Experimentation

Jaime Arguello
INLS 613: Text Data Mining
jarguell@email.unc.edu

Outline

Cross-Validation

Significance tests

Training and Testing

Training and Testing

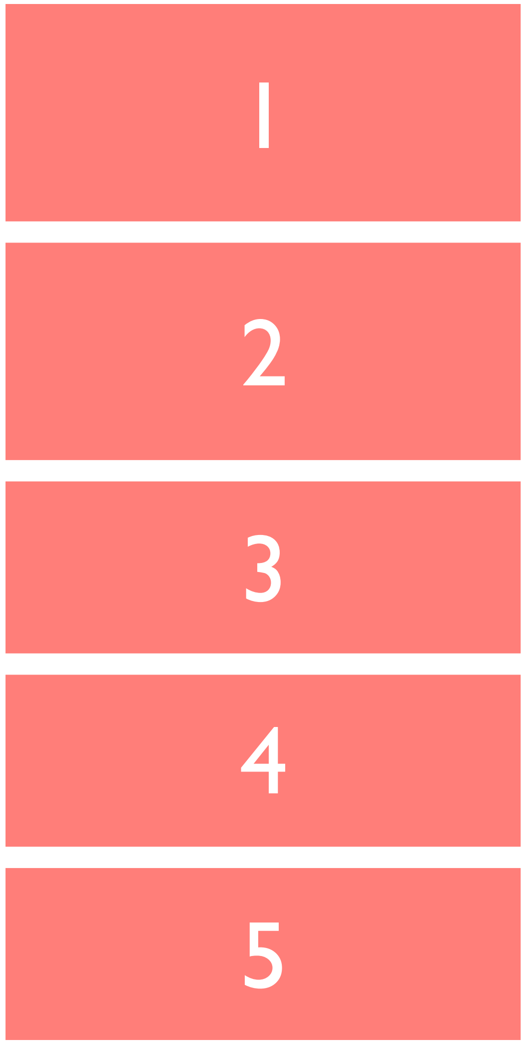
- Split the data into two sets.
- Find the parameter values that maximizes performance on the training set.
- Evaluate the system with that parameter value on the test set.

F = 0.50

Cross-Validation

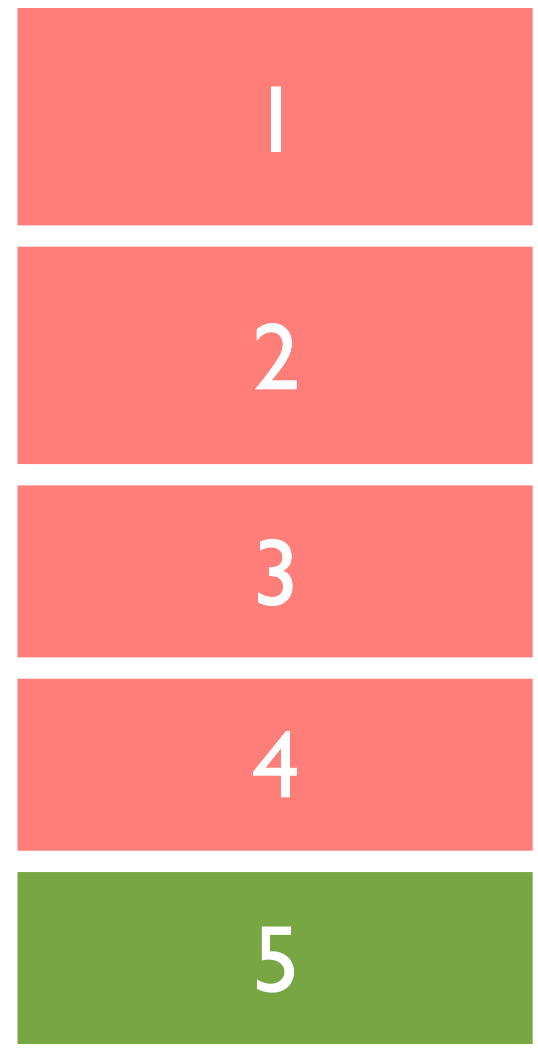
Cross-Validation

- Split the data into $N = 5$ folds



Cross-Validation

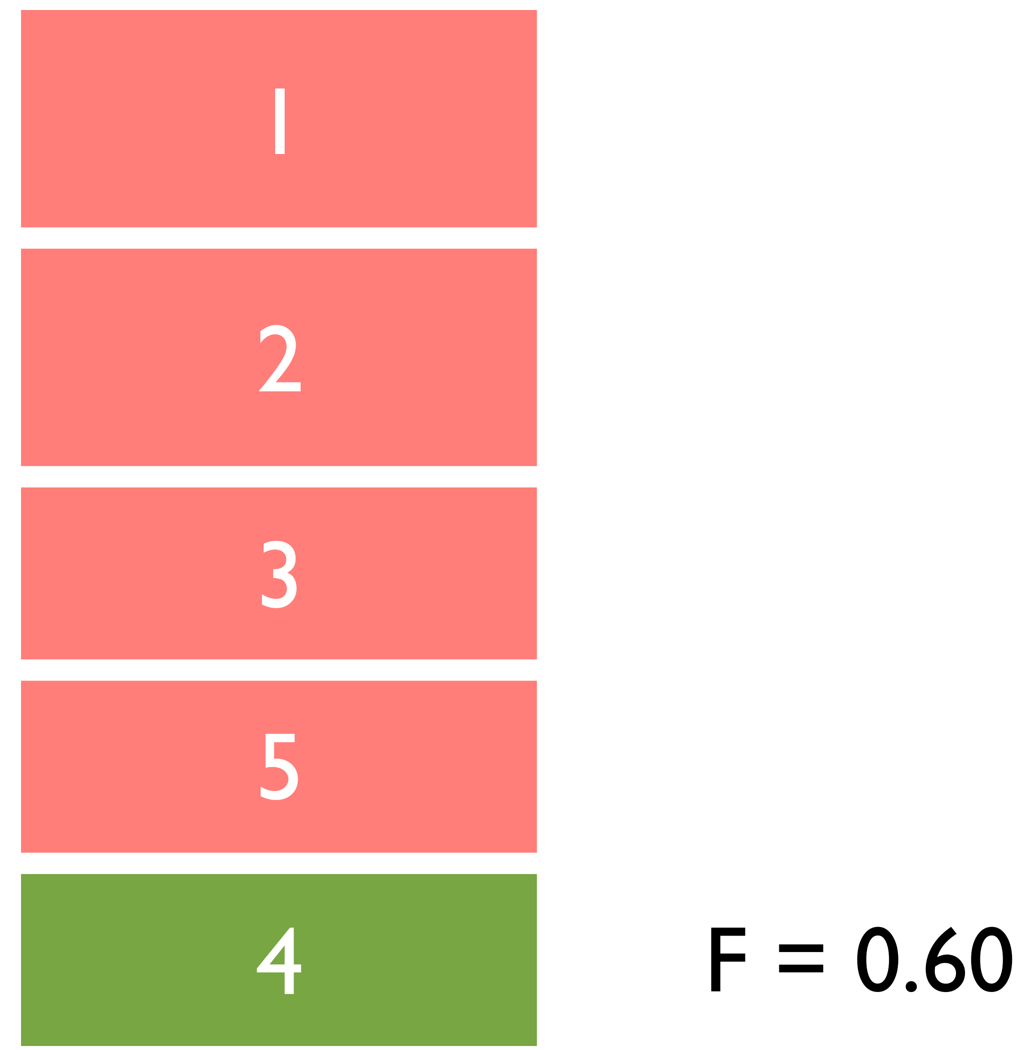
- For each fold:
- Train a model on the union of the other folds
- Test on the holdout fold



F = 0.50

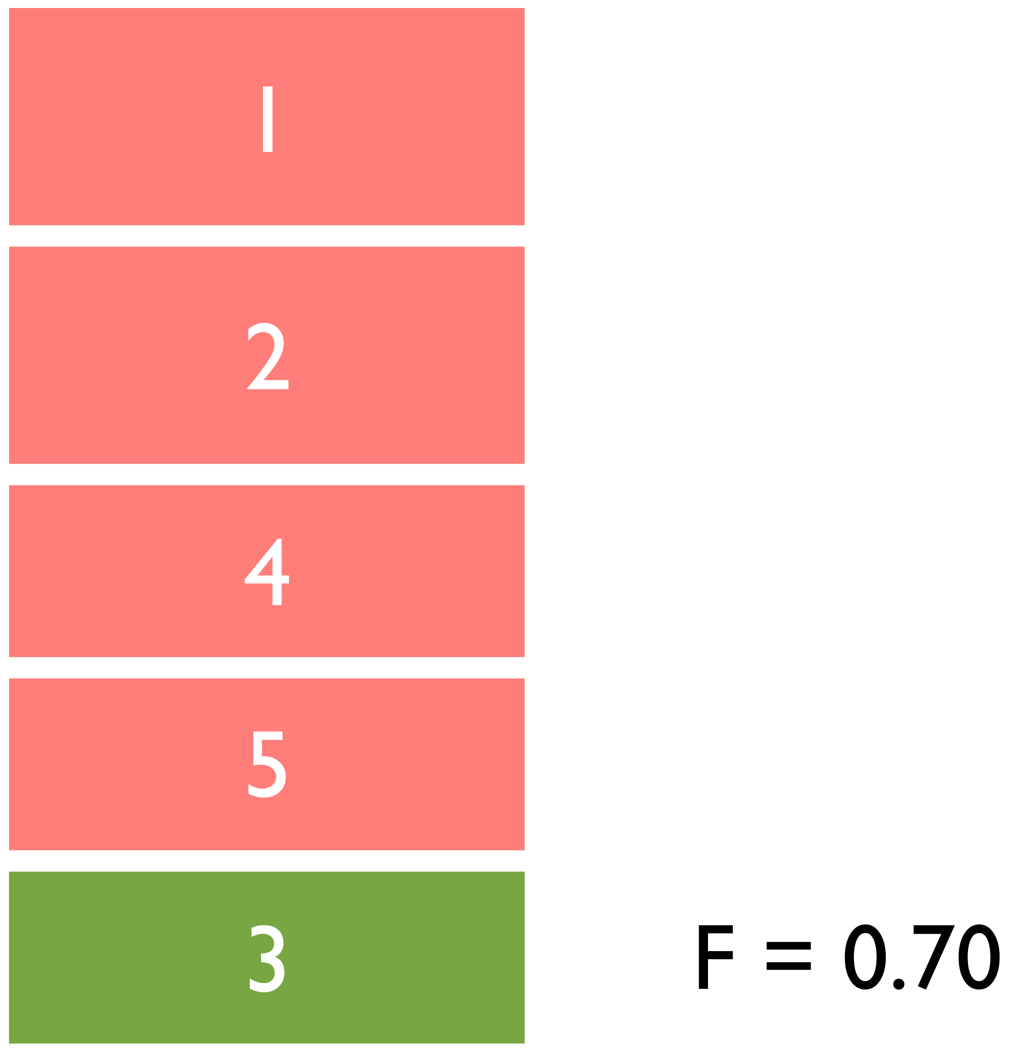
Cross-Validation

- For each fold:
- Train a model on the union of the other folds
- Test on the holdout fold



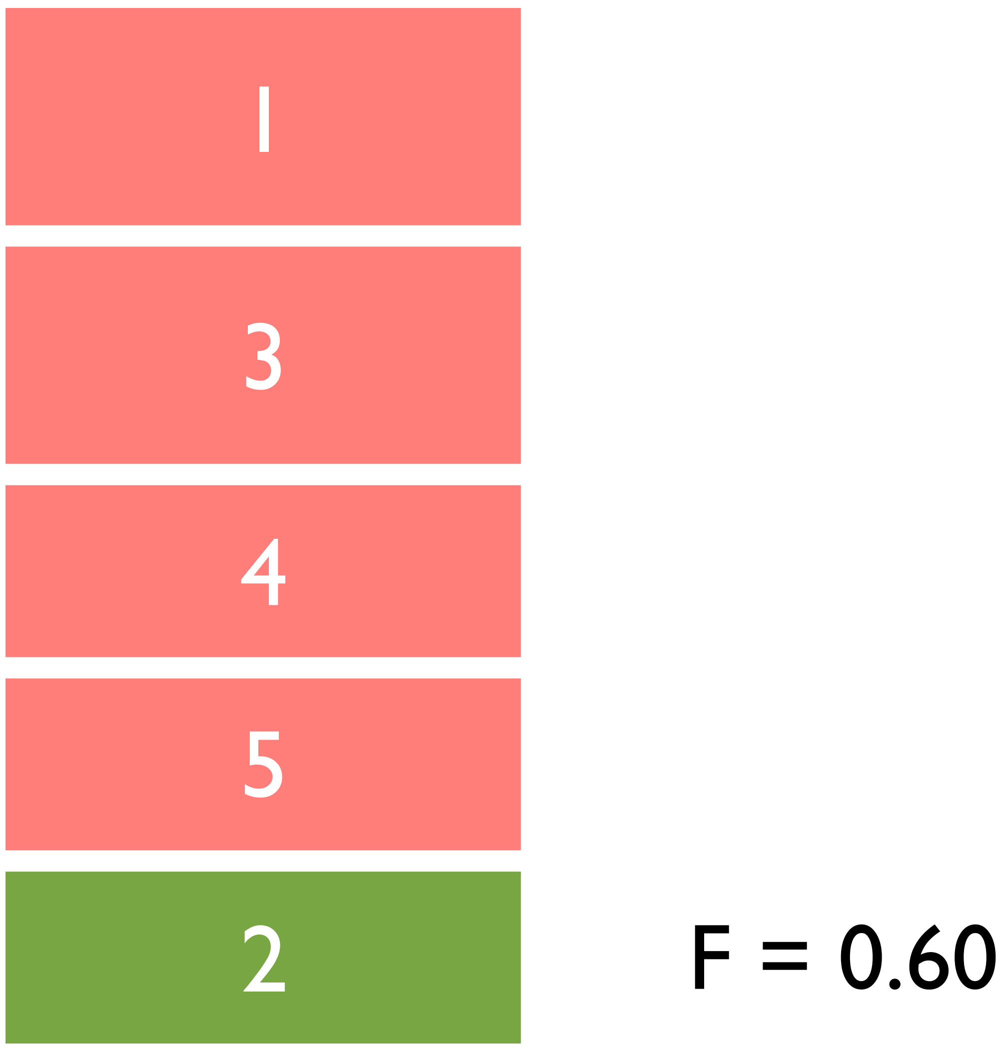
Cross-Validation

- For each fold:
- Train a model on the union of the other folds
- Test on the holdout fold



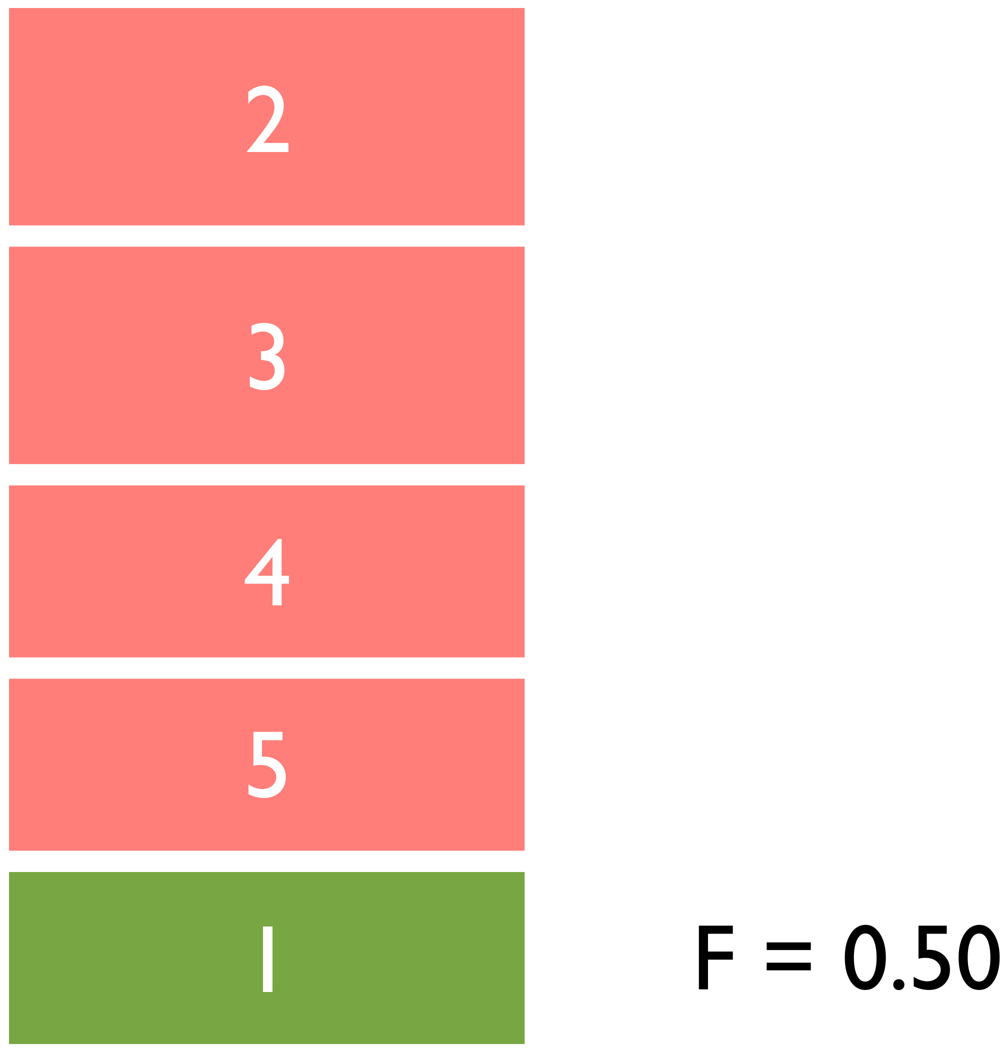
Cross-Validation

- For each fold:
- Train a model on the union of the other folds
- Test on the holdout fold



Cross-Validation

- For each fold:
- Train a model on the union of the other folds
- Test on the holdout fold



Cross-Validation

- For each fold:
- Train a model on the union of the other folds
- Test on the holdout fold

Fold	Score
1	F = 0.50
2	F = 0.60
3	F = 0.70
4	F = 0.60
5	F = 0.50
Average	**F = 0.58**

Cross-Validation

- Average the performance across held-out folds

Fold	F
1	F = 0.50
2	F = 0.60
3	F = 0.70
4	F = 0.60
5	F = 0.50
Average	**F = 0.58**

What should we set N to?

Outline

Cross-Validation

Significance tests

Comparing Systems

- Train and test both systems using 10-fold cross validation
- Use the same folds for both systems
- Compare the difference in average performance across held-out folds

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	0.07

Significance Tests
motivation

- Why would it be risky to conclude that System B is better System A?
- Put differently, what is it that we're trying to achieve?

Significance Tests
motivation

- In theory: that the average performance of System B is greater than the average performance of System A for all data.
- However, we don't have all data. We have a sample
- And, this sample may favor one system vs. the other!

Significance Tests
definition

- A significance test is a statistical tool that allows us to determine whether a difference in performance reflects a true pattern or just random chance

Significance Tests
ingredients

- Test statistic: a measure used to judge the two systems (e.g., the difference between their average F-measure)
- Null hypothesis: no “true” difference between the two systems
- P-value: take the value of the observed test statistic and compute the probability of observing a value that large (or larger) <u>under the null hypothesis</u>

Significance Tests
ingredients

- If the p-value is large, we cannot reject the null hypothesis
- That is, we cannot claim that one system is better than the other
- If the p-value is small (p<0.05), we can reject the null hypothesis
- That is, the observed test statistic is not due to random chance

Comparing Systems

- P-value: the probability of observing a difference **equal to or greater than** 0.07 under the null hypothesis (i.e., the systems are actually equally good).

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	0.07

Fisher's Randomization Test

procedure

- **Inputs:** counter = 0, N = 100,000
- Repeat N times:

  Step 1: for each fold, flip a coin and if it lands 'heads', flip the result between System A and B

  Step 2: see whether the test statistic is equal to or greater than the one observed and, if so, increment counter
- **Output:** counter / N

Fisher's Randomization Test

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	0.07

Fisher's Randomization Test

Fold	System A	System B
1	**0.5**	**0.2**
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	**0.9**	**0.8**
7	0.3	0.1
8	0.1	0.2
9	**0.5**	**0**
10	0.9	0.8
Average	0.5	0.39
	Difference	-0.11

at least 0.07?

iteration = 1 counter = 0

Fisher's Randomization Test

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	**0.1**	**0.1**
4	0.4	0.4
5	**1**	**1**
6	0.8	0.9
7	**0.1**	**0.3**
8	**0.2**	**0.1**
9	0	0.5
10	**0.08**	**0.9**
Average	0.318	0.5
	Difference	0.182

at least 0.07?

iteration = 2   counter = 1

Fisher's Randomization Test

Fold	System A	System B
1	**0.5**	**0.2**
2	0.3	0.3
3	**0.1**	**0.1**
4	**0.4**	**0.4**
5	1	1
6	**0.9**	**0.8**
7	0.3	0.1
8	0.1	0.2
9	**0.5**	**0**
10	0.9	0.8
Average	0.5	0.39
	Difference	-0.11

at least 0.07?

iteration = 100,000 counter = 25,678

Fisher’s Randomization Test

procedure

- **Inputs:** counter = 0, N = 100,000
- Repeat N times:

  Step 1: for each query, flip a coin and if it lands ‘heads’, flip the result between System A and B

  Step 2: see whether the test statistic is equal to or greater than the one observed and, if so, increment counter

- **Output:** counter / N = (25,678/100,00) = 0.25678

Fisher's Randomization Test

- Under the null hypothesis, the probability of observing a value of the test statistic of 0.07 or greater is about 0.26.
- Because $p > 0.05$, we cannot confidently say that the value of the test statistic is <u>not</u> due to random chance.
- A difference between the average F-measure values of 0.07 is not significant

Fisher's Randomization Test

procedure

- **Inputs:** counter = 0, N = 100,000
- Repeat N times:

  Step 1: for each query, flip a coin and if it lands 'heads', flip the result between System A and B

  Step 2: see whether the test statistic is equal to or greater than the one observed and, if so, increment counter
- **Output:** counter / N = (25,678/100,00) = 0.25678

This is a one-tailed test (B > A).
How can we modify it to be a two-tailed test (B != A)

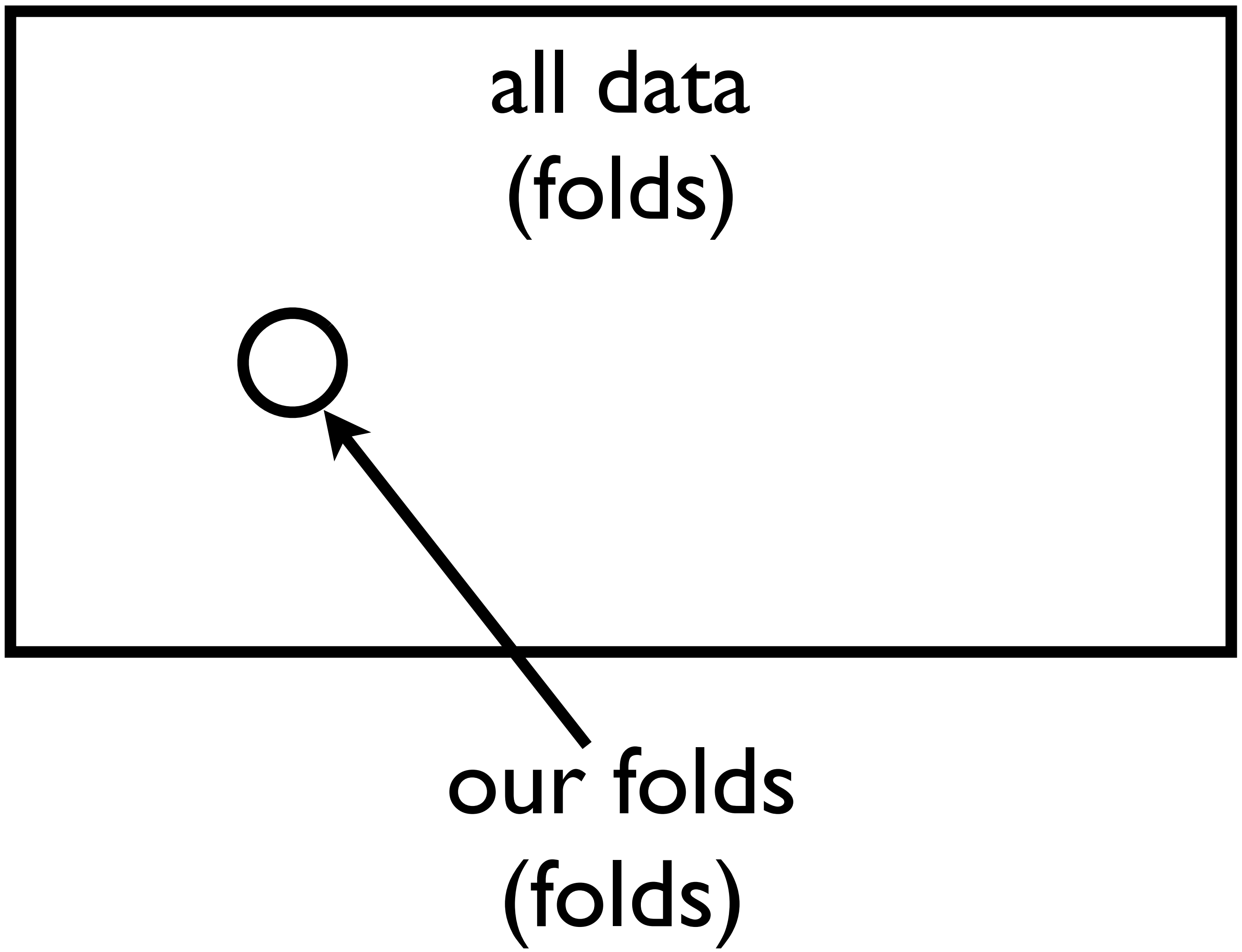
Fisher's Randomization Test
procedure

- P-value: the probability of observing a difference *in the absolute value* **equal to or greater than** 0.07 under the null hypothesis (i.e., the systems are actually equal).

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	$\boxed{0.07}$

Bootstrap-Shift Test
motivation

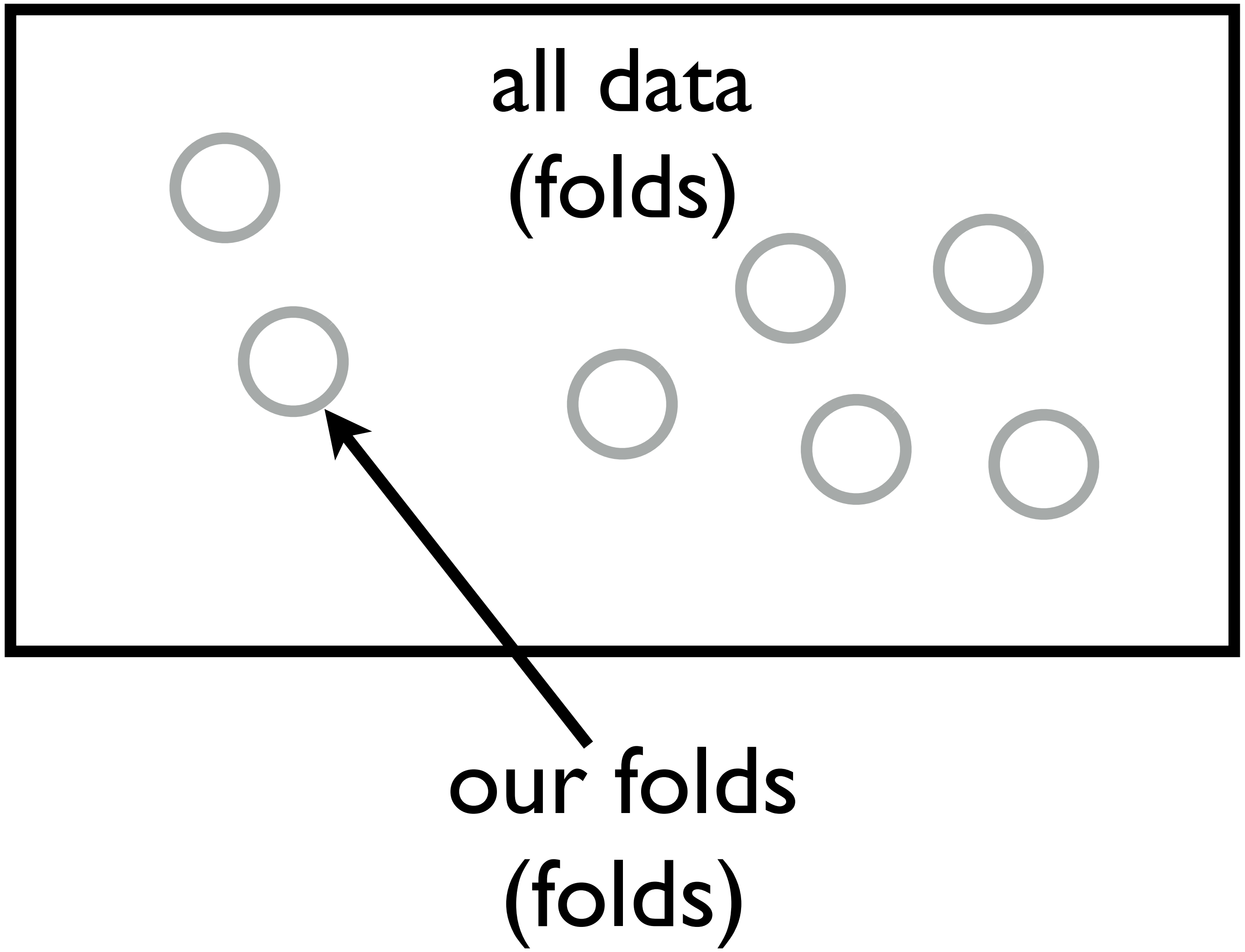
- Our sample is a representative sample of all data



Bootstrap-Shift Test

motivation

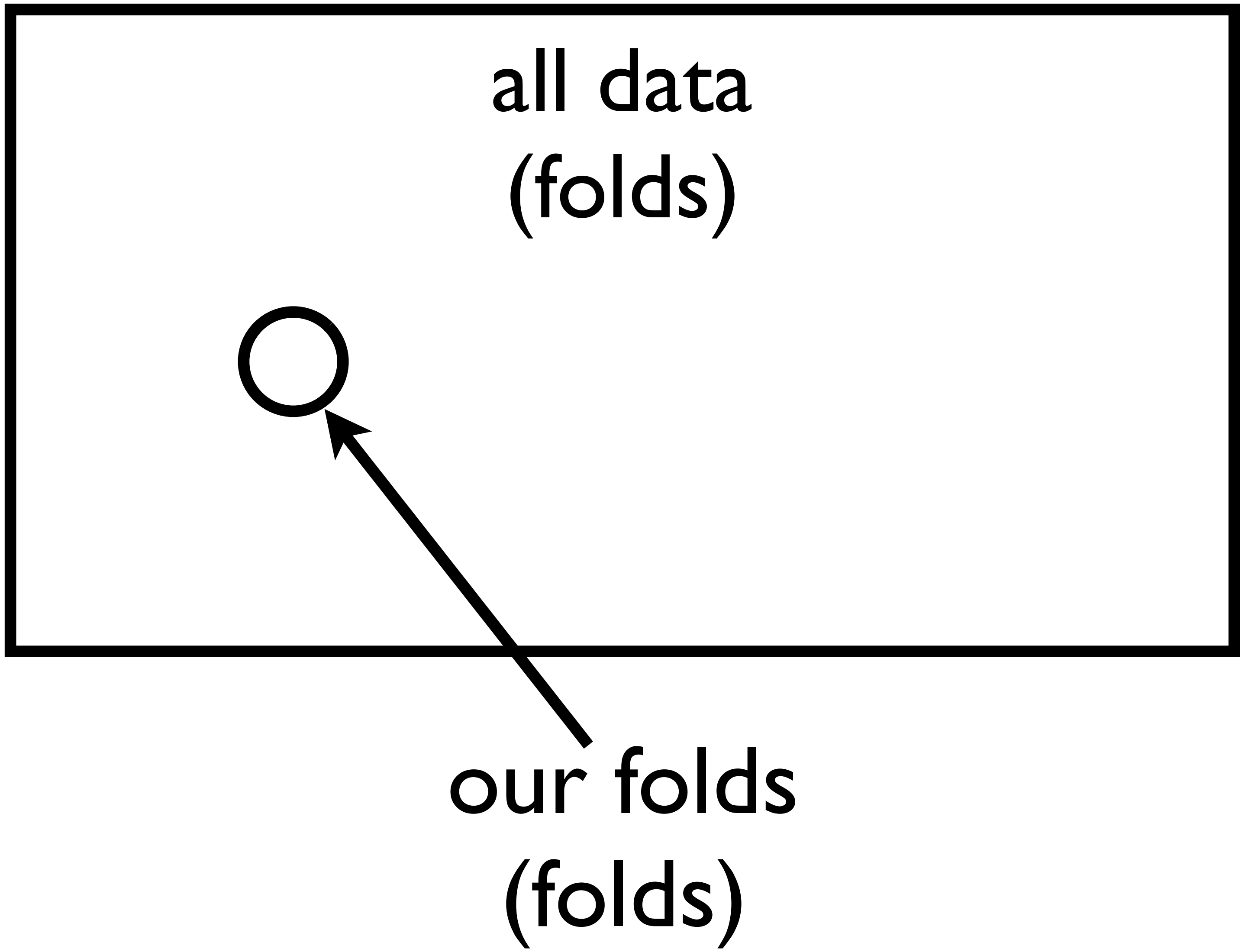
- Our sample is a representative sample of all data



Bootstrap-Shift Test

motivation

- If we sample (with replacement) from our sample, we can generate a new representative sample of all data



Bootstrap-Shift Test

procedure

- **Inputs:** Array T = {}, N = 100,000
- Repeat N times:

  Step 1: sample 10 folds (with replacement) from our set of 10 folds (called a subsample)

  Step 2: compute test statistic associated with new sample and add to T

- **Step 3:** compute <u>average</u> of numbers in T
- **Step 4:** reduce every number in T by <u>average</u>
- **Output:** % of numbers in T greater than or equal to the observed test statistic

Bootstrap-Shift Test

procedure

- **Inputs:** Array T = {}, N = 100,000
- Repeat N times:

  Step 1: sample 10 folds (with replacement) from our set of 10 folds (called a subsample)

  Step 2: compute test statistic associated with new sample and add to T

- **Step 3:** compute average of numbers in T
- **Step 4:** reduce every number in T by average
- **Output:** % of numbers in T greater than or equal to the observed test statistic

Bootstrap-Shift Test

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	0.07

Bootstrap-Shift Test

Fold	System A	System B	sample
1	0.2	0.5	**0**
2	0.3	0.3	**1**
3	0.1	0.1	**2**
4	0.4	0.4	**2**
5	1	1	**0**
6	0.8	0.9	**1**
7	0.3	0.1	**1**
8	0.1	0.2	**1**
9	0	0.5	**2**
10	0.9	0.8	**0**

iteration = 1

Bootstrap-Shift Test

Fold	System A	System B
2	0.3	0.3
3	0.1	0.1
3	0.1	0.1
4	0.4	0.4
4	0.4	0.4
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
9	0	0.5
Average	0.25	0.35
	Difference	**0.1**

$T = \{\mathbf{0.10}\}$

iteration = 1

Bootstrap-Shift Test

Fold	System A	System B	**sample**
1	0.2	0.5	**0**
2	0.3	0.3	**0**
3	0.1	0.1	**3**
4	0.4	0.4	**2**
5	1	1	**0**
6	0.8	0.9	**1**
7	0.3	0.1	**1**
8	0.1	0.2	**1**
9	0	0.5	**1**
10	0.9	0.8	**1**

$T = \{\mathbf{0.10}\}$

iteration = 2

Bootstrap-Shift Test

Fold	System A	System B
3	0.1	0.1
3	0.1	0.1
3	0.1	0.1
4	0.4	0.4
4	0.4	0.4
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.32	0.36
	Difference	**0.04**

T = {**0.10**, **0.04**}

iteration = 2

Bootstrap-Shift Test

Fold	System A	System B
1	0.2	0.5
1	0.2	0.5
4	0.4	0.4
4	0.4	0.4
4	0.4	0.4
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
8	0.1	0.2
10	0.9	0.8
Average	0.38	0.44
	Difference	**0.06**

$T = \{$**0.10**, **0.04**, ..., **0.06**$\}$

iteration = 100,000

Bootstrap-Shift Test

procedure

- **Inputs:** Array T = {}, N = 100,000
- Repeat N times:

  Step 1: sample 10 folds (with replacement) from our set of 10 folds (called a subsample)

  Step 2: compute test statistic associated with new sample and add to T

- **Step 3:** compute <u>average</u> of numbers in T
- **Step 4:** reduce every number in T by <u>average</u>
- **Output:** % of numbers in T' greater than or equal to the observed test statistic

Bootstrap-Shift Test

procedure

- For the purpose of this example, let's assume N = 10.

T = {**0.10**, **0.04**, **0.21**, **0.20**, **0.13**, **0.09**, **0.22**, **0.07**, **0.03**, **0.11**}

Step 3

Average = **0.12**

Step 4

T'= {**-0.02**, **-0.08**, **0.09**, **0.08**, **0.01**, **-0.03**, **0.10**, **-0.05**, -**0.09**, **-0.01**}

Bootstrap-Shift Test
procedure

- **Inputs:** Array T = {}, N = 100,000
- Repeat N times:

  Step 1: sample 10 folds (with replacement) from our set of 10 folds (called a subsample)

  Step 2: compute test statistic associated with new sample and add to T
- **Step 3:** compute <u>average</u> of numbers in T
- **Step 4:** reduce every number in T by <u>average</u>
- **Output:** % of numbers in T' greater than or equal to the observed test statistic

Bootstrap-Shift Test

procedure

- **Output:** (3/10) = **0.30**

T = {**0.10**, **0.04**, **0.21**, **0.20**, **0.13**, **0.09**, **0.22**, **0.07**, **0.03**, **0.11**}

Step 3

Average = **0.12**

Step 4

T'= {**-0.02**, **-0.08**, **0.09**, **0.08**, **0.01**, **-0.03**, **0.10**, **-0.05**, -**0.09**, **-0.01**}

Bootstrap-Shift Test
procedure

- **Output:** (3/10) = **0.30**

T = {**0.10**, **0.04**, **0.21**, **0.20**, **0.13**, **0.09**, **0.22**, **0.07**, **0.03**, **0.11**}

This is a one-tailed test. How can we modify it to be a two-tailed test?

Step 3

Step 4

T'= {**-0.02**, **-0.08**, **0.09**, **0.08**, **0.01**, **-0.03**, **0.10**, **-0.05**, -**0.09**, **-0.01**}

Average = **0.12**

Significance Tests

summary

- Significance tests help us determine whether the outcome of an experiment signals a “true” trend
- The null hypothesis is that the observed outcome is due to random chance (sample bias, error, etc.)
- There are many types of tests
- Parametric tests: assume a particular distribution for the test statistic under the null hypothesis
- Non-parametric tests: make no assumptions about the test statistic distribution under the null hypothesis
- The randomization and bootstrap-shift tests make no assumptions, are robust, and easy to understand