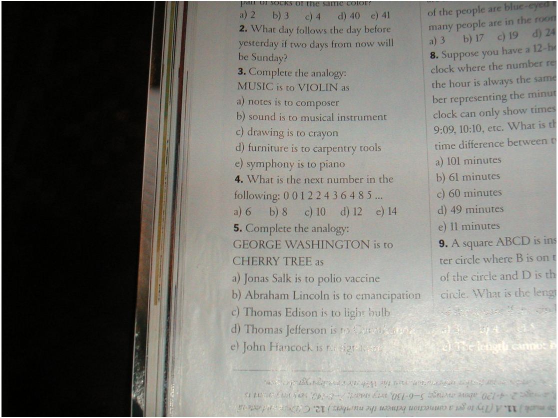
Sampling Case-Control Studies Over Time

http://math.usc.edu/∼larry

From the United Airlines In Flight Magazine



pair of socks of the same color?

a) 2 b) 3 c) 4 d) 40 e) 41

2. What day follows the day before yesterday if two days from now will be Sunday?

3. Complete the analogy:
MUSIC is to VIOLIN as

a) notes is to composer
b) sound is to musical instrument
c) drawing is to crayon
d) furniture is to carpentry tools
e) symphony is to piano

4. What is the next number in the following: 0 0 1 2 2 4 3 6 4 8 5 ...

a) 6 b) 8 c) 10 d) 12 e) 14

5. Complete the analogy:
GEORGE WASHINGTON is to CHERRY TREE as

a) Jonas Salk is to polio vaccine
b) Abraham Lincoln is to emancipation
c) Thomas Edison is to light bulb
d) Thomas Jefferson is to
e) John Hancock is to

of the people are blue-eyed
many people are in the room

a) 3 b) 17 c) 19 d) 24

8. Suppose you have a 12-h
clock where the number re
the hour is always the same
ber representing the minut
clock can only show times
9:09, 10:10, etc. What is th
time difference between t

a) 101 minutes
b) 61 minutes
c) 60 minutes
d) 49 minutes
e) 11 minutes

9. A square ABCD is ins
ter circle where B is on t
of the circle and D is th
circle. What is the lengt

e) The length cannot b

What is the next number in the sequence

$$0, 0, 1, 2, 2, 4, 3, 6, 4, 8, 5 \ldots ?$$

What is the next number in the sequence

$$0, 0, 1, 2, 2, 4, 3, 6, 4, 8, 5 \ldots ?$$

Two Alternating Sequences

$$0, 0, 1, 2, 2, 4, 3, 6, 4, 8, 5, 10, 6 \ldots$$

Logistic Model

In cohort $\mathcal{R}$, the failure indicator D_i for individual $i \in \mathcal{R}$ has distribution

$$P(D_i = 1) = \frac{\lambda_0 \exp(\boldsymbol{\beta}_0' \mathbf{Z}_i)}{1 + \lambda_0 \exp(\boldsymbol{\beta}_0' \mathbf{Z}_i)},$$

where λ_0 is a baseline factor, $\mathbf{Z}_i$ is a covariate vector, and $\boldsymbol{\beta}_0$ an unknown parameter.

Extend to a model for observing the cohort over time.

Cox Model 1972: Cohort over time

Failure rate for individual $i \in \mathcal{R}$ at time t is common baseline hazard function $\lambda_0(t)$ adjusted for covariates $\mathbf{Z}_i(t)$, accommodate censoring by time dependent indicator $Y_i(t)$:

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp(\boldsymbol{\beta}_0'\mathbf{Z}_i(t)).$$

Semi-parametric model: $\boldsymbol{\beta}_0$ finite dimensional parameter of interest, λ_0 is infinite dimensional 'nuisance' parameter.

Still of interest to estimate integrated baseline hazard,

$$\Lambda_0(t) = \int_0^t \lambda_0(s)ds$$

Cox Model: Analysis

Define the risk set

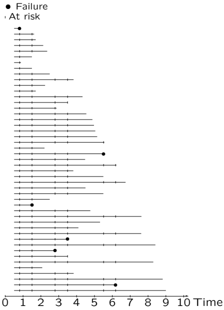
$$\mathcal{R}(t) = \{i : Y_i(t) = 1\}.$$

Cox partial likelihood estimator (MPLE) maximizes the product of probabilities that failure i_j was observed to fail at time t_j, given that there was one failure at that time from those in the risk set $\mathcal{R}_j = \mathcal{R}(t_j)$,

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{\text{failure times } t_j} \frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_{i_j}(t_j)}{\sum_{k\in\mathcal{R}_j} \exp(\boldsymbol{\beta}'\mathbf{Z}_k(t_j))}.$$

Consistent, asymptotically normal, though not a true likelihood: dependence, ignore information between failures.

RISK SETS IN CONTINUOUS TIME



Sampling Designs

The risk sets $\mathcal{R}(t)$ may be too large for the collection of complete information. Sampling designs specify the probability

$$\pi_t(\mathbf{r}|i)$$

of using $\mathbf{r}$ as the sampled risk set should i fail at t.

Weights cancel out common factors,

$$w_i(t, \mathbf{r}) = \frac{\pi_t(\mathbf{r} \mid i)}{n(t)^{-1} \sum_{l \in \mathbf{r}} \pi_t(\mathbf{r} \mid l)}.$$

Sampling Estimators: Cox Model

When $\pi_t(\mathbf{r}|i)$ selects sampled risk set $\tilde{\mathcal{R}}_j$ to use when i_j fails at time t_j, maximize

$$L(\boldsymbol{\beta}) = \prod_{\text{failure times } t_j} \frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_{i_j}(t_j))\pi_{t_j}(\tilde{\mathcal{R}}_j|i_j)}{\sum_{k\in\tilde{\mathcal{R}}_j}\exp(\boldsymbol{\beta}'\mathbf{Z}_k(t_j))\pi_{t_j}(\tilde{\mathcal{R}}_j|k)}.$$

Estimated integrated baseline hazard

$$\hat{\Lambda}_0(t) = \sum_{t_j\leq t} \frac{1}{\sum_{l\in\tilde{\mathcal{R}}_j}\exp(\hat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{Z}_l(t_j))w_l(t_j,\tilde{\mathcal{R}}_j)}.$$

Two Questions

1. For observations over times, what sampling designs might we consider and what are their asymptotic properties (e.g. relative efficiency)? (with B. Langholz and Ø. Borgan)

2. What about sampling for models with no time dependence ? (B. Langholz, R. Arratia)

1. Designs, over time

FC, Full Cohort Design: Analyzed by Andersen and Gill, 1982, using Counting Processes and Martingale methods for time varying covariates and general censoring.

NCCS, Nested Case Control Design, Thomas 1977: Sample $m-1$ controls for the failure at the failure time.

CC, Case Cohort Design, Prentice 1986: Choose subcohort $\tilde{C}$ from full cohort at time $t=0$, let the sampled risk set for failure i_j be $\tilde{C} \cup \{i_j\}$.

CM, Counter Matching Design, Langholz and Borgan 1995: Sample to obtain m_l individuals in each strata $l \in \mathcal{C}$.

Specification of Designs

Let $\pi_t(\mathbf{r}|i)$ be the probability of choosing $\mathbf{r}$ as the sampled risk set should i fail at time t.

NCCS: Sample over $\mathbf{r} \ni i, \mathbf{r} \subset \mathcal{R}(t), |\mathbf{r}| = m, n(t) = |\mathcal{R}(t)|$,

$$\pi_t(\mathbf{r}|i) = \binom{n(t) - 1}{m - 1}^{-1}$$

CM: Sample over $\mathbf{r} \ni i, \mathbf{r} \subset \mathcal{R}(t), |\mathbf{r} \cap \mathcal{R}_l(t)| = m_l;\ l \in \mathcal{C}$

$$\pi_t(\mathbf{r}|i) = \left[\prod_{l \in \mathcal{C}} \binom{n_l(t)}{m_l}\right]^{-1} \frac{n_{C_i(t)}(t)}{m_{C_i(t)}}.$$

Analysis by Martingales

The process $N_i(t)$ counting the number of events for i in $(0, t]$ has intensity $\lambda_i(t)$, and

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(s)ds$$

is a martingale (AG '82). For sampling, the process $N_{i,\mathbf{r}}(t)$ counting the number of events for $(i, \mathbf{r})$ in $(0, t]$ has intensity

$$\lambda_{i,\mathbf{r}}(t) = \lambda_i(t)\pi_t(\mathbf{r}|i)$$

and subtracting its integral gives the martingale

$$M_{i,\mathbf{r}}(t) = N_{i,\mathbf{r}}(t) - \int_0^t \lambda_{i,\mathbf{r}}(s)ds.$$

Likelihood Methods Apply

The score function $\partial \log L_t(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is given by

$$\mathbf{U}_t(\boldsymbol{\beta}) = \int_0^t \sum_{i,\mathbf{r}} \{\mathbf{Z}_i(u) - \mathbf{E}_{\mathbf{r}}(\boldsymbol{\beta}, u)\}\, \mathrm{d}N_{i,\mathbf{r}}(u)$$

is a martingale at the true $\boldsymbol{\beta}_0$ (and hence has mean zero) *even though* $L_t(\boldsymbol{\beta}_0)$ is a product of *dependent* terms.

Derive asymptotic properties of estimators by Lenglart's inequality and the Martingale Central Limit Theorem.

Need the covarariate and intensity to be 'predictable' processes w.r.t $\mathcal{F}_t \uparrow$; suffices if they are left continuous and adapted (cannot use these techniques for CC).

Asymptotics

The normal limits of

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{-1})$$

and

$$W(\cdot) = \sqrt{n}\left(\hat{\Lambda}_0(\cdot\,; \hat{\boldsymbol{\beta}}) - \Lambda_0(\cdot)\right) + \sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^{\mathsf{T}} B(\cdot\,; \boldsymbol{\beta}_0)$$

are asymptotically independent, and have limiting variances which can be consistently estimated.

Performance Relative to Full Cohort, $\beta_0 = 0$.

NCCS: Sampling $m-1$ controls for each failure,

$$\text{ARE}_{\text{NCCS}} = (m-1)/m,$$

so

$$\text{ARE}_{\text{NCCS}} \to 1 \quad \text{as } m \to \infty.$$

Performance Relative to Full Cohort, $\beta_0 = 0$.

CM: Consider $|\mathcal{C}| = 2$ strata made up of those surrogate exposed $X = 1$, and those surrogate unexposed $X = 0$. With sensitivity and specificity of the surrogate X for the true Z given by

$$\tau = P(X = 1|Z = 1),\ \gamma = P(X = 0|Z = 0),$$

we have

$$\text{ARE}_{\text{CM}} = \tau\gamma + (1 - \tau)(1 - \gamma).$$

If τ, γ are close to 1 (or 0 !) then we have close to full cohort efficiency, e.g. $(\tau, \gamma) = (0.95, 0.90)$ gives

$$\text{ARE}_{\text{CM}} = 0.86.$$

Fixed time Model

Probability for failure of individual i in group $\mathcal{R} = \{1, \ldots, N\}$,

$$\tilde{p}_i = \frac{\lambda_0 x(\mathbf{Z}_i, \boldsymbol{\beta}_0)}{1 + \lambda_0 x(\mathbf{Z}_i, \boldsymbol{\beta}_0)}, \quad i \in \mathcal{R},$$

where $\mathbf{Z}_i$ is covariate vector of i and $\boldsymbol{\beta}_0$ is unknown parameter, $x(0, \boldsymbol{\beta}) = x(\mathbf{z}, 0) = 1$,$\lambda_0$ baseline odds.

Taking $\quad x(\mathbf{z}, \boldsymbol{\beta}) = \exp(\mathbf{z}'\boldsymbol{\beta}) \quad$ gives the logistic model.

Set of cases

$$D = \{i : D_i = 1\}.$$

Logisitic Analysis: MLE

With individuals independent and D the set of failures, can apply likelihood methods (MLE) on

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}) &= \prod_{i \in D} \frac{\lambda x(\mathbf{Z}_i, \boldsymbol{\beta})}{1 + \lambda x(\mathbf{Z}_i, \boldsymbol{\beta})} \prod_{i \notin D} \frac{1}{1 + \lambda x(\mathbf{Z}_i, \boldsymbol{\beta})} \\
&= \lambda^{|D|} x_D Q_{\mathcal{R}},
\end{aligned}$$

where, abbreviating $x_i = x(\mathbf{Z}_i, \boldsymbol{\beta})$,

$$x_{\mathbf{r}} = \prod_{i \in \mathbf{r}} x_i \quad \text{and} \quad Q_{\mathcal{R}} = \prod_{i \in \mathcal{R}} \frac{1}{1 + \lambda x_i}.$$

Sampling and Likelihood

Need to sample when N is large. For a given set of cases D, a sampled risk set E is chosen according to a given (design) distribution $\pi(E|D)$ with

$$\sum_{E \subset \mathcal{R}} \pi(E|D) = 1.$$

Given a sampled risk set E, the likelihood $P_{\boldsymbol{\beta}}(D|E)$ is

$$L(\boldsymbol{\beta}) \quad = \quad \frac{\lambda^{|D|} x_D \pi(E|D)}{\sum_{\mathbf{w} \subset \mathcal{R}} \lambda^{|w|} x_{\mathbf{w}} \pi(E|\mathbf{w})}.$$

Frequency Matching

Choose a sampled risk set E uniformly over all sets of some fixed size which contain the case set D of size η; sampling probabilities are constant and cancel, yielding the likelihood

$$L_E(\boldsymbol{\beta}) = P_{\boldsymbol{\beta}}(D|E) \quad = \quad \frac{x_D}{\sum_{\mathbf{w}\subset E, |\mathbf{w}|=\eta} x_{\mathbf{w}}}.$$

'Rejective sampling' studied by Hájek in 1964,

$$S_{E,\eta}(r) \quad = \quad \frac{x_r}{\sum_{\mathbf{w}\subset E, |\mathbf{w}|=\eta} x_{\mathbf{w}}}.$$

Simple Random Sampling as Rejective Sampling

Generally, sample a set $\mathbf{r}$ of size η from E proportional to the product of weights

$$x_{\mathbf{r}} = \prod_{A \in \mathbf{r}} x_A.$$

When all weights $x_A = t$ are equal, $x_{\mathbf{r}} = t^{\eta}$ and

$$S_{E,\eta}(\mathbf{r}) = \frac{x_{\mathbf{r}}}{\sum_{\mathbf{w} \subset E, |\mathbf{w}|=\eta} x_{\mathbf{w}}} = \binom{|E|}{\eta}^{-1}.$$

Analysis of Estimators

For convenience take

$$x(\boldsymbol{\beta}, \mathbf{z}) = \exp(\boldsymbol{\beta}'\mathbf{z}), \quad \text{which gives} \quad \frac{\partial x_A}{\partial \boldsymbol{\beta}} = \mathbf{z}_A x_A;$$

otherwise, define the 'effective covariate'

$$\mathbf{z}_A = \frac{\partial \mathbf{x}_A}{\partial \boldsymbol{\beta}} x_A^{-1}.$$

Score for β

The score function for $\boldsymbol{\beta}$ is

$$\mathcal{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \left(\log \frac{x_D}{\sum_{\mathbf{w} \subset E, |\mathbf{w}| = \eta} x_{\mathbf{w}}} \right).$$

Using

$$\frac{\partial x_A}{\partial \boldsymbol{\beta}} = \mathbf{z}_A x_A \quad \text{so that} \quad \frac{\partial \log x_A}{\partial \boldsymbol{\beta}} = \mathbf{z}_A,$$

and letting $p_A = \mathbf{E}(I_A|E)$ (rejective), score simplifies to

$$\mathcal{U}(\boldsymbol{\beta}) = \sum_{A \in D} \mathbf{Z}_A - \sum_{A \in E} \mathbf{Z}_A p_A.$$

Information for β

With $p_A = \mathbf{E}(I_A|E)$, and $p_{AB} = \mathbf{E}(I_A I_B|E)$ the second derivative of the score (information) $\mathcal{I}(\boldsymbol{\beta})$ is,

$$\sum_{A\in E} \mathbf{Z}_A \mathbf{Z}'_A p_A q_A + \sum_{A,B\in E, A\neq B} \mathbf{Z}_A \mathbf{Z}'_B (p_{AB} - p_A p_B).$$

Note presence of the rejective sampling correlation

$$\text{Corr}(A, B) = p_{AB} - p_A p_B.$$

Asymptotics: Consistency

We want

$$|E|^{-1}\mathcal{I}(\boldsymbol{\beta}_0) \xrightarrow{p} \Sigma,$$

that is

$$\frac{1}{|E|}\sum_{A\in E}\mathbf{Z}_A\mathbf{Z}'_A p_A q_A + \frac{1}{|E|}\sum_{A,B\in E, A\neq B}\mathbf{Z}_A\mathbf{Z}'_B(p_{AB} - p_A p_B),$$

to converge. So need second order correlation

$$p_{AB} - p_A p_B = O(|E|^{-1}) = O(|E|^{-(2+2\,\mathrm{mod}\,2)/2}).$$

Higher Order Correlations

$$\mathsf{Corr}(H) = \mathbb{E}\left(\prod_{A\in H}(I_A - p_A)\right).$$

Asymptotics: Consistency

We want third derivative (remainder term)

$$|E|^{-1}\mathsf{Rem} = O_p(1).$$

But $|E|^{-1}$ Rem term has

$$\frac{1}{|E|}\sum_{A,B,C} \text{Products of covariates} \times \text{third order correlation},$$

need

$$\mathsf{Corr}(A,B,C) = O(|E|^{-2}) = O(|E|^{-(3+3\,\mathsf{mod}\,2)/2}).$$

Follows if for rejective sampling we have

$$\mathsf{Corr}(H) = O(|E|^{-(|H|+|H|\,\mathsf{mod}\,2)/2}).$$

Simple Random Sampling

Sample η out of set E, all $\binom{|E|}{\eta}$ subsets equally likely. For $H = \{A, B\} \subset E$, inclusion indicators I_A, I_B; $p_A = p_B = \eta/|E|$, covariance

$$\mathbb{E}\,(I_A - p_A)(I_B - p_B) = \frac{-1}{|E|}\frac{\eta(|E| - \eta)}{|E|(|E| - 1)}$$

so when $\eta/|E| \to d$,

$$|E|\mathbb{E}\,(I_A - p_A)(I_B - p_B) \to d(d-1).$$

$$Corr(|H|) = O(|E|^{-(|H|+|H| mod 2)/2})$$

$$\begin{array}{rcl}
|E|^1 \mathsf{Corr}(2) & \rightarrow & d(d-1) \\
|E|^2 \mathsf{Corr}(3) & \rightarrow & 2d(d-1)(2d-1) \\
|E|^2 \mathsf{Corr}(4) & \rightarrow & 3d^2(d-1)^2 \\
|E|^3 \mathsf{Corr}(5) & \rightarrow & 20d^2(d-1)^2(2d-1) \\
|E|^3 \mathsf{Corr}(6) & \rightarrow & 15d^3(d-1)^3 \\
|E|^4 \mathsf{Corr}(7) & \rightarrow & 210d^3(d-1)^3(2d-1) \\
|E|^4 \mathsf{Corr}(8) & \rightarrow & 105d^4(d-1)^4 \\
|E|^5 \mathsf{Corr}(9) & \rightarrow & 2520d^4(d-1)^4(2d-1) \\
 & \cdots &
\end{array}$$

What is the next number in the sequence

$$1, 2, 3, 20, 15, 210, 105, 2520, \ldots ?$$

Two alternating sequences

$$1, 2, 3, 20, 15, 210, 105, 2520, \ldots ?$$

$$1, 3, 15, 105, \ldots,$$

$$2, 20, 210, 2520, \ldots,$$

Generating functions, induction, difference equations, or....
another of our most sophisticated techniques:

The On-Line Encyclopedia of Integer Sequences

Enter a ◉ sequence, ○ word, or ○ sequence number:

2,20,210,2520

Search | Restore example | **Clear** | **Hints** | **Advanced look-up**

OEIS in other languages: Chinese, Chinese (simplified), Czech, Finnish, French, German, Hindi, Hungarian, Italian, Portuguese, Russian, Serbian, Spanish, Swedish.

For information about the Encyclopedia see the **Welcome** page.

Lookup | **Welcome** | **Francais** | **Demos** | **Index** | **Browse** | **More** | **WebCam**
Contribute new seq. or comment | **Format** | **Transforms** | **Puzzles** | **Hot** | **Classics**
More pages | **Superseeker** | Maintained by **N. J. A. Sloane** (**njas@research.att.com**)

[Last modified Sun Sep 22 22:30:28 EDT 2002. Contains 75236 sequences.]

---

home | people | projects | research areas | resources |

Copyright © 2002 AT&T
All rights reserved
Read the AT&T Privacy Policy
Send comments about these web pages to Webmaster@research.att.com

Greetings from the On-Line Encyclopedia of Integer Sequences!

Matches (up to a limit of 30) found for 2 20 210 2520 :
[It will take a few minutes to search the table (the second and later lookups are faster)]

```
ID Number: A000906 (Formerly M2124 and N0841)
Sequence:  2,20,210,2520,34650,540540,9459450,183783600,3928374450,
           91662070500,2319050383650,63246828645000,1849969737866250
Name:      Expansion of 2(1+3x)/(1-2x)^7/2.
References L. Comtet, Advanced Combinatorics, Reidel, 1974, p. 256.
           F. N. David and D. E. Barton, Combinatorial Chance. Hafner, NY, 1962, p.
              296.
           Charles Jordan, On Stirling's Numbers, Tohoku Math. J., 37 (1933),
              254-278.
           C. Jordan, Calculus of Finite Differences. Budapest, 1939, p. 152.
See also:  Equals 2*A000457.
Keywords:  nonn
Offset:    0
Author(s): njas
```

Show internal format for above sequence? Yes

Copyright © 2002 AT&T
All rights reserved
Read the AT&T Privacy Policy
Send comments about these web pages to Webmaster@research.att.com

Rewriting Generating Function

Odd k^{th} order coefficient is

$$\frac{1}{3}(k-1)EZ^{k+1} \quad \text{for } Z \sim \mathcal{N}(0,1).$$

Even k^{th} order coefficient is

$$EZ^{2k} = 1 \cdot 3 \cdot (2k-3)(2k-1) \quad \text{for} \quad Z \sim \mathcal{N}(0,1).$$

The Normal Distribution is Here
But where is a sum of independent variables?

Conditioning

With T_θ the (Bernoulli sampling) measure under which $I_A, A \in E$ are independent with success probability

$$\tilde{p}_A = \theta x_A/(1 + \theta x_A),$$

then for all $\theta > 0$,

$$S_{E,\eta}(\mathbf{r}) = T_\theta(\{A \in E : I_A = 1\} = \mathbf{r} | \sum_{A \in E} I_A = \eta).$$

To make $\tilde{p}_A$ and p_A close, choose θ so that

$$\mathbb{E}\left(\sum_{A \in E} I_A\right) = \eta.$$

Local Central Limit Theorem

For X_n the sum $I_1, I_2, \ldots$ of independent indicators with $p_j = \mathbb{E}\, I_j$, when $\exists\ \epsilon > 0, n_\epsilon$ so that $\sum_{j=1}^n p_j q_j \geq \epsilon n$ for all $n \geq n_\epsilon$, with $\mathbb{E}\, \exp(itX_n) = \phi_n(t)$, let

$$m_\nu(s) = \sum_{j=0}^{s} \frac{(-i\nu)^j}{j!} \mathcal{I}_{n,j},$$

where

$$\mathcal{I}_{n,j} = \frac{1}{2\pi} \int_{|t| \leq \pi} t^j \phi_n(t) dt.$$

Then for given κ and even s, for all $|\nu| \leq \kappa$ with $\mathbb{E}\, X_n + \nu \in \mathbf{N}$

$$P(X_n = \mathbb{E}\, X_n + \nu) = m_\nu(s) + \Theta_{\epsilon,\kappa,s}\left(n^{-(s+3)/2}\right).$$

Conditioning + LCLT

Relate p_A and $\tilde{p}_A$ by conditioning Bernoulli sample T,

$$
\begin{aligned}
p_A &= \frac{P(A \in T, |T| = \eta)}{P(A \in T, |T| = \eta) + P(A \notin T, |T| = \eta)} \\
&= \frac{\tilde{p}_A P(|T \setminus A| = \eta - 1)}{\tilde{p}_A P(|T \setminus A| = \eta - 1) + \tilde{q}_A P(|T \setminus A| = \eta)}.
\end{aligned}
$$

Applying LCLT expansion gives rates (and a more careful analysis for simple random sampling, 34650).

Frequency matching under the Null

Probability of failure is d. Full cohort has Information

$$d(1-d)\mathsf{Var}(Z).$$

Frequency matching $m-1$ controls for each failure has Information

$$\frac{(m-1)d}{m}\mathsf{Var}(Z).$$

$$\mathsf{ARE}_{\mathsf{FM}} = \frac{m-1}{m(1-d)}.$$

E.g. $d = .1, m = 5$, Matching/Full = 0.8/0.9 = 89%.

Counter Matching

Binary covariate Z, binary surrogate X. With 1:1 counter matching, so the sampled cohort has same number of surrogate exposed as unexposed,

$$\pi_{ij} = P(X = i, Z = j) \quad \text{and}$$

$$\tau = P(X = 1|Z = 1),\ \gamma = P(X = 0|Z = 0),$$

relative efficiency to Full Cohort is

$$\begin{aligned} \{\tau\gamma \quad &+ \quad (1-\tau)(1-\gamma)\} \\ &- \quad d\frac{\pi_{.1}\pi_{.0}}{\pi_{0.}\pi_{1.}}(\tau\gamma - (1-\tau)(1-\gamma))^2. \end{aligned}$$

Further Questions

1. Analysis of counter matching design away from the null, asymptotics.
2. Other designs in logistic type models.
3. Group time models.
4. And . . . ??

Further Questions

1. Analysis of counter matching design away from the null, asymptotics.
2. Other designs in logistic type models.
3. Group time models.
4. Other problems from the in flight magazine.