LC Confirm Phase Guidelines:
Approximate Expected InterQuartile-range of the Permutation LTD Distribution

Bob Obenchain, © Risk Benefit Statistics, 2015.

1. Introduction

The primary objective of the **Confirm** phase of Local Control strategy is to demonstrate that **X**-space microaggregation of experimental units truly matters. Specifically, the CDF of the observed LTD distribution from microaggregation needs to have an **adjusted** location and/or shape that is visually distinct from the CDF of its random permutation counterpart. After all, this counterpart LTD distribution ignores the numerical values of all baseline **X**-covariates (potential confounders) and randomly assigns experimental units to the same number of clusters, of the same sizes, as the observed LTD distribution. **If there are no obvious, important differences between these two LTD distributions, no meaningful adjustment has occurred.**

These notes actually focus upon a specific, secondary objective when comparing an observed LTD distribution to its corresponding permutation LTD distribution. This secondary objective is to monitor progress and/or potential completion of **variance-bias trade-offs** as clusters become more numerous and, thus, smaller. It's a mistake to use too many clusters (that are too small) when there is no clear evidence that additional bias is thereby being removed; **these notes show how the variances of LTD estimates are inflated by using clusters that are smaller than really needed.**

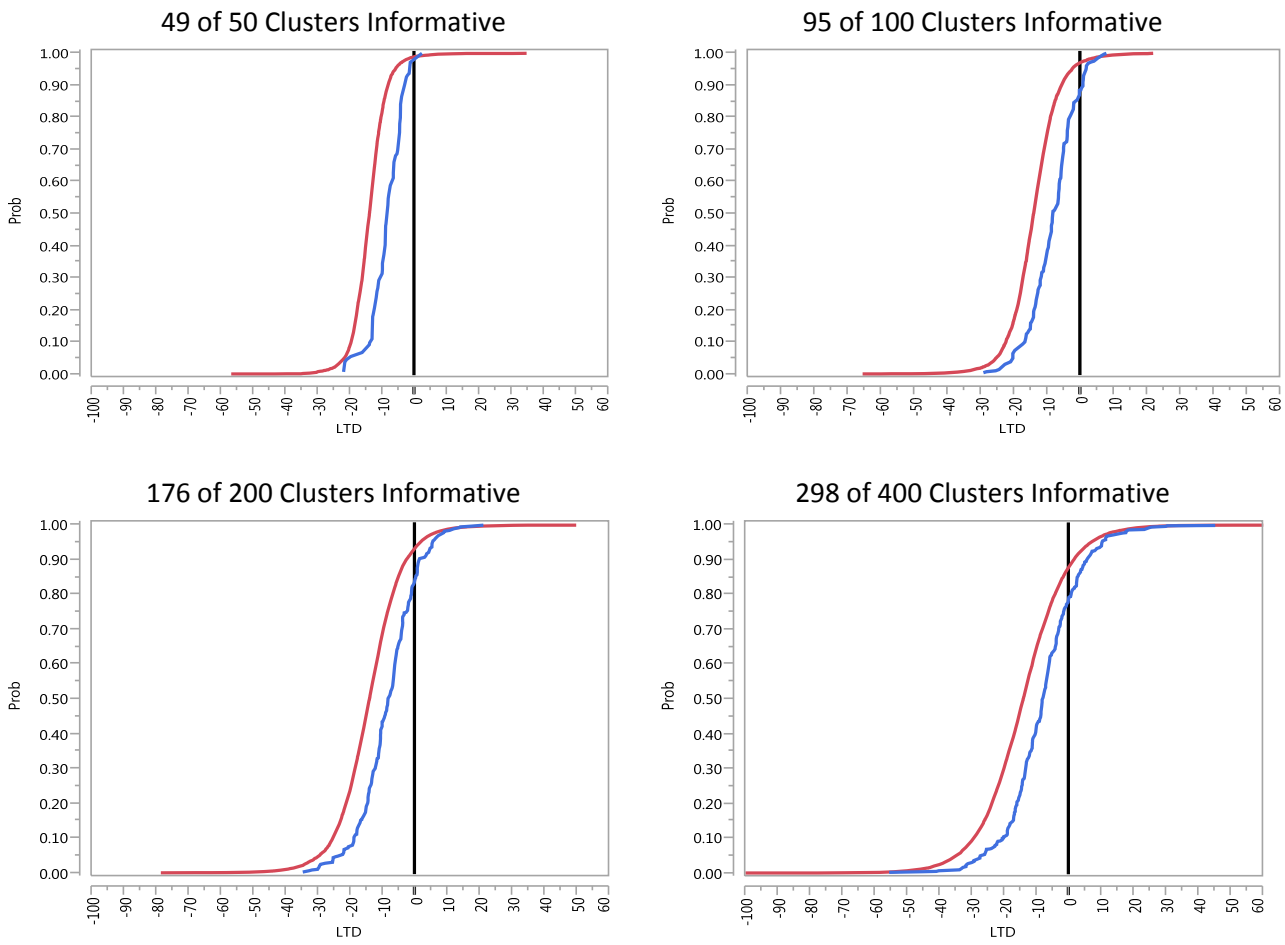
For example, Figure 1 below contains four pairs of observed and permutation CDFs for observational data on the effect of indoor Radon level (High minus Low) on Lung Cancer Mortality for 2,881 US counties. We will use these plots to explain basic concepts and to illustrate their potential interpretations. The total number of clusters requested are **K** = 50, 100, 200 and 400, respectively. With this increase in number of clusters requested, the average cluster size is thereby reduced from roughly 58 to 30, to 16 and finally to about 8 counties.

The four CDF displays in Figure 1 show that the "apparent" variance (spread) in these empirical distributions increases as the number of informative clusters[1] increases. Specifically, both the red and blue empirical CDFs tend to become less and less steep in the vicinity of their median values as the number of informative clusters increases.

Note that all four observed (adjusted) LTD distributions in Figure 1 are shifted somewhat to the right (towards zero) and tend to have somewhat shorter tails than their random permutation counterparts. Therefore, **X**-matching clearly "matters" in each of these four alternative LC micro-aggregations of US counties.

[1] To be informative about a LTD, a cluster must contain at least one treated experimental unit as well as at least one control experimental unit.

Figure 1. Four pairs of CDFs for observed and permutation LTD distributions.



Intuitively, the variances of individual LTD estimates (observed or permutation) are lowest when clusters are large. But, due to improved **X**-space matching, the biases of individual observed LTD estimates should be lowest when clusters are small (nearer to exact **X**-space matches.) These two competing effects suggest that a variance-bias trade-off occurs as the number of requested (and informative) clusters is increased. In other words, the variance of LTD estimates is allowed to increase in the hope that additional bias (increased separation between blue and red CDFs) will be removed from the observed LTD distribution.

The good news here is two-fold.

I. The CDF of the permutation LTD distribution can be estimated with arbitrary precision by simply increasing the number of "complete" replications used to compute it. Each complete replication provides a permutation LTD estimate for each informative cluster.

II. The very definition of the permutation LTD distribution makes it relatively simple to approximately predict the **inter-quartile range**, a well-known measure of variability, for one complete replication.

2. Basic concepts and notation.

Let **y** given either **t** = 0 or 1 be independent and Normally distributed random variables with

$$\text{mean} = \mu_t \quad [\text{where } \mu_1 = \boldsymbol{\mu} + \boldsymbol{\tau} \text{ and } \mu_0 = \boldsymbol{\mu}\,] \text{ and standard deviation} = \boldsymbol{\sigma}. \qquad \textbf{(2.1)}$$

Let $\mathbf{ybar_t}$ denote the average of $\mathbf{n_t}$ such observations. Thus $\mathbf{ybar_t}$ is also a Normally distributed random variable with

$$\text{mean} = \mu_t \quad \text{and} \quad \text{standard deviation} = \boldsymbol{\sigma} / \text{sqrt}(\,\mathbf{n_t}\,). \qquad \textbf{(2.2)}$$

The value of the permutation LTD for an informative cluster is then $\mathbf{pLTD} = \mathbf{ybar_1 - ybar_0}$ ...which is another Normally distributed random variable with

$$\text{mean} = \boldsymbol{\tau} \quad \text{and}$$
$$\text{standard deviation} = \boldsymbol{\sigma} \times \text{sqrt}(\,[\mathbf{n_1} + \mathbf{n_0}] / [\mathbf{n_1} \times \mathbf{n_0}]\,) = \boldsymbol{\sigma} / \text{sqrt}(\,\mathbf{np[1-p]}\,) \qquad \textbf{(2.3)}$$

where $\mathbf{n} \equiv [\mathbf{n_1} + \mathbf{n_0}]$ and $\mathbf{p} \equiv \mathbf{n_1} / [\mathbf{n_1} + \mathbf{n_0}]$ by definition. Since $\mathbf{n_1} > 0$ and $\mathbf{n_0} > 0$, it follows that both **p** and **[1–p]** are strictly positive.

Given **K** (mutually exclusive and informative) clusters, each containing **n** experimental units, the expected values of the order statistics from observed **pLTD** values are

$$\mathbf{pLTD}_{[j]} = \boldsymbol{\tau} + \mathbf{Z}_{[j]} \times \boldsymbol{\sigma} / \text{sqrt}(\mathbf{np[1-p]}) \text{ for } 1 \le j \le \mathbf{K}, \qquad \textbf{(2.4)}$$

where $\mathbf{Z}_{[1]} < \mathbf{Z}_{[2]} < \ldots < \mathbf{Z}_{[K]}$ are called **RankIts**, which are the expected values of order statistics of Normal(0,1) random deviates.

Thus our ***crude approximation*** to the expected inter-quartile range of the permutation LTD distribution resulting from **K** informative clusters, that are of average size **n** , and that contain an overall proportion **p** of treated experimental units is:

$$\text{eIQR}(\mathbf{K, n, p}) = \mathbf{2} \times |\mathbf{Z}_{[j(\mathbf{K})]}| \times \boldsymbol{\sigma} / \text{sqrt}(\mathbf{np[1-p]}) \qquad \textbf{(2.5)}$$

$$\text{where } j(\mathbf{K}) = \mathbf{K}/2 - \text{floor}(\mathbf{K}/4) \text{ when } \mathbf{K} \text{ is even, else}$$
$$= (\mathbf{K}-1)/2 - \text{floor}(\mathbf{K}/4) \text{ when } \mathbf{K} \text{ is odd.}$$

The approximation given by expression **(2.5)** is crude in at least two ways; clusters typically vary in size, **n**, and within-cluster treatment choice fractions, **p**, typically also vary between clusters. Finally, the **K** value used in expression **(2.5)** corresponds to a single "complete" replication,

while many such replications are typically used to depict a relatively smooth and precise permutation LTD distribution. Luckily, the value of **K** is the same for each complete replication.

Next, we use a numerical example to illustrate that all of the above potential shortcomings in the proposed approximation **(2.5)** tend to be either relatively unimportant or to cancel each other out.

3. Numerical Illustration: the IQ-range of the Perm LTD Distribution is Predictable

Table 1. Calculations for Lung Cancer Mortality and indoor Radon level

Clusters Requested	K = Inform Clusters	Tot Inform Exp Units (counties)	n = Avg Cluster Size	Total UnInform Clusters	UnInform Low Rn (counties)	UnInform High Rn (counties)	Inform Low Rn (counties)	Inform High Rn (counties)
50	49	2870	58.6	1	11	0	1650	1220
100	95	2848	30.0	5	21	12	1640	1208
200	176	2732	15.5	24	114	35	1547	1185
400	298	2512	8.4	102	272	97	1389	1123

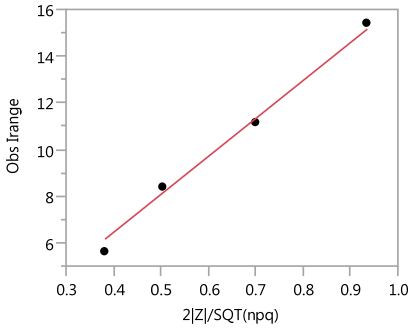
K = Inform Clusters	j of Quartile	n = Avg Cluster Size	p = Fraction High Rn (counties)	RankIt	2|Z|/SQRT(npq)	Obs IQ-range
49	12	58.6	0.4251	-0.7198	0.3805	5.6130
95	24	30.0	0.4242	-0.6810	0.5034	8.3720
176	44	15.5	0.4337	-0.6825	0.6991	11.1957
298	75	8.4	0.4471	-0.6739	0.9338	15.4630

Note: The above RankIts were computed using the normOrder() function from the R-package SuppDists. The tabulated value is rankit[j] from the assignment rankit <– normOrder(K).

The Observed values of IQ-range come from the data on indoor Radon level and Lung Cancer Mortality for 2,881 US Counties. The approximation of expression **(2.5)** is relatively good iff some fixed, positive scalar multiple (an estimate of the **σ** for lung cancer mortality) of the values in the **2|Z|/SQRT(npq)** column above adequately predict the values in the **Obs IQ-range** column.

Using linear regression (with intercept forced to zero) to predict σ for lung cancer mortality:

Response here is Obs IQ-range.



Root Mean Square Error	0.396687
Mean of Response	10.16093
Observations (or Sum Wgts)	4

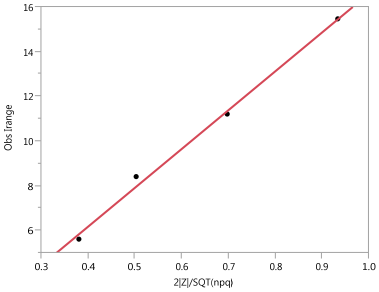
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	465.57214	465.572	2958.632
Error	3	0.47208	0.157	**Prob > F**
C. Total	4	466.04422		<.0001*

The implied estimate of σ is the fitted β-coefficient

Term	Estimate	Std Error	t Ratio	Prob>|t|
2|Z|/SQRT(npq)	**16.270027**	0.299118	54.39	<.0001*

CHECK: Next, verify that a fitted intercept would not have been significantly different from zero.



Obs IQ-range = -0.7601 + 17.3577 * 2|Z|/SQRT(npq)

RSquare	0.995446
RSquare Adj	0.993169
Root Mean Square Error	0.347622
Mean of Response	10.16093
Observations (or Sum Wgts)	4

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	52.824951	52.8250	437.1450
Error	2	0.241682	0.1208	**Prob > F**
C. Total	3	53.066633		0.0023*

Parameter Estimates

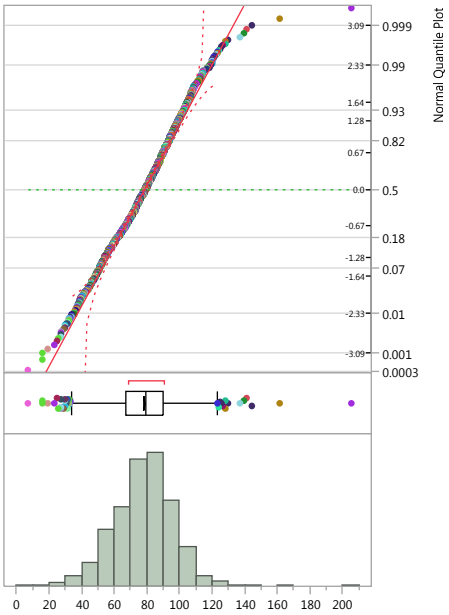
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	**-0.760134**	**0.550498**	**-1.38**	**0.3014**
2|Z|/SQRT(npq)	17.357733	0.830195	20.91	0.0023*

Finally, we need to verify that the observed mortality **y**-outcomes for 2,881 US counties are indeed consistent with the above **σ** estimates of roughly 16 or 17 deaths per 100,000 person-years:

[a] Estimated σ(county mortality rates) = 17.65 per 100K person-years

Mean	78.126974
Std Dev	**17.650253**
Std Err Mean	0.328836
N	2881

[b] The Normal probability plot and histogram at right also validate our assumption that mortality y-outcomes are approximately Normaly distributed.



This numerical example has illustrated that the approximation of expression **(2.5)** can actually be rather good.

4. Misleading visual impression of "Increasing Spread" in Permutation LTD Distributions.

The numerical example of Section §3 showed that the **2**|**Z**| term in the numerator of expression **(2.5)** is fairly constant as **K** increases. In fact, the **RankIt** column of Table 1 (page 4) shows that this numerator term may actually decrease slightly in absolute value as **K** increases.

On the other hand, the **n** = cluster size factor within the sqrt(**np**[1−**p**]) term in the denominator of expression **(2.5)** typically decreases rapidly as **K** increases. As a result, the **predicted inter-quartile range** typically increases steadily with increases in **K** = number of informative clusters.

This is unfortunate and potentially misleading because the true y-outcome standard deviation, **σ**, is actually fixed and unchanging here. A useful "standardization" of the horizontal scale for display of observed and permutation CDFs would thus be to assure that the **inter-quartile range** of the permutation LTD distribution is held constant (by choice of horizontal scale)! If nothing else, interpreters of these displays should be warned to disregard any appearances of increase in horizontal "spread" of the permutation and observed LTD distributions.

5. Variance-Bias Trade-Offs in Estimation of Heterogeneous Treatment Effects

The simple, approximate calculations outlined here do, however, provide some basic guidelines on interpreting a sequence of LC **Confirm** phase graphics like those of Figure 1 and picking a "most appropriate" number of LC clusters, **K**, to optimize variance-bias trade-offs.

Since the relative positions of **red** and **blue** CDFs are much the same in all four CDF plots, the plots for **K**= 95, 176 or 298 provide no clear indication that appreciable additional bias was removed beyond that achieved with **K**=49. The first of these four LC **Confirm** phase graphics must thus be the best depiction of optimal LTD adjustment. It uses the smallest number of informative clusters **K** = 49 (of largest average size **n** = 58.6) that make this particular LTD distribution most stable (least variance inflated.)