

# Thesis Proposal: Two-Sample Kernel Based Tests

Nelson Ray (joint work with Susan Holmes)

Stanford University

January 30, 2012

# Outline

- Motivation: breast cancer study with heterogeneous data

# Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [1]: leverage regression and classification techniques

# Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [1]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation  $t$ -test

# Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [1]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation  $t$ -test
- Permutation dependence: Stein's method for rates of convergence

# Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [1]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation  $t$ -test
- Permutation dependence: Stein's method for rates of convergence
- Simulations to inform bounds in proof (experimental mathematics)

# Outline

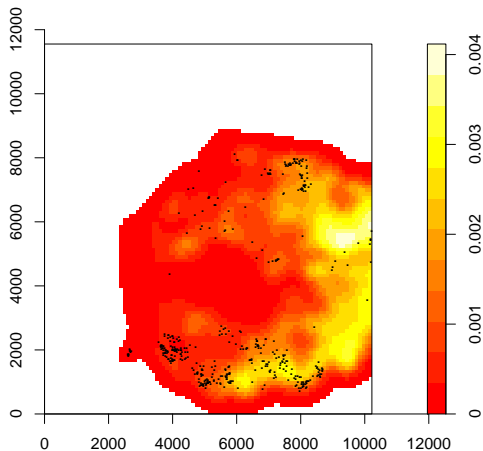
- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [1]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation  $t$ -test
- Permutation dependence: Stein's method for rates of convergence
- Simulations to inform bounds in proof (experimental mathematics)
- Twitter example for text data

# Outline

- Motivation: breast cancer study with heterogeneous data
- Friedman's two-sample test [1]: leverage regression and classification techniques
- Univariate data and linear scoring functions: permutation  $t$ -test
- Permutation dependence: Stein's method for rates of convergence
- Simulations to inform bounds in proof (experimental mathematics)
- Twitter example for text data
- Future work: theory for general case, heterogeneous data and combining kernels



# Breast Cancer Data: Spatial



# Breast Cancer Data: Survival

Pathology no.	Initial Diagnosis Date	Relapse or Disease Free	RDF (R=relapsed; F=DF)	Recurrence Date	Las
98_17969D	1997-08-25	Disease Free	F	Disease Free	
97_24046C8	1997-08-25	Disease Free	F	Disease Free	
98_8501C1	1998-04-03	Disease Free	F	Disease Free	
98_8501A1	1998-04-03	Disease Free	F	Disease Free	
98_9134D4	1998-04-09	Left in-situ BrCa in 1999 (2nd primary cancer, not a metastasis from the right BrCa in 1997)	F	Disease Free	
98_9134B	1998-04-09	Left in-situ BrCa in 1999 (2nd primary cancer, not a metastasis from the right BrCa in 1997)	F	Disease Free	
98_14783B1	1998-06-10	bone, brain, lymph nodes, pericardium, liver metastasis	R	2004-07-30	
98_14783A	1998-06-10	bone, brain, lymph nodes, pericardium, liver metastasis	R	2004-07-30	
98_16169C2	1998-06-24	Disease Free	F	Disease Free	
98_16169A	1998-06-24	Disease Free	F	Disease Free	
98_16169B	1998-06-24	Disease Free	F	Disease Free	
98_16253C1	1998-06-25	Disease Free	F	Disease Free	
60C1	1998-07-10	Disease Free	F	Disease Free	

# Breast Cancer Data: Medical

Pathology no.	Age at time of diagnosis	Gender	SLN tumor status	Diagnosis	ER status	PR status	Her-2 overexpression
98_17969D	68	F	+	Invasive ductal carcinoma (IDC)	-	-	-
97_24046C8	68	F	+	Invasive ductal carcinoma (IDC)	-	-	-
98_8501C1	51	F	+	IDC & DCIS	+	+	?
98_8501A1	51	F	+	IDC & DCIS	+	+	?
98_9134D4	70	F	+	IDC	+	+	n/a
98_9134B	70	F	+	IDC	+	+	n/a
98_14783B1	67	F	+	IDC & DCIS	+	+	+
98_14783A	67	F	+	IDC & DCIS	+	+	+
98_16169C2	79	F	+mic	IDC	+	+	+
98_16169A	79	F	+mic	IDC	+	+	+
98_16169B	79	F	+mic	IDC	+	+	+
98_16253C1	70	F	+mic	IDC & DCIS	+	-	-
60C1	51	F	- (rare keratin+ cells)	IDC & DCIS	+	+	+

# Breast Cancer Study

- How do you deal with the data integration problem?

# Breast Cancer Study

- How do you deal with the data integration problem?
- Kernel methods

# Breast Cancer Study

- How do you deal with the data integration problem?
- Kernel methods
- Are there any differences (spatial, medical) between women who relapse and those who remain disease free?

# Breast Cancer Study

- How do you deal with the data integration problem?
- Kernel methods
- Are there any differences (spatial, medical) between women who relapse and those who remain disease free?
- Two-sample tests

# Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$  from  $p(\mathbf{x})$  and  $\{\mathbf{z}_i\}_1^M$  from  $q(\mathbf{z})$  testing  
 $\mathcal{H}_A: p \neq q$  against  $\mathcal{H}_0: p = q$



# Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$  from  $p(\mathbf{x})$  and  $\{\mathbf{z}_i\}_1^M$  from  $q(\mathbf{z})$  testing

$\mathcal{H}_A: p \neq q$  against  $\mathcal{H}_0: p = q$

- 1 Pool the two samples  $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$ .

# Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$  from  $p(\mathbf{x})$  and  $\{\mathbf{z}_i\}_1^M$  from  $q(\mathbf{z})$  testing

$\mathcal{H}_A: p \neq q$  against  $\mathcal{H}_0: p = q$

- 1 Pool the two samples  $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$ .
- 2 Assign label  $y_i = 1$  to the first group and  $y_i = -1$  to the second group.

# Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$  from  $p(\mathbf{x})$  and  $\{\mathbf{z}_i\}_1^M$  from  $q(\mathbf{z})$  testing

$\mathcal{H}_A: p \neq q$  against  $\mathcal{H}_0: p = q$

- 1 Pool the two samples  $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$ .
- 2 Assign label  $y_i = 1$  to the first group and  $y_i = -1$  to the second group.
- 3 Apply a binary classification learning machine  $f$  to the training data to score the observations  $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$ .

# Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$  from  $p(\mathbf{x})$  and  $\{\mathbf{z}_i\}_1^M$  from  $q(\mathbf{z})$  testing

$\mathcal{H}_A: p \neq q$  against  $\mathcal{H}_0: p = q$

- 1 Pool the two samples  $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$ .
- 2 Assign label  $y_i = 1$  to the first group and  $y_i = -1$  to the second group.
- 3 Apply a binary classification learning machine  $f$  to the training data to score the observations  $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$ .
- 4 Calculate a univariate two-sample test statistic  
 $T = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$ .

# Friedman's Two-Sample Test

$\{\mathbf{x}_i\}_1^N$  from  $p(\mathbf{x})$  and  $\{\mathbf{z}_i\}_1^M$  from  $q(\mathbf{z})$  testing

$\mathcal{H}_A: p \neq q$  against  $\mathcal{H}_0: p = q$

- 1 Pool the two samples  $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$ .
- 2 Assign label  $y_i = 1$  to the first group and  $y_i = -1$  to the second group.
- 3 Apply a binary classification learning machine  $f$  to the training data to score the observations  $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$ .
- 4 Calculate a univariate two-sample test statistic  $T = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$ .
- 5 Determine the permutation null distribution of the above statistic to yield a p-value.

# Permutation T-test Connection

With univariate data and linear scoring functions, Friedman's test reduces to the permutation  $t$ -test.

# Permutation T-test Connection

With univariate data and linear scoring functions, Friedman's test reduces to the permutation  $t$ -test.

With multivariate data, the test is close to Hotelling's  $T^2$ -test.

# Permutation T-test Connection

With univariate data and linear scoring functions, Friedman's test reduces to the permutation  $t$ -test.

With multivariate data, the test is close to Hotelling's  $T^2$ -test.

Strategy: Analyze the simple case (univariate/linear) and attempt to generalize.



## Other Work

- Fisher (1935) [2] proposed distribution free randomization test.

## Other Work

- Fisher (1935) [2] proposed distribution free randomization test.
- Lehmann [3] proved a normal convergence result for the randomization distribution.

# Other Work

- Fisher (1935) [2] proposed distribution free randomization test.
- Lehmann [3] proved a normal convergence result for the randomization distribution.
- Bentkus et al. [4], Shao [5] proved Berry-Esseen bounds for Student's  $t$ -statistic in independent (but not i.i.d.) case.

# Stein's Method and the Randomization Distribution

Let  $\Phi(t)$  denote the standard normal CDF and  $T$  be a random variable that is distributed according to our permutation  $t$  null distribution. Can we get a bound on

$$\sup_{t \in \mathbb{R}} |P(T \leq t) - \Phi(t)|?$$

# Stein's Method and the Randomization Distribution

Let  $\Phi(t)$  denote the standard normal CDF and  $T$  be a random variable that is distributed according to our permutation  $t$  null distribution.

Can we get a bound on

$$\sup_{t \in \mathbb{R}} |P(T \leq t) - \Phi(t)|?$$

We are finishing up a proof using the method of exchangeable pairs where our bound is  $O(N^{-1/4})$ .

# Proof Ideas

Chen et al. [6]:

## Theorem

If  $T, T'$  are mean 0, variance 1 exchangeable random variables satisfying

$$\mathbb{E}[T - T' | T] = \lambda(T - R)$$

for some  $\lambda \in (0, 1)$  and some random variable  $R$ , then

$$\sup_{t \in \mathbb{R}} |P(T \leq t) - \Phi(t)| \leq B + (2\pi)^{-1/4} \sqrt{\frac{\mathbb{E}|T' - T|^3}{\lambda}} + \mathbb{E}|R|,$$

where  $B \leq \frac{\Theta}{2\lambda}$  and  $\Theta = \sqrt{\text{var}(\mathbb{E}[(T' - T)^2 | T])}$ .

# Proof Ideas

- Attempts to follow Stein's [7] proof of the Hoeffding combinatorial central limit theorem

# Proof Ideas

- Attempts to follow Stein's [7] proof of the Hoeffding combinatorial central limit theorem
- General contraction property, or “approximate case,” from Stein et al. [8] and Holmes [9]



# Proof Ideas

- Attempts to follow Stein's [7] proof of the Hoeffding combinatorial central limit theorem
- General contraction property, or “approximate case,” from Stein et al. [8] and Holmes [9]
- Simulation aided proof (Borwein [10]) with efficient  $t$ -statistic updates similar to Diaconis et al. [11]

# Exchangeable Pair

For simplicity, assume  $M = N$ . We have data  $\{u_1, \dots, u_N, u_{N+1}, \dots, u_{2N}\}$ . Take a uniformly random permutation  $\pi$ , and let

$$T = T\left(\{u_{\pi(i)}\}_{i=1}^N, \{u_{\pi(i)}\}_{i=N+1}^{2N}\right).$$

# Exchangeable Pair

For simplicity, assume  $M = N$ . We have data  $\{u_1, \dots, u_N, u_{N+1}, \dots, u_{2N}\}$ . Take a uniformly random permutation  $\pi$ , and let

$$T = T\left(\{u_{\pi(i)}\}_{i=1}^N, \{u_{\pi(i)}\}_{i=N+1}^{2N}\right).$$

Let  $(I, J)$  be a uniformly random transposition between groups: over the  $N^2$  cases where  $1 \leq I \leq N$  and  $N+1 \leq J \leq 2N$ . Then

$$T' = T\left(\{u_{\pi \circ (I, J)(i)}\}_{i=1}^N, \{u_{\pi \circ (I, J)(i)}\}_{i=N+1}^{2N}\right).$$

$T$  and  $T'$  form an exchangeable pair.

# Bound Calculations

$$\begin{aligned} \sup_{t \in \mathbb{R}} |P(T \leq t) - \Phi(t)| &\leq \underbrace{\frac{\sqrt{\text{var}(\mathbb{E}[(T' - T)^2 | T])}}{2\lambda}}_1 \\ &\quad + \underbrace{(2\pi)^{-1/4} \sqrt{\frac{\mathbb{E}|T' - T|^3}{\lambda}}}_{2} \\ &\quad + \underbrace{\mathbb{E} \left| -\frac{1}{\lambda} \mathbb{E}[T - T' | T] + T \right|}_{3} \end{aligned}$$

# Simulation Information

- 1 Draw samples  $\{\mathbf{x}_i\}_1^N$  and  $\{\mathbf{z}_i\}_1^N$ .

# Simulation Information

- 1 Draw samples  $\{\mathbf{x}_i\}_1^N$  and  $\{\mathbf{z}_i\}_1^N$ .
- 2 Pick a permutation  $\pi$  uniformly at random.

# Simulation Information

- 1 Draw samples  $\{\mathbf{x}_i\}_1^N$  and  $\{\mathbf{z}_i\}_1^N$ .
- 2 Pick a permutation  $\pi$  uniformly at random.
- 3 Calculate the two-sample  $t$ -statistic,  $T$ , on the permuted data.

# Simulation Information

- 1 Draw samples  $\{\mathbf{x}_i\}_1^N$  and  $\{\mathbf{z}_i\}_1^N$ .
- 2 Pick a permutation  $\pi$  uniformly at random.
- 3 Calculate the two-sample  $t$ -statistic,  $T$ , on the permuted data.
- 4 Calculate the  $N^2$  values of  $T'$  resulting from all allowable transpositions  $(I, J)$  that swap an  $x$  for a  $z$ .



# Simulation Information

- 1 Draw samples  $\{\mathbf{x}_i\}_1^N$  and  $\{\mathbf{z}_i\}_1^N$ .
- 2 Pick a permutation  $\pi$  uniformly at random.
- 3 Calculate the two-sample  $t$ -statistic,  $T$ , on the permuted data.
- 4 Calculate the  $N^2$  values of  $T'$  resulting from all allowable transpositions  $(I, J)$  that swap an  $x$  for a  $z$ .
- 5 Calculate conditional expectations with respect to  $T$ .

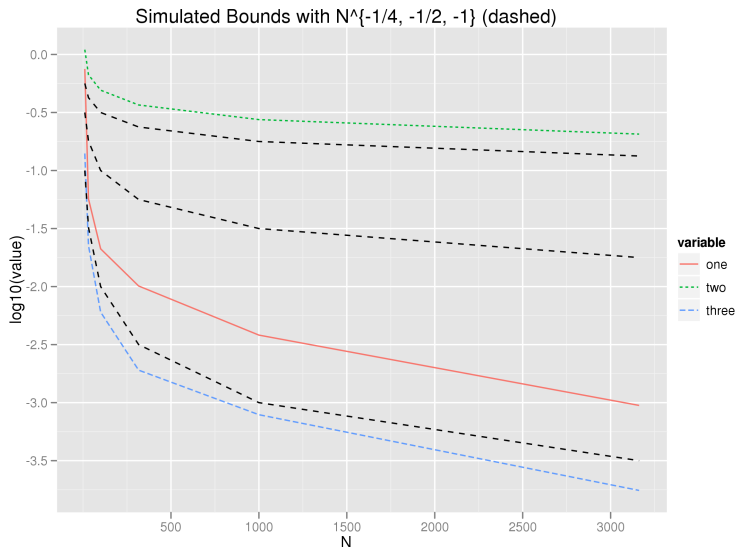
# Simulation Information

- 1 Draw samples  $\{\mathbf{x}_i\}_1^N$  and  $\{\mathbf{z}_i\}_1^N$ .
- 2 Pick a permutation  $\pi$  uniformly at random.
- 3 Calculate the two-sample  $t$ -statistic,  $T$ , on the permuted data.
- 4 Calculate the  $N^2$  values of  $T'$  resulting from all allowable transpositions  $(I, J)$  that swap an  $x$  for a  $z$ .
- 5 Calculate conditional expectations with respect to  $T$ .
- 6 Average over many values of  $T$ , and repeat for a sequence of  $N$ 's.

# Simulated Data

	T	Tprime	N	lambda
1	-1.6646969	-1.4150824	10	0.2000000000
2	-1.6646969	-2.8302749	10	0.2000000000
3	-1.6646969	-1.5975851	10	0.2000000000
4	-1.6646969	-2.1813520	10	0.2000000000
5	-1.6646969	-2.5914846	10	0.2000000000
6	-1.6646969	-1.9817233	10	0.2000000000
...				
88873283	0.2425782	0.3088987	3162	0.0006325111
88873284	0.2425782	0.2740881	3162	0.0006325111
88873285	0.2425782	0.2816923	3162	0.0006325111
88873286	0.2425782	0.2992468	3162	0.0006325111
88873287	0.2425782	0.2931195	3162	0.0006325111
88873288	0.2425782	0.2677967	3162	0.0006325111

# Bounds Comparison



# Twitter Example



**Barack Obama** ✓

@BarackObama Washington, DC  
44th President of the United States  
<http://www.barackobama.com>

+ Follow



Tweets

Favorites Following Followers Lists



**BarackObama** Barack Obama

We need to reward education reforms that are driven not by Washington, but by principals and teachers and parents.  
<http://OFA.BO/6p2EMy>

21 May



**BarackObama** Barack Obama

Speaking today about the United States' policy in the Middle East and North Africa. Watch live: <http://wh.gov/live>  
#MEdspeech

19 May



**BarackObama** Barack Obama

Delivering the commencement address at the United States Coast Guard Academy. Watch live at 11:30am ET:  
[www.wh.gov/live](http://www.wh.gov/live)

18 May



**Sarah Palin** ✓

@SarahPalinUSA Alaska  
Former Governor of Alaska and GOP Vice Presidential Nominee  
<http://www.facebook.com/sarahpalin>

+ Follow



Tweets

Favorites Following Followers Lists



**SarahPalinUSA** Sarah Palin

You betcha!! MT "@AlaskaAces: Alaska Aces are 2011 Kelly Cup Champs w/ 5-3 win over Kalamazoo Wings! Aces win ECHL Championship series 4-1"

21 May



**SarahPalinUSA** Sarah Palin

Yes, they did & we couldn't be any more blessed! RT" @C4Palin: Track Palin and Britta Hanson Married  
<http://bit.ly/jCkT3i> #tcot #palin"

19 May



**SarahPalinUSA** Sarah Palin

I'm jealous! RT"@secupp: At the Wasilla Sportsman's Warehouse w/Joe the Plumber, Colorado Buck, Ken Onion and Sarah's parents. Good people."

19 May

# Twitter Data

Raw:

```
"BarackObama: We need to reward education reforms that are  
driven not by Washington, but by principals and teachers and  
parents. http://OFA.B0/6p2EMy"
```

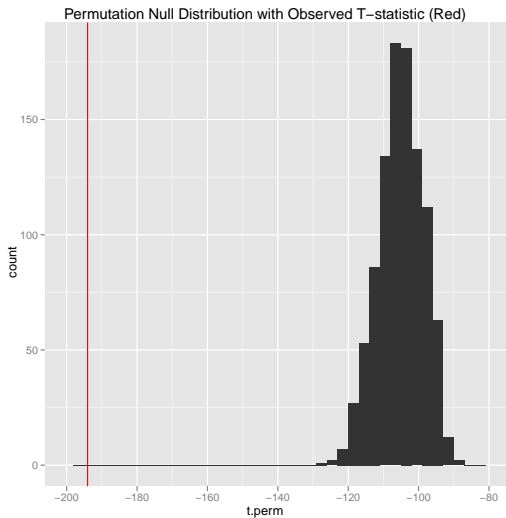
```
"SarahPalinUSA: You betcha!! MT \"@AlaskaAces: Alaska Aces  
are 2011 Kelly Cup Champs w/ 5-3 win over Kalamazoo Wings!  
Aces win ECHL Championship series 4-1\""
```

After pre-processing:

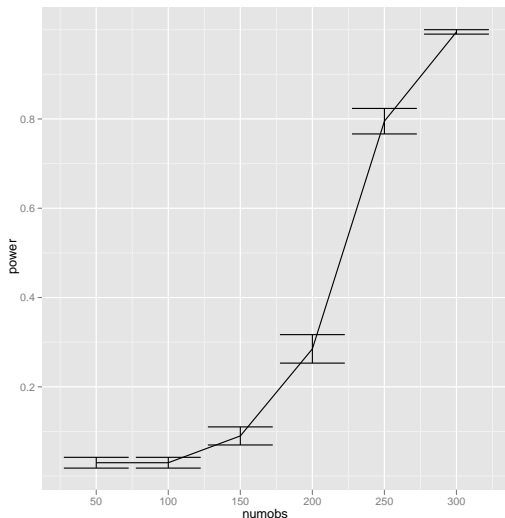
```
"we need to reward education reforms that are driven not by  
washington but by principals and teachers and parents "  
"you betcha mt alaskaaces alaska aces are kelly cup champs  
w win over kalamazoo wings aces win echl championship  
series "
```

# Twitter Example

$p < .001$ :



# Power Simulations at .05 Level





# Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions

# Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's  $T^2$ -test

# Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's  $T^2$ -test
- Explore performance on different types of data, in particular, unstructured data such as images






# Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's  $T^2$ -test
- Explore performance on different types of data, in particular, unstructured data such as images
- Heterogeneous data: optimal combinations of kernels via SDPs, KL divergence






# Future Work

- Generalize theory for higher dimensional settings and/or non-linear scoring functions
- Develop similarities with Hotelling's  $T^2$ -test
- Explore performance on different types of data, in particular, unstructured data such as images
- Heterogeneous data: optimal combinations of kernels via SDPs, KL divergence

# References I

-  J. Friedman, “On Multivariate Goodness-of-Fit and Two-Sample Testing,” *Proceedings of Phystat2003*, <http://www.slac.stanford.edu/econf/C>, vol. 30908, 2003.
-  R. Fisher, “The design of experiments.,” 1935.
-  E. Lehmann, *Elements of large-sample theory*. Springer Verlag, 1999.
-  V. Bentkus and F. Götze, “The berry-esseen bound for student’s statistic,” *The Annals of Probability*, vol. 24, no. 1, pp. 491–503, 1996.
-  Q. Shao, “An explicit berry-esseen bound for students t-statistic via steins method,” *Steins Method and Applications (AD Barbour and LHY Chen eds). Lecture Notes Series, Institute for Mathematical Sciences, NUS*, vol. 5, pp. 143–155, 2005.

## References II

-  L. Chen, L. Goldstein, and Q. Shao, *Normal Approximation by Stein's Method*.  
Springer Verlag, 2010.
-  C. Stein, “Approximate computation of expectations,” *Lecture Notes-Monograph Series*, vol. 7, 1986.
-  C. Stein, P. Diaconis, S. Holmes, and G. Reinert, “Use of exchangeable pairs in the analysis of simulations,” *Lecture Notes-Monograph Series*, pp. 1–26, 2004.
-  S. Holmes, “Stein's method for birth and death chains,” *Lecture Notes-Monograph Series*, pp. 45–67, 2004.
-  J. Borwein and D. Bailey, *Mathematics by Experiment: Plausible reasoning in the 21st century*.  
AK Peters, 2004.

## References III



P. Diaconis and S. Holmes, “Gray codes for randomization procedures,” *Statistics and Computing*, vol. 4, no. 4, pp. 287–302, 1994.