Mathematical Statistics Concepts

Data: $X_1, .., X_n$ – result of some random experiment (lab experiment, survey sampling, ...)

Each X_i is random variable

Any transformation of the data $S = f(X_1, ..., X_n)$ is called a $\boxed{\text{statistic}}$.

Examples: $S_1 = \min\{X_1, ..., X_n\}$, $S_2 = \frac{1}{n}\sum_{i=1}^{n} X_i$ – are also random variables

Often we can assume that $X_1, .., X_n$ are iid and have cdf $F(x) = P(X_1 \le x)$. Given iid data $X_1, ..., X_n$ we want to learn something about F.

Example Lady tasting tea

Your friend claims that they can tell the difference between two similar drinks. You prepare n unlabeled cups, randomize the drinks and record $X_i \in \{0, 1\}$ – indicator of a successfull drink identification at the ith trial.

$$X_1, ..., X_n \overset{iid}{\sim} \text{Bernoulli}(p),$$

where $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$. Let

statistics

$\boxed{K = \sum_{i=1}^{n} X_i}$ – # of successes

$\boxed{\hat{p} = \frac{K}{n}}$ is a reasonable estimator of p

$$E(\hat{p}) = E\left(\frac{K}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{E(X_i)}_{p} =$$

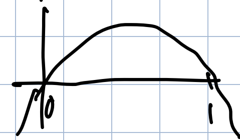
$$\frac{1}{n} \cdot n \cdot p = p$$

$Var(X) = E[(X - E(X))^2]$

$Var(aX) = E[(aX - E(aX))^2] = E(a^2 (X - E(X))^2)$

$$Var(\hat{p}) = Var\left(\frac{K}{n}\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = [\text{independence}] =$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \underbrace{Var(X_i)}_{p(1-p)} = \frac{1}{n^2} \cdot n\, p(1-p) = \frac{p(1-p)}{n}$$



$Var(\hat{p}) \xrightarrow[n \to \infty]{} 0$ — ✓

Def Suppose $\hat{\theta}$ is an estimator of true quantity θ. $Bias(\hat{\theta}) \stackrel{def}{=} E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$

If $E(\hat{\theta}) = \theta$, then $\hat{\theta}$ is called an unbiased estimator.

Monte Carlo Methods in Statistical Inference

When we observe data we summarize them into a set of statistics (e.g., $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$). These statistics are random variables and we want to know their distributions. We ask ourselves "what if we repeated our experiment and obtained a new set of observations $x_1^*, \dots, x_n^*$. If we assume that we know cdf of $X_1, \dots, X_n$, we can "repeat" our experiment via Monte Carlo simulations.

Example MSE estimation

MSE = mean squared error

$$MSE(\hat\theta) = E\left[(\hat\theta - \theta)^2\right]$$

Pretending we know F - cdf of our data

we simulate m fake data sets

$$x_{11}^*, \dots, x_{1n}^* \to \hat\theta^{(1)}$$
$$\vdots$$
$$x_{m1}^*, \dots, x_{mn}^* \to \hat\theta^{(m)}$$

Monte Carlo estimate of MSE:

$$\widehat{MSE} = \frac{1}{n} \sum_{j=1}^{m} (\hat\theta^{(j)} - \theta)^2$$

Example Poisson rate estimation

$$X \sim \text{Poisson}(\underset{\substack{\uparrow \\ \text{rate}}}{\lambda}) \quad \text{if} \quad \Pr(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}, \; k=0,1,2,\dots$$

$$E(X) = \lambda, \quad Var(X) = \lambda.$$

Two estimators: $x_1, \dots, x_n \sim \text{Poisson}(\lambda)$
↑ objective

$$\hat\lambda_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat\lambda_2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Let's compare their MSEs

Monte Carlo Estimator of Confidence Interval Coverage.

Recall that $(1-\alpha)\%$ confidence interval (CI) promises to contain the true value of the parameter with probability $(1-\alpha)\%$.

We can check whethe a CI lives up to its definition by Monte Carlo.

We generate m fake data sets and use them to form m CIs:

$$\left(\hat\theta_\ell^{(1)}, \hat\theta_u^{(1)}\right)$$
$$\vdots$$
$$\left(\hat\theta_\ell^{(m)}, \hat\theta_u^{(m)}\right)$$

Coverage: $\widehat{\text{Coverage}} = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\hat\theta_\ell^{(j)} < \theta < \hat\theta_u^{(j)}\}$

$\widehat{\text{Coverage}}$ has a Monte Carlo error, so 0.95 ± 0.03 does not discredit our CI

Example: Poisson rate estimation

$x_1, \dots, x_n \sim \text{Poisson}(\lambda)$

$$\hat\lambda_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\hat\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ 95% CI: $\hat\lambda_1 \pm 1.96\frac{\hat\sigma}{\sqrt{n}}$

How do we estimate Monte Carlo error of $\widehat{\text{Coverage}}$. Suppose q = Coverage and $\hat q = \widehat{\text{Coverage}}$

$\hat q = \frac{1}{m}\sum_{j=1}^{n} \mathbb{1}_{\{\ \}}$ Monte Carlo error

$$\sqrt{\frac{\hat q(1-\hat q)}{m}}$$

$$\hat{q} \pm 1.96 \sqrt{\frac{\hat{q}(1-\hat{q})}{m}}$$

Monte Carlo for Hypothesis Testing

	H_0 true	H_0 false
reject H_0	Type I error α	Power $1-\beta$
fail to reject H_0	$1-\alpha$	Type II error β

We can estimate Type I error prob α or Power $1-\beta$ via Monte Carlo:

Type I error: simulate from H_0

Power: simulate from an alternative H_a

generate m data sets from H_0:

$$x_{11}^*, \ldots, x_{1n}^* \to T_1 \to I_1$$

test statistic

decision (reject H_0 or not)

$$x_{m1}^*, \ldots, x_{mn}^* \to T_m \to I_m$$

$$\hat{\alpha} = \frac{1}{m} \sum_{j=1}^{m} I_j, \quad I_j = \begin{cases} 1 & \text{if rejected } H_0 \\ 0 & \text{if not} \end{cases}$$

if we change to H_a: $1-\hat{\beta} = \frac{1}{m} \sum_{j=1}^{m} I_j$

Bootstrap

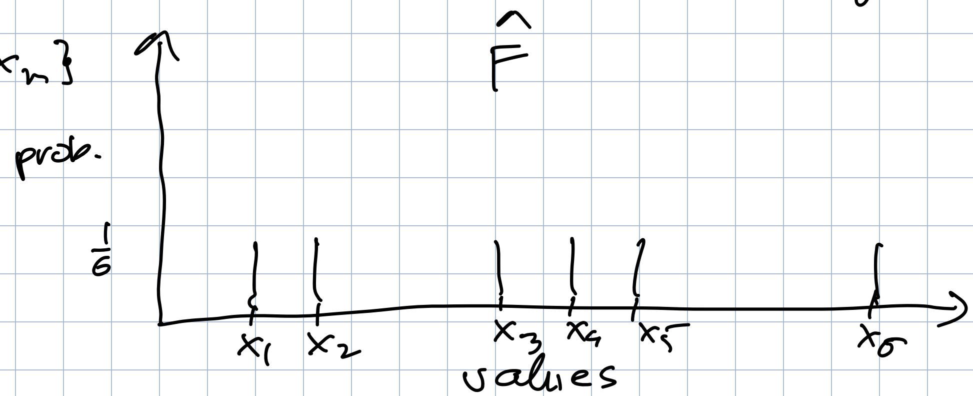
Data: $x_1, \dots, x_n \overset{iid}{\sim} F \leftarrow$ unknown

estimator: $\hat{\theta} = s(x_1, \dots, x_n)$

The main idea of bootstrap is to study the distribution of $\hat{\theta}^* = s(x_1^*, \dots, x_n^*)$, where

$x_1^*, \dots, x_n^* \sim \hat{F}$, where

$\hat{F}$ is the empirical cdf defined as a cdf of a uniform random variable X taking values on $\{x_1, \dots, x_n\}$



$\hat{F} \to F$ as $n \to \infty$

If $X \sim \hat{F}$ → simulating from $\hat{F}$ is easy - sampling with replacement from $\{x_1, \dots, x_n\}$

Bootstrap estimation of estimator's bias

Given $x_1, \dots, x_n$ and an estimator $\hat{\theta} = s(x_1, \dots, x_n)$ we would like to compute bias $(\hat{\theta})$

We construct B bootstrap samples by sampling $x_1, \dots, x_n$ with replacement n times

$$x_{11}^*, \dots, x_{1n}^* \leftarrow \text{sample } 1$$

$$\vdots$$

$$x_{B1}^*, \dots, x_{Bn}^* - \text{sample } B$$

Notice that original x_i can appear more than once in each row above (i.e., we can have

$$x_{11}^* = x_3 \ , \quad x_{15}^* = x_3 \ )$$

Evaluate $\hat{\theta}^{(b)} = s(x_{b1}^*, \dots, x_{bn}^*)$ for $b = 1, \dots, B$

$$\overline{\hat{\theta}} \approx E(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}$$

$$\hat{\theta}_{obs} = s(x_1, \dots, x_n)$$

$$\widehat{\text{bias}}(\hat{\theta}) = \overline{\hat{\theta}} - \hat{\theta}_{obs}$$

<u>Example</u>: see lab quarto