Statistical CSI: An Attempt to Detect Fraud in Papers Published From a Medical Biochemistry Department

Mark Hudes, PhD
Senior Statistician
Children's Hospital Oakland

Assumption:
The observed data are normally distributed with finite mean μ and finite variance σ^2.

The population CV (expressed below as a proportion), CV_{pop}, would then be

$$CV_{pop} = \sigma/\mu.$$

This is estimated by the sample CV,

$$CV_{sample} = s/\bar{x}$$

where $\bar{x}$ and s are the sample mean and sample sd.

For a sample of size n from a normal distribution,

$$T = \sqrt{n}/CV_{sample}$$

has a noncentral Student *t* distribution with *n* -1 degrees of freedom and noncentrality parameter

$$ncp = \sqrt{n}/CV_{pop}$$

Construction of Boundaries Within Which 50% of CVs Are Likely to Fall

$$P(t'_{0.25} < T < t'_{0.75}) = 0.50$$

$$P\left[\frac{1}{t'_{0.25}} > \frac{CV_{sample}}{\sqrt{n}} > \frac{1}{t'_{0.75}}\right] = 0.50$$

$$P\left[\frac{\sqrt{n}}{t'_{0.75}} < CV_{sample} < \frac{\sqrt{n}}{t'_{0.25}}\right] = 0.50$$

Hudes ML et al 2008 *Faseb J 23:689-703 (2009)*

The median CV from each journal article was used to estimate CV_{pop}.

The construction of these intervals makes the following assumptions:

a) All measurements are on the same variable, such as a particular enzyme activity;

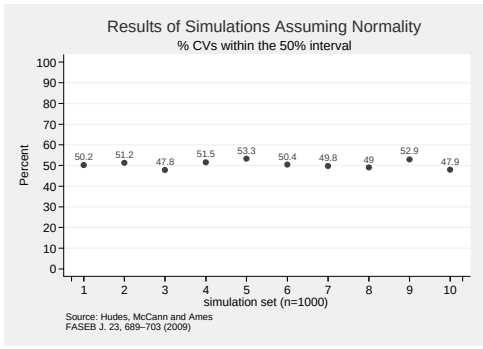
b) All measurements are from the same treatment group, such as the control group or a group treated with lipoic acid;

c) The median CV is an adequate estimate of CVpop.

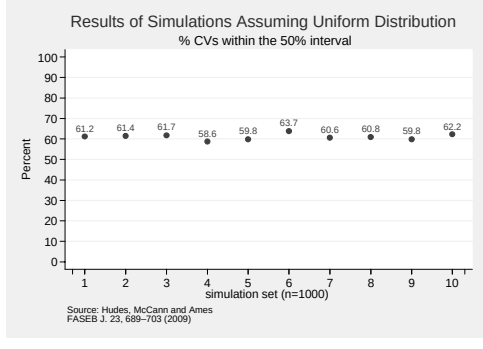
d) Each variable being examined has an approximate normal distribution.

Calculation of an approximate "P value"

The binomial distribution was used to compute the likelihood of obtaining *k* or more (out of a total of *N*) CVs **inside** the constructed 50% limits.

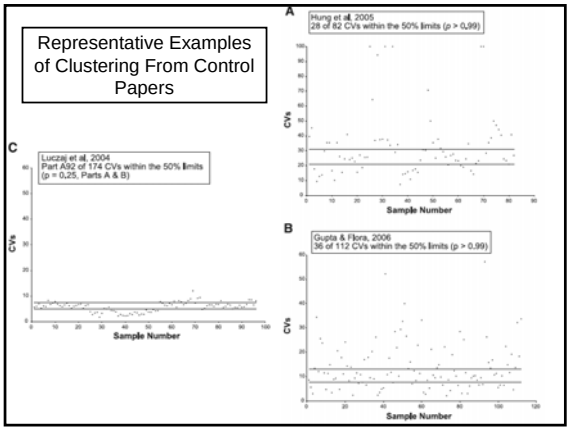


7

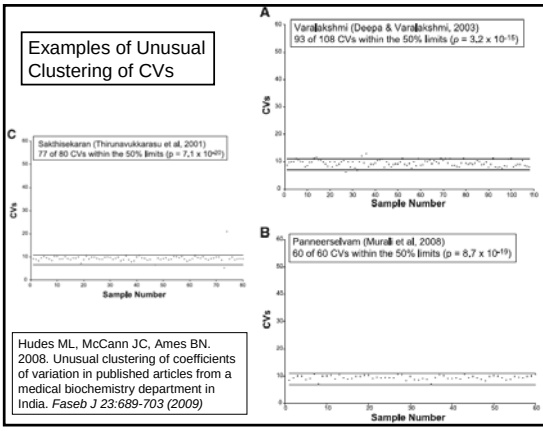


8

Representative Examples of Clustering From Control Papers



Examples of Unusual Clustering of CVs



Hudes ML, McCann JC, Ames BN. 2008. Unusual clustering of coefficients of variation in published articles from a medical biochemistry department in India. *Faseb J 23:689-703 (2009)*

Examples of Clustering From Very Similar Experiments

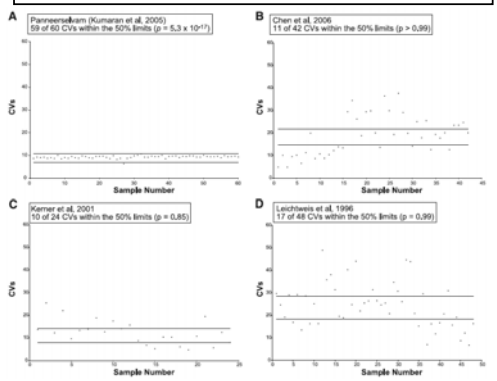


TABLE 1. *Aberrant clustering (P<0.01) of CVs in articles from VPS compared to other laboratories*

		Laboratories of VPS	
"*P* value"	Other laboratories	Published before January 2000	Published after January 2000
$<10^{-8}$			16
$<10^{-4}$	1[a]		17
0.0001–0.0009			3
0.001–0.009			4
0.01–0.049		1	5
0.05–0.09			1
0.1–0.49	4	2	7
>0.5	21	15	13

"*P* values" were calculated as described in Hudes *et al.* (1).
[a] From a laboratory on the same campus as VPS.

12