we don’t know the population distribution exactly, we can’t simulate samples from the population

Idea:

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

Simulation-based approximation to sampling distribution, if population distribution is *not* known:

1) Simulate many samples by resampling from the observed data
2) For each simulated sample, calculate the estimate $\hat{\theta}_{MLE}$
3) The distribution of estimates from different simulated samples approximates the sampling distribution of the estimator $\hat{\theta}_{MLE}$.


μ
x̄
Estimate of population= original data $\hat{F}$
−6
x̄
8
Bootstrap Samples
−6
$\bar{x}^*$
8
−6
$\bar{x}^*$
8
−6
$\bar{x}^*$
8
Bootstrap distribution of $\hat{\mu}^* = \bar{x}^*$
−3
μ
x̄
3

Example with Poisson data (one last time!)

Recall our Poisson data, one last time:

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: $\bar{x} = 24.9$

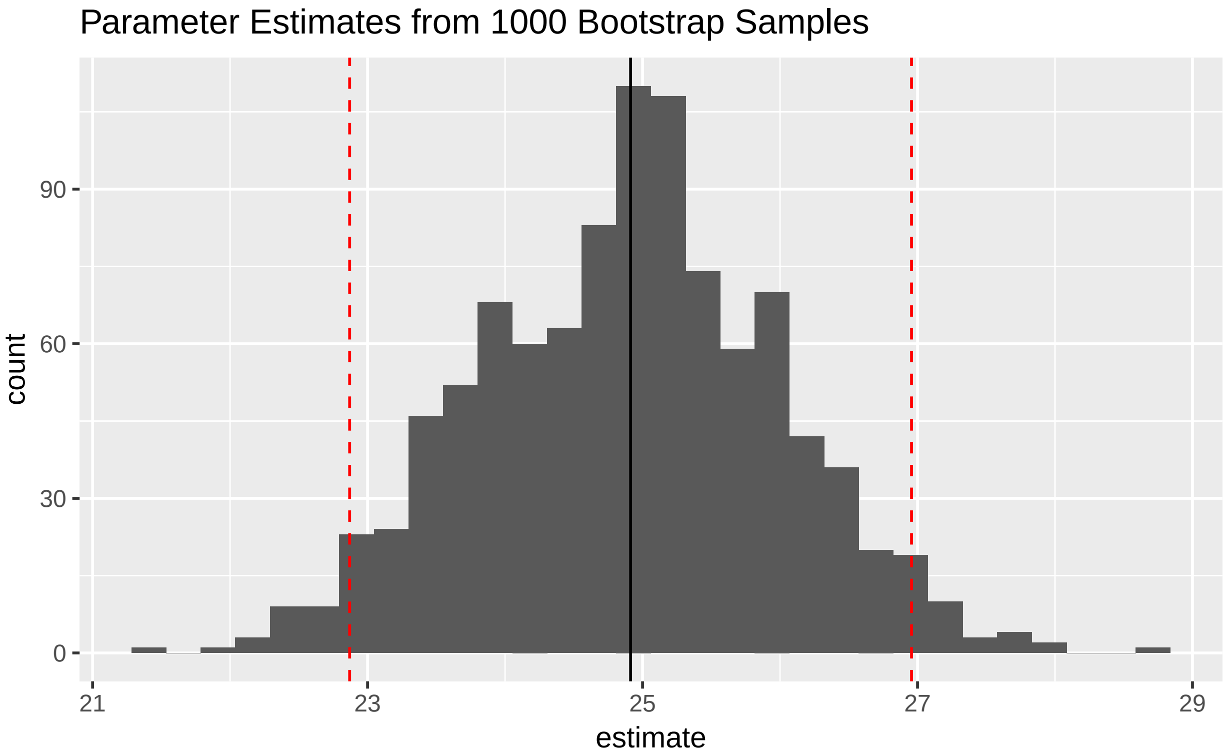
Model: $X_i \sim \text{Poisson}(\lambda)$

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

A bootstrap interval estimate:

```r
# 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)
}

# find 2.5th percentile and 97.5th percentile; endpoints of a 95% Bootstrap Percentile Interval
quantile(bootstrap_estimates$estimate, prob = c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 22.86957 26.95652
```


Parameter Estimates from 1000 Bootstrap Samples
count
90
60
30
0
21
23
25
27
29
estimate