13: INTRO TO THE BOOTSTRAP

Stat250 S25

Prof Amanda Luby

1 Roadmap

Observe $X_1, \ldots, X_n \sim F(\theta)$ with θ unknown. Estimate $\hat{\theta} = g(X_i)$.

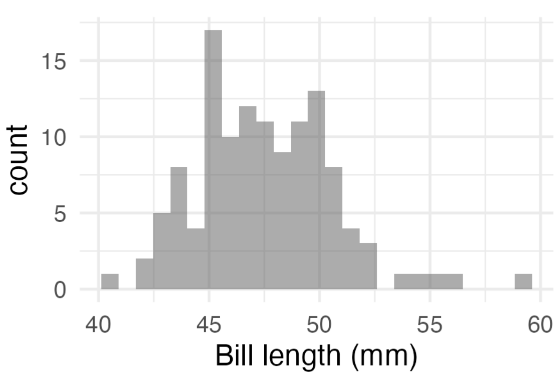
Do we expect $\hat{\theta}$ to be **exactly equal** to θ?

What are **plausible values** for θ given an observed $\hat{\theta}$?

We want to develop an interval estimate of a population parameter:

1. *Exact method:* Find the sampling distribution in closed form (Ch 4). REquires knowledge of the distribution of the data
2. **Bootstrap Method**: Use the sample to approximate the population and simulate a sampling distribution (Ch5)
3. *Asymptotic method:* Use large-sample theory to approximate the sampling distribution (e.g., appeal to the CLT; Ch7)

2 Example: gentoo penguin bill length



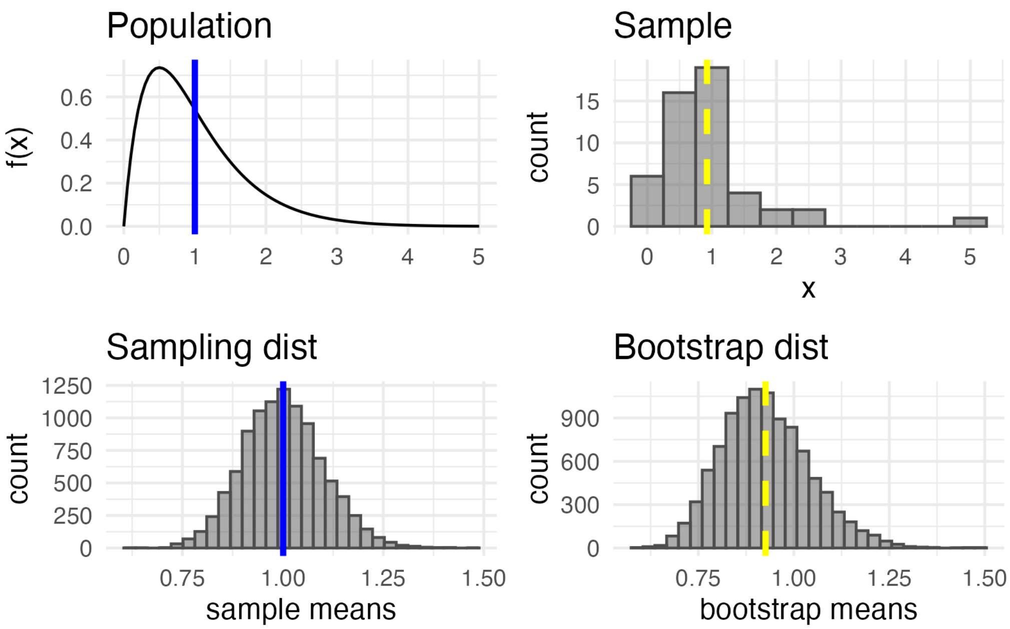
min	40.90
Q1	45.30
median	47.30
Q3	49.55
max	59.60
mean	47.50
sd	3.08
n	123.00
missing	1.00

2.1 The one-sample bootstrap algorithm

Given a sample of size n from a population,

1. Draw a resample of size n, **with replacement**, from the sample.
2. Compute the statistic of interest.
3. Repeat this resampling process (steps 1-2) many times, say 10,000.
4. Construct the bootstrap distribution of the statistic.

3 How does the bootstrap work?



	Mean	SD	Bias
Population	1	0.5	
Sample			
Sampling distribution			
Bootstrap distribution			

4 Why does the bootstrap work?

First, recall the definition of the CDF:

$$F_x(x_0) = P(X \leq x_0)$$

In other words, F_x is the probability of the event $\{X \leq x_0\}$. If we observe a sample of $X_1, ..., X_n \sim F_x$, a natural estimator for this probability is the observed proportion of observations where $\{X_i \leq x_0\}$.

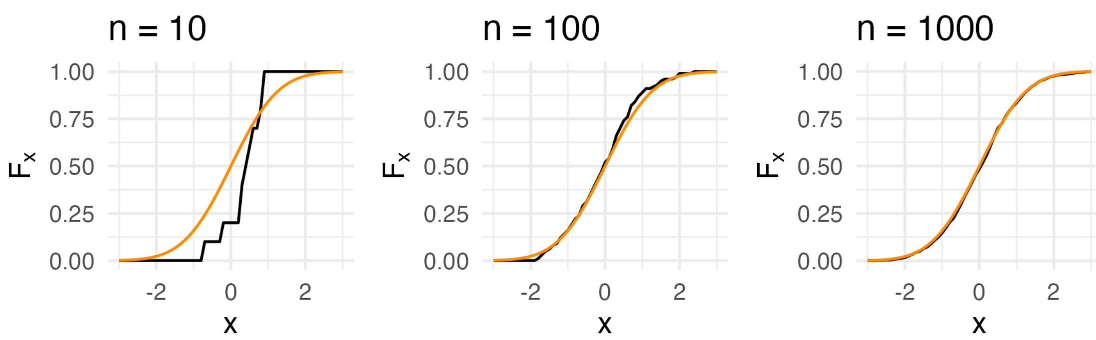
$$\widehat{F}_n = \frac{\sum \mathbb{I}(X_i \leq x_0)}{n}$$

so __________ is an estimator for _________

Each bootstrap sample is drawn from $\widehat{F}$:

$X_1^{*(1)}, X_2^{*(1)}, ... X_n^{*(1)} \sim \widehat{F}_n$

As n increases, $\widehat{F}_n$ gets closer to true F



It turns out that $\widehat{F}_n$ is an ______________________ and ______________________ estimator for F!

- When n is large, $\widehat{F}_n$ is very close to F
- So any statistic that is based on $\widehat{F}_n$ is very similar to the same statistic based on F
- Re-sampling from our original sample results in a sampling distribution that is very similar to the theoretical sampling distribution
- This is true even if we don't know what the theoretical sampling distribution is!

5 R Implementation

```r
y <- gentoo$bill_length_mm # original sample
n <- nrow(gentoo)          # sample size
N <- 10^4                  # desired no. resamples
boot_means <- numeric(N)   # a place to store the bootstrap stats

# Resampling from the sample
for (i in 1:N) {
  x <- sample(y, size = n, replace = TRUE)
  boot_means[i] <- mean(x, na.rm = TRUE)  # you can choose other statistics
}
# Calculate a 95% percentile interval
quantile(boot_means, probs = c(0.025, 0.975))
```

```
    2.5%    97.5%
46.97258 48.06372
```