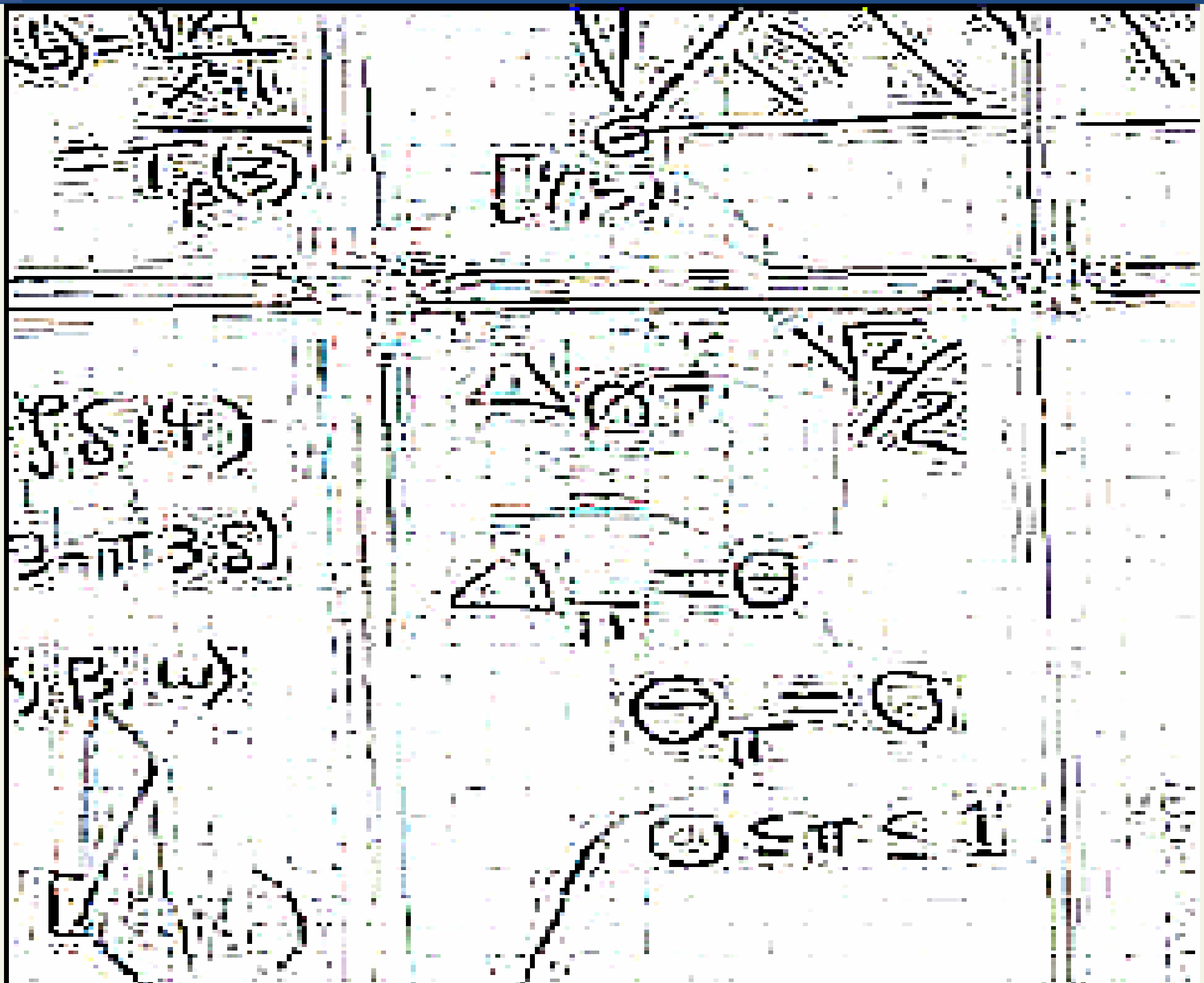
THE WEIGHTING GAME

Ciprian M. Crainiceanu
Thomas A. Louis
Department of Biostatistics

http://commprojects.jhsph.edu/faculty/bio.cfm?F=Ciprian&L=Crainiceanu

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Getting rid of the superfluous information



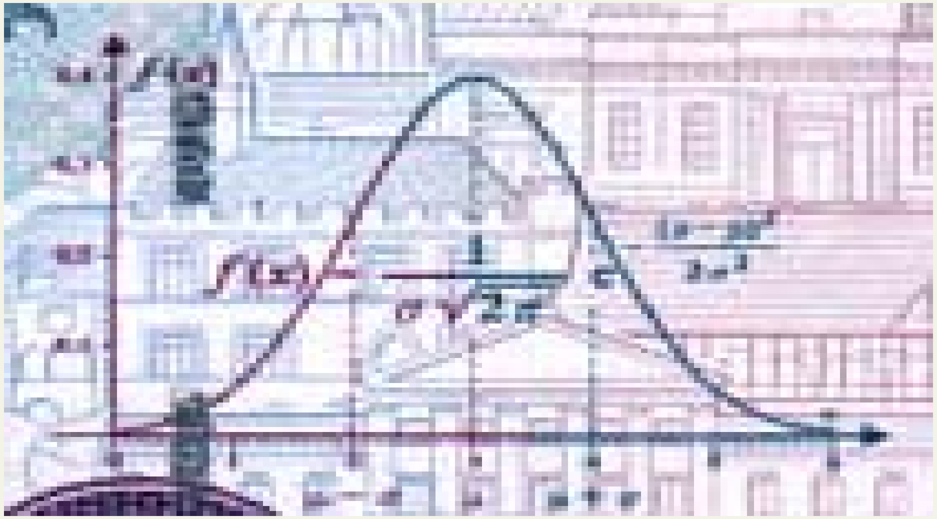
How the presentation could have started, but didn't: Proof that statisticians can speak alien languages

- Let (Ω,K,P) be a probability space, where (Ω,K) is a measurable space and

  P: K ↦ [0,1]

  is a probability measure function from the σ-algebra K. It is perfectly natural to ask oneself what a σ-algebra or σ-field is.

  Definition. A σ-field is a collection of subsets K of the sample space Ω with



$$\Phi \in \mathrm{K}\,,\ A \in \mathrm{K} \Rightarrow A^{C} \in \mathrm{K}\,,\ A_1, A_2, \ldots \in \mathrm{K} \Rightarrow \bigcup A_i \in \mathrm{K}$$

Of course, once we mastered the σ-algebra or σ-field concept it is only reasonable to wonder what a probability measure is

Definition. A probability measure P has the following properties

Where do all these fit in the big picture?

Every sample space is a particular case of probability space and weighting is intrinsically related to sampling

$$P(\Omega) = 1,\ \ A_1, A_2, \ldots \in \mathrm{K} \text{ disjoint sets} \Rightarrow P\left(\bigcup A_i\right) = \sum P\left(A_i\right)$$

Why simple questions can have complex answers?

Question: What is the average length of in-hospital stay for patients?

Complexity: The original question is imprecise.

New question: What is the average length of stay for:

- Several hospitals of interest?
- Maryland hospitals?
- Blue State hospitals? …

"Data" Collection & Goal

Survey, conducted in 5 hospitals

- Hospitals are selected
- $n_{hospital}$ patients are sampled at random
- Length of stay (LOS) is recorded
- **Goal:** Estimate the population mean

Procedure

- Compute hospital specific means
- "Average" them
  - For simplicity assume that the population variance is known and the same for all hospitals
- **How should we compute the average?**
- Need a (good, best?) way to **combine information**

DATA

Hospital	# sampled n_{hosp}	Hospital size	% of Total size: $100\pi_{hosp}$	Mean LOS	Sampling variance
1	30	100	10	25	$\sigma^2/30$
2	60	150	15	35	$\sigma^2/60$
3	15	200	20	15	$\sigma^2/15$
4	30	250	25	40	$\sigma^2/30$
5	15	300	30	10	$\sigma^2/15$
Total	150	1000	100		

Weighted averages

Examples of various weighted averages: $\sum w_i \overline{X}_i$

Weighting strategy	Weights x100	Mean	Variance Ratio
Equal	20 20 20 20 20	25.0	130
Inverse variance	20 40 10 20 10	29.5	100
Population	10 15 20 25 30	23.8	172

Variance using inverse variance weights is smallest

What is weighting? (via Constantine)

- Εσσενχε: α γενεραλ ωαψ οφ χομπυτινγ αϖεραγεσ
- Τηερε αρε μυλτιπλε ωειγητινγ σχηεμεσ
- Μινιμιζε ϖαριανχε βψ υσινγ ινϖερσε ϖαριανχε ωειγητσ
- Μινιμιζε βιασ φορ τηε ποπυλατιον μεαν
- Πολιχψ ωειγητσ

What is weighting?

- **The Essence**: a general (fancier?) way of computing averages
- There are multiple weighting schemes
- Minimize variance by using inverse variance weights
- Minimize bias for the population mean by using population weights (“survey weights”)
- Policy weights
- “My weights,” ...

Weights and their properties

- Let (μ_1, μ_2, μ_3, μ_4, μ_5) be the TRUE hospital-specific LOS
- Then $\sum w_i \bar{X}_i$ estimates $\sum w_i \mu_i$
- If $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_\pi = \Sigma\, \mu_i\, \pi_i$ **ANY** set of weights that add to 1 estimate μ_π .
- So, it's best to minimize the variance

- But, if the TRUE hospital-specific E(LOS) are not equal
  - Each set of weights estimates a different target
  - Minimizing variance might not be "best"
  - An unbiased estimate of μ_π sets $\boldsymbol{w_i = \pi_i}$

General idea

Trade-off variance inflation & bias reduction

Mean Squared Error

General idea

Trade-off variance inflation & bias reduction

MSE = Expected**(Estimate - True)^2**

= **Variance + Bias^2**

- Bias is unknown unless we know the μ_i (the true hospital-specific mean LOS)
- But, we can study MSE (μ, w, π)
- Consider a **true** value of the variance of the between hospital means
- Study BIAS, Variance, MSE for various **assumed** values of this variance

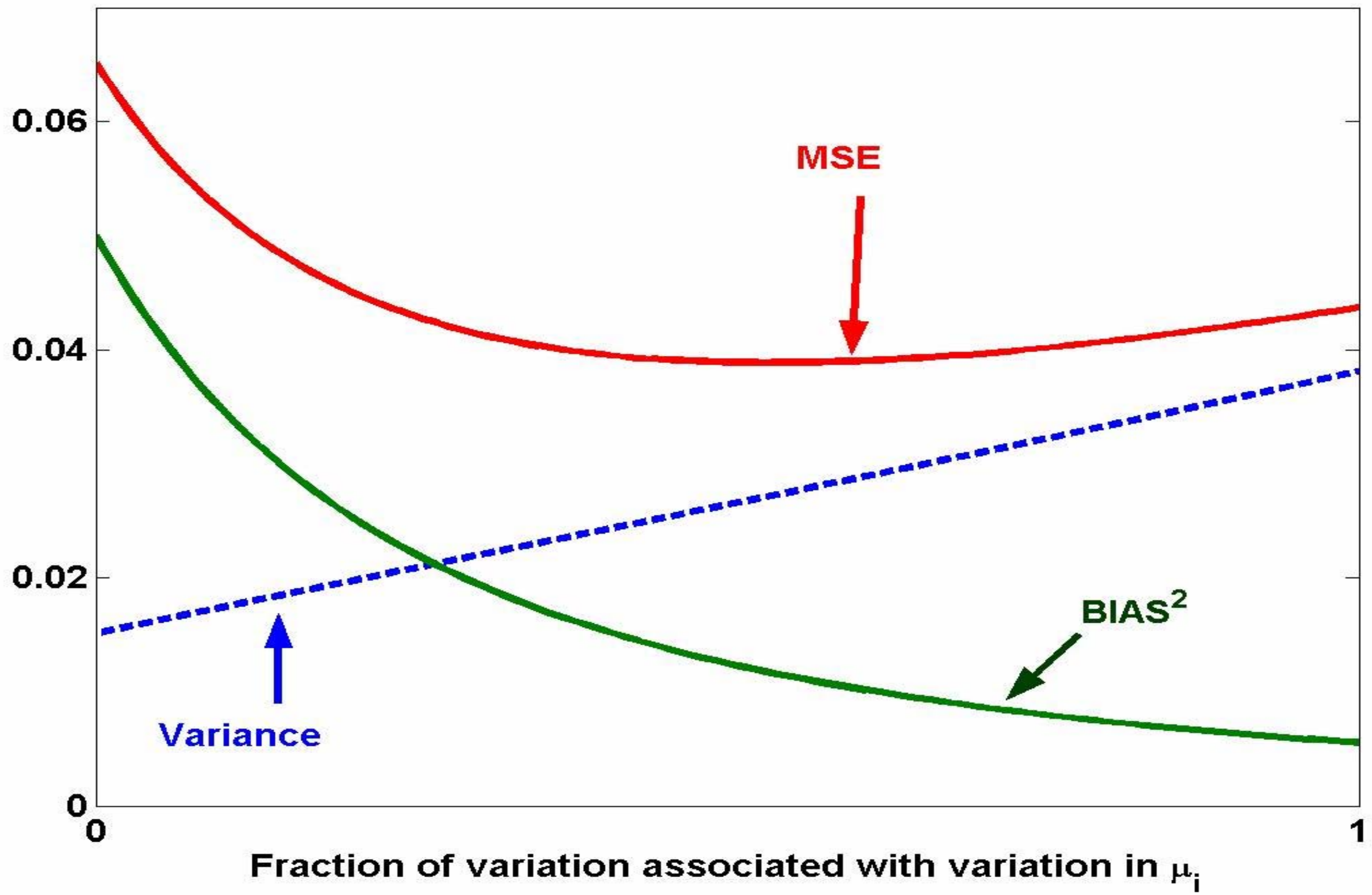
Mean Squared Error

- Consider a **true** value of the variance of the between hospital means

$$\mathbf{T} = \Sigma(\mu_i - \mu^*)^2$$

- Study BIAS, Variance, MSE for optimal weights based on **assumed** values (**A**) of this variance
- When **A = T,** MSE is minimized
- Convert T and A to fraction of total variance

The bias-variance trade-off
X is assumed variance fraction
Y is performance computed under the true fraction



Summary

- Much of statistics depends on weighted averages
- Choice of weights should depend on assumptions and goals
- If you **trust** your (regression) model,
  - Then, minimize the variance, using "optimal" weights
  - This generalizes the equal μs case
- If you **worry** about model validity (bias for μ_π),
  - Buy full insurance, by using population weights
  - You pay in variance (efficiency)
  - Consider purchasing only what you need

Using compromise weights

Statistics is/are everywhere!

EURO: our short wish list