Metode Resampling
(Bootstrap & Jackknife)

Komputasi Statistika MAS246

Resampling Procedures

Metode resampling membebaskan peneliti dari dua batasan statistik konvensional: asumsi bahwa data sesuai dengan kurva berbentuk lonceng (*normally-distributed*) dan kebutuhan untuk fokus pada ukuran statistik yang sifat teoretisnya dapat dianalisis secara matematis.

Diaconis, P., and B. Efron. (1983). Computer-intensive methods in statistics. Scientific American, May, 116-130.

Resampling Procedures

Metode resampling, mengatasi masalah utama dalam statistik: bagaimana menyimpulkan 'kebenaran' dari sampel data yang mungkin tidak lengkap atau diambil dari populasi yang tidak jelas.

Peterson, I. (July 27, 1991). Pick a sample. Science News, 140, 56-58.

Resampling Procedures

Using resampling methods, “you're trying to get something for nothing. You use the same numbers over and over again until you get an answer that you can't get any other way. In order to do that, you have to assume something, and you may live to regret that hidden assumption later on”

Statement by Stephen Feinberg, cited in:

Peterson, I. (July 27, 1991). Pick a sample. Science News, 140, 56-58.

Resampling Methods

- Cross-Validation

Used to estimate test set prediction error rates associated with a given machine learning method to evaluate its performance, or to select the appropriate level of model flexibility.

- Bootstrap

Used most commonly to provide a measure of accuracy of a parameter estimate or of a given machine learning method.

Bootstrap Introduction

- Bootstrap adalah metode umum untuk melakukan analisis statistik tanpa membuat asumsi parametrik yang kuat.
- Tahun 1979, Efron menamai dan menerbitkan makalah pertama tentang bootstrap yang bertepatan dengan munculnya personal computer.
- Efron melakukan resampling dari data asli dengan pengembalian (sampling replacement).
- Bootstrap pada awalnya dirancang untuk memperkirakan bias dan standar error untuk mengestimasi statistik seperti halnya dengan jackknife.
- Bootstrap kemudian diperluas kegunaannya untuk:

  (1) confidence intervals and hypothesis tests,

  (2) linear and nonlinear regression,

  (3) time series analysis and other problems

Bootstrap Introduction

- Efron (1979) menyatakan metode bootstrap merupakan suatu metode berbasis resampling data sampel dengan syarat pengembalian pada datanya dalam menyelesaikan statistik ukuran suatu sampel dengan harapan sampel tersebut mewakili data populasi sebenarnya.
- Jadi, bootstrap adalah metode **resampling** atau pengambilan n sampel **dengan pengembalian** dari n data sampel asli yang dilakukan secara berulangkali untuk mendapatkan distribusi sampling dari suatu penduga parameter.

Metode Resampling

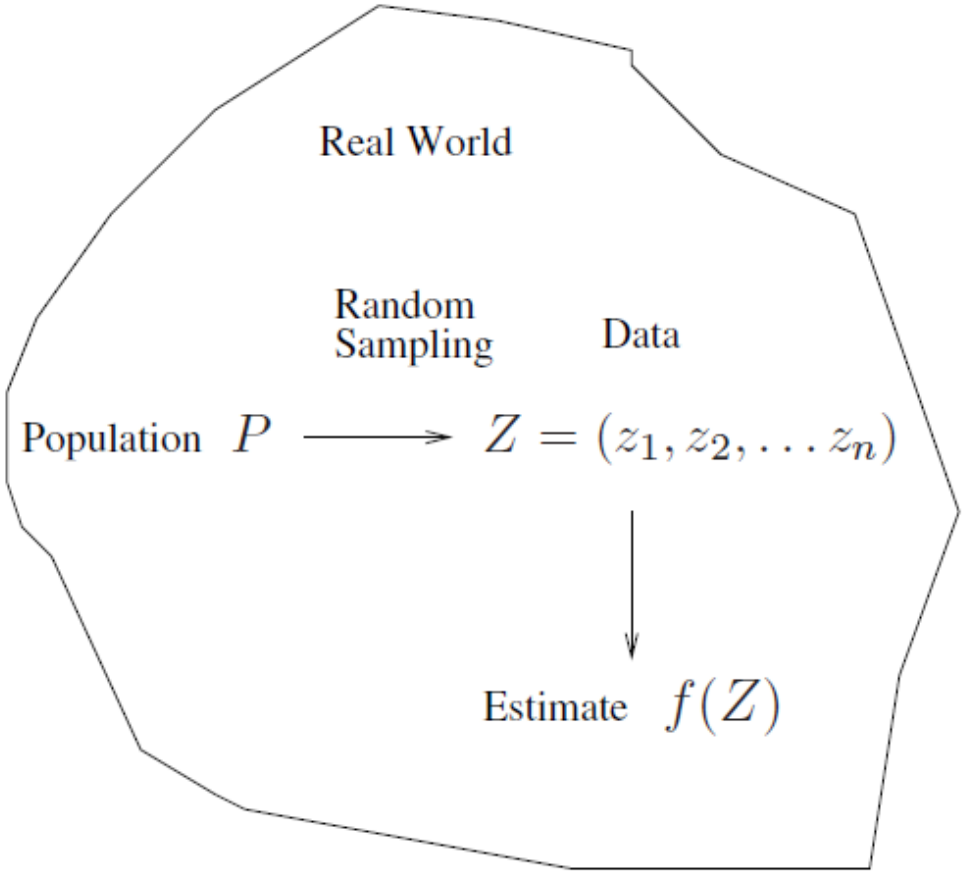
- Permutation
- Bootstrap
- Jackknife
- Cross validation

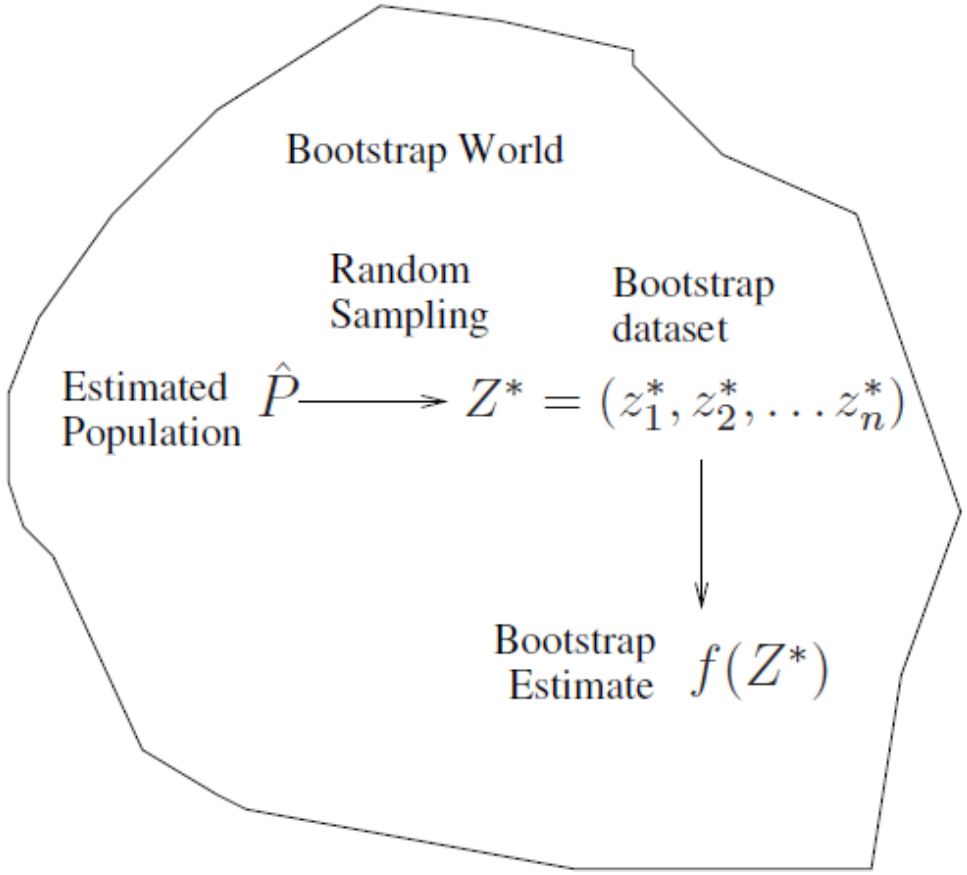
Resampling Method	Application	Sampling procedure used
Bootstrap	Standard deviation, confidence interval, hypothesis testing, bias	Samples drawn at random, with replacement
Jackknife	Standard deviation, confidence interval, bias	Samples consist of full data set with one observation left out
Permutation	Hypothesis testing	Samples drawn at random, without replacement.
Cross-validation	Model validation	Data is randomly divided into two or more subsets, with results validated across sub-samples.

Bootstrap Parametrik VS Bootstrap Nonparameterik

- Bootstrap parametrik adalah metode bootstrap yang memiliki asumsi bahwa suatu penduga parameter berasal dari distribusi tertentu.
- Bootstrap nonparametric adalah metode bootstrap yang tidak memerlukan asumsi distribusi.
- Jika asumsi tentang distribusi populasi benar, maka bootstrap parametrik akan bekerja lebih baik daripada bootstrap nonparametrik. Sebaliknya, maka bootstrap nonparametrik akan bekerja lebih baik.

The Bootstrap: Overview



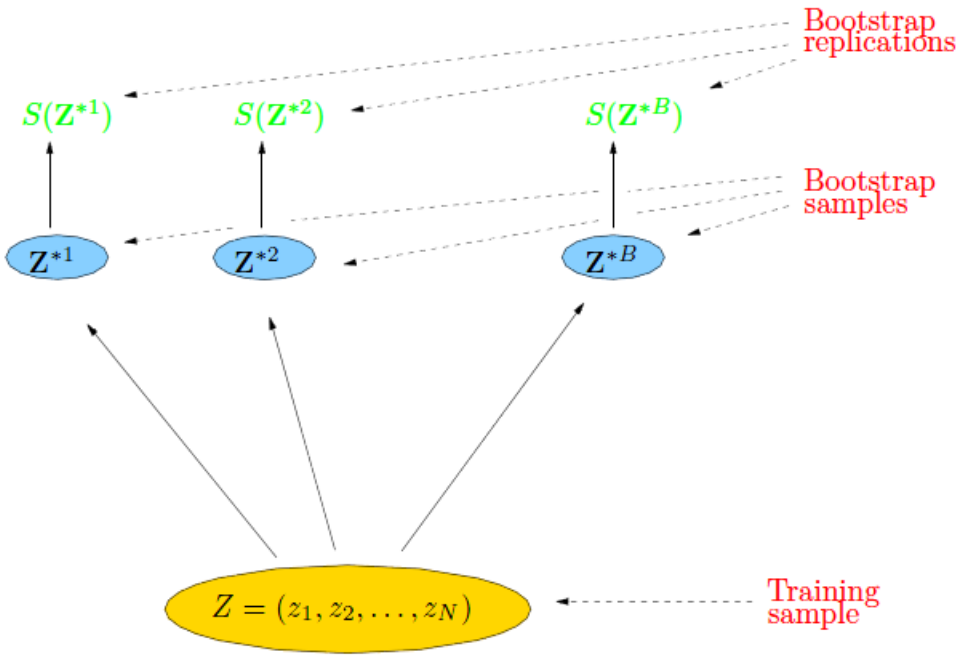


Tahapan Bootstrap

1) Sampel data x didefinisikan sebagai data sampel asli berukuran n dari populasi dengan distribusi diketahui (parametrik) atau tidak (nonparametrik) yang terdiri dari $x = x_1, x_2, \dots, x_n$ dengan x sebagai vektor data pengamatan.
2) Tentukan banyaknya bootstrap resample B dengan $b = 1,2, \dots, B$.
3) Ambil $b = 1$ dan lakukan iterasi (i)-(ii) sampai $b = B$.
   i. Bangkitkan sampel acak x^* sebanyak n dari x dan beri notasi $x_1^*, x_2^*, \dots, x_n^*$
   ii. Lakukan pendugaan parameter titik yaitu $\hat{\theta}$ dan beri notasi sebagai $\hat{\theta}_b$. Misalnya $\hat{\theta}_1 = \frac{x_1^*, x_2^*, \dots, x_n^*}{n}$ untuk $b = 1$.
4) Sehingga diperoleh statistic bootstrap $\hat{\theta}^* = \sum_{i=1}^{B} \frac{\hat{\theta}_i^*}{B}$ dari nilai bootstrap $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$ yang akan membentuk distribusi sampling bagi $\hat{\theta}^*$.

Tahapan Bootstrap (another illustration)

- Suppose we have a model fit to a set of training data. We denote the training set by $\mathbf{Z} = (z_1, z_2, \ldots, z_N)$ where $z_i = (x_i, y_i)$.
- The basic idea is to randomly draw datasets **with replacement** from the training data, each sample the same size as the original training set.
- This is done B times, producing B bootstrap datasets. Then we refit the model to each of the bootstrap datasets, and examine the behavior of the fits over the B replications.
- $S(\mathbf{Z})$ is any quantity computed from the data $\mathbf{Z}$, for example, the prediction at some input point.
- From the bootstrap sampling we can estimate any aspect of the distribution of $S(\mathbf{Z})$, for example, its variance:

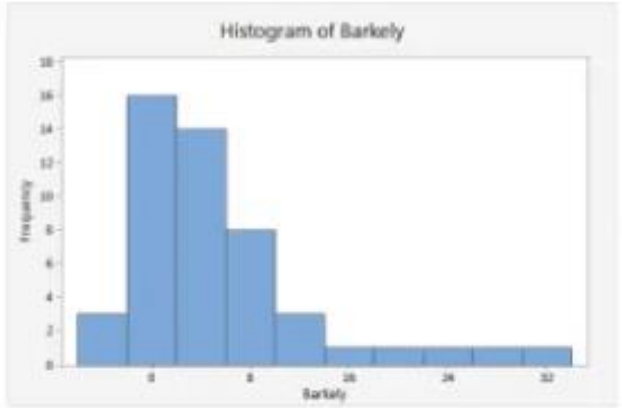


$$\widehat{\mathrm{Var}}[S(\mathbf{Z})] = \frac{1}{B-1}\sum_{b=1}^{B}(S(\mathbf{Z}^{*b}) - \bar{S}^*)^2$$

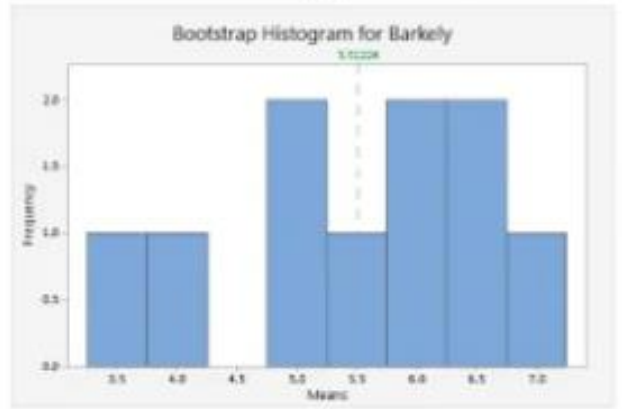
$$\bar{S}^* = \sum_b S(\mathbf{Z}^{*b})/B$$

Gambar di samping merupakan ilustrasi perbandingan data asli, data dengan boostrapping 10 sampel, dan data dengan boostrapping 1.000 sampel.

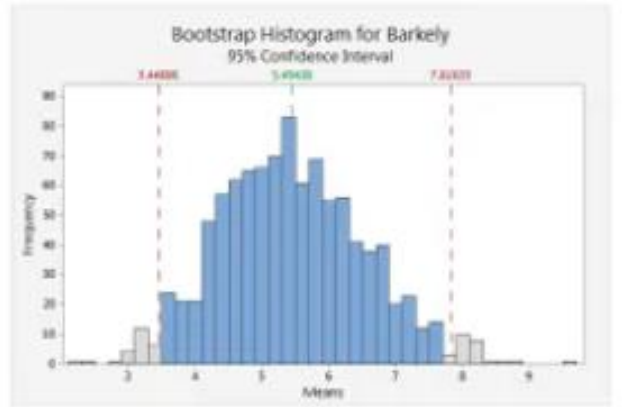
Pada data original, distribusi data terlihat menceng ke kanan. Seperti yang terlihat pada gambar, semakin besar resample yang dilakukan, distribusi sampel akan semakin mendekati bentuk distribusi normal.



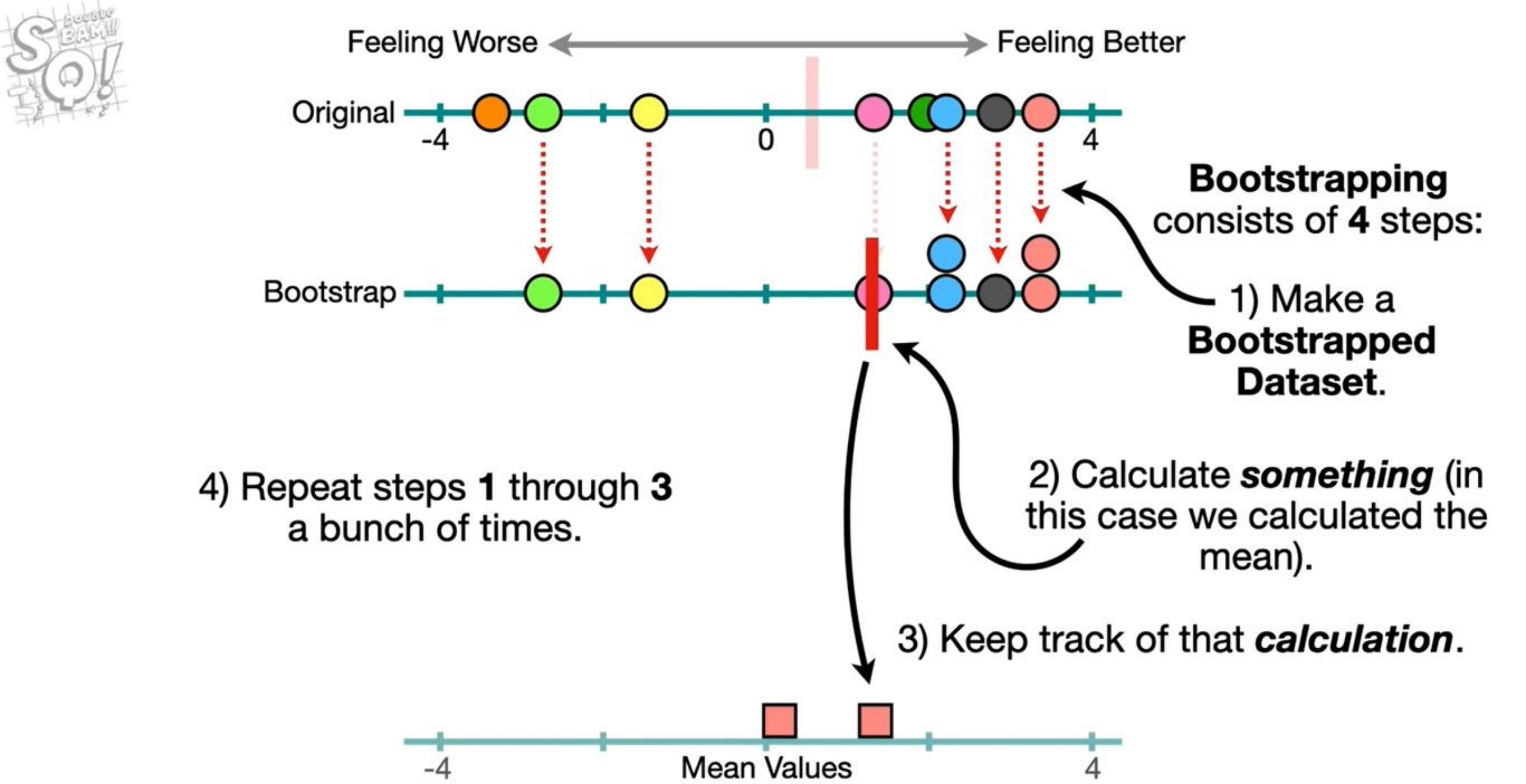
Data original



Bootstrap 10 sample



Bootstrap 1.000 sample



https://youtu.be/Xz0x-8-cgaQ

How many Bootstrap Replications, B?

- ✓ A fairly small number, $B = 25$, is sufficient to be "informative" (Efron)
- ✓ $B = 50$ is typically sufficient to provide a crude estimate of the SE, but $B > 200$ is generally used.
- ✓ CIs require larger values of B, B no less than 500, with $B = 1000$ recommended.

Bootstrap Pros and Cons

Advantages

- All-purpose, computer-intensive method useful for statistical inference.
- Bootstrap estimates of precision do not require knowledge of the theoretical form of an estimator's standard error, no matter how complicated it is.

Disadvantages

- Typically not useful for correlated (dependent) data.
- Missing data, censoring, data with outliers are also problematic
- Often used incorrectly

Ilustrasi

- Misal ingin mengestimasi rata-rata tinggi badan bayi baru lahir di Indonesia. Maka kita bisa menghitung nilai CI untuk μ?
- Semua inferensi untuk CI (Confidence Interval) memerlukan asumsi tentang distribusi sampel atau ukuran sampel
- Dengan bootstrap maka hal ini bisa dihindari dengan cara melakukan resampling yang merepresentasikan populasi
- Kalau sampel yang diambil secara acak berdistribusi tertentu (distribusi normal) dan/atau ukuran sampel besar, maka metode estimasi parameter yang lalu bisa digunakan. Bagaimana jika sebaliknya?
- Gunakanlah Bootstrap!

Ilustrasi

- Misalkan kita ingin mengestimasi berat badan bayi baru lahir di Indonesia, kemudian kita ambil random sampel ($n = 1009$).
- Berapakah nilai CI untuk μ dengan bootstrap (dengan $B = 10{,}000$) ?
- $\hat{\theta} \pm d$, parameter yang akan diestimasi adalah θ dan d adalah *margin of error.*
- Ingat! CI untuk mean: $\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
- Maka,
- Lakukan poin-poin berikut 10,000 kali:
  1. Randomly sample 1009 observation with replacement
  2. Calculate $\bar{x}$
  3. Store it for later (keseluruhan hasil dari $\bar{x}$ sebanyak 10,000 dapat dibuat histogram)
  4. Cek bootstrap distribution untuk $\bar{x}$