Define the Response Variables of Robustness DOE

Charles Tan

2025-09-12

Notations

For an assay robustness DOE:

- Let N denote total number of runs, say 24
- Let n denote total number of center runs, say 8
- Let $i = 0$ index the center point condition
- Let $i = 1, 2, ..., N - n$ index the non center point conditions
- Let X_i denote the row for condition i in the design matrix
- Let K donate total number of samples, say 14 or 22
- Let $T_{i,k}$ denote the observed log titer or log concentration of kth sample under ith condition
- Let μ_k denote the mean log titer or log concentration of kth under center condition, $\mu_k = E(T_{0,k})$

Primary Response

For each non center point condition, $T_{i,k} - \mu_k$ is the deviation from center point. Define $Y_i = E(T_{i,k} - \mu_k)^2$ where the expectation is over all samples k. This Y_i is on the same scale as variance on the log scale, which has a one-to-one translation to %RSD scale. We can build a model $Y \sim X$ using SVEM to predict Y given condition X, then impose a criterion on the Y surface on the %RSD scale. If the whole Y surface is below the criterion, we can conclude the assay is robust within the range of the design of the experiment.

The definition of Y_i involves K unknown (nuisance) parameters μ_k. The simplest estimator of μ_k is $m_k = \frac{1}{n} \sum T_{0,k}$, then the simple naive estimator of Y_i is $\frac{1}{K} \sum_k (T_{i,k} - m_k)^2$.

Secondary Response

If the predicted Y in some part of the design space is above the acceptance criterion, the natural followup question is why. The logical place to start the investigation is to look at bias induced by conditions. Define $Z_i = E(T_{i,k} - \mu_k)$, again, the expectation is over all samples k. It can serve as the secondary response variable useful for investigation.

Plugging in m_k, we have the simple naive estimator of Z_i as $\frac{1}{K}\sum_k (T_{i,k} - m_k)$.

Relationship

Under the assumption that the bias and variability are constant across samples at condition i, (but not necessarily across i,) there is a relationship between Y_i and Z_i: $Y_i = Z_i^2 + Var(T_{i,.})$. This relationship allows us to breakdown the reasons for Y to be large: bias or random measurement variability or both.

Model Averaging

At each condition i, there are at least two different scenarios:

The first scenario is that all the differences of $T_{i,k} - \mu_k$ are just measurement variability, i.e., $T_{i,k} - \mu_k = \epsilon_{i,k}$. Under this model, $\hat{Y}_i = \frac{1}{K}\sum_k (T_{i,k} - m_k)^2$.

The second scenario is that there are both bias and variability in $T_{i,k} - \mu_k$, i.e., $T_{i,k} - \mu_k = \delta_i + \epsilon_{i,k}$. Under this model,

The response variables Y and Z need to be estimated. Under different assumptions about the homogeneity of the bias and variability across different condition i, the naive estimators of Y_i and Z_i could be improved if we have sufficiently large K at condition i. Four different model assumptions are described below, and their associated estimators of Y_i and Z_i are listed.

Furthermore, we don't have to "pick a model". We can do "model averaging" by weighting different estimators according to their Akaike weights.

At Center Point

By definition, $Z_0 = 0$ and $Y_0 = \frac{1}{K}\sum_k Var(T_{0,k})$, which is the average measurement variability (precision).