Math 341: Probability
Twenty-third Lecture (12/3/09)

Steven J Miller
Williams College

Steven.J.Miller@williams.edu
http://www.williams.edu/go/math/sjmiller/public_html/341/

Bronfman Science Center
Williams College, December 3, 2009

Summary for the Day

Summary for the day

- More Sum Than Difference Sets:
  - Review.
  - Inputs (Chebyshev's Theorem).
  - Proofs.
- Sabermetrics (baseball math):

More Sums Than Differences: Introduction

Statement

A finite set of integers, $|A|$ its size. Form

- Sumset: $A + A = \{a_i + a_j : a_j, a_j \in A\}$.
- Difference set: $A - A = \{a_i - a_j : a_j, a_j \in A\}$.

Definition

We say A is difference dominated if $|A - A| > |A + A|$, balanced if $|A - A| = |A + A|$ and sum dominated (or an MSTD set) if $|A + A| > |A - A|$.

Clicker Question

Binomial Model

Consider the 2^N subsets of $\{1, 2, \dots, N\}$. As $N \to \infty$, what can you say about the percentage that are MSTD?

1. It tends to 1.
2. It tends to 1/2.
3. It tends to a small positive constant.
4. It tends to 0.

Questions

Expect generic set to be difference dominated:

- addition is commutative, subtraction isn't:
- Generic pair (x, y) gives 1 sum, 2 differences.

Questions

- Do there exist sum-dominated sets?
- If yes, how many?

Examples

Examples

- Conway: $\{0, 2, 3, 4, 7, 11, 12, 14\}$.
- Marica (1969): $\{0, 1, 2, 4, 7, 8, 12, 14, 15\}$.
- Freiman and Pigarev (1973): $\{0, 1, 2, 4, 5, 9, 12, 13, 14, 16, 17, 21, 24, 25, 26, 28, 29\}$.
- Computer search of random subsets of $\{1, \dots, 100\}$: $\{2, 6, 7, 9, 13, 14, 16, 18, 19, 22, 23, 25, 30, 31, 33, 37, 39, 41, 42, 45, 46, 47, 48, 49, 51, 52, 54, 57, 58, 59, 61, 64, 65, 66, 67, 68, 72, 73, 74, 75, 81, 83, 84, 87, 88, 91, 93, 94, 95, 98, 100\}$.
- Recently infinite families (Hegarty, Nathanson).

Probability Review

X random variable with density $f(x)$ means

- $f(x) \geq 0$;
- $\int_{-\infty}^{\infty} f(x) = 1$;
- $\mathrm{Prob}(X \in [a, b]) = \int_a^b f(x)dx$.

Key quantities:

- Expected (Average) Value: $\mathbb{E}[X] = \int xf(x)dx$.
- Variance: $\sigma^2 = \int (x - \mathbb{E}[X])^2 f(x)dx$.

Binomial model

Binomial model, parameter $p(n)$

Each $k \in \{0, \dots, n\}$ is in A with probability $p(n)$.

Consider uniform model ($p(n) = 1/2$):

- Let $A \in \{0, \dots, n\}$. Most elements in $\{0, \dots, 2n\}$ in $A + A$ and in $\{-n, \dots, n\}$ in $A - A$.
- $\mathbb{E}[|A + A|] = 2n - 11$, $\mathbb{E}[|A - A|] = 2n - 7$.

Martin and O'Bryant '06

Theorem

Let A be chosen from $\{0, \dots, N\}$ according to the binomial model with constant parameter p (thus $k \in A$ with probability p). At least $k_{\mathrm{SD};p} 2^{N+1}$ subsets are sum dominated.

- $k_{\mathrm{SD};1/2} \geq 10^{-7}$, expect about 10^{-3}.
- Proof ($p = 1/2$): Generically $|A| = \frac{N}{2} + O(\sqrt{N})$.
  - ⋄ about $\frac{N}{4} - \frac{|N-k|}{4}$ ways write $k \in A + A$.
  - ⋄ about $\frac{N}{4} - \frac{|k|}{4}$ ways write $k \in A - A$.
  - ⋄ Almost all numbers that can be in $A \pm A$ are.
  - ⋄ Win by controlling fringes.

Notation

- $X \sim f(N)$ means $\forall \epsilon_1, \epsilon_2 > 0$, $\exists N_{\epsilon_1,\epsilon_2}$ st $\forall N \geq N_{\epsilon_1,\epsilon_2}$
$$\mathrm{Prob}\left(X \not\in [(1-\epsilon_1)f(N), (1+\epsilon_1)f(N)]\right) \ < \ \epsilon_2.$$
- $\mathcal{S} = |A+A|$, $\mathcal{D} = |A-A|$,
$\mathcal{S}^{\mathrm{c}} = 2N+1-\mathcal{S}$, $\mathcal{D}^{\mathrm{c}} = 2N+1-\mathcal{D}$.

New model: Binomial with parameter $p(N)$:

- $1/N = o(p(N))$ and $p(N) = o(1)$;
- $\mathrm{Prob}(k \in A) = p(N)$.

Conjecture (Martin-O'Bryant)

As $N \to \infty$, A is a.s. difference dominated.

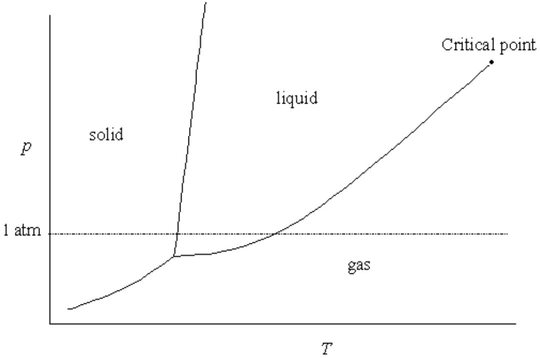
Main Result

Theorem (Hegarty-Miller)

$p(N)$ as above, $g(x) = 2\frac{e^{-x}-(1-x)}{x}$.

- $p(N) = o(N^{-1/2})$: $\mathcal{D} \sim 2\mathcal{S} \sim (Np(N))^2$;
- $p(N) = cN^{-1/2}$: $\mathcal{D} \sim g(c^2)N$, $\mathcal{S} \sim g\left(\frac{c^2}{2}\right)N$ ($c \to 0$, $\mathcal{D}/\mathcal{S} \to 2$; $c \to \infty$, $\mathcal{D}/\mathcal{S} \to 1$);
- $N^{-1/2} = o(p(N))$: $\mathcal{S}^c \sim 2\mathcal{D}^c \sim 4/p(N)^2$.

Critical Thresholds



Can generalize Hegarty-Miller to binary linear forms, still have critical threshold.

Inputs

Key input: recent strong concentration results of Kim and Vu (Applications: combinatorial number theory, random graphs, ...).

Example (Chernoff): t_i iid binary random variables, $Y = \sum_{i=1}^{n} t_i$, then

$$\forall \lambda > 0 : \ \mathrm{Prob}\left(|Y - \mathbb{E}[Y]| \ge \sqrt{\lambda n}\right) \ \le \ 2e^{-\lambda/2}.$$

Need to allow dependent random variables.
Sketch of proofs: $\mathcal{X} \in \{\mathcal{S}, \mathcal{D}, \mathcal{S}^c, \mathcal{D}^c\}$.

1. Prove $\mathbb{E}[\mathcal{X}]$ behaves asymptotically as claimed;
2. Prove $\mathcal{X}$ is strongly concentrated about mean.

Proofs

Setup

Note: only need strong concentration for $N^{-1/2} = o(p(N))$.

Setup

Note: only need strong concentration for $N^{-1/2} = o(p(N))$.

Will assume $p(N) = o(N^{-1/2})$ as proofs are elementary (i.e., Chebyshev: $\text{Prob}(|Y - \mathbb{E}[Y]| \geq k\sigma_Y) \leq 1/k^2$)).

Setup

Note: only need strong concentration for $N^{-1/2} = o(p(N))$.

Will assume $p(N) = o(N^{-1/2})$ as proofs are elementary (i.e., Chebyshev: $\text{Prob}(|Y - \mathbb{E}[Y]| \geq k\sigma_Y) \leq 1/k^2$)).

For convenience let $p(N) = N^{-\delta}$, $\delta \in (1/2, 1)$.

IID binary indicator variables:

$$X_{n;N} = \begin{cases} 1 & \text{with probability } N^{-\delta} \\ 0 & \text{with probability } 1 - N^{-\delta}. \end{cases}$$

$X = \sum_{i=1}^{N} X_{n;N}$, $\mathbb{E}[X] = N^{1-\delta}$.

Proof

Lemma

$P_1(N) = 4N^{-(1-\delta)}$,
$\mathcal{O} = \#\{(m,n) : m < n \in \{1,\dots,N\} \bigcap A\}$.
With probability at least $1 - P_1(N)$ have

1. $X \in \left[\frac{1}{2}N^{1-\delta}, \frac{3}{2}N^{1-\delta}\right]$.
2. $\frac{\frac{1}{2}N^{1-\delta}(\frac{1}{2}N^{1-\delta}-1)}{2} \le \mathcal{O} \le \frac{\frac{3}{2}N^{1-\delta}(\frac{3}{2}N^{1-\delta}-1)}{2}$.

Proof

Lemma

$P_1(N) = 4N^{-(1-\delta)}$,
$\mathcal{O} = \#\{(m,n) : m < n \in \{1,\dots,N\} \bigcap A\}$.
With probability at least $1 - P_1(N)$ *have*

1. $X \in \left[\frac{1}{2}N^{1-\delta}, \frac{3}{2}N^{1-\delta}\right]$.
2. $\frac{\frac{1}{2}N^{1-\delta}(\frac{1}{2}N^{1-\delta}-1)}{2} \le \mathcal{O} \le \frac{\frac{3}{2}N^{1-\delta}(\frac{3}{2}N^{1-\delta}-1)}{2}$.

Proof:

- (1) is Chebyshev: $\mathrm{Var}(X) = N\mathrm{Var}(X_{n;N}) \le N^{1-\delta}$.
- (2) follows from (1) and $\binom{r}{2}$ ways to choose 2 from r.

Concentration

Lemma

- $f(\delta) = \min\left(\frac{1}{2}, \frac{3\delta-1}{2}\right)$, $g(\delta)$ *any function st* $0 < g(\delta) < f(\delta)$.
- $p(N) = N^{-\delta}$, $\delta \in (1/2, 1)$, $P_1(N) = 4N^{-(1-\delta)}$, $P_2(N) = CN^{-(f(\delta)-g(\delta))}$.

With probability at least $1 - P_1(N) - P_2(N)$ *have* $\mathcal{D}/\mathcal{S} = 2 + O(N^{-g(\delta)})$.

Concentration

Lemma

- $f(\delta) = \min\left(\frac{1}{2}, \frac{3\delta - 1}{2}\right)$, $g(\delta)$ *any function st* $0 < g(\delta) < f(\delta)$.
- $p(N) = N^{-\delta}$, $\delta \in (1/2, 1)$, $P_1(N) = 4N^{-(1-\delta)}$, $P_2(N) = CN^{-(f(\delta) - g(\delta))}$.

With probability at least $1 - P_1(N) - P_2(N)$ *have* $\mathcal{D}/\mathcal{S} = 2 + O(N^{-g(\delta)})$.

Proof: Show $\mathcal{D} \sim 2\mathcal{O} + O(N^{3-4\delta})$, $\mathcal{S} \sim \mathcal{O} + O(N^{3-4\delta})$.

As $\mathcal{O}$ is of size $N^{2-2\delta}$ with high probability, need $2 - 2\delta > 3 - 4\delta$ or $\delta > 1/2$.

Analysis of $\mathcal{D}$

Contribution from ‘diagonal’ terms lower order, ignore.

Analysis of $\mathcal{D}$

Contribution from ‘diagonal’ terms lower order, ignore.

Difficulty: (m, n) and (m', n') could yield same differences.

Analysis of $\mathcal{D}$

Contribution from 'diagonal' terms lower order, ignore.

Difficulty: (m, n) and (m', n') could yield same differences.

Notation: $m < n$, $m' < n'$, $m \leq m'$,

$$Y_{m,n,m',n'} = \begin{cases} 1 & \text{if } n - m = n' - m' \\ 0 & \text{otherwise.} \end{cases}$$

Analysis of $\mathcal{D}$

Contribution from 'diagonal' terms lower order, ignore.

Difficulty: (m, n) and (m', n') could yield same differences.

Notation: $m < n$, $m' < n'$, $m \leq m'$,

$$Y_{m,n,m',n'} = \begin{cases} 1 & \text{if } n - m = n' - m' \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{E}[Y] \leq N^3 \cdot N^{-4\delta} + N^2 \cdot N^{-3\delta} \leq 2N^{3-4\delta}$. As $\delta > 1/2$,
Expected number bad pairs $\lll |\mathcal{O}|$.

Analysis of $\mathcal{D}$

Contribution from 'diagonal' terms lower order, ignore.

Difficulty: (m, n) and (m', n') could yield same differences.

Notation: $m < n$, $m' < n'$, $m \leq m'$,

$$Y_{m,n,m',n'} = \begin{cases} 1 & \text{if } n - m = n' - m' \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{E}[Y] \leq N^3 \cdot N^{-4\delta} + N^2 \cdot N^{-3\delta} \leq 2N^{3-4\delta}$. As $\delta > 1/2$,
Expected number bad pairs $\lll |\mathcal{O}|$.

Claim: $\sigma_Y \leq N^{r(\delta)}$with $r(\delta) = \frac{1}{2}\max(3 - 4\delta, 5 - 7\delta)$. This and Chebyshev conclude proof of theorem.

Proof of claim

Cannot use CLT as $Y_{m,n,m',n'}$ are not independent.

Proof of claim

Cannot use CLT as $Y_{m,n,m',n'}$ are not independent.

Use $\mathrm{Var}(U+V) \leq 2\mathrm{Var}(U) + 2\mathrm{Var}(V)$.

Proof of claim

Cannot use CLT as $Y_{m,n,m',n'}$ are not independent.

Use $\mathrm{Var}(U+V) \le 2\mathrm{Var}(U) + 2\mathrm{Var}(V)$.

Write

$$\sum Y_{m,n,m',n'} \;=\; \sum U_{m,n,m',n'} + \sum V_{m,n,n'}$$

with all indices distinct (at most one in common, if so must be $n = m'$).

$$\mathrm{Var}(U) = \sum \mathrm{Var}(U_{m,n,m',n'}) + 2 \sum_{\substack{(m,n,m',n') \neq \\ (\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}')}} \mathrm{CoVar}(U_{m,n,m',n'}, U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'})$$

Analyzing $\mathrm{Var}(U_{m,n,m',n'})$

At most N^3 tuples.

Each has variance $N^{-4\delta} - N^{-8\delta} \le N^{-4\delta}$.

Thus $\sum \mathrm{Var}(U_{m,n,m',n'}) \le N^{3-4\delta}$.

Analyzing $\mathrm{CoVar}(U_{m,n,m',n'}, U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'})$

- All 8 indices distinct: independent, covariance of 0.
- 7 indices distinct: At most N^3 choices for first tuple, at most N^2 for second, get

$$\mathbb{E}[U_{(1)}U_{(2)}] - \mathbb{E}[U_{(1)}]\mathbb{E}[U_{(2)}] = N^{-7\delta} - N^{-4\delta}N^{-4\delta} \le N^{-7\delta}.$$

- Argue similarly for rest, get $\ll N^{5-7\delta} + N^{3-4\delta}$.