Bootstrap

A Key Technique in Random Forest

References

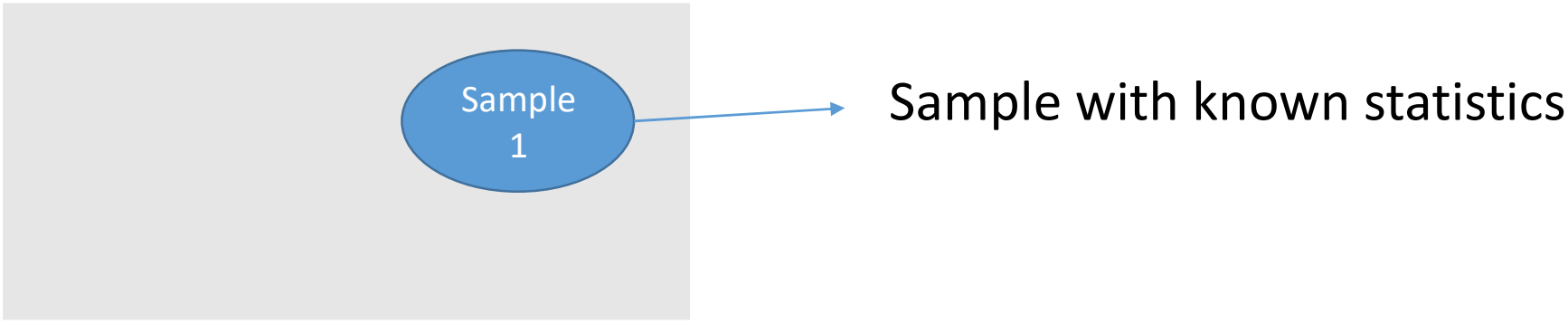
- Gareth James et. Al. (2017) An Introduction to Statistical Learning.
  - Section 5.2 The Bootstrap, pp. 187 – 190.
  - eTextbook download: https://www.statlearning.com/
- Chew C.H. (2022) Artificial Intelligence, Analytics and Data Science, Vol. 2.
  - Estimated Q4 2022

Bootstrap as an advanced underlying statistical technique for Machine Learning

- Bootstrap is a general statistical technique, not a model.
- An extremely creative way to use the given data sample.
- Allows one to infer from sample to unknown population without those assumptions.
- The basis for some advanced Machine Learning methods.
  - Essential for Random Forest (next topic).

Inference from Sample to Population (without bootstrap)

Target Population with unknown parameters



Example: Election

- Population: All eligible voters as listed in the voter registry
- Population Parameter: Proportion of all voters who voted for party A. Unknown constant.
- Sample: 100 eligible voters
- Sample Statistic: Proportion of all voters who voted for party A in the sample. Known, but varies from sample to sample.

Inference from Sample to Population (without bootstrap)

Population Parameter	Sample Statistic	Inference via Confidence Interval	Key Inference Assumption
Mean: μ	$\bar{x}$	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$, or $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$	CLT applies (with sufficient sample size), or x has normal dist.
Standard Dev: σ	s	$\sqrt{\frac{(n-1)s^2}{\chi^{2*}_{df,\frac{\alpha}{2}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^{2*}_{df,1-\frac{\alpha}{2}}}}$	x has normal dist.

For 95% CI: $z^* = 1.96$, $t^* = 2.776$, $\chi^{2*}_{df=n-1=39,0.975} = 23.65$, $\chi^{2*}_{df=39,0.025} = 58.12$

Source: Chew C.H. (2022) A.I., Analytics & Data Science, Vol. 2.

Inference from Sample to Population (without bootstrap)

Population Parameter	Sample Statistic	Inference via Confidence Interval	Key Inference Assumption
Proportion: π	p	$p \pm z^* \sqrt{\frac{p(1-p)}{n}}$	π is a binomial dist. parameter & np ≥ 5, n(1−p) ≥ 5
Correlation: ρ	r	Fisher Transform: $\mathrm{F} = \frac{1}{2}\frac{1+r}{1-r}$ $F_l = F - \frac{z^*}{\sqrt{n-3}}$ $F_u = F + \frac{z^*}{\sqrt{n-3}}$ $\frac{e^{2F_l} - 1}{e^{2F_l} + 1} < \rho < \frac{e^{2F_u} - 1}{e^{2F_u} + 1}$	x_A and x_B have iid Bivariate Normal Dist.

Source: Chew C.H. (2022) A.I., Analytics & Data Science, Vol. 2.

Model-Based Inference

- Linear Regression
  - P-values of betas
  - Confidence Interval of betas
- Logistic Regression
  - P-values of betas
  - Confidence Interval of betas
  - P-values of Odds Ratios
  - Confidence Interval of Odds Ratios
- Inference assumptions typically involve z, t or χ^2
- Other models have inference results too

Non-standard Inference

- What if:
  - You are not sure whether assumptions are valid and do not want to assume.
  - Sample or domain knowledge suggests assumptions are not valid. E.g. not independent, not Normal dist.,…
  - There are no (or you do not know) standard inference formulas e.g. median, 99 percentile, 10% trimmed mean, parameters of newly invented model, etc.
- Then, how do you do inference e.g. confidence interval?
  - Ans: Bootstrap

Bootstrap Animation Video at YouTube

- https://www.youtube.com/watch?v=Xz0x-8-cgaQ