Evaluating generalization
(validation)

Topics

- Validation of biomedical models
- Data-splitting
- Resampling
  - Cross-validation
  - Jackknife
  - Bootstrap

Generalization Problems

- Different population characteristics
  - Descriptive statistics on populations help determine discrepancies
- Overfitting management
  - Regression: shrinkage (keep coefficients small)
  - Neural networks: weight decay (keep weights small)

  - Regression: variable selection
  - Neural networks: weight elimination
  - Classification trees: pruning
  - Association rules: rule elimination

  - Neural networks: early stopping

  - General: penalty for adding parameters

Types of validation

- External
  - Different data sets for building model (including tuning parameters) and testing
  - Can be achieved with data splitting of same sample (random or chronological) or finding a new sample
- Internal
  - Resampling

Data Splitting

- Training set is used to build the model
- Test set left aside for evaluation purposes

- Training set also known as construction set, which can contain a subset to estimate initial parameters and another to tune them (hold-out set)
- Rationale: If data are abundant, then there is no need to “recycle” cases
- This is the most accepted form of validation in medicine!

Resampling

- When sample is small (relative to the number of parameters) one cannot afford to "loose" cases to the test set
- Cases are used both to build and to test the models
- There will be some "optimism" in the performance of the model
- Types: cross-validation and bootstrap

Cross-validation

- Several training and test set pairs are created
- Results are pooled from all test sets to estimate performance of the "full" model (the one built using all cases)
- Each case is used once in evaluation

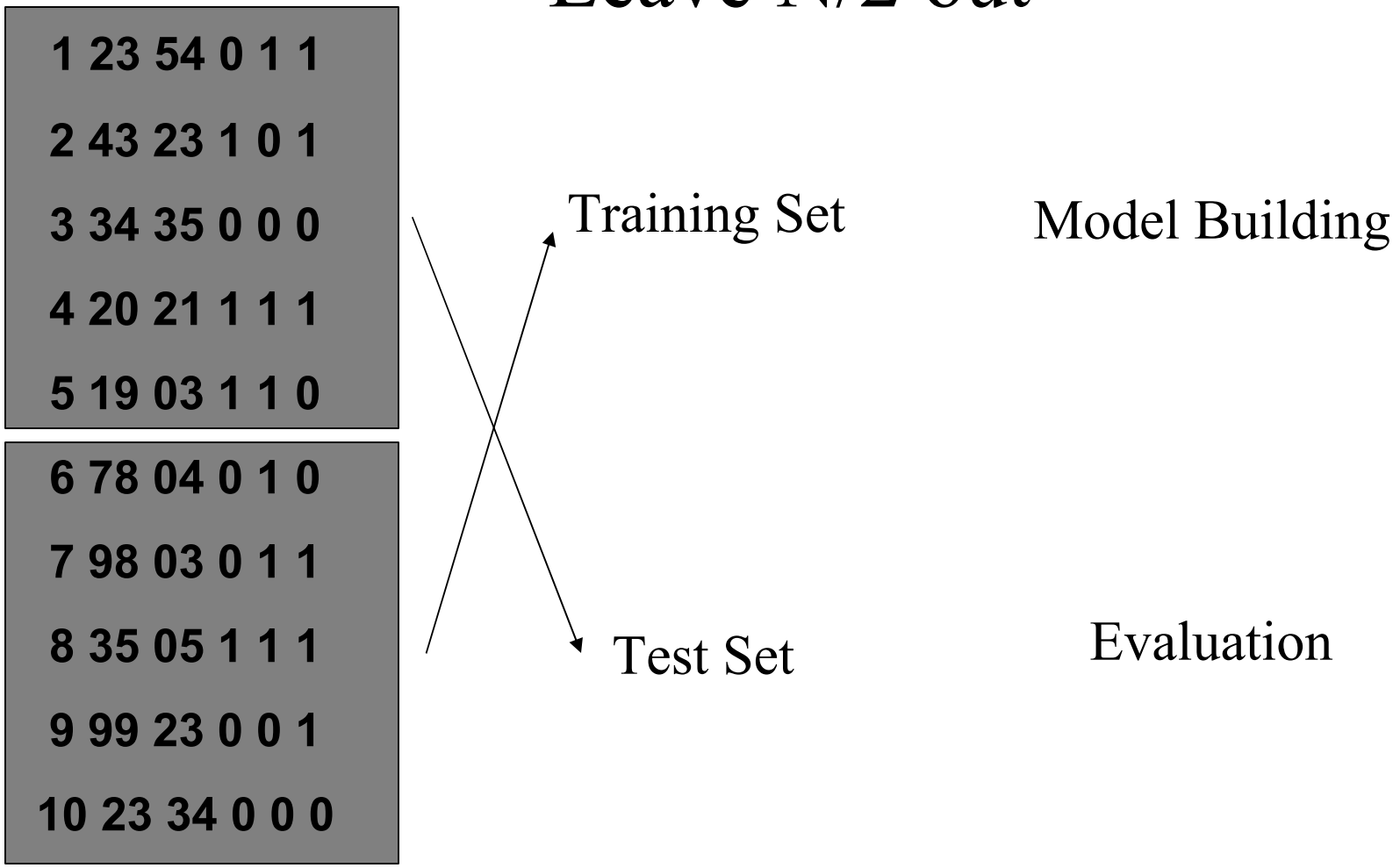
Types

- "Leave-*n*-out"
- Jackknife ("Leave-1-out")

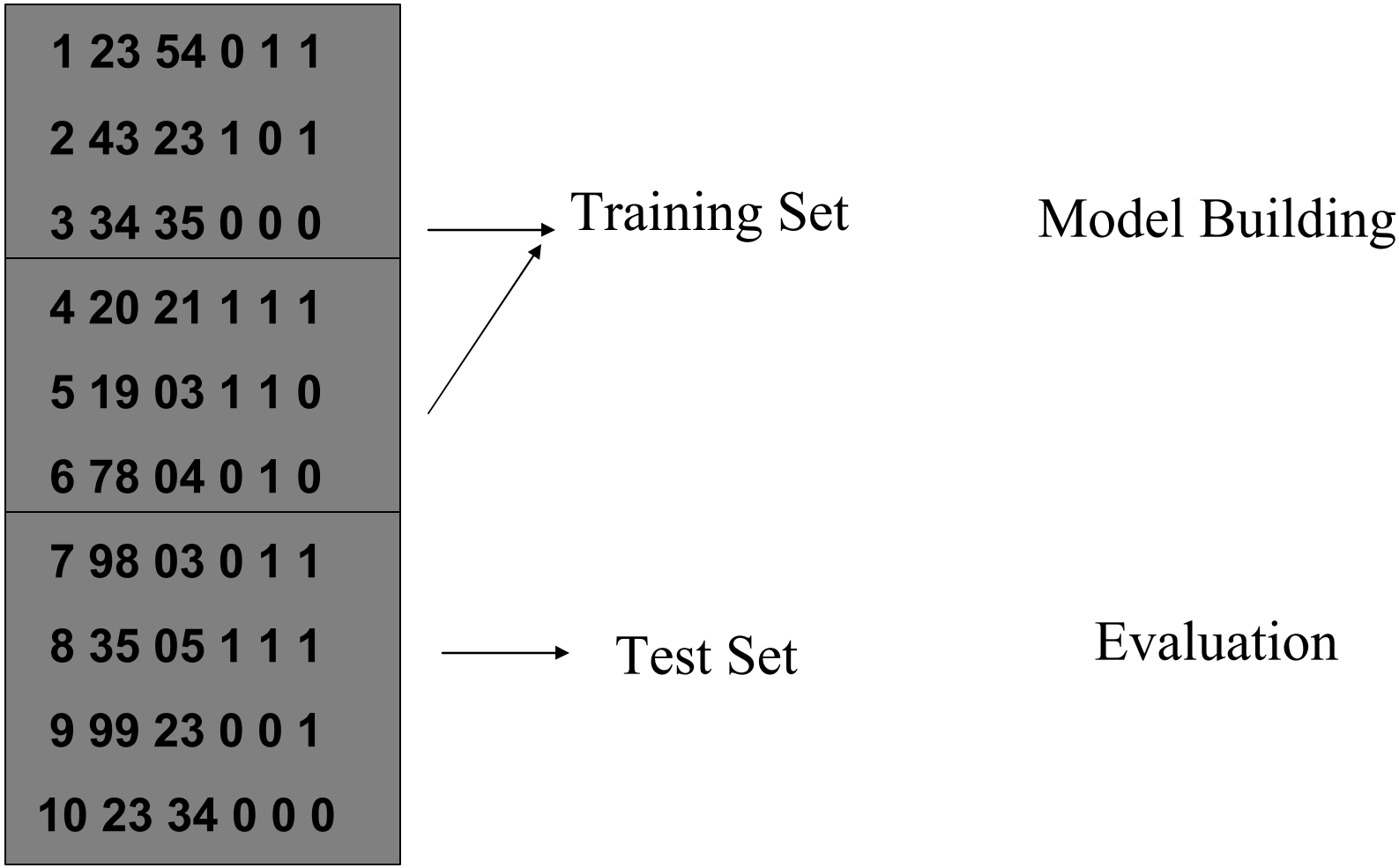
Cross-validation
Leave N/2 out

1	23	54	0	1	1	
2	43	23	1	0	1	
3	34	35	0	0	0	→ Training Set — Model Building
4	20	21	1	1	1	
5	19	03	1	1	0	
6	78	04	0	1	0	
7	98	03	0	1	1	
8	35	05	1	1	1	→ Test Set — Evaluation
9	99	23	0	0	1	
10	23	34	0	0	0	

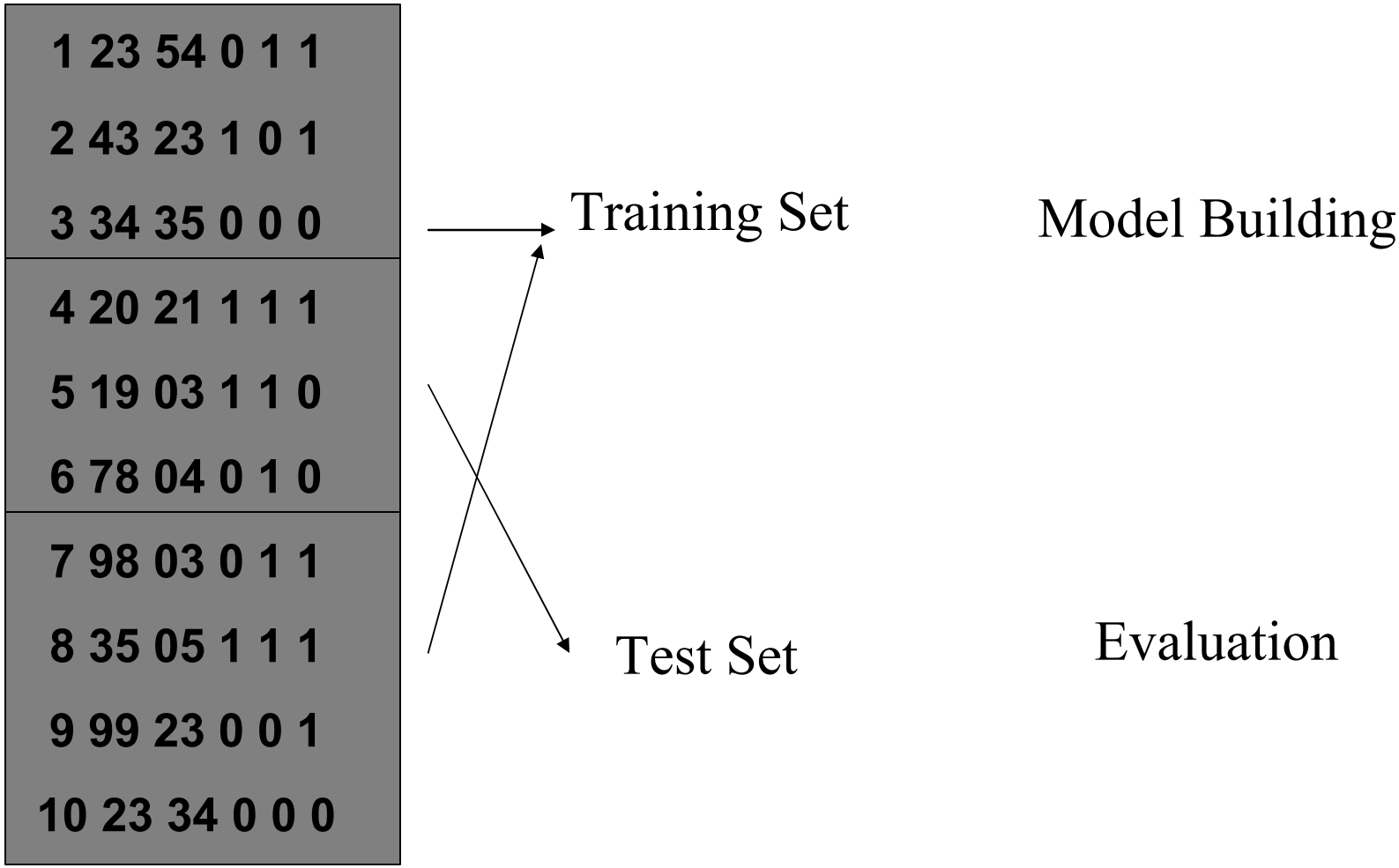
Cross-validation
Leave N/2 out



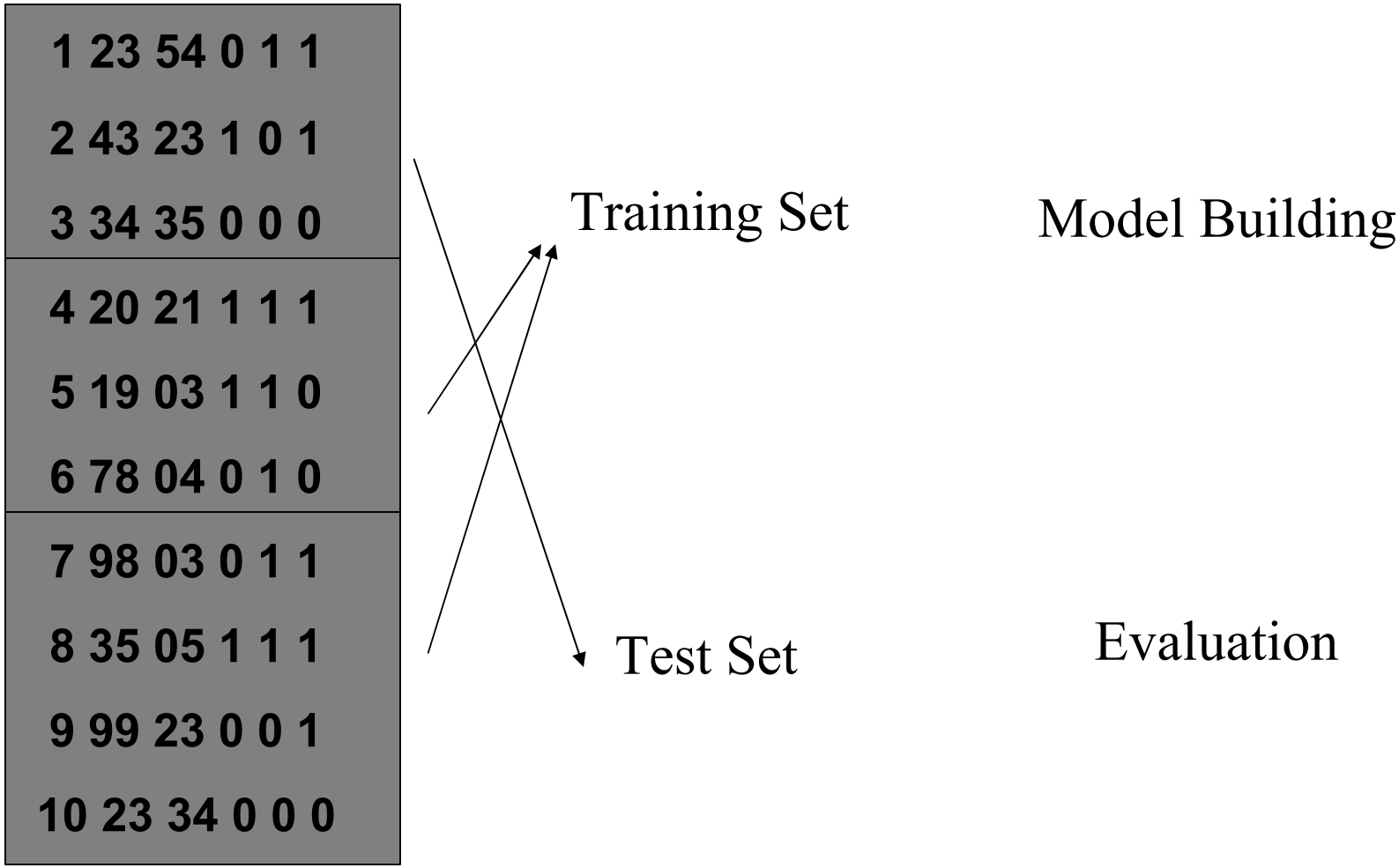
Leave-N/3-out


1 23 54 0 1 1
2 43 23 1 0 1
3 34 35 0 0 0
4 20 21 1 1 1
5 19 03 1 1 0
6 78 04 0 1 0
7 98 03 0 1 1
8 35 05 1 1 1
9 99 23 0 0 1
10 23 34 0 0 0
Training Set
Model Building
Test Set
Evaluation

Leave-N/3-out


1 23 54 0 1 1
2 43 23 1 0 1
3 34 35 0 0 0
4 20 21 1 1 1
5 19 03 1 1 0
6 78 04 0 1 0
7 98 03 0 1 1
8 35 05 1 1 1
9 99 23 0 0 1
10 23 34 0 0 0
Training Set
Test Set
Model Building
Evaluation

Leave-N/3-out


1 23 54 0 1 1
2 43 23 1 0 1
3 34 35 0 0 0
4 20 21 1 1 1
5 19 03 1 1 0
6 78 04 0 1 0
7 98 03 0 1 1
8 35 05 1 1 1
9 99 23 0 0 1
10 23 34 0 0 0
Training Set
Test Set
Model Building
Evaluation

Bootstrap for Validation

- Difference with cross-validation
  - Each case can be used more than once in the evaluation
- Bootstrap is not just used for validation
- It can be used to estimate other quantities such as confidence intervals, median, etc.

Bootstrap Motivation

- Sometimes it is not possible to collect many samples from a population
- Sometimes it is not correct to assume a certain distribution for the population

- Goal: Assess sampling variation and use the measurement to assess population

Bootstrap

- Efron (Stanford biostats) late 80's
  - "Pulling oneself up by one's bootstraps"
- Nonparametric approach to statistical inference
- Uses *computation* instead of traditional distributional assumptions and asymptotic results
- Can be used to estimate non-parametrically standard errors, confidence intervals, and other statistics

Bootstrap

- General idea is to reuse cases in the sample to artificially create samples (from the sample).
- Then measure properties of a statistic in the artificially created (“bootstrap”) samples

Example

- Adapted from Fox (1997) “Applied Regression Analysis”
- Goal: Estimate mean difference between Male and Female finding X
- Four pairs of observations are available:

Observ.	Male	Female	Differ. Y
1	24	18	6
2	14	17	-3
3	40	35	5
4	44	41	3

Mean Difference

- Sample mean $\overline{Y}$ is (6-3+5+3)/4 = 2.75
- If Y were normally distributed, 95% CI

$$\mu = \overline{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- But we do not know σ

Estimates

- Estimate of σ is $S = \sqrt{\frac{\sum \left(Y_i - \overline{Y}\right)^2}{(n-1)}}$
- Estimate of standard error is $S\hat{E}\left(\overline{Y}\right) = \frac{S}{\sqrt{n}}$
- Assuming population is normally distributed, we can use t-distribution as

$$\mu = \overline{Y} \pm t_{n-1,0.025} \frac{S}{\sqrt{n}}$$

Confidence Interval

$$\mu = \overline{Y} \pm t_{n-1,0.025} \frac{S}{\sqrt{n}}$$

$$\mu = 2.75 \pm 4.30\ (2.015) = 2.75 \pm 8.66$$

$$-5.91 < \mu < 11.41$$

Sample with Replacement

- Use distribution Y* of bootstrap samples to estimate distribution Y in population

Sample	Y_1*	Y_2*	Y_3*	Y_4*	Y *
1	6	6	6	6	6.00
2	6	6	6	-3	3.75
3	6	6	6	5	5.75
..					
100	-3	5	6	3	2.75
101	-3	-3	-3	-3	-3
…					
255	-3	3	3	5	3.5
256	3	3	3	3	3.00

Sampling with replacement

- Expected fraction of data points that will make it into a bootstrap sample is

$$1 - e^{-1} = \mathbf{0.632}$$

P(chosen) = 1/n; P(not chosen) = 1 – 1/n

P(not chosen n times) = $(1 - 1/n)^n$

If n is large, this approaches $e^{-1} = 0.368$

Calculating the CI

- Mean of 256 bootstrap means is 2.75, but SE is

$$SE^*(\bar{Y}^*) = \sqrt{\frac{\sum_{b=1}^{n^n} \left(\bar{Y}_b^* - \bar{Y}\right)^2}{n^n}} = 1.745$$

(Other estimate for SE was 2.015)

- One can assume normality and use new SE
- For 95% CI, one can look up sorted table and find 2.5^{th} and 97.5^{th} percentiles directly

Procedure

- 1. Specify data-collection scheme that results in observed sample

  Collect(population) -> sample

- 2. Use sample as if it were population (with replacement)

  Collect(sample) -> bootstrap sample1

  bootstrap sample 2

  etc…

Cont.

- 3. For each bootstrap sample, calculate the estimate you are looking for
- 4. Use the distribution of the bootstrap estimates to estimate the properties of the sample

Bootstrapping Regression

Observed estimate is usually the coefficient(s)

- (at least) 2 ways of doing this

- Resample observations (usual) and re-regress (X will vary)
- Resample residuals (X are fixed, Y*=Y+E* is new dependent variable, re-regress X fixed)

The population is to the sample

as

the sample is to the bootstrap samples

In practice (as opposed to previous example), not all bootstrap samples are selected, as n^n may be high

Usually the size of the bootstrap sample is the same as the size of the original sample

Bootstrap Validation

- Simple:
  - build models on bootstrap samples (training sets) and evaluate them in the full sample (test sets)
  - Average the test set indices
- Enhanced:
  - Use bootstrap to calculate optimism (index from bootstrap sample (training set) minus the index from the original sample (test set))
  - Subtract optimism from the index of the model built on the original sample to come up with a bias-corrected index

Simple Bootstrap Example

1. One model with original sample is built
2. 100 bootstrap samples serve as training sets for 100 logistic regression models
3. The area under the ROC curve (c-index) is calculated for each of the 100 models and averaged (e.g., 0.75). This is assumed to be the c-index for the model built with original samples (from step 1)

Enhanced Bootstrap Example

1. One model with original sample is built. Calculate C-index for training set. (e.g., 0.80)
2. 100 bootstrap samples serve as training sets for 100 logistic regression models
3. For each model in step 2, calculate the difference between c-index on bootstrap sample (training set) and original sample (test set).
4. Average all differences from step 3. This is the "optimism". (e.g., 0.20)
5. Subtract "optimism" from c-index obtained in step 1. This is the bias-corrected (or overfitting-corrected) c-index. (e.g., 0.6)

0.632 method

- Bias-corrected estimate does not use average "optimism"
- Error is

  $[0.368\ \alpha + 0.632 \times \varepsilon]$, where ε is weighted average of errors on observations omitted from bootstrap samples; α is training set error.
- So more weight on test observations (previously unseen)

0.632 example

- Apparent accuracy is 0.80
- Weighted average of error in non-used samples is 0.30
- Error is
- [0.368 (0.2)] x [0.632(0.30)] = 0.2632

0.632 example

- Breiman et al (1984) example
  - Assume no relationship between independent and dependent variables
  - Binary outcome
  - One-nearest neighbor will have error = 0
  - 0.632 bootstrap will give an error of .632 x 0.5 = 0.316
  - What should have been the true error?

0.632+

- Consider the error if independent and dependent variables were not associated

  γ is the no-information error rate, estimated by evaluating the prediction model on all possible combinations of targets y_i and predictors x_i

0.632+

- Relative overfitting rate is
- $R = (\varepsilon - \alpha)/(\gamma - \alpha)$, where ε is weighted average of errors on observations omitted from bootstrap samples; α is the training data error
- $\text{Error}^{0.632+} = (1 - w)\,\alpha + w\,\varepsilon$, where
- $w = 0.632/(1 - 0.368\,R)$

0.632+

- For the same example, one-nearest neighbor model in data for which independent and dependent variables were not associated
- $w = R = 1$, since
- $R = (\varepsilon - \alpha)/(\gamma - \alpha) = (0.5 - 0)/(0.5 - 0)$
- $\text{Error}^{0.632+} = \gamma = 0.5$

Bootstrap for predictive models

- Used in other classification methods (neural networks, classification trees, etc.)
- Usually useful when sample size is small and no distribution assumptions can be made
- Same principles apply

Summary

- Always more convincing to test performance in previously unseen cases
- Several ways of doing this: split sample, cross-validation, bootstrap

- General indices not very informative
- Combination of indices describing discrimination and calibration are more informative
- Hard to conclude that one system is better than another one in terms of classification performance alone: explanation, variable selection, and acceptability by clinicians are key