Oddity #5 Increase sample size; StdErr($\bar{X}$) increases!

Actual problem: Alex B was measuring the effect on gum tissue of a particular method of doing a crown preparation. This was a pilot dataset

- Upper right first molar of volunteers:
  - make a cast of tooth and gum before crown prep
  - do crown prep; wait a little while
  - make a second cast of tooth and gum
  - make digital 3-D images of "before" and "after" casts,
  - align digital images using fixed surface of tooth
  - compute change in gum height, (after) - (before)
- This measurement is, for practical purposes, without error
- Alex B considered a 46.5 mm length of gum between two landmarks.
- Design Question: At how many locations in this 46.5 mm length should Alex measure?



4/19/08, rev 3/6/12

→

Design Question: At how many locations in this 46.5 mm length should Alex measure?

Other facts: - Comparing crown prep vs. no prep
⇒ comparison is between persons

- Although measurements are, in effect, without error, they are costly in time.
- Alex provided a dataset with a few teeth (?10?) measured before and after at (?) 11 locations
- I fit a bunch of spatial models to this dataset and ended up with this model:

Change at location s in person i $\quad y_{si} = \mu_i + \varepsilon_s \qquad s = \left(\frac{i-1}{n-1}\right) 46.25$

$i = 1, \ldots, n$

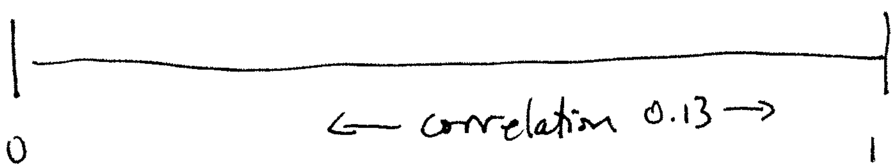
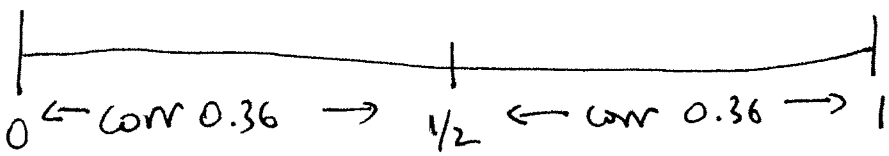
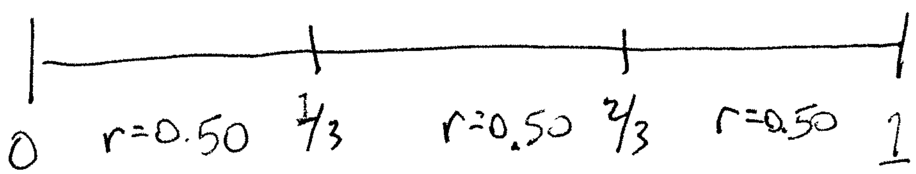
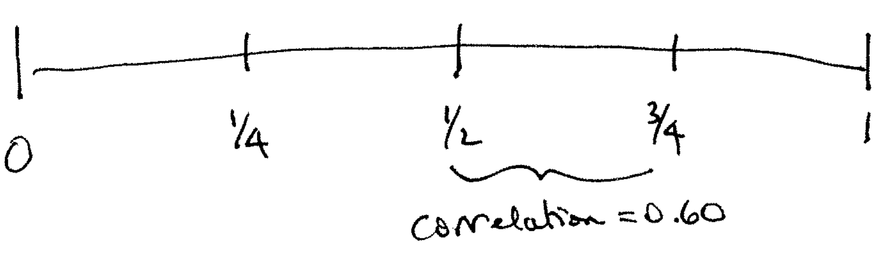
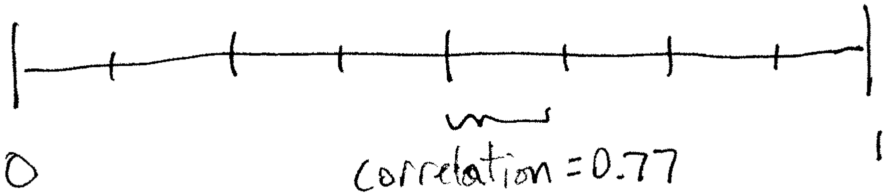
$$\mathrm{cov}(\underset{\sim}{\varepsilon}) = \sigma^2 \left(\left( \exp(-d_{ij}/22.5) \right)\right) \qquad \sigma^2 = 95^2 \ \mu m$$

$d_{ij}$ = distance (mm) between two measurements = $[46.25(i-j)/(n-1)]$ mm.

Very important: As you take more measurements (as n increases), adjacent measurements are closer together

II E/1B 4/19/08 →

- Because I like to do dumb simple things for power calculations, I chose to use $\bar{y}_{i,n}$ as the summary of the $n$ measures from subject $i$.
- Easy to show: $Var(\bar{y}_{i,n}) = \frac{\sigma^2}{n^2}\left(n + 2\sum_{i>j} \exp\left(\frac{-46.25}{22.5}\frac{(i-j)}{n-1}\right)\right)$

| Design | # measures | $Var(\bar{y}_{i,n})$ |
|---|---|---|
|   | 2 | $0.560\sigma^2$ |
|   | 3 | $0.5208\sigma^2$ |
|   | 4 | $0.5185\sigma^2$ ↕ !?! |
|   | 5 | $0.5218\sigma^2$ |
|   | 9 | $0.5345\sigma^2$ !!!??! |

- This result does not depend on the choice of constants (22.5, 46.25)
- It is not specific to this correlation function $e^{-d/\theta}$
  - I got the same qualitative result using:
    - $\text{corr}(i,j) = e^{-d_{ij}^2/\theta^2}$
    - $\text{corr}(i,j) = 1 - d_{ij}/46.25$
- Unable to find any errors in my work, I asked several colleagues and was directed to

Morris MD, Ebey SF (1984) The Amer. Stat. 38: 127-129

which proves this result for the covariance function

$\text{cov}(y_s, y_{s'}) = \sigma^2 e^{|s'-s|}$

- They also show that if you leave out $y_0$, this result no longer holds, i.e. $\text{Var}(\bar{X}_n)$ decreases monotonically in $n$, though as $n$ becomes large $\text{Var}(\bar{X}_n)$ becomes effectively flat (Hoel PG (1961), Ann. Math. Stat. 32: 1042-1047).
- Jargon (e.g. N. Cressie's Spatial book): "Infill Asymptotics"

II E / 20 4/19/08

# This is not just an oddity – it could happen to you

Example:

- You are designing a clinical trial comparing two groups.
- You will take measurements at times 0 and 12 months.
- Design decision: Should you take a measurement at 6 months?
- Fact: If the within-person correlation is high enough, the SE of the group main effect ***increases*** if you use the 6 month measurement.

# This is not just an oddity (continued)

Each group has $n$ subjects; error variance of one measurement is $\sigma^2$.

Person $i$ in group 1:

$\text{Corr}(X_{1i,0}, X_{1i,12}) = \rho$, $\text{Corr}(X_{1i,0}, X_{1i,6}) = \text{Corr}(X_{1i,6}, X_{1i,12}) = \sqrt{\rho}$

Person $i$ in group 2: same model.

Including the 6-mo measurement: $\text{var}(\bar{X}_{1.} - \bar{X}_{2.}) = \frac{2\sigma^2}{9n}(3 + 4\sqrt{\rho} + 2\rho)$

Excluding the 6-mo measurement: $\text{var}(\bar{X}_{1.} - \bar{X}_{2.}) = \frac{2\sigma^2}{4n}(2 + 2\rho)$

The variance of the group main effect is

lower using the 6-mo measurement for $\rho < 0.36$

*higher* using the 6-mo measurement for $\rho \geq 0.36$.

The ratio var(with)/var(without) peaks at $\rho \approx 0.61$.

The rest of this lecture is taken from Lavine ML, Hodges JS

"An Old Curiosity, Some Intuition for It, and a Modestly Interesting Implication"

which is under review at *The American Statistician*.

# More examples, from LH

Variance of $\bar{\mu}$ with $\sigma^2 = 1$ and $\mathrm{Cov}[Y_t, Y_{t'}] = \rho^{|t-t'|}$, with equally-spaced measurement locations.

| $\rho$ | 2 | 3 | 4 | 5 | 10 | 20 | 30 | 40 | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|
| .001 | .501 | .348 | .290 | .265 | .241 | .241 | .242 | .243 | .248 |
| .010 | .505 | .380 | .344 | .331 | .325 | .331 | .334 | .335 | .341 |
| .100 | .550 | .496 | .490 | .492 | .506 | .516 | .520 | .522 | .529 |
| .400 | .700 | .703 | .712 | .719 | .735 | .744 | .747 | .749 | .753 |

Why does this happen? What's the intuition?

# Intuition (Michael Lavine)

Due to autocorrelation, an observation $Y_t$ at time $t$ provides some information about the process at times *surrounding* $t$.

When enough observations have been taken in the fixed interval $[0, 1]$, adding observations in $[0, 1]$ provides little more information about the process in that interval.

But the observations at $t = 0$ and $t = 1$ also provide information about the process *outside* of $[0, 1]$.

As more observations are added inside $[0, 1]$, and all observations are given equal weight, the information about the process *outside* $[0, 1]$ becomes diluted to an extent that outweighs the additional information from *inside* $[0, 1]$, which in turn causes $\text{Var}(\bar{\mu})$ to increase.

# Support for the intuition, part 1: Optimal weights

For fixed $n$ and equally spaced measurements, consider estimators

$$\hat{\mu} = \sum_{i=1}^{n} w_i Y_i$$

where $\sum_{i=1}^{n} w_i = 1$ but the $w_i$'s are not necessarily equal.

This is the model $\mathbf{y} = \mathbf{1}_n \mu + \epsilon$, where $\text{Cov}(\epsilon) = \sigma^2 C$, $\mathbf{1}_n$ is the $n$-vector of 1's, and $C$ is the correlation matrix of $(Y_1, \ldots, Y_n)$.

What vector of weights $w = (w_1, \ldots, w_n)$ minimizes $Var(\hat{\mu})$?

$$w = k\mathbf{1}^T C^{-1}$$

where $k = (\mathbf{1}_n' C^{-1} \mathbf{1}_n)^{-1}$ is a scalar constant.

# Support for the intuition, part 1: Optimal weights

For equally-spaced observations from an AR(1) process,

$$\text{Corr}[Y_t, Y_{t'}] = \rho^{|t-t'|}$$

and the correlation matrix $C$ has a simple closed-form inverse.

Given $n$, the correlation between adjacent observations is $\gamma = \rho^{\frac{1}{n-1}}$.

The optimal weights are

$w_1 = w_n = k(1-\gamma)$ and

$w_2 = w_3 = \cdots = w_{n-1} = k(1 - 2\gamma + \gamma^2) = k(1-\gamma)^2$.

The ratio of an endpoint weight to an interior weight is $1/(1-\gamma)$, so

- an endpoint carries more information than an interior point, and
- the effect is stronger as $\gamma$ increases.

# Intuition, part 2: Unequally-spaced observations

How does unequal spacing of observations affect Var($\bar{Y}$)?

To study this, locate internal points at quantiles of a *Beta*$(\alpha, \alpha)$.

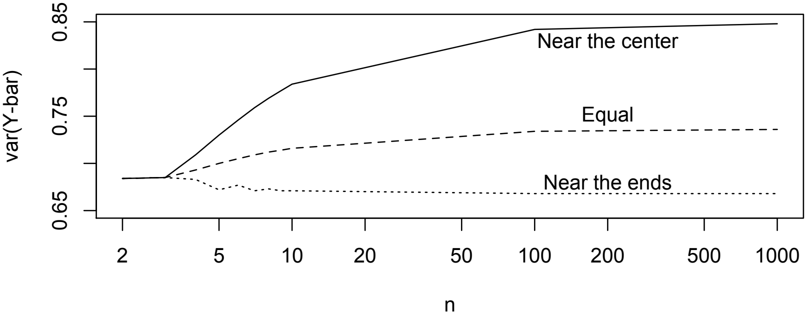
For a given *n*,

- $\alpha < 1$ gives locations that are more closely spaced near the unit interval's endpoints and more distantly spaced at the center,
- $\alpha = 1$ gives equally spaced locations, and
- $\alpha > 1$ gives locations more closely spaced near the interval's center.

# Intuition, part 2: Unequally-spaced observations

Variance of $\bar{Y}$ for $\sigma^2 = 1$, $\text{Cov}[Y_t, Y_{t'}] = \rho^{|t-t'|}$, $\rho = \exp(-1) \approx 0.37$.

Measurement locations are quantiles of $Beta(\alpha, \alpha)$, as indicated.




"Near the ends" diminishes the info loss near the ends of the intervals.

“A modestly interesting implication” (produced by M. Lavine):

MCMC draws are autocorrelated, so iterations after burn-in are like Morris & Ebey's interval $[0, 1]$.

Is it possible to get *smaller* MCMC variance for estimating $\mathrm{E}(h(X)|\mathrm{data})$ by dropping every second observation?

Yes: ML produced an example with an extremely high one-lag autocorrelation.

But the size of the effect is quite small $\Rightarrow$ no practical implication.