

# The Bootstrap

- The *bootstrap* is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

Prof. Tibshirani was one of the inventors of the Bootstrap method. It was a big statistical development of the 1990s.  
So not all Stats is 100 years old!

[1] R. Tibshirani and K. Knight, "Model Search by Bootstrap 'Bumping,'" Journal of Computational and Graphical Statistics, vol. 8, no. 4, pp. 671–686, Dec. 1999, <https://doi.org/10.1080/10618600.1999.10474842>.

[2] B. Efron and R. Tibshirani, "Improvements on Cross-Validation: The 632+ Bootstrap Method," J. Amer. Stat. Assoc., vol. 92, no. 438, pp. 548–560, Jun. 1997, <https://doi.org/10.1080/01621459.1997.10474007>.

[3] Bradley Efron and Robert J Tibshirani, An introduction to the bootstrap, vol. 57. Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 1993.

## Where does the name came from?

- The use of the term bootstrap derives from the phrase *to pull oneself up by one's bootstraps*, widely thought to be based on one of the eighteenth century “The Surprising Adventures of Baron Munchausen” by Rudolph Erich Raspe:

*The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.*

- It is not the same as the term “bootstrap” used in computer science meaning to “boot” a computer from a set of core instructions, though the derivation is similar.

## A simple example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of  $X$  and  $Y$ , respectively, where  $X$  and  $Y$  are random quantities.
- We will invest a fraction  $\alpha$  of our money in  $X$ , and will invest the remaining  $1 - \alpha$  in  $Y$ .
- We wish to choose  $\alpha$  to minimize the total risk, or variance, of our investment. In other words, we want to minimize  $\text{Var}(\alpha X + (1 - \alpha)Y)$ .
- One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

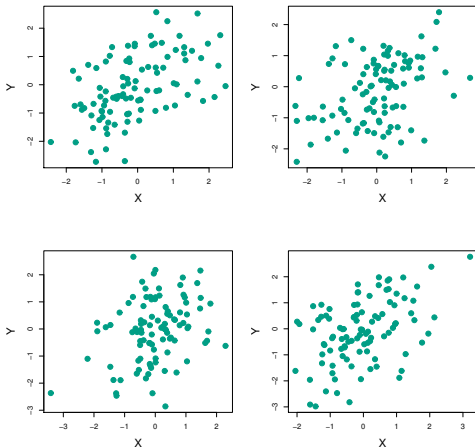
where  $\sigma_X^2 = \text{Var}(X)$ ,  $\sigma_Y^2 = \text{Var}(Y)$ , and  $\sigma_{XY} = \text{Cov}(X, Y)$ .

## Example continued

- But the values of  $\sigma_X^2$ ,  $\sigma_Y^2$ , and  $\sigma_{XY}$  are unknown.
- We can compute estimates for these quantities,  $\hat{\sigma}_X^2$ ,  $\hat{\sigma}_Y^2$ , and  $\hat{\sigma}_{XY}$ , using a data set that contains measurements for  $X$  and  $Y$ .
- We can then estimate the value of  $\alpha$  that minimizes the variance of our investment using

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

## Example continued



*Each panel displays 100 simulated returns for investments  $X$  and  $Y$ . From left to right and top to bottom, the resulting estimates for  $\alpha$  are 0.576, 0.532, 0.657, and 0.651.*

## Example continued

- To estimate the standard deviation of  $\hat{\alpha}$ , we repeated the process of simulating 100 paired observations of  $X$  and  $Y$ , and estimating  $\alpha$  1,000 times.
- We thereby obtained 1,000 estimates for  $\alpha$ , which we can call  $\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{1000}$ .
- The left-hand panel of the Figure on slide 29 displays a histogram of the resulting estimates.
- For these simulations the parameters were set to  $\sigma_X^2 = 1, \sigma_Y^2 = 1.25$ , and  $\sigma_{XY} = 0.5$ , and so we know that the true value of  $\alpha$  is 0.6 (indicated by the red line).

## Example continued

- The mean over all 1,000 estimates for  $\alpha$  is

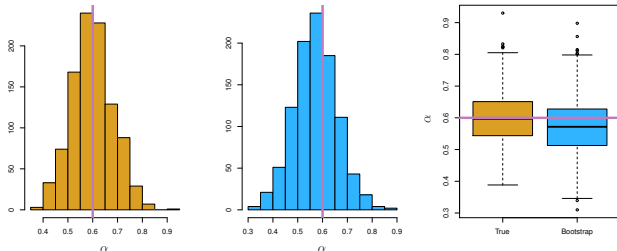
$$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996,$$

very close to  $\alpha = 0.6$ , and the standard deviation of the estimates is

$$\sqrt{\frac{1}{1000 - 1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083.$$

- This gives us a very good idea of the accuracy of  $\hat{\alpha}$ :  $\text{SE}(\hat{\alpha}) \approx 0.083$ .
- So roughly speaking, for a random sample from the population, we would expect  $\hat{\alpha}$  to differ from  $\alpha$  by approximately 0.08, on average.

# Results



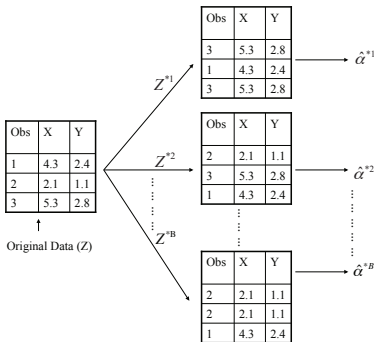
*Left:* A histogram of the estimates of  $\alpha$  obtained by generating 1,000 simulated data sets from the true population. *Center:* A histogram of the estimates of  $\alpha$  obtained from 1,000 bootstrap samples from a single data set. *Right:* The estimates of  $\alpha$  displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of  $\alpha$ .



## Now back to the real world

- The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set *with replacement*.
- Each of these “bootstrap data sets” is created by sampling *with replacement*, and is the *same size* as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

## Example with just 3 observations



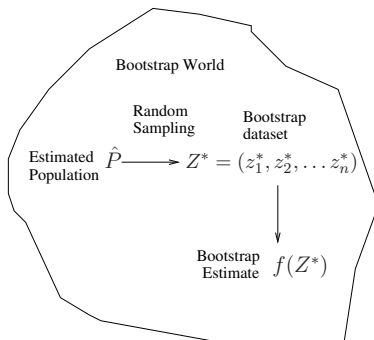
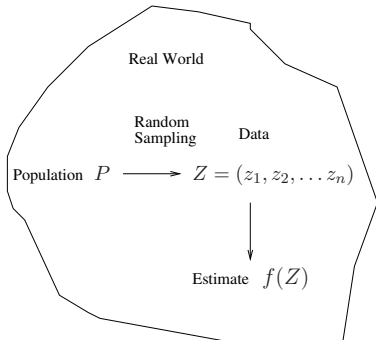
A graphical illustration of the bootstrap approach on a small sample containing  $n = 3$  observations. Each bootstrap data set contains  $n$  observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of  $\alpha$

- Denoting the first bootstrap data set by  $Z^{*1}$ , we use  $Z^{*1}$  to produce a new bootstrap estimate for  $\alpha$ , which we call  $\hat{\alpha}^{*1}$
- This procedure is repeated  $B$  times for some large value of  $B$  (say 100 or 1000), in order to produce  $B$  different bootstrap data sets,  $Z^{*1}, Z^{*2}, \dots, Z^{*B}$ , and  $B$  corresponding  $\alpha$  estimates,  $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \dots, \hat{\alpha}^{*B}$ .
- We estimate the standard error of these bootstrap estimates using the formula

$$\text{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^B (\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^*)^2}.$$

- This serves as an estimate of the standard error of  $\hat{\alpha}$  estimated from the original data set. See center and right panels of Figure on slide 29. Bootstrap results are in blue. For this example  $\text{SE}_B(\hat{\alpha}) = 0.087$ .

## A general picture for the bootstrap



## The bootstrap in general

- In more complex data situations, figuring out the appropriate way to generate bootstrap samples can require some thought.
- For example, if the data is a time series, we can't simply sample the observations with replacement (*why not?*).
- We can instead create blocks of consecutive observations, and sample those with replacements. Then we paste together sampled blocks to obtain a bootstrap dataset.

## Other uses of the bootstrap

- Primarily used to obtain standard errors of an estimate.
- Also provides approximate confidence intervals for a population parameter. For example, looking at the histogram in the middle panel of the Figure on slide 29, the 5% and 95% quantiles of the 1000 values is (.43, .72).
- This represents an approximate 90% confidence interval for the true  $\alpha$ . *How do we interpret this confidence interval?*
- The above interval is called a *Bootstrap Percentile* confidence interval. It is the simplest method (among many approaches) for obtaining a confidence interval from the bootstrap.

## Can the bootstrap estimate prediction error?

- In cross-validation, each of the  $K$  validation folds is distinct from the other  $K - 1$  folds used for training: *there is no overlap*. This is crucial for its success.

## Can the bootstrap estimate prediction error?

- In cross-validation, each of the  $K$  validation folds is distinct from the other  $K - 1$  folds used for training: *there is no overlap*. This is crucial for its success. *Why?*
- To estimate prediction error using the bootstrap, we could think about using each bootstrap dataset as our training sample, and the original sample as our validation sample.
- But each bootstrap sample has significant overlap with the original data. About two-thirds of the original data points appear in each bootstrap sample. *Can you prove this?*
- This will cause the bootstrap to seriously underestimate the true prediction error. *Why?*
- The other way around— with original sample = training sample, bootstrap dataset = validation sample— is worse!



## Removing the overlap

- Can partly fix this problem by only using predictions for those observations that did not (by chance) occur in the current bootstrap sample.
- But the method gets complicated, and in the end, cross-validation provides a simpler, more attractive approach for estimating prediction error.

## Pre-validation

- In microarray and other genomic studies, an important problem is to compare a predictor of disease outcome derived from a large number of “biomarkers” to standard clinical predictors.
- Comparing them on the same dataset that was used to derive the biomarker predictor can lead to results strongly biased in favor of the biomarker predictor.
- *Pre-validation* can be used to make a fairer comparison between the two sets of predictors.

## Motivating example

An example of this problem arose in the paper of van't Veer *et al.* *Nature* (2002). Their microarray data has 4918 genes measured over 78 cases, taken from a study of breast cancer. There are 44 cases in the good prognosis group and 34 in the poor prognosis group. A “microarray” predictor was constructed as follows:

1. 70 genes were selected, having largest absolute correlation with the 78 class labels.
2. Using these 70 genes, a nearest-centroid classifier  $C(x)$  was constructed.
3. Applying the classifier to the 78 microarrays gave a dichotomous predictor  $z_i = C(x_i)$  for each case  $i$ .

## Results

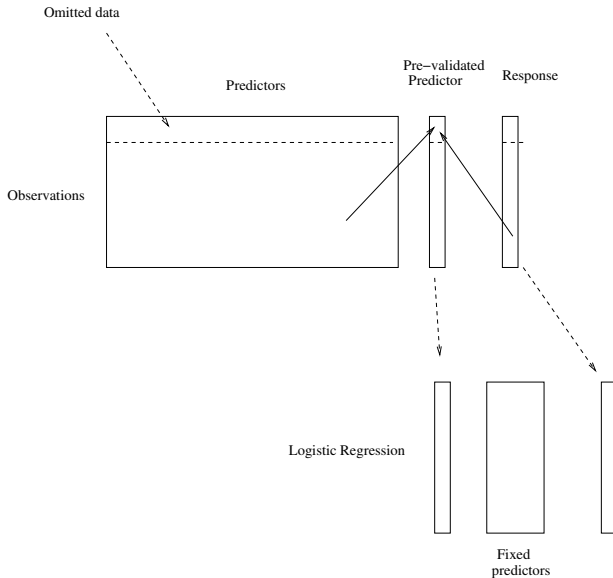
Comparison of the microarray predictor with some clinical predictors, using logistic regression with outcome **prognosis**:

Model	Coef	Stand. Err.	Z score	p-value
Re-use				
microarray	4.096	1.092	3.753	0.000
angio	1.208	0.816	1.482	0.069
er	-0.554	1.044	-0.530	0.298
grade	-0.697	1.003	-0.695	0.243
pr	1.214	1.057	1.149	0.125
age	-1.593	0.911	-1.748	0.040
size	1.483	0.732	2.026	0.021
Pre-validated				
microarray	1.549	0.675	2.296	0.011
angio	1.589	0.682	2.329	0.010
er	-0.617	0.894	-0.690	0.245
grade	0.719	0.720	0.999	0.159
pr	0.537	0.863	0.622	0.267
age	-1.471	0.701	-2.099	0.018
size	0.998	0.594	1.681	0.046

## Idea behind Pre-validation

- Designed for comparison of adaptively derived predictors to fixed, pre-defined predictors.
- The idea is to form a “pre-validated” version of the adaptive predictor: specifically, a “fairer” version that hasn’t “seen” the response  $y$ .

# Pre-validation process



## Pre-validation in detail for this example

1. Divide the cases up into  $K = 13$  equal-sized parts of 6 cases each.
2. Set aside one of parts. Using only the data from the other 12 parts, select the features having absolute correlation at least .3 with the class labels, and form a nearest centroid classification rule.
3. Use the rule to predict the class labels for the 13th part
4. Do steps 2 and 3 for each of the 13 parts, yielding a “pre-validated” microarray predictor  $\tilde{z}_i$  for each of the 78 cases.
5. Fit a logistic regression model to the pre-validated microarray predictor and the 6 clinical predictors.

## The Bootstrap versus Permutation tests

- The bootstrap samples from the estimated population, and uses the results to estimate standard errors and confidence intervals.
- Permutation methods sample from an estimated *null* distribution for the data, and use this to estimate p-values and False Discovery Rates for hypothesis tests.
- The bootstrap can be used to test a null hypothesis in simple situations. Eg if  $\theta = 0$  is the null hypothesis, we check whether the confidence interval for  $\theta$  contains zero.
- Can also adapt the bootstrap to sample from a null distribution (See Efron and Tibshirani book “An Introduction to the Bootstrap” (1993), chapter 16) but there’s no real advantage over permutations.