Statistics for Astronomers: Lecture 11, 2020.11.17

Prof. Sundar Srinivasan

IRyA/UNAM

Review

Confidence intervals

Asymmetric CIs: central, symmetric, shortest.

The empirical distribution function (CDF from sample)

Given a dataset of N points $X_i (i = 1, 2, \cdots, N) \sim F_X(x)$ (unknown CDF),

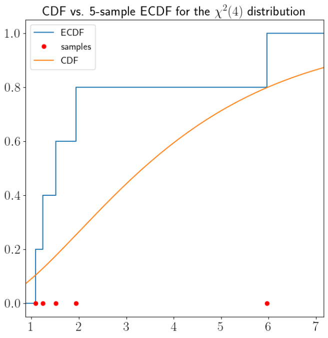
Empirical distribution: $\widehat{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{X_i \leq x}(x)$, with $\mathbb{I}_{X_i \leq x}(x) = \begin{cases} 1 & X_i \leq x \\ 0 & \text{otherwise} \end{cases}$

The empirical distribution function (CDF from sample)

Given a dataset of N points $X_i (i = 1, 2, \cdots, N) \sim F_X(x)$ (unknown CDF),

Empirical distribution: $\widehat{F}_N(x) = \dfrac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{X_i \le x}(x)$, with $\mathbb{I}_{X_i \le x}(x) = \begin{cases} 1 & X_i \le x \\ 0 & \text{otherwise} \end{cases}$

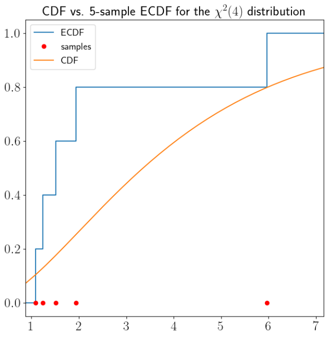
Probability mass increases by $1/N$ at each sample point.



The empirical distribution function (CDF from sample)

Given a dataset of N points $X_i (i = 1, 2, \cdots, N) \sim F_X(x)$ (unknown CDF),

Empirical distribution: $\widehat{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{X_i \leq x}(x)$, with $\mathbb{I}_{X_i \leq x}(x) = \begin{cases} 1 & X_i \leq x \\ 0 & \text{otherwise} \end{cases}$



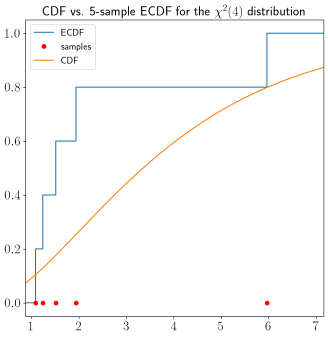
Probability mass increases by $1/N$ at each sample point.

For fixed x and for a given i, $\mathbb{I}_{X_i \leq x}(x)$ is a Bernoulli variable.

The empirical distribution function (CDF from sample)

Given a dataset of N points $X_i (i = 1, 2, \cdots, N) \sim F_X(x)$ (unknown CDF),

Empirical distribution: $\widehat{F}_N(x) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbb{I}_{X_i \le x}(x)$, with $\mathbb{I}_{X_i \le x}(x) = \begin{cases} 1 & X_i \le x \\ 0 & \text{otherwise} \end{cases}$



Probability mass increases by $1/N$ at each sample point.

For fixed x and for a given i, $\mathbb{I}_{X_i \le x}(x)$ is a Bernoulli variable.

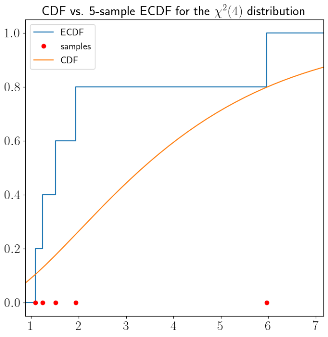
$$P(\text{"success"}) = P(X_i \le x) = \mathbb{E}\left[\mathbb{I}_{X_i \le x}(x)\right] = F_X(x).$$

Variance: $F_X(x)\Big(1 - F_X(x)\Big)$.

The empirical distribution function (CDF from sample)

Given a dataset of N points $X_i (i = 1, 2, \cdots, N) \sim F_X(x)$ (unknown CDF),

Empirical distribution: $\widehat{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{X_i \leq x}(x)$, with $\mathbb{I}_{X_i \leq x}(x) = \begin{cases} 1 & X_i \leq x \\ 0 & \text{otherwise} \end{cases}$



Probability mass increases by $1/N$ at each sample point.

For fixed x and for a given i, $\mathbb{I}_{X_i \leq x}(x)$ is a Bernoulli variable.

$$P(\text{"success"}) = P(X_i \leq x) = \mathbb{E}\left[\mathbb{I}_{X_i \leq x}(x)\right] = F_X(x).$$

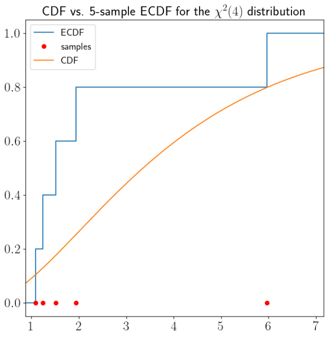
Variance: $F_X(x)\Big(1 - F_X(x)\Big)$.

$\widehat{F}_N(x) =$ mean of N Bernoulli variables $\Longrightarrow$ a binomial variable.

The empirical distribution function (CDF from sample)

Given a dataset of N points $X_i (i = 1, 2, \cdots, N) \sim F_X(x)$ (unknown CDF),

Empirical distribution: $\widehat{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{X_i \le x}(x)$, with $\mathbb{I}_{X_i \le x}(x) = \begin{cases} 1 & X_i \le x \\ 0 & \text{otherwise} \end{cases}$



Probability mass increases by $1/N$ at each sample point.

For fixed x and for a given i, $\mathbb{I}_{X_i \le x}(x)$ is a Bernoulli variable.

$$P(\text{“success”}) = P(X_i \le x) = \mathbb{E}\left[\mathbb{I}_{X_i \le x}(x)\right] = F_X(x).$$

Variance: $F_X(x)\Big(1 - F_X(x)\Big)$.

$\widehat{F}_N(x) =$ mean of N Bernoulli variables $\Longrightarrow$ a binomial variable.

$\mathbb{E}\left[\widehat{F}_N(x)\right] = F_X(x)$ (ECDF = unbiased estimator of CDF).

$$Var\left[\widehat{F}_N(x)\right] = \frac{\widehat{F}_N(x)\Big(1 - \widehat{F}_N(x)\Big)}{N} \to 0 \text{ as } N \to \infty$$

(The ECDF is a consistent estimator of the CDF).

Empirical distribution function (contd.)

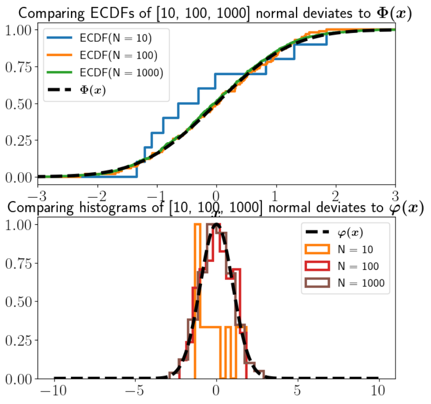
Draw $N = [10, 100, 1000]$ values from the standard normal. Compare $\widehat{F}_N(x)$ to $\Phi(x)$.

In `Python`: `statsmodels.distributions.empirical_distribution.ECDF`.

Empirical distribution function (contd.)

Draw $N = [10, 100, 1000]$ values from the standard normal. Compare $\widehat{F}_N(x)$ to $\Phi(x)$.

In Python: `statsmodels.distributions.empirical_distribution.ECDF`.



Resampling: The Bootstrap

Why resampling?

Sometimes we only have a (small) sample from an unknown distribution.

The sample is the best (only) information you have about the population.

Performing inference on this sample – point estimates, interval estimates, confidence intervals?

Why resampling?

Sometimes we only have a (small) sample from an unknown distribution.

The sample is the best (only) information you have about the population.

Performing inference on this sample – point estimates, interval estimates, confidence intervals?

Resampling – using multiple subsets of the existing data to infer the underlying distribution.

Use the data alone – an example of nonparametric statistics/inference.

Why resampling?

Sometimes we only have a (small) sample from an unknown distribution.

The sample is the best (only) information you have about the population.

Performing inference on this sample – point estimates, interval estimates, confidence intervals?

Resampling – using multiple subsets of the existing data to infer the underlying distribution.

Use the data alone – an example of nonparametric statistics/inference.

Some resampling methods: jackknife, bootstrap, cross-validation.

Bootstrap

Introduction: here and here. ("Wild About Statistics", Chris Wild, University of Auckland.)

Bootstrap

Introduction: here and here. (“Wild About Statistics”, Chris Wild, University of Auckland.)

Sampling (from dataset) with replacement.

No assumptions about the underlying distribution! Preserves characteristics of original data, including selection effects such as truncation/censoring.

Bootstrap

Introduction: here and here. ("Wild About Statistics", Chris Wild, University of Auckland.)

Sampling (from dataset) with replacement.

No assumptions about the underlying distribution! Preserves characteristics of original data, including selection effects such as truncation/censoring.

Can estimate sampling distribution of almost any statistic. Inference, hypothesis testing.

Bootstrap

Introduction: here and here. ("Wild About Statistics", Chris Wild, University of Auckland.)

Sampling (from dataset) with replacement.

No assumptions about the underlying distribution! Preserves characteristics of original data, including selection effects such as truncation/censoring.

Can estimate sampling distribution of almost any statistic. Inference, hypothesis testing.

Maximum # distinct bootstrap samples from an N-point dataset: $B_{\max} = \binom{2N-1}{N}$.

Bootstrap

Introduction: here and here. ("Wild About Statistics", Chris Wild, University of Auckland.)

Sampling (from dataset) with replacement.

No assumptions about the underlying distribution! Preserves characteristics of original data, including selection effects such as truncation/censoring.

Can estimate sampling distribution of almost any statistic. Inference, hypothesis testing.

Maximum # distinct bootstrap samples from an N-point dataset: $B_{\rm max} = \binom{2N-1}{N}$.

Increasing the number of bootstrap samples cannot increase the amount of information in the original data; but reduces effects of random sampling errors which can arise from a bootstrap procedure itself (suggested: $B \gtrsim 50$).

Bootstrap

Introduction: here and here. ("Wild About Statistics", Chris Wild, University of Auckland.)

Sampling (from dataset) with replacement.

No assumptions about the underlying distribution! Preserves characteristics of original data, including selection effects such as truncation/censoring.

Can estimate sampling distribution of almost any statistic. Inference, hypothesis testing.

Maximum # distinct bootstrap samples from an N-point dataset: $B_{\max} = \binom{2N-1}{N}$.

Increasing the number of bootstrap samples cannot increase the amount of information in the original data; but reduces effects of random sampling errors which can arise from a bootstrap procedure itself (suggested: $B \gtrsim 50$).

Useful when one of the following is unknown: underlying distribution, statistical properties (*e.g.*, machine learning classification, principal component analysis), or standard way to calculate (*e.g.*, 95% CI on correlation coefficient between two observables, or on slope/intercept of regression line).

Bootstrap

Introduction: here and here. ("Wild About Statistics", Chris Wild, University of Auckland.)

Sampling (from dataset) with replacement.

No assumptions about the underlying distribution! Preserves characteristics of original data, including selection effects such as truncation/censoring.

Can estimate sampling distribution of almost any statistic. Inference, hypothesis testing.

Maximum # distinct bootstrap samples from an N-point dataset: $B_{\rm max} = \binom{2N-1}{N}$.

Increasing the number of bootstrap samples cannot increase the amount of information in the original data; but reduces effects of random sampling errors which can arise from a bootstrap procedure itself (suggested: $B \gtrsim 50$).

Useful when one of the following is unknown: underlying distribution, statistical properties (*e.g.*, machine learning classification, principal component analysis), or standard way to calculate (*e.g.*, 95% CI on correlation coefficient between two observables, or on slope/intercept of regression line).

Requires: (a) iid sample (b) finite population variance (~~heavy-tailed distributions~~).

Bootstrap: intuition

Let $X_i (i = 1, \cdots, N)$ be iid variables drawn from an unknown distribution.
Suppose we want to compute a statistic $\hat{\theta}$ whose true value is θ.

Bootstrap: intuition

Let $X_i (i = 1, \cdots, N)$ be iid variables drawn from an unknown distribution.
Suppose we want to compute a statistic $\hat{\theta}$ whose true value is θ.

Construct ECDF: $\widehat{F}_N(t) \equiv P(X \leq t) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbb{I}_{X_i \leq t} = \left(\text{Fraction of points } \leq t\right)$.

Bootstrap: intuition

Let $X_i (i = 1, \cdots, N)$ be iid variables drawn from an unknown distribution.
Suppose we want to compute a statistic $\hat{\theta}$ whose true value is θ.

Construct ECDF: $\widehat{F}_N(t) \equiv P(X \leq t) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbb{I}_{X_i \leq t} = \left(\text{Fraction of points } \leq t\right).$

Draw single value from $\widehat{F}_N(t)$:

each data point from original sample has probability $\dfrac{1}{N}$ of being selected.

Bootstrap: intuition

Let $X_i (i = 1, \cdots, N)$ be iid variables drawn from an unknown distribution.
Suppose we want to compute a statistic $\hat{\theta}$ whose true value is θ.

Construct ECDF: $\widehat{F}_N(t) \equiv P(X \leq t) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbb{I}_{X_i \leq t} = \Big(\text{Fraction of points } \leq t\Big)$.

Draw single value from $\widehat{F}_N(t)$:

each data point from original sample has probability $\dfrac{1}{N}$ of being selected.

Generating a new size-N sample from ECDF $\equiv$ resampling from original data with replacement.
(X_i are independent)

Bootstrap: intuition

Let $X_i (i = 1, \cdots, N)$ be iid variables drawn from an unknown distribution.
Suppose we want to compute a statistic $\hat{\theta}$ whose true value is θ.

Construct ECDF: $\widehat{F}_N(t) \equiv P(X \le t) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mathbb{I}_{X_i \le t} = \left(\text{Fraction of points } \le t\right)$.

Draw single value from $\widehat{F}_N(t)$:

each data point from original sample has probability $\dfrac{1}{N}$ of being selected.

Generating a new size-N sample from ECDF $\equiv$ resampling from original data with replacement.
(X_i are independent)

Procedure:

Generate B size-N samples from original dataset. (`np.random.choice`)

Compute $\hat{\theta}$ for each of the B samples.

Compute mean and variance of the B values of $\hat{\theta}$.

Bootstrap: intuition

Let $X_i (i = 1, \cdots, N)$ be iid variables drawn from an unknown distribution.
Suppose we want to compute a statistic $\hat{\theta}$ whose true value is θ.

Construct ECDF: $\widehat{F}_N(t) \equiv P(X \leq t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{X_i \leq t} = \left(\text{Fraction of points } \leq t\right)$.

Draw single value from $\widehat{F}_N(t)$:

each data point from original sample has probability $\frac{1}{N}$ of being selected.

Generating a new size-N sample from ECDF $\equiv$ resampling from original data with replacement.
(X_i are independent)

Procedure:

Generate B size-N samples from original dataset. (`np.random.choice`)

Compute $\hat{\theta}$ for each of the B samples.

Compute mean and variance of the B values of $\hat{\theta}$.

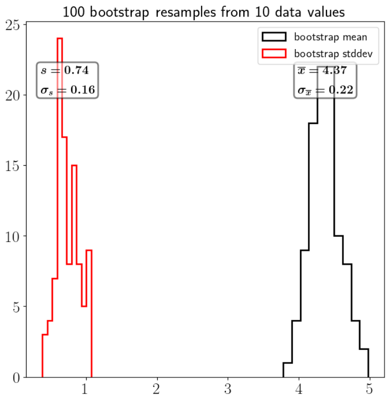
Central Limit Theorem: $\overline{\hat{\theta}} \sim \mathcal{N}\left(\mathbb{E}[\hat{\theta}], \mathrm{Var}[\hat{\theta}]/B\right)$. If $\hat{\theta}$ unbiased, $\overline{\hat{\theta}} \sim \mathcal{N}\left(\theta, \mathrm{Var}[\hat{\theta}]/B\right)$.

Bootstrap example

You are given the following 10-point dataset:

[3.55875989, 6.02903508, 3.63978782, 5.1328453, 3.72245259, 4.21030686, 3.56197579, 4.96159969, 4.95257256, 4.43649666]

Your mission: use $B = 100$ bootstrap resamples on this dataset to plot the bootstrap resampled distribution of the sample mean.



Code for plot available here.

From CLT, $\overline{x} \sim \mathcal{N}(\mu, \sigma^2/N) \implies$

$$\mathrm{mean}(\overline{x}) = \mu \text{ and } \sigma_{\overline{x}} = \sigma/\sqrt{N}.$$

Similarly, using the theoretical mean and variance for the χ distribution, we can estimate s and σ_s:

$$s = \frac{\sigma}{\sqrt{N-1}}\chi(N-1) \implies$$

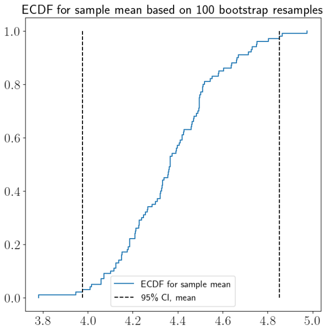
$$\bar{s} \approx 0.99\sigma \text{ and } \sigma_s \approx 0.16\sigma.$$

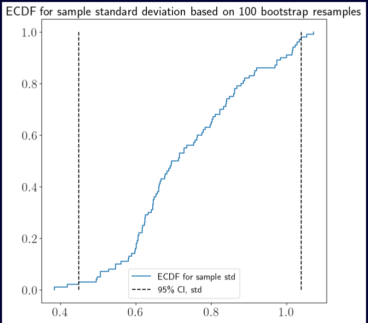
Results from the simulation are consistent with the above theoretical estimates.

Bootstrap example, contd.

How to construct a 95% central CI for the same problem:

Resampling generates B values for the bootstrap mean and bootstrap standard deviation. Generate ECDF for sample mean and sample standard deviation.






```
95% central CI on mean:  [4.0, 4.9]
```

```
95% central CI on std:  [0.4, 1.0]
```