Exact Inference using Stratified Random Samples

and the 0-1 Knapsack Problem

Michael J. Higgins

Based on work with
Ronald L. Rivest and Philip B. Stark

August 4, 2011

Problem Set-up

- There are N units divided into C groups.
  Each group c has N_c units, $c = 1, \ldots, C$.
- Each unit j has a value e_j: only known if unit j is sampled.
  A priori upper bound u_j on e_j: known for each unit j.
- In each group c, take simple random sample of n_c units.
  Stratified random sample; strata are groups.
- Want inference on $\sum_{j=1}^{N} e_j$.

  Specifically, want to test the null hypothesis $\sum_{j=1}^{N} e_j \geq K$, for a specified K.
- Use maximum value of the sample as the test statistic.
  (Why maximum and not sample sum? To be explained later.)

Applications

- Auditing problems: Election [Sta08], [HRS11], financial, etc.
  - Election auditing: Votes are counted by machines. True outcome is what a full hand count would show.
  - Ensure that machine-count and hand-count outcomes agree.
  - Sample: a set of precincts. Stratified random samples common—strata are counties.
- e_j is a measure of misstatement.
  - If nothing "wrong," e_j should be close to 0.
  - Election auditing: e_j is difference in margin between machine-count and hand-count subtotals in precinct j. (e_j positive = machine count favored reported winner.)
- Want to ensure that the amount of misstatement is not large.
  - Election auditing: Hand-count outcome and machine-count outcome do not agree if

$$\sum_{j=1}^{N} e_j \geq \text{Margin of Victory.}$$

Applications Cont'd

- Audit SHOULD protect against fraud.

Applications Cont'd

- Audit SHOULD protect against fraud.
- Problem: If misstatement is deliberate, it is placed to make detection by audit hard.
- For audits using a stratified random sample: place large misstatements in a small number of units.

No Asymptotics Allowed!

- Problem: Sampling from very skewed distributions.
- Most items j have value $e_j = 0$, small number have very large values—bad normal approximation!

No Asymptotics Allowed!

- Problem: Sampling from very skewed distributions.
- Most items j have value $e_j = 0$, small number have very large values—bad normal approximation!
- If T is the sample maximum, for any fixed t, $P(T \leq t) = P(T \leq t; (e_j)_{j=1}^N)$ is easy to compute, given the values $(e_j)_{j=1}^N$ (hypergeometric for each stratum).
- Smaller maximum = stronger evidence that amount of misstatement is small.

No Asymptotics Allowed!

- Problem: Sampling from very skewed distributions.
- Most items j have value $e_j = 0$, small number have very large values—bad normal approximation!
- If T is the sample maximum, for any fixed t, $P(T \leq t) = P(T \leq t; (e_j)_{j=1}^N)$ is easy to compute, given the values $(e_j)_{j=1}^N$ (hypergeometric for each stratum).
- Smaller maximum = stronger evidence that amount of misstatement is small.
- Solution: Suppose the largest value we observe is t^*. Obtain p-value by maximizing $P(T \leq t^*)$ over all possible values $(e_j)_{j=1}^N$ satisfying null hypothesis $\sum_{j=1}^N e_j \geq K$.

Relation to 0-1 Knapsack

- Finding maximum of $P(T \leq t^*; (e_j)_{j=1}^N)$ through an exhaustive search of $(e_j)_{j=1}^N$ is bad.
  Too many possibilities for $(e_j)_{j=1}^N$.

Relation to 0-1 Knapsack

- Finding maximum of $P(T \leq t^*; (e_j)_{j=1}^N)$ through an exhaustive search of $(e_j)_{j=1}^N$ is bad.
  Too many possibilities for $(e_j)_{j=1}^N$.
- For stratified random samples, finding an $(e_j)_{j=1}^N$ that maximizes $P(T \leq t^*)$ is equivalent to solving a 0-1 knapsack problem.
- 0-1 knapsack problem: well-studied problem in Operations Research.

0-1 Knapsack: Minimum Version

- There are N items, each item j has a price ν_j and a cost γ_j.
- Choose a subset of items so that, given that the sum of prices is at least κ, the sum of costs is minimized.

0-1 Knapsack: Minimum Version

- There are N items, each item j has a price ν_j and a cost γ_j.
- Choose a subset of items so that, given that the sum of prices is at least κ, the sum of costs is minimized.
- IP formulation: Find $(x_j)_{j=1}^N \in \{0,1\}^N$ that minimizes

$$\sum_{j=1}^{N} \gamma_j x_j$$

under the constraint that

$$\sum_{j=1}^{N} \nu_j x_j \geq \kappa.$$

0-1 Knapsack: Example

- Example:
  - A thief wants to go into a store and steal $\$\kappa$ worth of items.
  - Each item has a price. Stealing an item will increase the chance the thief gets caught (cost).
  - 72" Flat screen TV: Big price, but hard to steal covertly (big cost).
  - Gum: Small price, easy to steal (small cost).
  - Thief will want to steal expensive items (big price) that can be stolen easily (small cost).

0-1 Knapsack: Example

- Example:
  - A thief wants to go into a store and steal $\$\kappa$ worth of items.
  - Each item has a price. Stealing an item will increase the chance the thief gets caught (cost).
  - 72" Flat screen TV: Big price, but hard to steal covertly (big cost).
  - Gum: Small price, easy to steal (small cost).
  - Thief will want to steal expensive items (big price) that can be stolen easily (small cost).
- For auditing:
  - Price = amount of misstatement an item can hold (u_j).
  - Cost = increase in chance of audit discovering misstatement.
  - Steal item = hide misstatement in item.

Main Theorem:

- Suppose we are testing the null hypothesis $\sum_{j=1}^{N} e_j \geq K$. We have taken a stratified random sample of items and have observed a maximum value of t^*.
- If item j is the item in stratum c with the kth largest value of u, let

$$q_j \equiv -\log\left(\frac{(N_c - n_c - k + 1) \vee 0}{N_c - k + 1}\right).$$

- Obtain solution of 0-1 knapsack $(x_j)_{j=1}^{N}$ with $\nu_j = u_j - t^*$, $\gamma_j = q_j$, and $\kappa = K - Nt^*$.

0-1 Knapsack: Minimum Version

- There are N items, each item j has a price with a price ν_j and a cost γ_j.
- Choose a subset of items so that, given that the sum of prices is at least M, the sum of costs is minimized.
- IP formulation: Find $(x_j)_{j=1}^N \in \{0,1\}^N$ that minimizes

$$\sum_{j=1}^{N} \gamma_j x_j$$

under the constraint that

$$\sum_{j=1}^{N} \nu_j x_j \geq \kappa.$$

Main Theorem:

- Suppose we are testing the null hypothesis $\sum_{j=1}^{N} e_j \geq K$. We have taken a stratified random sample of items and have observed a maximum value of t^*.
- If item j is the item in stratum c with the kth largest value of u,

$$q_j \equiv -\log\left(\frac{(N_c - n_c - k + 1) \vee 0}{N_c - k + 1}\right).$$

- Obtain solution of 0-1 knapsack $(x_j)_{j=1}^{N}$ with $\nu_j = u_j - t^*$, $\gamma_j = q_j$, and $\kappa = K - Nt^*$.
- $P(T \leq t^*)$ is maximized by placing u_j misstatement in unit j when $x_j = 1$, and t^* misstatement when $x_j = 0$.
- P-value: $\exp(-\sum_{j=1}^{N} q_j x_j)$ [HRS11].

Sketch of proof

- For a set of values $(e_j)_{j=1}^N$, for a given t,
  $\#_c t(e)$: Number of items in stratum c with value greater than t.
  If $1 \leq \#_c t(e) \leq n_c$,

$$P(T \leq t) = \prod_{c=1}^{C} \frac{\binom{N_c - \#_c t(e)}{n_c}}{\binom{N_c}{n_c}} = \prod_{c=1}^{C} \prod_{k=1}^{\#_c t(e)} \frac{\binom{N_c - k}{n_c}}{\binom{N_c - k + 1}{n_c}}$$
$$= \prod_{c=1}^{C} \prod_{k=1}^{\#_c t(e)} \left(\frac{N_c - n_c - k + 1}{N_c - k + 1} \right).$$

- Stratified random sample: Every item within a stratum has an equal chance of being selected.
  Hardest to detect: put maximum misstatement in as few items as possible.
- Can obtain a 1-1 correspondence between units $j = 1, \ldots, N$ and costs $(q_j)_{j=1}^N$.

Additional Uses

- Select sample sizes to audit effectively.
  - Minnesota 2006 U.S. Senate race: same power of audit while sampling 40% fewer precincts. [HRS11].
- Obtain upper-confidence bound on the total amount of misstatement.

Bibliography I

M.J. Higgins, R.L. Rivest, and P.B. Stark.
Sharper p-Values For Stratified Election Audits.
Submitted to Statistics, Politics, and Policy, 2011.

P.B. Stark.
Conservative statistical post-election audits.
Ann. Appl. Stat, 2(2):550–581, 2008.