Recap: Confidence Sets

A $1 - \alpha$ confidence interval for θ is an interval C_n computed from the data such that $P_\theta(\theta \in C_n) \geq 1 - \alpha$ for all θ.

$1 - \alpha$ is called the coverage of the interval.

Note that the probability statement is about C_n, not θ, which is fixed. To emphasize this, we could write $P(C_n \ni \theta) \geq 1 - \alpha$ for all θ.

Suppose $\hat{\theta}_n \approx N(\theta, \hat{\sigma}_n^2)$. Then we can form an approximate $1-\alpha$ confidence interval for θ of

$$C_n = \hat{\theta}_n \pm z_{\alpha/2}\hat{\sigma}_n,$$

where $z_{\alpha/2}$ is chosen such that $P(Z > z_{\alpha/2}) = \alpha/2$ for $Z \sim N(0, 1)$.

Confidence Interval of the Empirical CDF

Dvoretzky-Kiefer-Wolfowitz Inequality: Let $X_1, \ldots, X_n \overset{iid}{\sim} F$. For any $\epsilon > 0$,

$$P\left(\sup_x |F(x) - \hat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

It follows that the functions

$$\begin{aligned} L(x) &= \max\{\hat{F}_n(x) - \epsilon_n, 0\} \\ U(x) &= \min\{\hat{F}_n(x) + \epsilon_n, 1\} \\ &\text{for } \epsilon_n = \sqrt{\log(2/\alpha)/(2n)} \end{aligned}$$

form a global $1 - \alpha$ confidence band for F. That is,

$$P\left(L(x) \leq F(x) \leq U(x) \text{ for all } x\right) \geq 1 - \alpha$$

Often we have $T(\hat{F}_n) \approx N(T(F), \widehat{se}^2)$, which allows us to form an approximate $1 - \alpha$ confidence interval for $T(F)$ of

$$T(\hat{F}_n) \pm z_{\alpha/2}\widehat{se}$$

Example: Verify that the `R` expression

```
mean(x) + c(-2, 2) * sd(x)/sqrt(length(x))
```

produces an approximate 95% confidence interval for the mean waiting time for Old Faithful Geyser Data (built-in data in R).

The Bootstrap

The bootstrap is a computer-intensive method for estimating measures of uncertainty in problems for which no analytical solution is available. There are technically two classes of bootstrap methods: parametric and nonparametric.

The nonparametric bootstrap uses two main ideas:

- Monte Carlo (MC) integration
  - MC is named after the Monte Carlo Casino in Monaco (1940s).
  - MC refers to computational methods that rely on random sampling to approximate numerical results.
- The empirical CDF

Monte Carlo integration is based on the following approximation:

$$
\begin{aligned}
E[h(Y)] &= \int h(y) dF_Y(y) \\
&\approx \frac{1}{B} \sum_{j=1}^{B} h(Y_j)
\end{aligned}
$$

where $Y_1, \ldots, Y_B \overset{iid}{\sim} F_Y$. Note that if $E[|h(Y)|] < \infty$,

$$
\frac{1}{B} \sum_{j=1}^{B} h(Y_j) \overset{as}{\to} E[h(Y)]
$$

as $B \to \infty$. Typically we have control over B, so we can make the approximation arbitrarily good.

A simple example: Use Monte Carlo integration to approximate

$$\int_{-\infty}^{\infty} \sin^2(x) e^{-x^2} dx$$

Solution: We can write this as $\sqrt{\pi} \int_{-\infty}^{\infty} \sin^2(x) f(x) dx$, where $f(x)$ is the PDF of a $N(0, 1/2)$ r.v. Therefore, we can

1. Draw $Y_1, \ldots, Y_B \overset{iid}{\sim} N(0, 1/2)$.

```
> B <- 10000; y <- 1/sqrt(2) * rnorm(B)
```

2. Approximate $\sqrt{\pi} \int_{-\infty}^{\infty} \sin^2(x) f(x) dx \approx \frac{\sqrt{\pi}}{B} \sum_{j=1}^{B} \sin^2(Y_j)$.

```
> sqrt(pi) * mean(sin(y)^2)
[1] 0.5509956
```

Importance sampling is an adaptation to the usual Monte Carlo integration that allows us to sample from an "importance function" g rather than the target density h. Note that

$$\begin{aligned} E_h[q(\theta)] &= \int q(\theta)h(\theta)d\theta \\ &= \int q(\theta)\frac{h(\theta)}{g(\theta)}g(\theta)d\theta \\ &\approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\frac{h(\theta_i)}{g(\theta_i)} \end{aligned}$$

where $\theta_1, \ldots, \theta_B \stackrel{iid}{\sim} g(\theta)$.

A more complicated example: Use Monte Carlo integration to approximate $V_\lambda[median(X_1, \ldots, X_n)]$ when $X_1, \ldots, X_n \stackrel{iid}{\sim} Exp(\lambda)$.

This is more complicated in two ways:

1. Unlike an analytical calculation, on the computer we need particular values of n and λ. To see how $V_\lambda[median(X_1, \ldots, X_n)]$ changes with n and λ, we need to use Monte Carlo integration many times for different combinations.

2. For each combination, we need to sample B times from the *sampling distribution* of $median(X_1, \ldots, X_n)$. That is, for each $j = 1, \ldots, B$, we need to sample $X_1, \ldots, X_n \stackrel{iid}{\sim} Exp(\lambda)$ and calculate the median. Don't confuse n and B: n is the sample size, while B is the number of MC samples.

One combination: Let $n = 10$ and $\lambda = 5$. Then

- Draw $Y_1, \ldots, Y_B \overset{iid}{\sim} F_Y$, where F_Y is the CDF of $median(X_1, \ldots, X_n)$.

```
> n <- 10; lambda <- 5; B <- 10000
> samples <- matrix(rexp(n*B, rate = 1/lambda),
+     nrow = B, ncol = n)
> y <- apply(samples, MARGIN = 1, FUN = median)
```

- Approximate $V_\lambda[median(X_1, \ldots, X_n)] \approx \frac{1}{B}\sum_{j=1}^{B}(Y_j - \bar{Y})^2$.

```
> var(y)
[1] 2.402400
```

Back to the bootstrap...

Suppose we have data $X_1, \ldots, X_n$ and we compute statistic $T_n = g(X_1, \ldots, X_n)$.

It's not always possible to calculate $V_F[T_n]$ analytically, which is where the bootstrap comes in.

If we knew F, we could use MC integration to approximate $V_F[T_n]$. However, we don't in practice, so we make an initial approximation of F with the empirical CDF $\hat{F}_n$.

$$\underbrace{V_F[T_n] \quad \approx \quad V_{\hat{F}_n}(T_n)}_{\text{ECDF; depends on } n} \qquad \underbrace{\approx \quad \widehat{V}_{\hat{F}_n}(T_n)}_{\text{MC integration; depends on } B}$$

Sampling from $\hat{F}_n$ is easy: just draw one observation at random from $X_1, \ldots, X_n$. Repeated sampling is "with replacement."

The algorithm:

1. Repeat the following B times to obtain $T^*_{n,1}, \ldots, T^*_{n,B}$, an iid sample from the sampling distribution for T_n implied by $\hat{F}_n$.
   (a) Draw $X^*_1, \ldots, X^*_n \sim \hat{F}_n$.
   (b) Compute $T^*_n = g(X^*_1, \ldots, X^*_n)$.

2. Use this sample to approximate $V_{\hat{F}_n}(T_n)$ by MC integration. That is, let

$$v_{boot} = \widehat{V}_{\hat{F}_n}(T_n) = \frac{1}{B}\sum_{j=1}^{B}\left(T^*_{n,j} - \frac{1}{B}\sum_{k=1}^{B} T^*_{n,k}\right)^2$$

Confidence intervals can also be constructed from the bootstrap samples.

Method 1: Normal-based interval

$$C_n = T_n \pm z_{\alpha/2}\widehat{se}_{boot}$$

where $\widehat{se}_{boot} = \sqrt{v_{boot}}$; this only works well if the distribution of T_n is close to Normal. Note that asymptotic normality of T_n is a property involving n, not B.

Method 2: Quantile intervals

$$C_n = \left(T^*_{\alpha/2}, T^*_{1-\alpha/2}\right)$$

where T^*_β is the β quantile of the bootstrap sample $T^*_{n,1}, \ldots, T^*_{n,B}$.

Bootstrapping method for estimating bias

$X_1, ..., X_n \sim F_0$. Let F_1 be the corresponding empirical CDF (i.e., $\hat{F}_n$). Then $\theta(F_1)$ is an empirical Plug–In estimator of $\theta(F_0)$. How to estimate the following bias?

$$t_0 = E_{F_0}(\theta(F_0) - \theta(F_1))$$

Answer: We draw a sample $Y_1, ..., Y_m$ from F_1 and derive the empirical CDF F_2. We can estimate t_0 by

$$\hat{t}_0 = E_{F_1}(\theta(F_1) - \theta(F_2))$$

Example (Bias correction). We want to estimate $\theta(F_0) = (E_{F_0}X)^2 = \mu^2$, where X follows F_0 with mean μ and variance σ^2. The EPI estimator is $\theta(F_1) = (E_{F_1}Y)^2 = \bar{X}^2$, where Y follows F_1. The bias is

$$t_0 = E_{F_0}(\theta(F_0) - \theta(F_1)) = \theta(F_0) - E_{F_0}[\theta(F_1)] = -\sigma^2/n.$$

Now we consider the estimator

$$\tilde{\theta} = \theta(F_1) + \hat{t}_0 = \theta(F_1) + [\theta(F_1) - E_{F_1}[\theta(F_2)]]$$

Note that $Y_1, ..., Y_m \sim F_1$ with mean $\bar{X}$ and variance $\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}$, and $\theta(F_2) = (E_{F_2}[Z])^2 = (\bar{Y})^2$, where $\bar{Y} = \sum_{i=1}^m \frac{Y_i}{m}$ and Z follows F_2.

$$E_{F_1}[\theta(F_2)] = E_{F_1}[(\bar{Y})^2] = (E_{F_1}[(\bar{Y})])^2 + Var_{F_1}(\bar{Y}) = (\bar{X})^2 + \frac{1}{m}\left(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}\right)$$

Then for the corrected estimator

$$\tilde{\theta} = \theta(F_1) + \hat{t}_0 = \theta(F_1) + [\theta(F_1) - E_{F_1}[\theta(F_2)]] = (\bar{X})^2 - \frac{1}{m}\left(\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}\right),$$

We have

$$\begin{aligned} E_{F_0}(\tilde{\theta}) &= \left(\mu^2 + \frac{\sigma^2}{n}\right) - E_{F_0}\left[\sum_{i=1} \frac{(X_i - \bar{X}^2)}{mn}\right] \\ &= \mu^2 + \frac{\sigma^2}{n} - \frac{(n-1)\sigma^2}{mn} = \mu^2 + \frac{(m-(n-1))\sigma^2}{mn} \end{aligned}$$

Note that $E_{F_0}[\theta(F_1)] = E_{F_0}[\bar{X}^2] = \mu^2 + \sigma^2/n$. Also note that when $m = n-1$, $\tilde{\theta}$ is an unbiased estimator of $\theta(F_0) = (E_{F_0}X)^2 = \mu^2$.

Parametric Inference

A parametric model has the form $\mathcal{F} = \{F(x; \theta) : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$ is the parameter space. We typically choose a class $\mathcal{F}$ based on knowledge about the particular problem.

Sufficient statistics and **likelihood functions** are two key principles of data reduction under a parametric model.

- Sufficient statistics compress the data while retaining all information about the parameters.
- Likelihood functions summarize the data into a parameter-based function that drives inference.

Both replace large datasets with compact objects that preserve the essential information for inference.