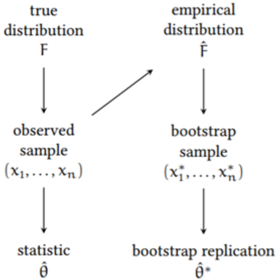
Statistical Methods for Data Science

Lesson 20 - Parametric bootstrap. Hypotheses testing.

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

Parametric bootstrap principle



- Let $X_1, \ldots, X_n \sim F(\gamma)$ be a random sample
  - with known F but *unknown* parameter γ
- Estimator $T = h(X_1, \ldots, X_n)$, e.g., $\bar{X}_n = (X_1 + \ldots + X_n)/n$
- From a dataset $x_1, \ldots, x_n$, we can
  - derive a point estimate $\hat{\theta} = h(x_1, \ldots, x_n)$
  - or, derive an estimate $\hat{\gamma}$ of γ
- From $F(\hat{\gamma})$ we can generate (a lot of) *bootstrap samples* $x_1^*, \ldots, x_n^*$
  - as realizations of $X_1^*, \ldots, X_n^* \sim F(\hat{\gamma})$

  and then (a lot of) bootstrap point estimates $\hat{\theta}^* = h(x_1^*, \ldots, x_n^*)$
- By the LLN, the empirical distribution of $\hat{\theta}^*$ will approximate the distribution of $T^* = h(X_1^*, \ldots, X_n^*)$ and then of T

Parametric bootstrap

Parametric bootstrap simulation (for $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, compute an estimate $\hat{\theta}$ for θ. Determine $F_{\hat{\theta}}$ as an estimate for F_θ, and compute the expectation $\mu^* = \mu_{\hat{\theta}}$ corresponding to $F_{\hat{\theta}}$.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from $F_{\hat{\theta}}$.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \mu_{\hat{\theta}},$$

where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

- Use the empirical distribution of $\delta^* = \bar{x}_n^* - \mu_{\hat{\theta}}$ for estimating
  - confidence interval (c_l, c_u) for $\delta = \bar{x}_n - \mu$ as $(q_{\alpha/2}, q_{1-\alpha/2})$ of δ^* distribution
  - $c_l \leq \delta = \bar{x}_n - \mu \leq c_u$ implies $\bar{x}_n - c_u \leq \mu \leq \bar{x}_n - c_l$, i.e. c.i. for μ is $(\bar{x}_n - c_u, \bar{x}_n - c_l)$

See R script

Application: distribution fitting

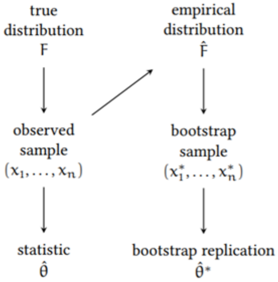
- Consider a dataset $x_1, \ldots, x_n \sim F$
- Is the dataset from an $Exp(\lambda)$ for some λ? I.e., is it $F = Exp(\lambda)$?
- We estimate $\hat{\lambda} = 1/\bar{x}_n$
- We measure how close is the dataset to the distribution as:

$$t_{ks} = \sup_{a \in \mathbb{R}} |F_n(a) - F_{\hat{\lambda}}(a)|$$

  where:
  - $F_n(a)$ is the empirical cumulative distribution of $x_1, \ldots, x_n$
  - $F_{\hat{\lambda}}(a) = 1 - e^{\hat{\lambda} a}$, for $a \geq 0$, is the distribution function of $Exp(\hat{\lambda})$
  - t_{ks} is called the *Kolmogorov-Smirnov* distance
- if $F = Exp(\lambda)$ then both $F_n \approx F$ and $F_{\hat{\lambda}} \approx F$, and then $F_n \approx F_{\hat{\lambda}}$, so that t_{ks} is small
- if $F \neq Exp(\lambda)$ then $F_n \approx F \neq Exp(\lambda) \approx F_{\hat{\lambda}}$, so that t_{ks} is large

See R script

Application: distribution fitting



- For the software dataset from the textbook
  - $\hat{\lambda} = 0.0015$ and $t_{ks} = 0.17$
- Is $t_{ks} = 0.17$ expected or an extreme value?
- Let's study the distribution of the bootstrap estimator:

$$T_{ks} = \sup_{a \in \mathbb{R}} |F_n^*(a) - F_{\hat{\Lambda}^*}(a)|$$

  where:
  - $X_1^*, \ldots, X_n^* \sim Exp(\hat{\lambda})$ is a bootstrap sample
  - $F_n^*(a)$ is the empirical cumulative distribution of the bootstrap sample
  - $\hat{\Lambda}^* = 1/\bar{X}_n^*$
- It turns out $P(T_{ks} > 0.17) \approx 0$, unlikely that $Exp()$ is the right model

See R script

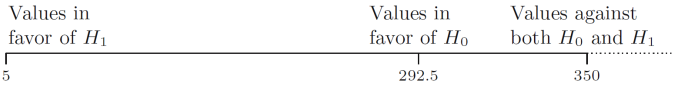
Hypothesis testing

- In the previous application, we tested how likely is *Exp*() for the given dataset
- In general, hypotheses testing consists of contrasting two conflicting theories (hypotheses) based on observed data
- Consider the German tank problem:
  - Military intelligence states that $N = 350$ tanks were produced *[H0 or null hypothesis]*
  - Alternative hypothesis: *[H1 hypothesis]*
    $N < 350$ (*one-tailed or one-sided test*), or $N \neq 350$ (*two-tailed or two-sided test*)
  - Observed serial tank id's: 61 19 56 24 16
- Statistical test: How likely is the observed data under the null hypothesis?
  - If it is NOT (sufficiently) likely, we reject the null hypothesis in favor of H1
  - If it is (sufficiently) likely, we cannot reject the null hypothesis
- Why '*we cannot reject the null hypothesis*' and not instead '*we accept the null hypothesis*'?
  - Other hypotheses, e.g., $N = 349$ or $N = 351$, could also not be rejected
  - We cannot say which of $N = 349$ or $N = 350$ or $N = 351$ is actually true

Test statistic

Test Statistic. Suppose the dataset is modeled as the realization of random variables $X_1, X_2, \ldots, X_n$. A *test statistic* is any sample statistic $T = h(X_1, X_2, \ldots, X_n)$, whose numerical value is used to decide whether we reject H_0.

- In the German tank example:
  - $H_0 : N = 350$
  - $H_1 : N < 350$
  - Observed serial tank id's: 61 19 56 24 16
- We use $T = \max\{X_1, X_2, X_3, X_4, X_5\}$
- If H_0 is true, i.e., $N = 350$, then $E[T] = \frac{5}{6}(N+1) = \frac{5}{6}351 = 292.5$



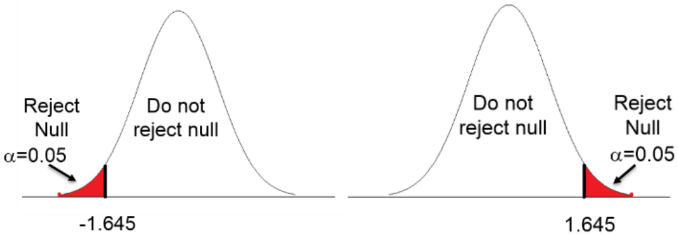
- If H_0 is true, we have:

$$P(T \leq 61) = P(\max\{X_1, X_2, X_3, X_4, X_5\} \leq 61) = \frac{61}{350} \cdot \frac{60}{349} \cdots \frac{57}{346} = 0.00014$$

very unlikely: either we are unfortunate, or H_0 can be rejected

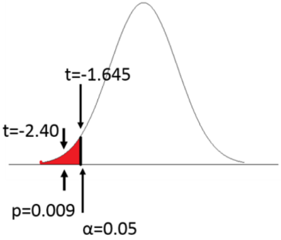
Statistical test of hypothesis: one-tailed

- H_0: $\theta = v$ *[Null hypothesis]*
- H_1: $\theta < v$ (resp. H_1: $\theta > v$) *[Left-tailed/Right-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset
- c_l s.t. $P(T \leq c_l) = \alpha$ (resp. c_u s.t. $P(T \geq c_u) = \alpha$) *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $h(x_1, \ldots, x_n) \leq c_l$ (resp. $h(x_1, \ldots, x_n) \geq c_u$): H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected



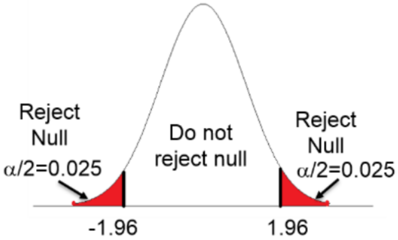
Statistical test of hypothesis: one-tailed

- H_0: $\theta = v$ *[Null hypothesis]*
- H_1: $\theta < v$ (resp. H_1: $\theta > v$) *[Left-tailed/Right-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset
- $p = P(T \leq h(x_1, \ldots, x_n))$ (resp. $p = P(T \geq h(x_1, \ldots, x_n))$) *[p-value]*
  - evidence against H_0 - the smaller the stronger evidence
- Output of the test at confidence level $100(1-\alpha)\%$ using p-values
  - $p \leq \alpha$: H_0 is rejected
  - otherwise: H_0 cannot be rejected



Statistical test of hypothesis: two-tailed

- H_0: $\theta = v$ *[Null hypothesis]*
- H_1: $\theta \neq v$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset
- c_l s.t. $P(T \leq c_l) = \alpha/2$ and c_u s.t. $P(T \geq c_u) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $h(x_1, \ldots, x_n) \leq c_l$ or $h(x_1, \ldots, x_n) \geq c_u$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected



Type I and Type II errors

		True state of nature	
		H_0 is true	H_1 is true
Our decision on the basis of the data	Reject H_0	*Type I error*	Correct decision
	Not reject H_0	Correct decision	*Type II error*

- Type I error: we falsely reject H_0 *[α-risk, false positive rate]*
  - E.g., convicting an innocent defendant
  - we reject H_0 when $p < \alpha$, so this error occur with probability $100\alpha\%$
  - this error can be controlled by setting the significance level α to the largest acceptable value
  - how much is an *acceptable value*?
  - A possible solution is to solely report the *p*-value, which conveys the maximum amount of information and permits decision makers to choose their own level
- Type II error: we falsely do not reject H_0 *[β-risk, false negative rate]*
  - E.g., acquitting a criminal
  - $1 - \beta = P(\text{Reject} H_0 | H_1 \text{ is true})$ is called the *power* of the test