# Inference on the AUC with clustered data

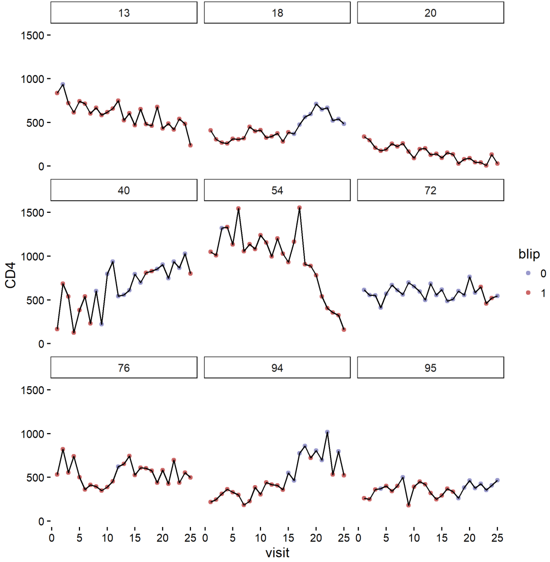
Haben Michael

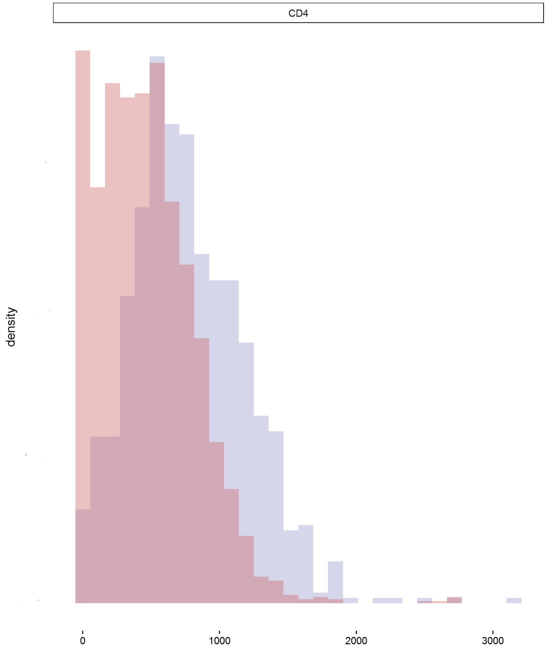
October 2021

# A motivating example

Data: The Yale Prospective Longitudinal HIV Cohort

Problem: Evaluate CD4 as a predictor of blip status




CD4
density
0
1000
2000
3000

| | control | case |
|---|---|---|
| obs. # 1 | $X_1$ | |
| ⋮ | ⋮ | |
| obs. # k | $X_k$ | |
| obs. # k+1 | | $Y_{k+1}$ |
| ⋮ | | ⋮ |
| obs. # N | | $Y_N$ |

The AUC is the probability that an observation drawn from a negative/control/non-diseased subject is less than an independent observation from a positive/case/diseased subject.

$$\text{AUC} = P(X < Y) = E(F_X(Y))$$

$$\widehat{\text{AUC}} = \frac{1}{k(N-k)} \sum_{i,j} \{X_i < Y_j\}$$

| | control | case |
|---|---|---|
| obs. # 1 | $(X_{11}, \ldots, X_{1m_1})$ | $(Y_{11}, \ldots, Y_{1n_1})$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| obs. # N | $(X_{N1}, \ldots, X_{Nm_N})$ | $(Y_{N1}, \ldots, Y_{Nn_N})$ |

- Assume iid observations. Denote AUC between cluster $i$ controls and cluster $j$ cases as

$$\phi_{ij} = \frac{1}{m_i n_k} \sum_{j=1}^{m_i} \sum_{l=1}^{n_k} \{X_{ij} < Y_{kl}\} \qquad \phi = \begin{pmatrix} \phi_{11} & \cdots & \phi_{1N} \\ \vdots & \vdots & \vdots \\ \phi_{N1} & \cdots & \phi_{NN} \end{pmatrix}$$

- Generalize AUC to clustered data as

$$\theta = \text{AUC} = E(\phi_{ij}), i \neq j$$

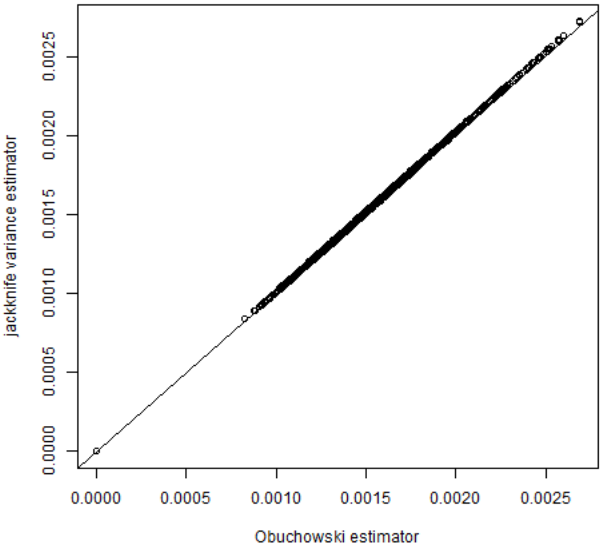
$$\hat{\theta} = \widehat{\text{AUC}} = \overline{\phi}_{..} = \frac{1}{N^2} \sum_{i,j} \phi_{ij}$$

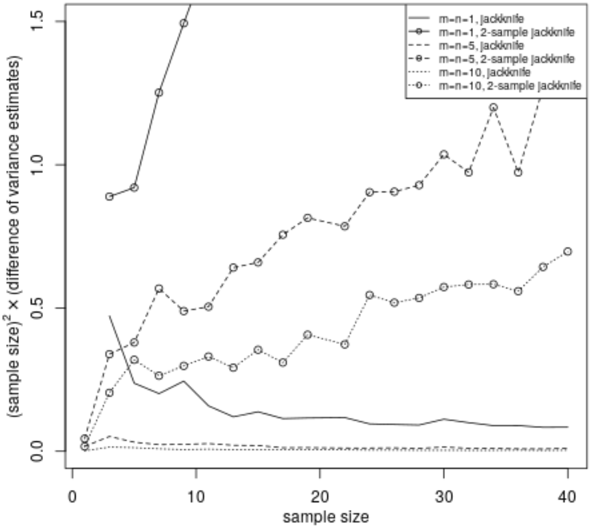
- markers: longitudinal measurements of tumour antigens (CEA, CA15-3, TPS) as markers, response: progression/non-progression of breast cancer (Emir 2000)
- markers: two measurements of the distortion product otoacoustic emissions taken from the left and right eyes ((think this should be ears)) of each patient, response: neonatal hearing impairment (Wu 2019)
- markers: longitudinal measurements of levels of vascular enothelial growth factor and a soluble fragment of Cytokeratin 19, response: progression/non-progression of non-small cell lung cancer (Wu, Wang 2011)

- Fix cluster sizes $m_i = m, n_j = n, 1 \leq i, j \leq N$
- Obuchowski '97 estimates the variance of $\overline{\phi}_{..} = \widehat{\text{AUC}}$ as

$$
\begin{aligned}
\hat{\sigma}^2_{obu} &= \frac{1}{N(N-1)} \sum_i (\overline{\phi}_{i.} + \overline{\phi}_{.i} - 2\overline{\phi}_{..})^2 \\
&= \frac{1}{N} \widehat{var}(\overline{\phi}_{i.} + \overline{\phi}_{.i}) \\
(\overline{\phi}_{..} - E(\overline{\phi}_{..}))/\hat{\sigma}_{obu} &\rightsquigarrow N(0,1)
\end{aligned}
$$

- Alternatively: Use the bootstrap. Sample the iid clusters without replacement, compute $\overline{\phi}_{..}$ on the sample, repeat. Take the sample variance of the resulting $\overline{\phi}_{..}$'s.


jackknife variance estimator
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
Obuchowski estimator



m=n=1, jackknife
m=n=1, 2-sample jackknife
m=n=5, jackknife
m=n=5, 2-sample jackknife
m=n=10, jackknife
m=n=10, 2-sample jackknife
(sample size)$^2$ × (difference of variance estimates)
sample size
0.0
0.5
1.0
1.5
0
10
20
30
40

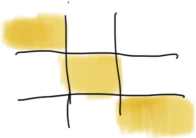
Objective is

$$
\begin{aligned}
&\propto \frac{2N-1}{N^2}\sum_j(\overline{\phi}_{j\cdot}+\overline{\phi}_{\cdot j}-2\hat{\theta})^2\\
&-\frac{2}{N}\sum_j\left(\overline{\phi}_{j\cdot}+\overline{\phi}_{\cdot j}-2\hat{\theta}\right)\left(\phi_{jj}-\frac{\mathrm{tr}(\phi)}{N}\right)\\
&+\text{lower order terms}
\end{aligned}
$$

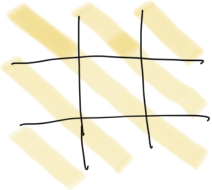
- When $m = n = 1$, clusters can be ordered by $y_i$ ("case") observations
- $\phi_{ij} = \{x_i < y_j\}$, e.g., $\phi = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$
- $a_i := \phi_{i\cdot}$ the row sums, $0 \le a_i \le N$ $F_a(x) = \sum_{i=1}^{N} \{a_i \le x\}$ the observed CDF of the $a_i$ objective is

$$\begin{aligned}
&\propto \frac{1}{N^2} \sum_i a_i^2 + \frac{2}{N} a_i (1 - F_a(N - i)) \\
&+ \sum_i (1 - F_a(N - i))^2 + \frac{2}{N^2} (\sum_i a_i)(\sum_i (1 - F_a(N - k))) \\
&- \sum_i (1 - F_a(N - k))\{a_i > N - i\} - \frac{4}{N^3} (\sum_i a_i)^2 \\
&- \frac{1}{N} \sum_i a_i \{a_i > N - i\}
\end{aligned}$$

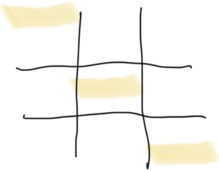
View the objective as a quadratic form $Q$ in the $N^2$ entries of $\phi$



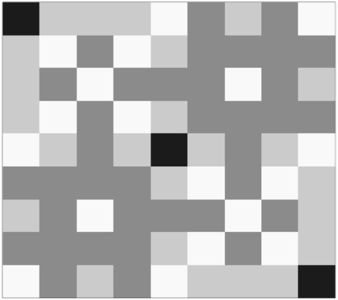
$\sum \phi_{i\cdot}^2$



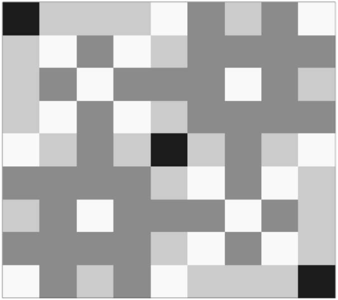
$\sum \phi_{\cdot j}^2$



$-\sum \phi_{ii}\phi_{i\cdot}$



- $N^2$ $N \times N$ blocks
- $Q_{pqrs} = \frac{1}{N^2}(\{p=q\}+\{r=s\}+\{q=r\}+\{p=s\}+\{s=q\}+\{p=r\}) - \frac{1}{2N}(\{p=q=r\}+\{p=q=s\}+\{p=r=s\}+\{q=r=s\}) - \frac{4}{N^3}$ is symmetric in $p, q, r, s$
- $Q$ is symmetric about the diagonal, anti-diagonal, and $180^\circ$ rotations



- $Q = (P \otimes P)^t Q (P \otimes P)$ for a permutation matrix $P$ (shuffling iid clusters)

- characteristic polynomial $x^{N^2-2(N-1)}(2x^2 - \frac{N-2}{N^2})^{N-1}$
- $Q^{2k+1} = \lambda^{2k} Q$ and $Q^{2k} = \lambda^{2(k-1)} Q^2$
- $Q$ is the scaled difference of two orthogonal projection matrices onto two $N-1$ dimensional subspaces

$$Q = \lambda(Q_1 - Q_2)$$

(aside) Relate the traces of powers $t_j = \mathrm{tr}(Q^j)$ to the coefficients $c_j$ of the characteristic polynomial

$$\begin{aligned}
c_1 &= t_1 \\
c_2 &= \frac{1}{2}(t_1^2 - t_2) \\
c_3 &= -\frac{1}{6}t_1^3 + \frac{1}{2}t_1 t_2 - \frac{1}{3}t_3 \\
&\vdots
\end{aligned}$$

With our traces and coefficients,

$$-\frac{N-1}{k} + \frac{(N-1)^2}{2!}\sum_{i_1+i_2=k}\frac{1}{i_1 i_2} - \frac{(N-1)^3}{3!}\sum_{i_1+i_2+i+3=k}\frac{1}{i_1 i_2 i_3} + \dots$$

$$+(-1)^{k-1}\frac{(N-1)^{k-1}}{(k-1)!} = (-1)^{N+k}\binom{N-1}{k}$$

(aside) i.e. $(-1)^k\binom{N-1}{k}$ is the coefficient of $x^k$ in

$$\sum_{m=1}^{N-1}(-1)^m\frac{(N-1)^m}{m!}(x+x^2/2+x^3/3+\ldots+\frac{x^{N-1}}{N-1})^m.$$

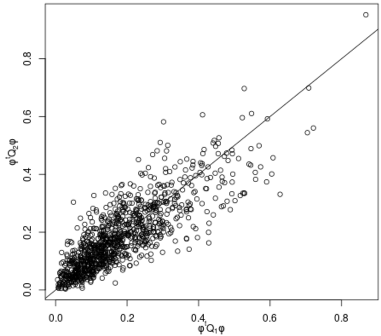
$$\sum_{i_1+\ldots+i_m}\frac{1}{i_1\cdot\ldots\cdot i_m}=\frac{|s_m^k|}{k!}$$

with $|s_m^k|$ the Stirling number of the first kind

$$x^{\underline{k}}=\frac{x!}{(x-k)!}=\sum_{m=1}^{k}s_m^k x^m$$

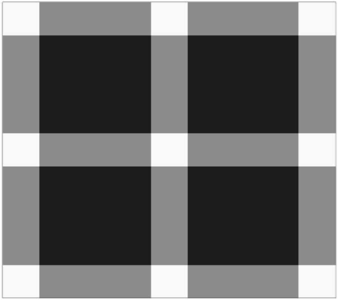
e.g.

$$x^{\underline{4}}=x(x-1)(x-2)(x-3)=x^4-6x^3+11x^2-6x$$




- nullspace has dimension $N^2 - 2(N-1)$
- need $Q$-null vectors $v$, $(v, Qv) = 0$, and relate these to the structure of $\phi$

(Aside)



- Element-wise ratio $Q_2/Q_1$ only takes 3 values: $-1, r_N, 1/r_N$

| $N$ | $r_N$ |
|---|---|
| 3 | $17+12\sqrt{2}=\epsilon_0(2)^4$ |
| 4 | *Inf* |
| 5 | $49+20\sqrt{6}=\epsilon_0(6)^2$ |
| 6 | $17+12\sqrt{2}=\epsilon_0(2)^4$ |
| 7 | 19.7270... |
| 8 | $7+4\sqrt{3}=\epsilon_0(3)^2$ |
| 9 | 10.8679... |
| 10 | 9 |
| 12 | $7/2+3/2\sqrt{5}=\epsilon_0(5)^4$ |
| 20 | 4 |

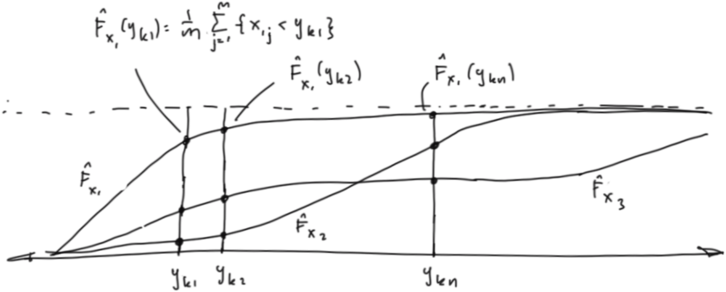
Mutually $Q$-orthogonal vectors include (vectorizations of)

- constant row/column $\phi$ e.g. $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$
- constant diagonal $\phi$ e.g. $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
- upper triangular $\phi$ in $N(Q)$ $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$

i.e., stack bases for the above to form a matrix $B$, then $B^t QB = 0$

- structure of matrix given by $\phi_{ij} = \mathbb{E} F_{X_j}(Y_j)$
- can't order $y$ (case) clusters when $n > 1$
- but $\phi$ is a mixture of totally ordered vectors $0 \leq v_1 \leq v_2 \leq \ldots \leq v_n \leq m$

$$\phi = \frac{1}{mn} \begin{pmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_{Nn} \\ \vdots & \vdots & & \vdots \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & & 1 \end{pmatrix}}_{\text{col sums} = n} \text{row sums} = 1$$




$$\frac{1}{n}\sum_{l=1}^{n} \hat{F}_{x_1}(y_{kl}) = \frac{1}{mn}\sum_{l=1}^{n}\sum_{j=1}^{m}\{x_{1j} < y_{kl}\} = \varphi_{1k}$$

$mn\phi =$

$$\text{row sums}{=}m \underbrace{\begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 0 \end{pmatrix}}_{\text{col sums}=1} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & & 1 \end{pmatrix}}_{\text{col sums}=n} \text{row sums}{=}1$$

If the 1st and 3rd matrices are permutation matrices, $\phi$ is in the nullspace of $Q$

- Project $\phi$ onto a subset of the $Q$-null vectors and examine residual
- Let $P = (1, \ldots, N)$ and $\overline{\phi}_{\cdot p} = (\overline{\phi}_{\cdot 1}, \ldots, \overline{\phi}_{\cdot N})$

$$Var(\overline{\phi}_{\cdot p}) - Cov^2(\overline{\phi}_{\cdot p}, \frac{P}{sd(P)})$$

- $D(x) = |Cov((x_{(1)}, \ldots, x_{(N)}), \frac{P}{sd(P)})|$ can be viewed as a measure of statistical dispersion of $(x_1, \ldots, x_N)$

$$D(ax + b) = |a|D(x) + b$$

- $Var(x) - D^2(x) \geq 0$, equals 0 when $x \propto (1, \ldots, N)$ i.e., $(x_1, \ldots, x_N)$ are evenly spaced
- conjecture it is maximized for $0 \leq x \leq 1$ when $x = (0, 0, \ldots, 0, 1)$, where the value is $O(1/N)$
-
    - Maximizing a postive semi-definite function over a polytope $0 \leq x_1 \leq \ldots \leq x_N \leq 1$
    - Maximum occurs at a corner point given by the intrsection of $N$ hyperplanes
    - $N$ of the $N + 1$ restrictions $0 \leq x_1, x_1 \leq x_2, \ldots, x_{N-1} \leq x_N, x_N \leq 1$ are active
    - Only need to check $x$ of the form $(0, \ldots, 0, 1, \ldots, 1)$.

The difference between the jackknife and Obuchowski variance estimates is $O(1/N^2)$ when $\phi$ is lower right triangular

Future work

- stochastic result needed or sharpen constants
- varying cluster sizes