Outline: **Resampling**
(F. Chiaromonte)

Introduction to Statistical Learning Chapter 5 Section 2 Lab 3

COMPUTATIONAL ASSESSMENT/~~TUNING~~ OF STATISTICAL PROCEDURES

For a very long time, the properties of statistical procedures were assessed using mathematical manipulations, facts about probability distributions and asymptotic results.

Mathematical tractability defined a narrow scope for statistics:

- Consider only simple procedures
- Introduce strong assumptions on the stochastic mechanism generating the data, and/or
- Prove properties only for large samples

’70s onward: replacing math with computation has substantially expanded the scope.

HOW DO WE EVALUATE THE ACCURACY AND SAMPLING VARIABILITY OF PROCEDURES THAT WE CANNOT TACKLE ANALYTICALLY?

A computational approach to evaluate the accuracy of a statistic used as estimator for a quantity of interest. We can estimate its standard error; in fact, we can approximate its sampling distribution by ***simulating*** its variability across samples.

Setup:

- Sample of n independent observations from a stochastic mechanism

$$x \quad = (x_1 \dots x_n) \ , \ \ x_i \ iid \sim F$$

- Statistics (function of the observations) produces a real valued estimate of θ

$$\hat{\theta} = g(x)$$

- What is its accuracy for θ? Standard error

$$se(\hat{\theta}) = \sqrt{E[(\hat{\theta} - \theta)^2]}$$

equal to the standard deviation of the sampling distribution of the estimator, if it is unbiased for θ.

How do we estimate this standard error?

We know the answer for the sample mean as an estimator of the population mean:

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

$$\mathrm{E}(\bar{x}) = \mu$$
$$se(\bar{x}) = sd(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$
$$\widehat{se}(\bar{x}) = \frac{s}{\sqrt{n}}$$

This is the skeleton of the traditional approach…

- We can proceed similarly for other estimators based on averaging, e.g. estimators of the slope parameters of a regression model.
- We can also extend this logic to estimators that are smooth (differentiable) functions of averages; the Delta method (based on Taylor expansions).

But what if the picture is more complex? e.g. estimating

- A quantile of a univariate population
- The correlation coefficient of a bivariate population
- The covariance eigen-ratio (largest/sum) for a multivariate population…

(note here we are still referring to populations in Euclidian spaces)

- An index defined on a population of curves

which may be hard to do with estimators based on averaging or smooth functions of averages… Whatever

- The nature of the statistics and its domain space
- The nature of the stochastic process F generating the data, and/or
- The size of the sample n (need not resort to limit theorems)

We can use the (one-sample, non-parametric) ***Bootstrap to estimate the standard error***.

The bootstrap algorithm:

- Generate B bootstrap samples of size n drawing with replacement from the data

$$x^*_{(b)} \quad b = 1 \dots B$$

- On each, compute the statistic – producing B bootstrap values; “copies” that mimic its sampling variability

$$\hat{\theta}^*_{(b)} = g(x^*_{(b)}) \quad b = 1 \dots B$$

- Estimate the standard error as the standard deviation of the bootstrap values

$$\widehat{se}_{BT} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(\cdot)})^2} \qquad \hat{\theta}^*_{(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{(b)}$$

Already rather good with **B=200**.

<u>Rationale 1</u>: since we cannot take more samples of size n from F, we do the next best thing, i.e. generate them from the empirical distribution $\hat{F}_n$. This is our best estimate of F based on the available sample of n observations (in fact, its non-parametric MLE).

<u>Rationale 2</u>: the bootstrap lets us gauge the variability of $\hat{\theta}$ creating perturbations of the original data by resampling (with replacement). These are less local of the perturbations created with the Jackknife (create n samples of size n-1 deleting one observation at a time).

$\hat{\theta}$ computed on the original sample of size n has true standard error $se(\hat{\theta}) = \rho(F; n)$.

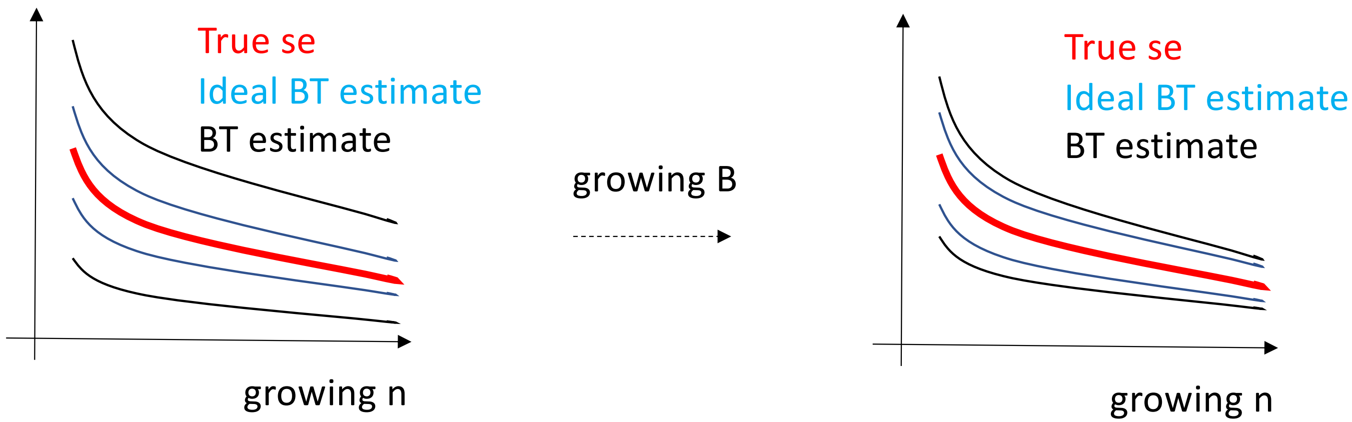
The “ideal” bootstrap estimate would be $\rho(\hat{F}_n; n)$ replacing F with $\hat{F}_n$.

As n increases

- The true standard error $\rho(F; n)$ likely gets smaller, closer to 0
- The empirical distribution $\hat{F}_n$ gets closer to F
- The ideal bootstrap estimate $\rho(\hat{F}_n; n)$ gets closer to the (shrinking) true $\rho(F; n)$.

However, for any given n, $\widehat{se}_{BT}$ is an approximation of the ideal $\rho(\hat{F}_n; n)$ based on B bootstrap samples. As B increases

- The bootstrap estimate $\widehat{se}_{BT}$ gets closer to the ideal $\rho(\hat{F}_n; n)$.



One step forward, from estimating standard error to producing ***Confidence Intervals***.

Pivot-based construction of the (1-α) coverage CI for the population mean:

$$CI(\alpha) = \bar{x} \pm m_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \qquad\qquad m_{(\cdot)} = \Phi^{-1}(\cdot)$$

inverse of the cdf of a N(0,1)

More generally, if we can assume that the sampling distribution is approximately $\hat{\theta} \sim N(\theta; sd(\hat{\theta}))$ (symmetric around quantity of interest, bell shaped, with very fast vanishing tails), we construct the interval as

$$CI(\alpha) = \hat{\theta} \pm m_{\frac{\alpha}{2}} \widehat{se}(\hat{\theta})$$

This works if the estimator is based on averaging or is a smooth function of averages because of the Central Limit Theorem. But if the picture is more complex the interval will be bad; coverage not guaranteed.

Traditional approach:

- Figure out an invertible transformation $\varphi(\theta)$ and an estimator for it such that $\hat{\varphi} \sim N\left(\varphi; sd(\hat{\varphi})\right)$

- Build a pivot-based (1-α) coverage interval for the transformation

$$LOW = \hat{\varphi} - m_{\frac{\alpha}{2}} \widehat{se}(\hat{\varphi}) \qquad UP = \hat{\varphi} + m_{\frac{\alpha}{2}} \widehat{se}(\hat{\varphi})$$

- Back-transform (invariance of coverage) to a (1-a) coverage interval for the quantity of interest

$$LOW = \varphi^{-1}(\hat{\varphi} - m_{\frac{\alpha}{2}} \widehat{se}(\hat{\varphi})) \qquad UP = \varphi^{-1}(\hat{\varphi} + m_{\frac{\alpha}{2}} \widehat{se}(\hat{\varphi}))$$

But what if it is hard or impossible to figure out an appropriate “normalizing” transformation?

In the bootstrap world, we could naively take (**<u>Bootstrap Standard Confidence Interval</u>**)

$$CI(\alpha) = \hat{\theta} \pm m_{\frac{\alpha}{2}} \widehat{se}_{BT}$$

but this is unlikely to be a good choice as it relies on assumptions on the nature of the sampling distribution that may not be met.

if we are willing to use computer power and generate **B=1000, 2000** bootstrap samples (instead of B=200) we can approximate the traditional approach without having to figure out the normalizing transformation.

Once again, replace math with computation and expand the scope!

Bootstrap Percentile Confidence Interval:

- Generate B bootstrap samples of size n drawing with replacement from the data

  $x^*_{(b)} \quad b = 1 \dots B$

- On each, compute the statistic – producing B bootstrap values; "copies" that mimic its sampling variability

  $\hat{\theta}^*_{(b)} = g(x^*_{(b)}) \quad b = 1 \dots B$

- Build the corresponding empirical cdf $\hat{G}$

- Use its percentiles to define a non-pivot-based (1-α) coverage CI as

$$LOW = G^{-1}\left(\frac{\alpha}{2}\right) \qquad UP = G^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Not the ultimate answer; this can be further improved introducing a correction for bias and potential non-constant variance in "tracking" θ with $\hat{\theta}^*_{(\cdot)}$.

Some additions:
variations on the bootstrap and random permutations

Multi-sample Bootstrap:

Samples from two univariate populations, e.g. cholesterol for individuals with/out metabolic syndrome:

$x^{(1)} = (x_1^{(1)} \dots x_{n1}^{(1)})$, $x_i^{(1)}$ $iid \sim F^{(1)}$ $\qquad$ $x^{(2)} = (x_1^{(2)} \dots x_{n2}^{(2)})$, $x_i^{(2)}$ $iid \sim F^{(2)}$

Of interest θ = shift in medians between the populations

Estimator: $\hat{\theta} \quad = g\left(x^{(1)}, x^{(2)}\right) = \mathrm{Med}\left(x^{(2)}\right) - \mathrm{Med}\left(x^{(1)}\right)$

- Generate B pairs of bootstrap samples of size n_1 and n_2 drawing with replacement from the data in $x^{(1)}$ and $x^{(2)}$, respectively

  $x_{(b)}^{(1)*}, x_{(b)}^{(2)*} \quad b = 1 \dots B$

- On each pair, compute the statistic – producing B bootstrap values; ”copies” that mimic its sampling variability

  $\hat{\theta}_{(b)}^{*} = g(x_{(b)}^{(1)*}, x_{(b)}^{(2)*}) \quad b = 1 \dots B$

The bootstrap must reproduce the original sampling; if we drew samples of size $n_1 + n_2$ with replacement from $x^{(1)} \cup x^{(2)}$, we could get more/less observations of each type, adding inappropriately to the variability of $\hat{\theta}$.

Parametric Bootstrap:

Sample from a population for which we can postulate a parametric form, e.g. income following a Pareto distribution with $\tau=(\eta,\alpha)$; *minimum* (scale) and *shape* parameters.

$x \quad =(x_1 \;\ldots x_n\,)\;, x_i \;\; iid \sim F(\tau) \in \mathcal{F}(\tau)$

Of interest θ = 90th percentile ,

Estimator: $\hat{\theta} \quad = g(x\,) = q_{0.95}(x)$

Compute the MLE $\hat{\tau} = \left(\min\{x_i\}\,;\frac{n}{\sum \ln(x_i) - \ln(\min\{x_i\})}\right)$ and use $F(\hat{\tau})$ instead of $\hat{F}_n$.

- Generate B bootstrap samples of size n from $F(\hat{\tau})$

  $x^*_{(b)} \quad b = 1 \ldots B$

- On each, compute the statistic – producing B bootstrap values; ”copies” that mimic its sampling variability

  $\hat{\theta}^*_{(b)} = g(x^*_{(b)}) \quad b = 1 \ldots B$

If we drew samples of size n with replacement from x, we would not utilize our knowledge of $\mathcal{F}(\tau)$, adding inappropriately to the variability of $\hat{\theta}$.

<u>Random permutations</u>:

Studying a relationship, e.g., Y (continuous or categorical) as a function of X (or X's)
A statistic that estimates an association parameter ρ (e.g., correlation coefficient or regression coefficient)
Instead of simulating the sampling distribution of the statistic, simulate the null distribution of that statistic under *H_o: there is no association*

<u>Setup</u>:

- Sample of n independent paired measurements from a stochastic mechanism

$$(y, x) \quad = ((y_1, x_{1)} \ldots (y_n, x_n)) \;\; , \quad (y_i, x_i) \; iid \sim F$$

- Statistics (function of the observations) produces a real valued estimate of ρ , $\hat{\rho} = r(y, x)$

<u>Algorithm</u>:

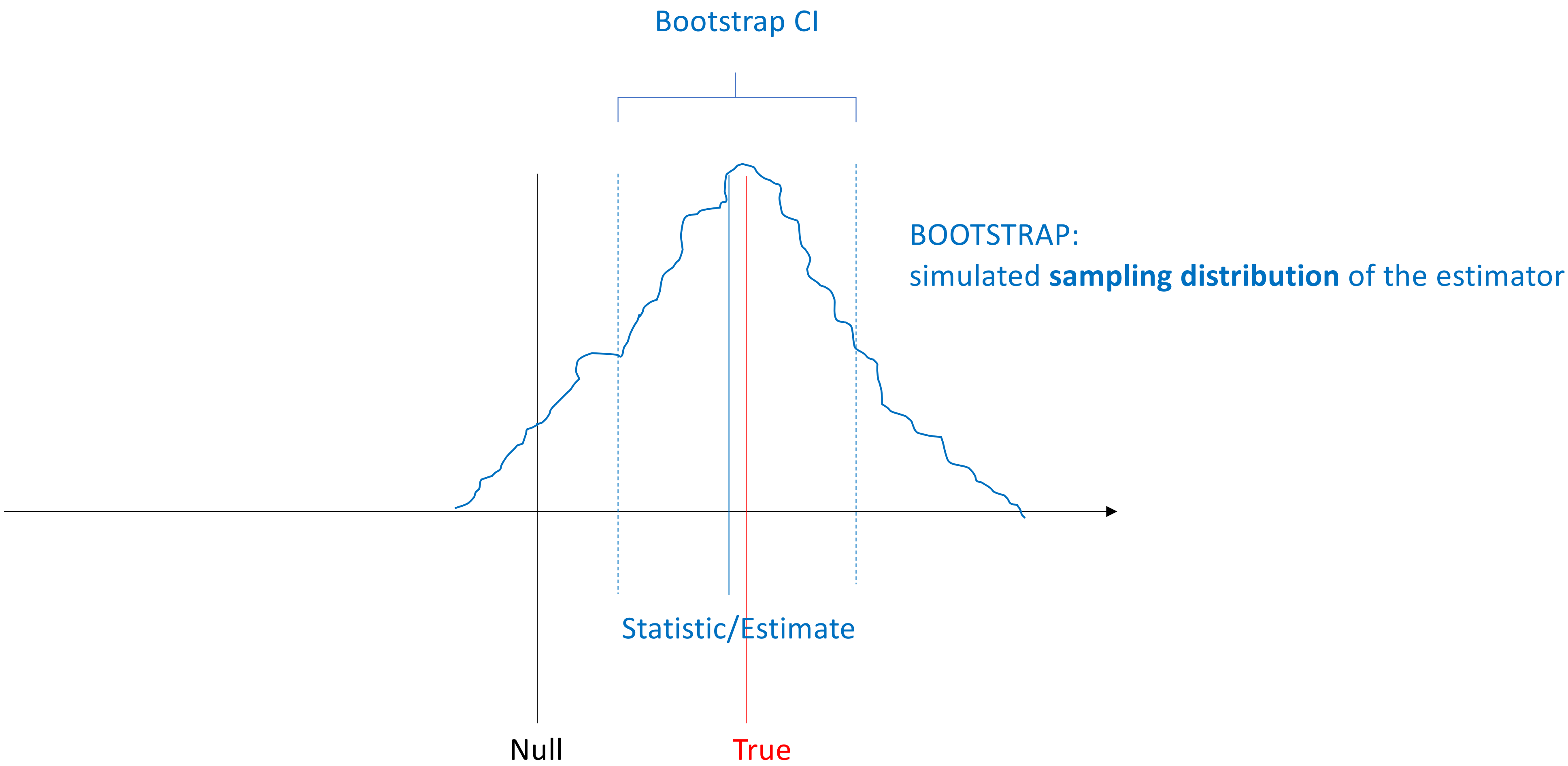
- Generate B no-association samples permuting the n entries of y at random

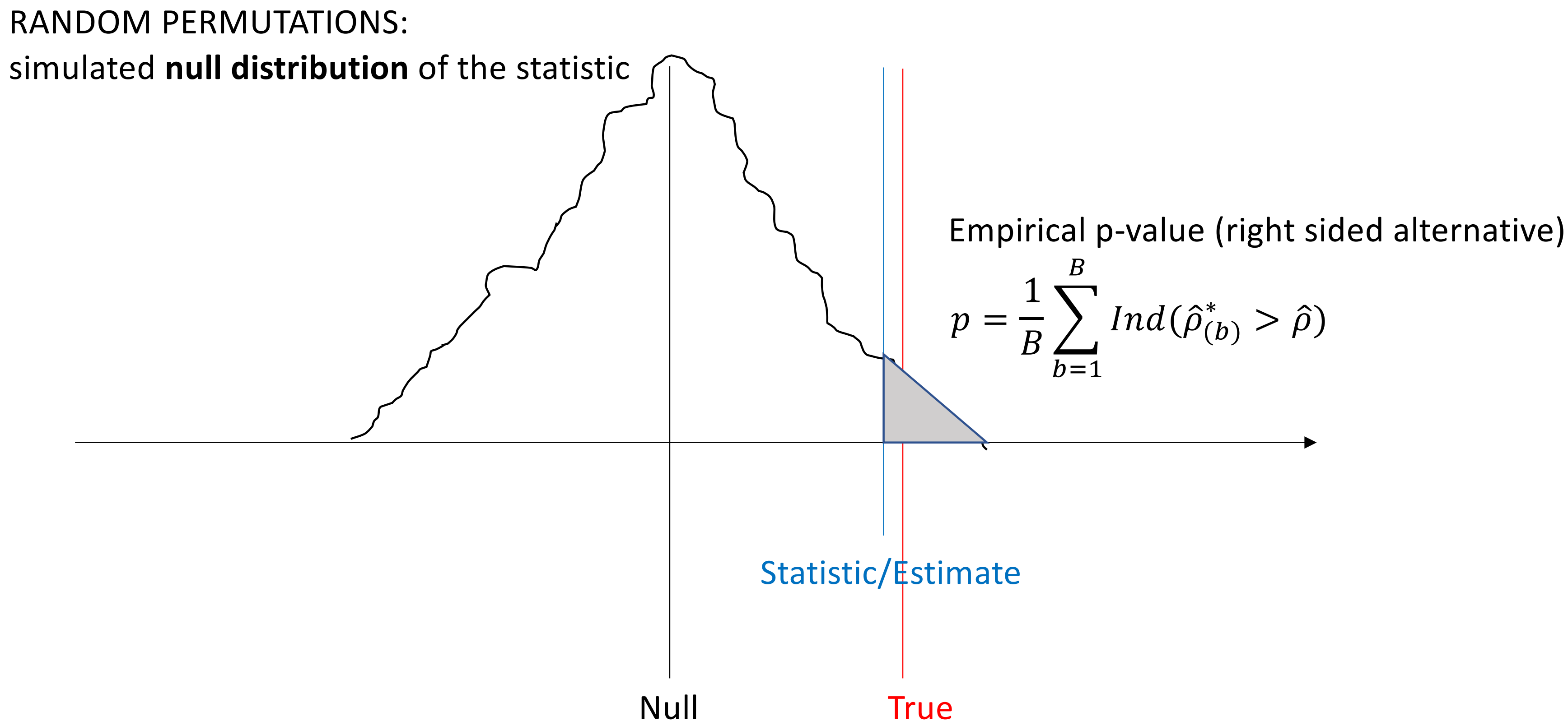
$$(y^*_{(b)}, x) \quad b = 1 \ldots B$$

- On each, compute the statistic – producing B "copies" that mimic its null distribution

$$\hat{\rho}^*_{(b)} = r(y^*_{(b)}, x) \quad b = 1 \ldots B$$

X_1	X_2	…	X_p	Y
x_{11}	x_{12}		x_{1p}	y_1
x_{21}	x_{22}		x_{2p}	y_2
x_{n1}	x_{n2}		x_{np}	y_n


Bootstrap CI
BOOTSTRAP:
simulated **sampling distribution** of the estimator
Statistic/Estimate
Null
True


RANDOM PERMUTATIONS:
simulated **null distribution** of the statistic
Empirical p-value (right sided alternative)
$$p = \frac{1}{B}\sum_{b=1}^{B} Ind(\hat{\rho}^{*}_{(b)} > \hat{\rho})$$
Statistic/Estimate
Null
True