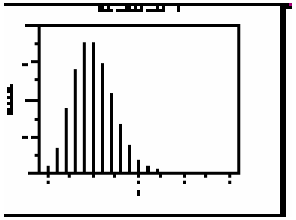
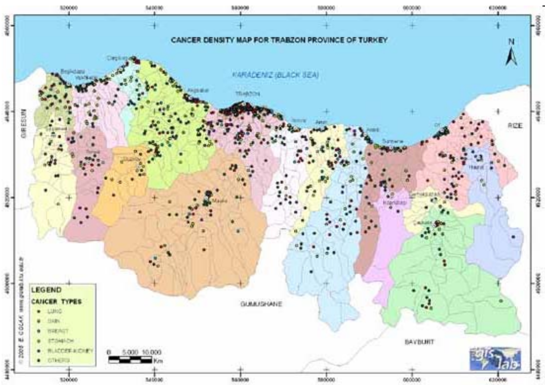
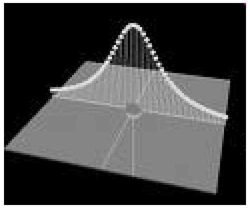
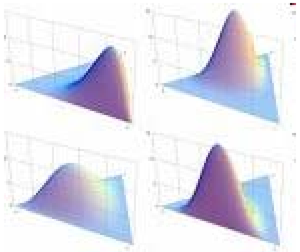
Testing properties of distributions

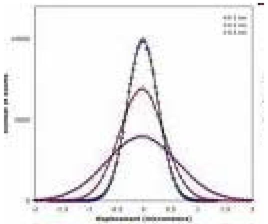
Ronitt Rubinfeld
MIT and Tel Aviv University



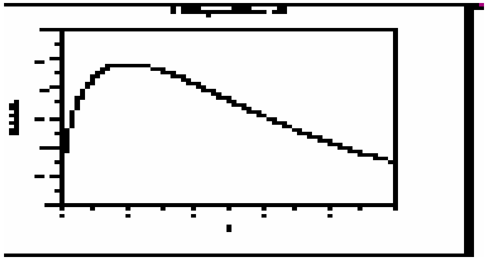


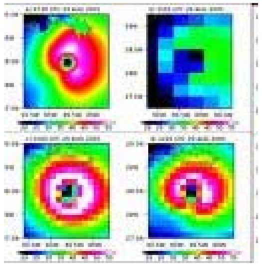


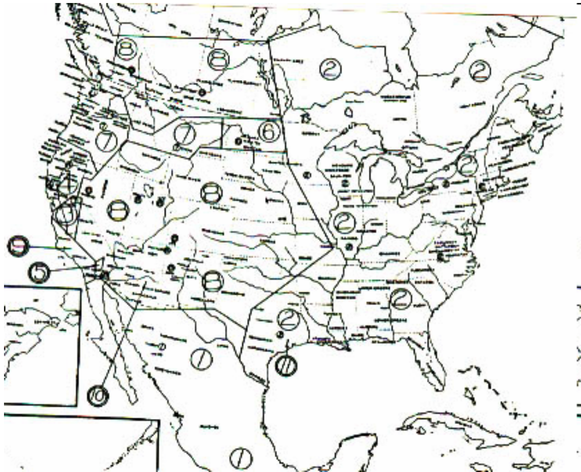




Distributions are everywhere



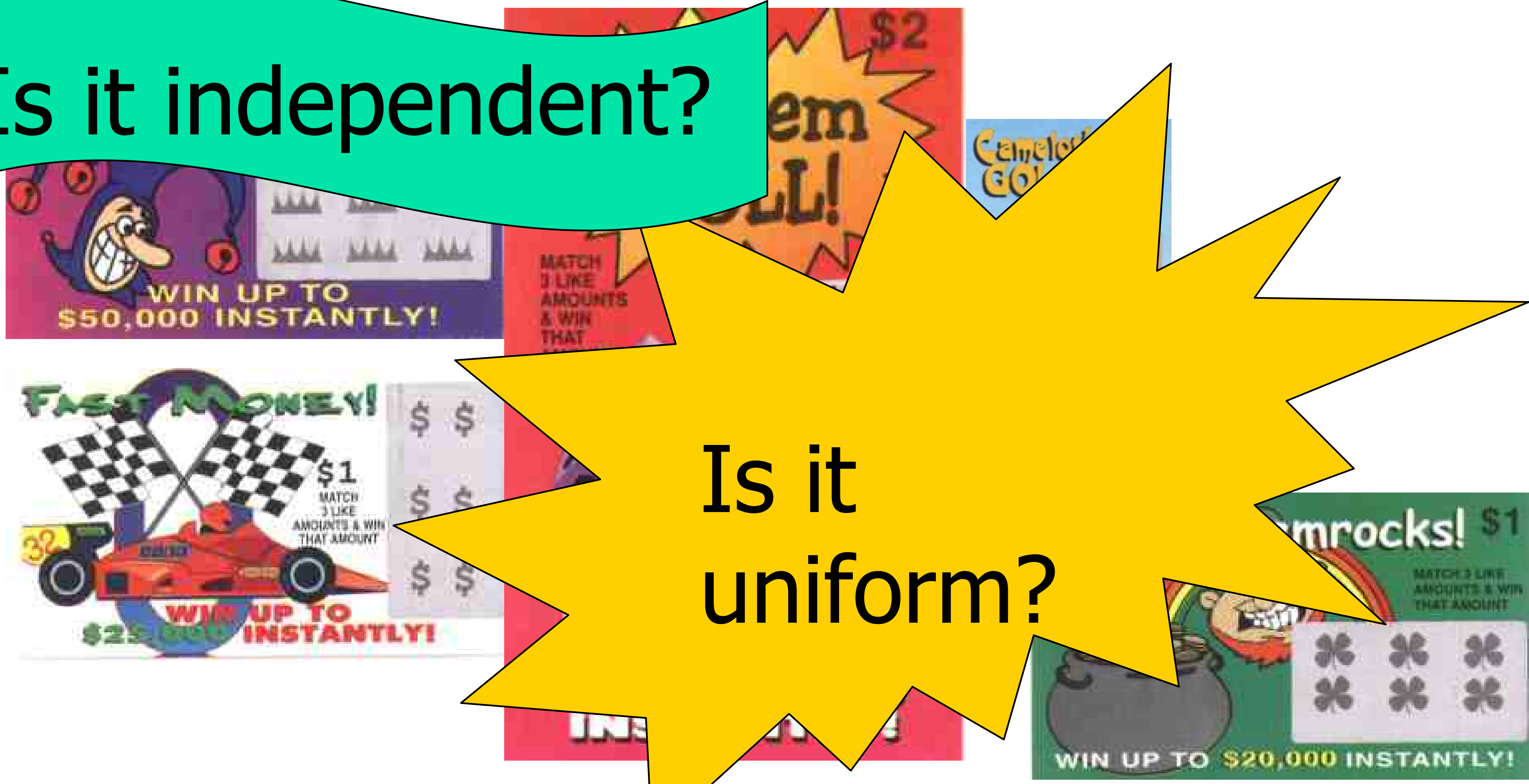




What properties do your distributions have?

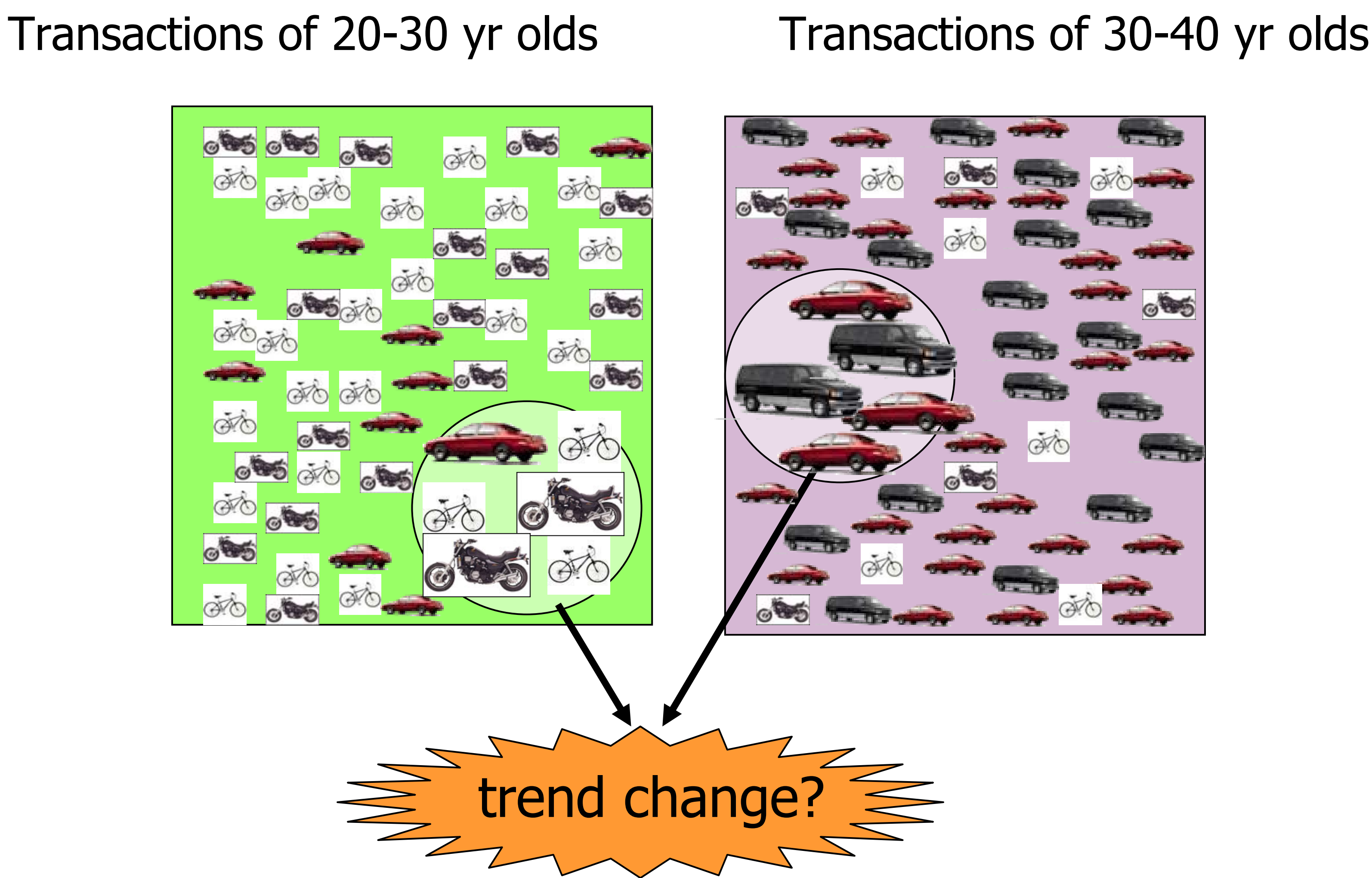
Play the lottery?

Is it independent?

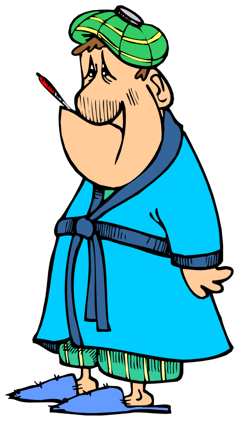


Is it uniform?

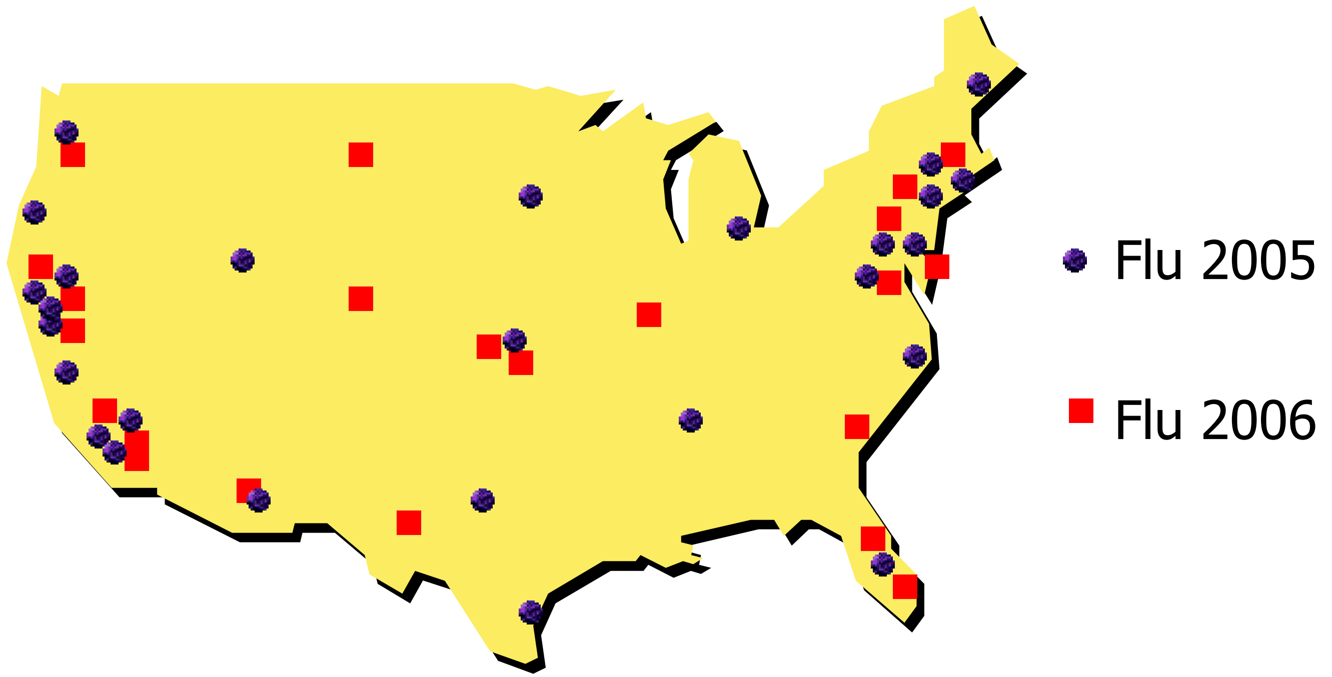
Testing closeness of two distributions:



Outbreak of diseases

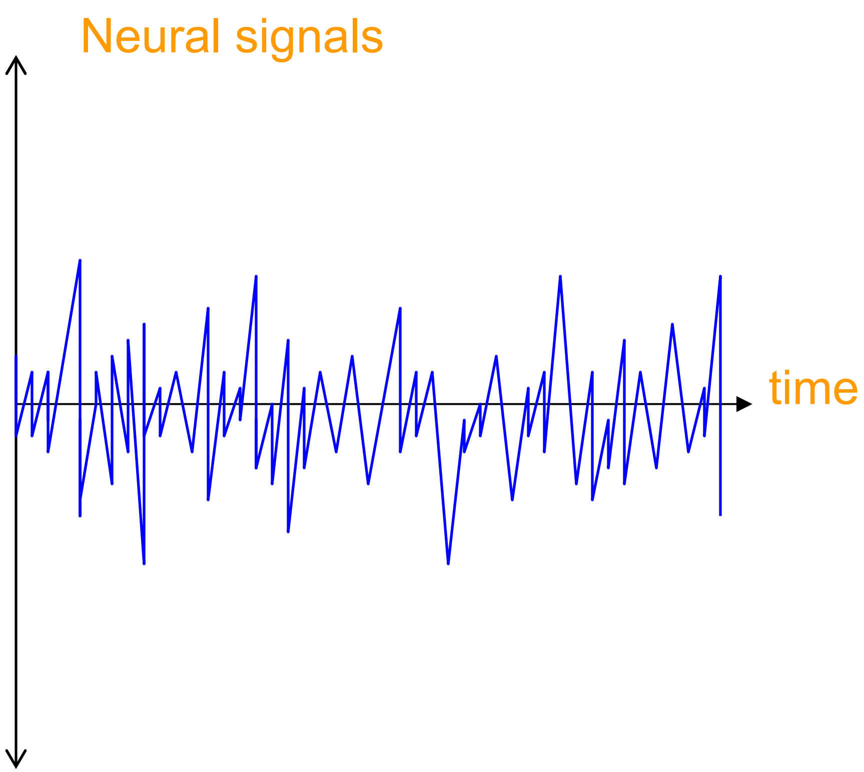


- Similar patterns?
- Correlated with income level?
- More prevalent near large airports?



Information in neural spike trails

[Strong, Koberle, de Ruyter van Steveninck, Bialek '98]

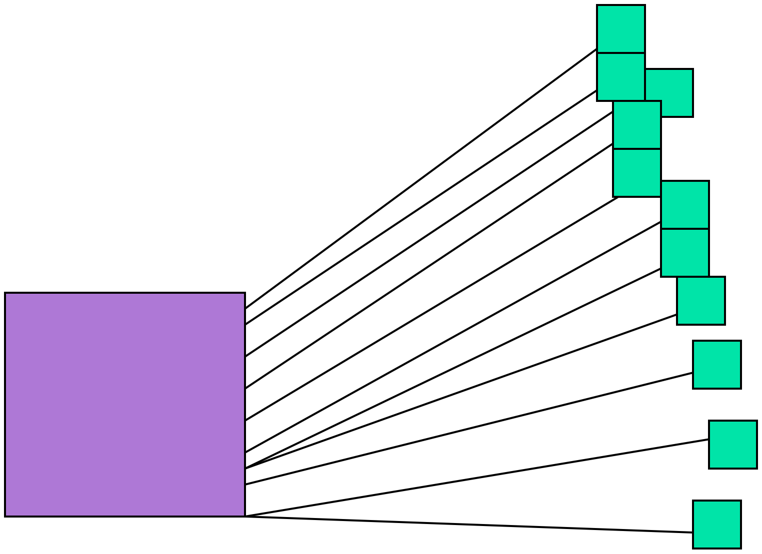


- Each application of stimuli gives sample of signal (spike trail)
- Entropy of (discretized) signal indicates which neurons respond to stimuli

Compressibility of data

Worm detection

- find ``heavy hitters'' – nodes that send to many distinct addresses

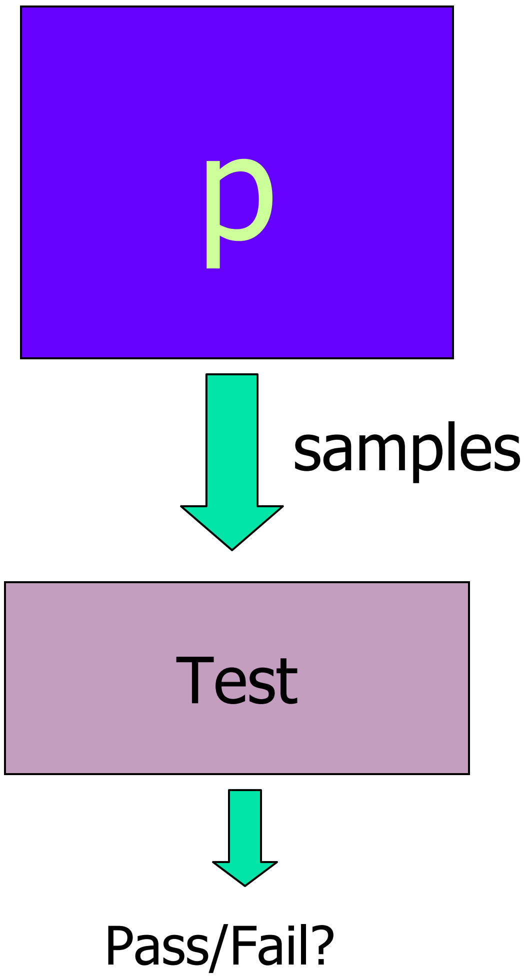


Testing properties of distributions:

- Decisions based on samples of distribution
- Focus on large domains
  - Can sample complexity be *sublinear* in size of the domain?

Rules out standard statistical techniques, learning distribution

Model:



- p is arbitrary black-box distribution over $[n]$, generates iid samples.
- p_i = Prob[p outputs i]
- Sample complexity in terms of n?

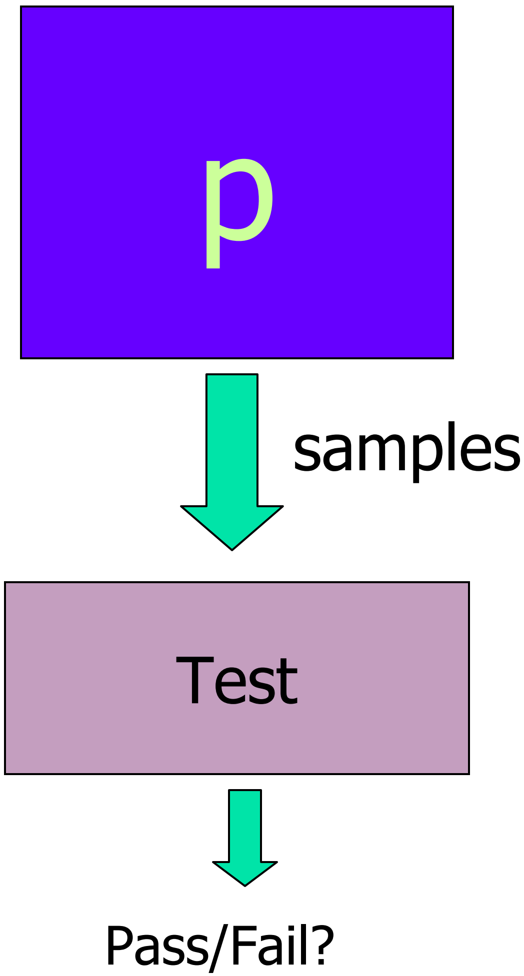
Some properties

- Similarities of distributions:
  - Testing uniformity
  - Testing identity
  - Testing closeness
- Entropy estimation
- Support size
- Independence properties
- Monotonicity

Similarities of distributions

- Are *p* and *q* close or far?
  - *q* is known to the tester
    - *q* is uniform
  - *q* is given via samples

Is *p* uniform?



- Theorem: ([Goldreich Ron][Batu Fortnow R. Smith White] [Paninski]) Sample complexity of distinguishing

  $p=U$

  from $|p-U|_1>\varepsilon$ is $\theta(n^{1/2})$

  $|p-q|_1=\Sigma|p_i-q_i|$

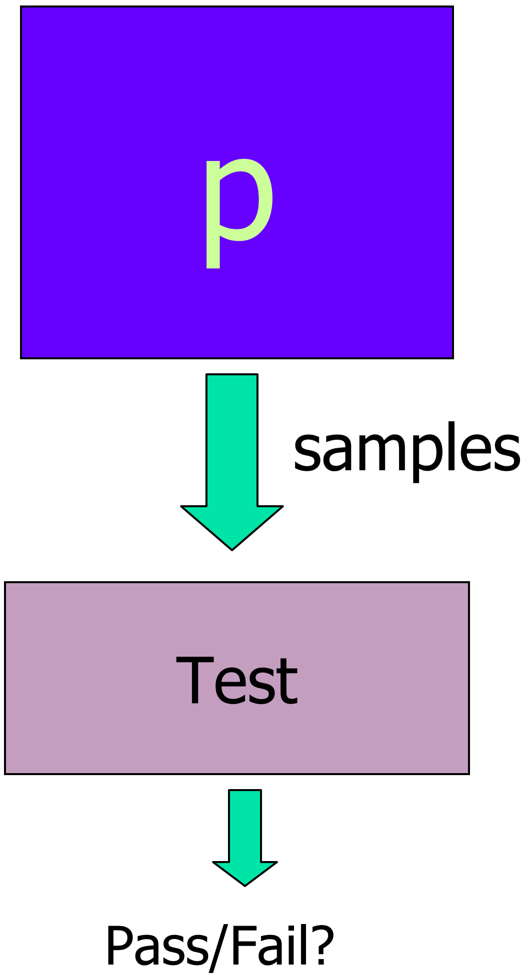
- Nearly ... test if p ... distribution [Batu Fischer Fortnow Kumar R. White]: "Testing identity"

Testing uniformity [GR][BFRSW]

- Upper bound: Estimate collision probability + bound L_∞ norm
  - Issues:
    - Collision probability of uniform is 1/n
    - Pairs not independent
    - Relation between L_1 and L_2 norms
  - Comment: [P] uses different estimator

- Easy lower bound: $\Omega(n^{1/2})$
  - Can get $\Omega(n^{1/2}/\varepsilon^2)$ [P]

Is *p* uniform?



- Theorem: ([Goldreich Ron][Batu Fortnow R. Smith White] [Paninski]) Sample complexity of distinguishing

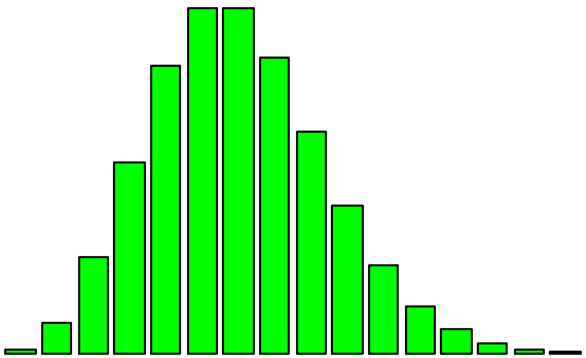
  $p=U$

  from $|p-U|_1>\varepsilon$ is $\theta(n^{1/2})$

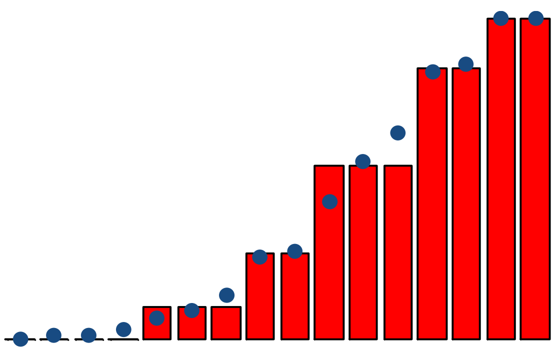
- Nearly same complexity to test if p is any *known* distribution [Batu Fischer Fortnow Kumar R. White]: "Testing identity"

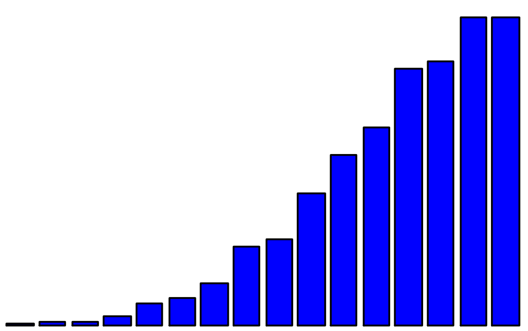
Testing identity via testing uniformity on subdomains:

q (known)

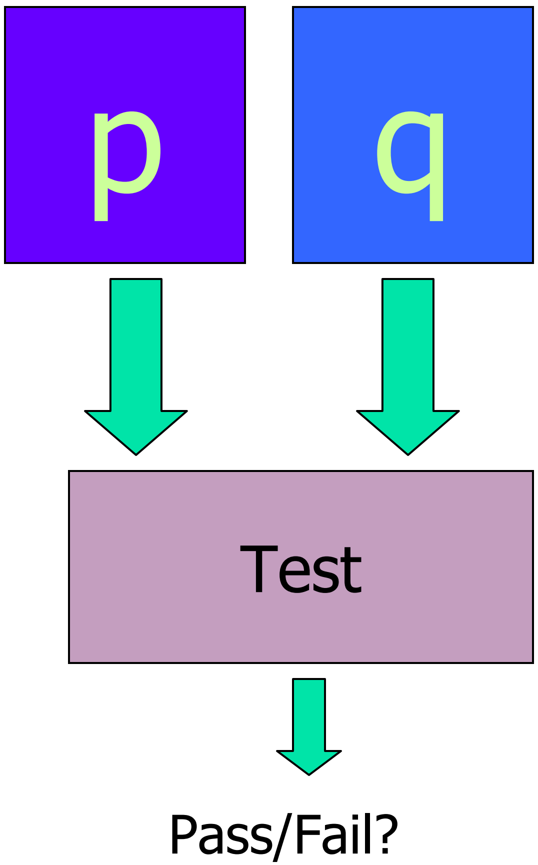


- *(Relabel domain so that q monotone)*
- Partition domain into *O(log n)* groups, so that each group almost “flat” --
  - differ by $<(1+\varepsilon)$ multiplicative factor
  - *q* close to uniform over each group
- Test:
  - Test that *p* close to uniform over each group
  - Test that *p* assigns approximately correct total weights to each group





Testing closeness



Theorem: ([BFRSW] [P. Valiant])

Sample complexity of distinguishing

$p=q$

from $|p-q|_1 > \varepsilon$

is $\tilde{\theta}(n^{2/3})$

A historical note:

- Interest in [GR] and [BFRSW] sparked by search for property testers for expanders
  - Eventual success! [Czumaj Sohler, Kale Seshadri, Nachmias Shapira]
  - Used to give $O(n^{2/3})$ time property testers for rapidly mixing Markov chains [BFRSW]
    - Is this optimal?

Approximating the distance between two distributions?

Distinguishing *whether* $|p-q|_1 < \varepsilon$ or $|p - q|_1$ is $\Theta(1)$ requires nearly linear samples [P. Valiant 08]

Can we approximate the entropy? [Batu Dasgupta R. Kumar]

- In general, not to within a multiplicative factor...
  - ≈ 0 entropy distributions are hard to distinguish (even in superlinear time)
- What if entropy is big (i.e. $\Omega(\log n)$)?
  - Can γ-multiplicatively approximate the entropy with $\tilde{O}(n^{1/\gamma^2})$ samples (when entropy $>2\gamma/\varepsilon$)
  - requires $\Omega(n^{1/\gamma^2})$ [Valiant]
  - better bounds in terms of support size [Brautbar Samorodnitsky]

Estimating Compressibility of Data

[Raskhodnikova Ron Rubinfeld Smith]

- General question undecidable
- Run-length encoding
- Huffman coding
  - Entropy
- Lempel-Ziv
  - ``Color number'' = Number of elements with probability at least $1/n$
  - Can weakly approximate in sublinear time
  - Requires nearly linear samples to approximate well [Raskhodnikova Ron Shpilka Smith]

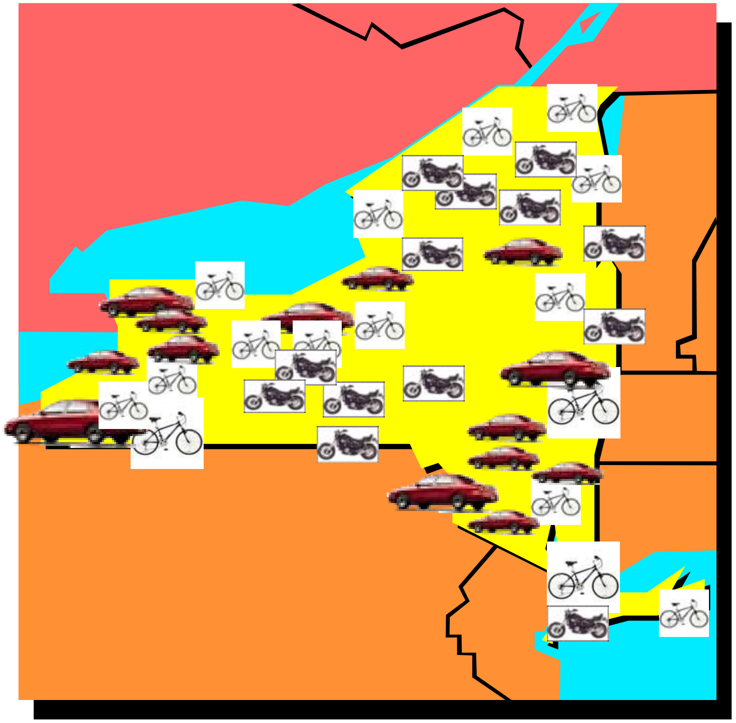
P. Valiant's characterization:

- Collisions tell all!
  - Canonical tester identifies if there is a distribution with the property that expects observed collision statistics
  - Difficulty in analysis:
    - Collision statistics aren't independent
    - Low frequency collision statistics can be ignored?
  - Applies to symmetric properties with "continuity" condition
    - Unifies previous results

- What about non-symmetric properties?

Testing Independence:

Shopping patterns:



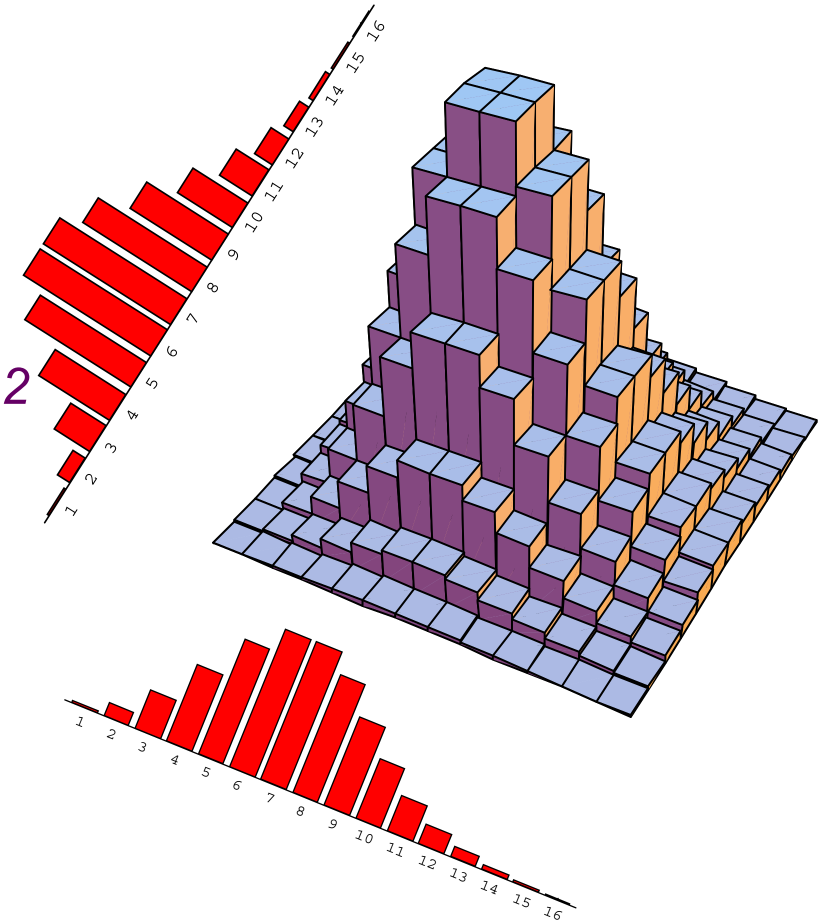
Independent of zip code?

Independence of pairs

- p is joint distribution on pairs $\langle a,b \rangle$ from $[n] \times [m]$ (wlog $n \geq m$)

- Marginal distributions p_1, p_2

- p *independent* if $p = p_1 \times p_2$, that is $p_{(a,b)} = (p_1)_a \, (p_2)_b$ for all a,b



Independence vs. product of marginals

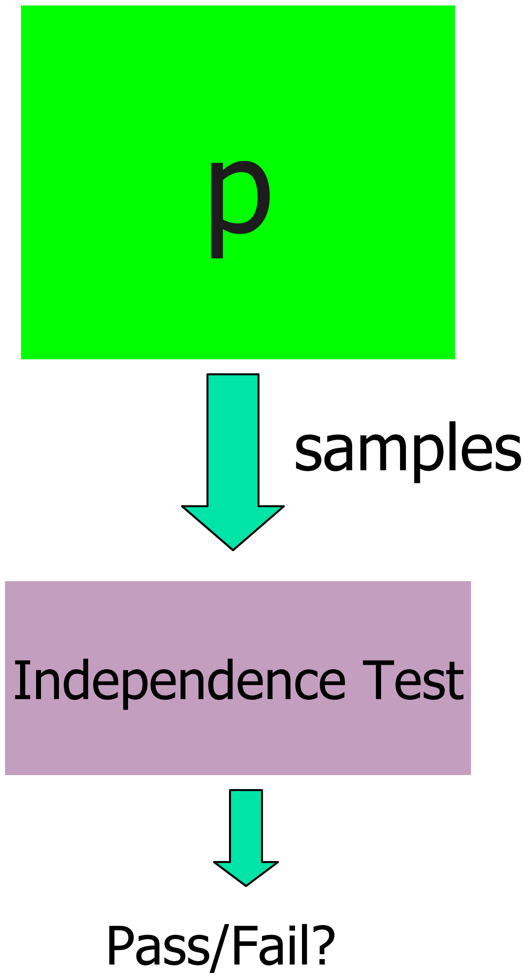
Lemma: [Sahai Vadhan]

If $\exists$ A,B, such that $\|p - A \times B\|_1 < \varepsilon/3$

then $\|p - p_1 \times p_2\|_1 < \varepsilon$

Testing Independence

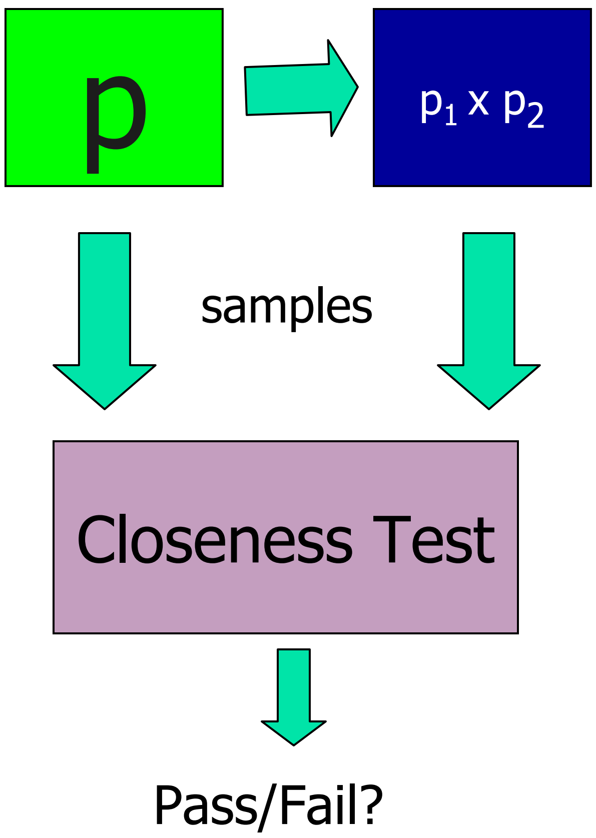
[Batu Fischer Fortnow Kumar R. White]



Goal:

- If $p = p_1 \times p_2$ then PASS
- If $\|p - p_1 \times p_2\|_1 > \varepsilon$ then FAIL

1st try: Use closeness test



- Simulate p_1 and p_2, and check $||p - p_1 \times p_2||_1 < \varepsilon$.
- Behavior:
  - If $||p - p_1 \times p_2||_1 < \varepsilon / n^{1/3}$ then PASS
  - If $||p - p_1 \times p_2||_1 > \varepsilon$ then FAIL
  - Sample complexity: $\tilde{O}((nm)^{2/3})$

2nd try: Use identity test

- Algorithm:
  - Approximate marginal distributions $f_1 \approx p_1$ and $f_2 \approx p_2$
  - Use Identity testing algorithm to test that $p \approx f_1 \times f_2$
- *Comments:*
  - use care when showing that good distributions pass
  - Sample complexity: $\tilde{O}(n+m + (nm)^{1/2})$
  - Can combine with previous using filtering ideas—
    - identity test works well on distribution restricted to ``heavy prefixes'' from p_1
    - closeness test works well if max probability element is bounded from above

Theorem: [Batu Fischer Fortnow Kumar R. White]

There exists an algorithm for testing independence with sample complexity $O(n^{2/3} m^{1/3} \mathrm{poly}(\log n, \varepsilon^{-1}))$ s.t.

- If $p = p_1 \times p_2$, it outputs PASS
- If $\|p-q\|_1 > \varepsilon$ for any independent q, it outputs FAIL

An open question:

- What is the complexity of testing independence of distributions over k-tuples from $[n_1] \times \ldots \times [n_k]$?

- Easy $\Omega(\prod n_i^{1/2})$ lower bound

k-wise Independent Distributions (binary case)

- p is distribution over $\{0,1\}^N$
- p is *k-wise independent* if restricting to any k coordinates yields the uniform distribution
- support size might only be $O(N^k)$
  - $\Omega(2^{N/2})$ lower bound for total independence doesn't apply

Bias

- Definition : For any $S \subseteq [N]$,

  $bias_p(S) = Pr_{x \varepsilon p}[\Sigma_{i \varepsilon S} x_i = 0] - Pr_{x \varepsilon p}[\Sigma_{i \varepsilon S} x_i = 1]$

  (Fourier coeff of p corresponding to S = $bias_p (S)/2^N$ *)*

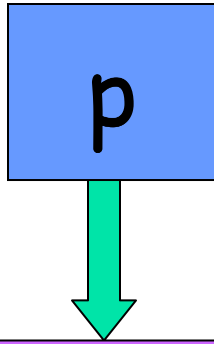
- distribution is *k*-wise independent

  iff all biases over sets S of size $1 \le i \le k$ are 0

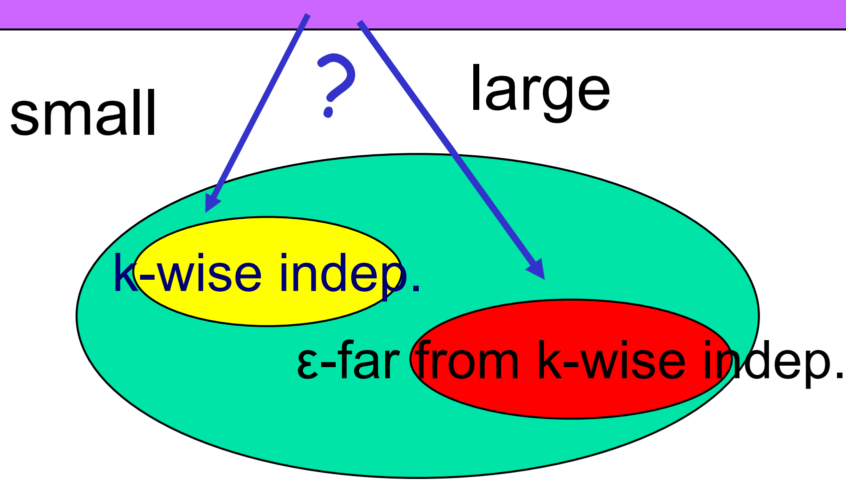
  (*iff* all degree $1 \le i \le k$ Fourier coefficients are 0)

- XOR Lemma [Vazirani 85] relates max bias to distance from uniform dist.

Proposed Testing algorithm



1. Take *O(?)* samples
2. Estimate all the biases up to size *k*
3. Consider the maximum |*bias(S)*|



Relation between p's distance to *k*-wise independence and biases:

Thm: [Alon Goldreich Mansour]

p's distance to closest *k*-wise independent distribution is bounded above by

$$O(\Sigma_{|S| \leq k}|bias_p(S)|)$$

- yields $\tilde{O}(N^{2k}/\varepsilon^{2})$ testing algorithm
- Proof idea:
  - "fix" each degree ≤ k Fourier coefficient by mixing *p* with uniform distribution over strings of "other" parity on S

Another relation between p's distance to *k*-wise independence and biases:

Thm: [Alon Andoni Kaufman Matulef R. Xie]

p's distance to closest *k*-wise independent distribution bounded above by

$$O((\log N)^{k/2}\, sqrt(\Sigma_{|S| \leq k} bias_p(S)^2))$$

- yields $\tilde{O}(N^k/\varepsilon^2)$ testing algorithm

Proof idea:

Let p_1 be p with all degree $1 \le i \le k$ Fourier coefficients zeroed out

- good news:
  - p_1 is k-wise independent
  - p and p_1 very close
  - sum of p_1 over domain is 1
- bad news:
  - p_1 might not be a distribution (some values not in [0,1])

Proof idea (cont.):

- fix negative values of p_1 by mixing with other *k*-wise independent distributions:
  - small negative values
    - removed in "one shot" by mixing p_1 with uniform distribution
  - larger negative values
    - removed "one by one" by mixing with small support *k*-wise independent distribution based on BCH codes
    - [Beckner, Bon Ami] + higher moment inequalities imply that not too many large
- values >1 work themselves out

Extensions [R. Xie 08]

- Larger alphabet case
  - Main issue: fixing procedure
- Arbitrary marginals

(δ,k)-wise Independent Distributions

- [Naor Naor] A distribution D is *(δ, k)-wise independent* if for all $i_1,\ldots,i_k$ and $v_1,\ldots,v_k$

$$|Pr[x_{i1}\ldots x_{ik}=v_1,\ldots,v_k] - 2^{-k}| \leq \delta$$

- (δ,k)-wise independent distributions even smaller!
  - require only $O(2^k \log N)$ support size
- How do the testing problems compare?

Sample complexity bounds

[AAKMRX]

- Testing independence

  lower bound: $\Omega(2^{N/2})$

- Testing k-wise independence

  upper bound: $\tilde{O}(N^k/\varepsilon^2)$

  lower bound: $\Omega(N^{(k-1)/2}/\varepsilon)$

- Testing (δ,k)-wise independence

  upper bound: $O(k \log N/ \delta^2 \varepsilon^2)$

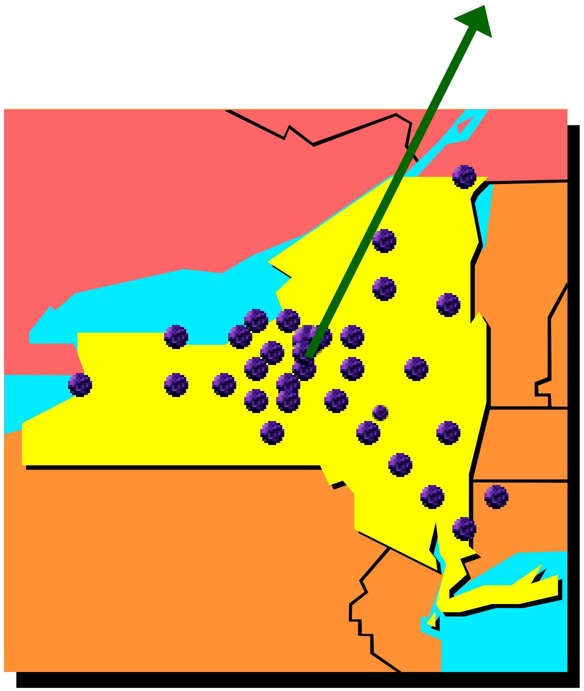
  lower bound: $\Omega(\mathrm{sqrt}(k \log N)/ (\varepsilon+\delta))$

Time complexity of Testing (ε,k)-wise independence

- Bad news: unlikely in polynomial time in terms of $(N, 1/\varepsilon, 1/\delta)$ [AAKMRX]
  - for $k=\theta(\log N)$
  - assuming hardness of finding planted clique of size t in $G(N,1/2,t)$ for $t(N) \approx \log^3 N$

Testing the monotonicity of distributions:

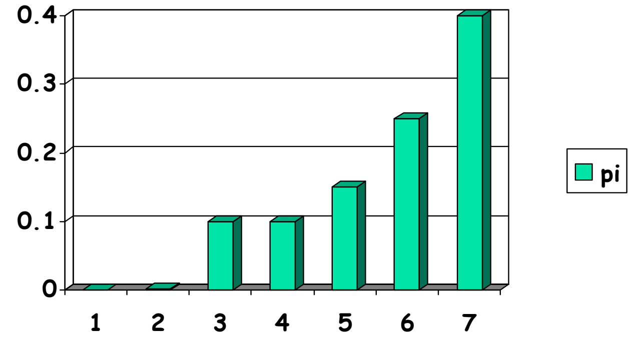
Does the occurrence of cancer decrease with distance from the nuclear reactor?



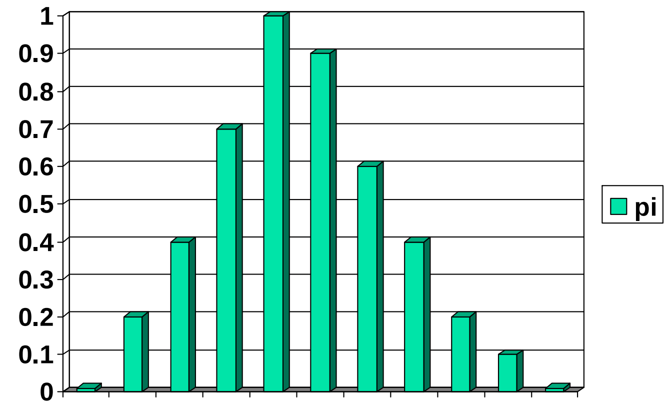
Monotone distributions

- p is monotone if

  $i<j$ implies $p_i \le p_j$

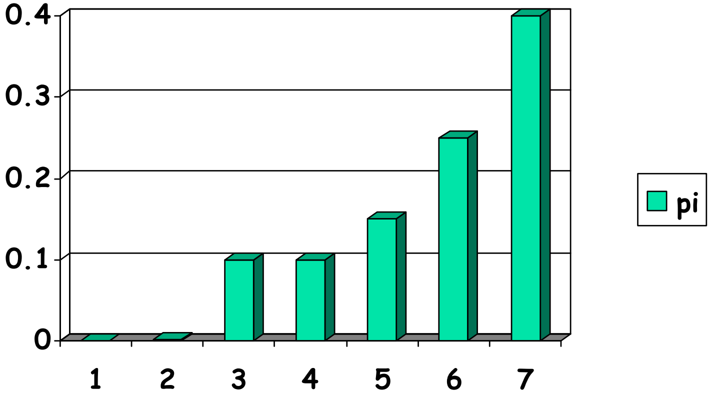


- Many distributions are monotone or are "made of" small number of monotone distributions



First…

Monotone distributions over *totally ordered domains [1..n]*



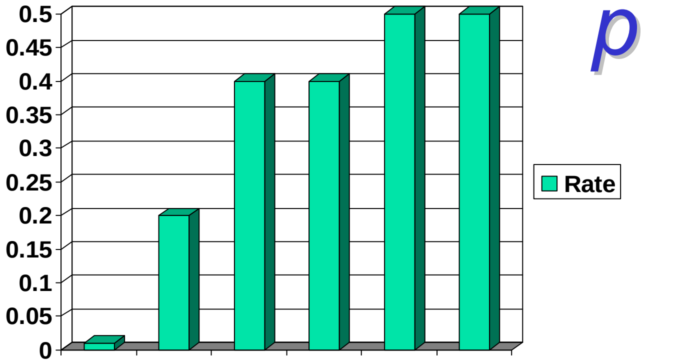
Form of test?

Idea: test that average weight of distribution in range *[i..j]* less than average weight of distribution in *[i'…j']* for various choices of $i<i', j<j'$

Problem: uniform distribution on even numbers passes such tests

Lower bound [Batu Kumar R.]

Lemma: Testing monotonicity requires $\Omega(\sqrt{n})$ samples

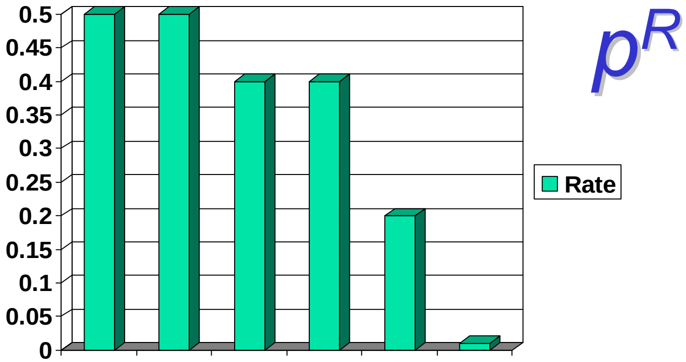


Proof:

p close to uniform

iff

p, p^R = "reversal" of p, are both close to monotone



Algorithm idea:

- Approximate distribution by *k-flat* distribution:
  - Properties:
    - Partition domain into k intervals
    - Conditional distribution uniform in each
  - Questions:
    - Does it exist for $k=O(polylog(n))$?
    - How do you find interval boundaries?
- Check if k-flat distribution close to monotone
  - Solve linear program on $O(polylog(n))$ variables

Upper bound [Batu Kumar R.]

- Lemma: There is an algorithm for testing monotonicity over totally ordered domains which uses $\tilde{O}(n^{1/2}\varepsilon^{-2})$ samples s.t. (with probability 2/3)
  - If p monotone, outputs PASS
  - If ε–far from monotone, outputs FAIL

- Can also test unimodal distributions

Monotonicity over general posets [Bhattacharyya Fischer R. Valiant]

- Can test distributions over poset decomposable into union of w disjoint chains of length at most c with $\tilde{O}(wc^{1/2}poly(1/\varepsilon))$ samples
  - Algorithm: approximate each chain by k-flat distribution and check if resulting distribution close to monotone
  - Implications:
    - $\tilde{O}(N^{3/2})$ bound for NxN grid (simplifying and slightly more efficient than in [BKR])
    - $\tilde{O}(2^N/N^{1/2})$ bound for N-dimensional hypercube
- There are posets for which monotonicity testing requires nearly linear samples

Other properties?

- *K*-flat distributions
- Mixtures of *k* Gaussians
- “Junta”-distributions
- Generated by a small Markovian process
- …

Getting past the lower bounds

- Special distributions
  - e.g, uniform on a subset, monotone
- Other query models
  - Queries to probabilities of elements
- Other distance measures

Flat distributions

Entropy can be estimated somewhat faster when distribution is uniform on a subset of the elements [Batu Dasgupta Kumar R.][Brautbar Samorodnitsky]

Monotone distributions over totally ordered domains

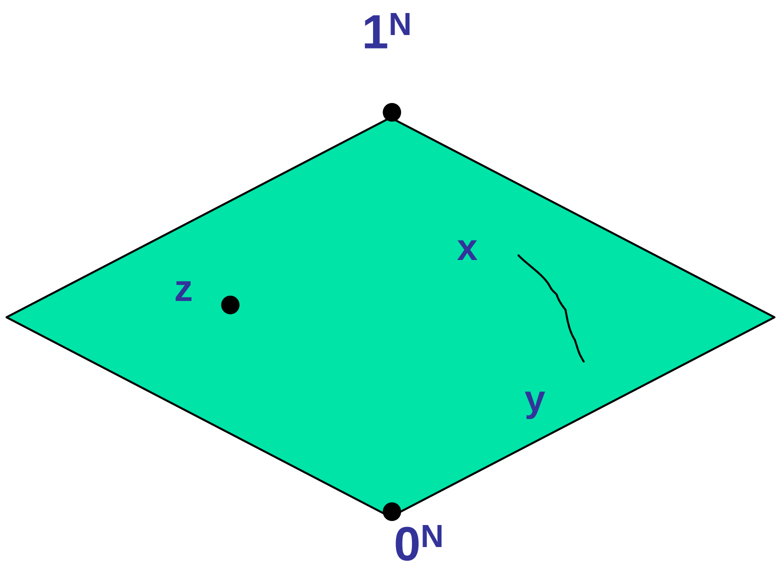
- Test uniformity with *O(1)* samples [Batu Kumar R.]
- Other tasks doable with polylogarithmic samples: [Batu Dasgupta Kumar R.][BKR]
  - Examples:
    - Testing closeness
    - Testing independence
    - Estimating entropy
  - Algorithm:
    - Use *k*-flat partitions to approximate distributions
    - Test property on approximation

- Do these big wins carry over to partial orders?

Monotone high-dimensional distributions

Domain: Boolean cube $\{0,1\}^N$

Are there testing algorithms with sample complexity polylogarithmic in domain size, i.e. *poly(N)*?



Testing Uniformity

Theorem [R. Servedio][Adamaszek Czumaj Sohler]**:** There is an $\tilde{O}(N/\varepsilon^2)$ sample complexity tester which given an unknown monotone distribution p over $\{0,1\}^N$ ($[0,1]^N$) satisfies (with probability 2/3):

- If $p=U$, algorithm outputs **"uniform"**
- If $||p - U||_1 > \varepsilon$, algorithm outputs **"far from uniform"**

- Comment: Nearly best possible

Bad news for Boolean cube

[R. Servedio]

- Technique for sample complexity lower bounds: monotone subcube decomposition
  - $2^{\Omega(N)}$ lower bound for testing equivalence to a known distribution (even product distributions!)
  - $2^{\Omega(N)}$ lower bound for approximating entropy

Open question for Boolean cube

Can one test monotone distributions over $\{0,1\}^N$ for any of the following properties

- equivalence to a known distribution
- approximating entropy
- independence

with fewer samples than for arbitrary distributions?

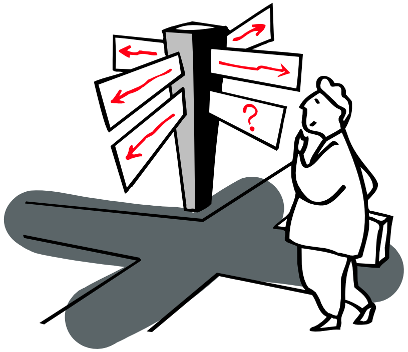
Other query models:

- Distribution given explicitly [BDKR]
- Distribution given both by samples and oracle for p_i's [BDKR][RS]
  - Can estimate entropy in polylog(n) time

Other distance measures:

- Earth Mover Distance [Doba Nguyen[2] R.]
  - Measures min weight matching to some distribution with the property
  - Can estimate distance between distributions, independence over $[0,1]^N$, in time *independent* of domain size
  - Still exponential in N
    - Can improve over highly clusterable distributions

Conclusions and Future Directions



- Distribution property testing problems are everywhere
- Several useful techniques known
- Other properties for which sublinear tests exist?
- Special classes of distributions?
- Time vs. query complexity
- Other query models?
- Non-iid samples?

Thank you