
Large sample inference for a screen quality measure in High-Throughput Screening assays

Antara Majumdar and David Stock

Bristol-Myers Squibb

Introduction

- The Z' factor, introduced by Zhang et al. (1999), is used extensively in drug discovery for evaluating the performance of High Throughput Screening (HTS) assays.
- Important decisions regarding HTS assay development, validation and quality are often based solely on point estimates of Z' .
- Although it would be beneficial to have a confidence interval for Z' , it appears that a formal inferential procedure has not yet been proposed.

Interval Estimator for Z'

- We propose a confidence interval for Z' based on large sample theory.
- Simulation studies found that the proposed confidence interval performed well with both independent and moderately correlated data.
- Our confidence interval is algebraically simple, and amenable to spreadsheet programming.

Quality of an HTS Assay

- The quality of an HTS assay is directly related to how well it signals the presence, absence, or degree, of a biochemical interaction.
- Often, this interaction is signaled by either the production, or reduction, of a luminescent signal.
- The ends of the luminescence range are usually empirically defined by positive controls (aka “totals”), and negative controls (aka “blanks”), that are run in a subset of the wells on the micro-titer assay plates.
- In a good assay the signals generated by the totals are clearly distinguishable from the signals generated by the blanks.

“Upper” and “Lower” Controls

- Depending on how an assay has been configured, either the positive or negative controls may produce the higher levels of luminescence, while the other will produce the lower levels.
- We will refer to the “upper controls” as the controls that produce the higher levels of assay signal, and the “lower controls” as the controls that produce the lower levels of assay signal.
- Z' measures the separation between the upper and lower controls as functions of their location and spread.

Z' Factor

- If the data are normally distributed, this implies that the data from each control group would be almost entirely contained within three standard deviations of the group mean.
- Let μ_u and σ_u be the mean and standard deviation of the “upper” controls, and let μ_l and σ_l be the mean and standard deviation of the “lower” controls. Then, Z' , as defined by Zhang, et al. (1999), is

$$Z' = \frac{(\mu_u - 3\sigma_u) - (\mu_l + 3\sigma_l)}{\mu_u - \mu_l}$$

Assay Acceptance Criterion

- Values for Z' can range between $-\infty$ to 1.
- Zhang, et al. (1999) provided cutoff criteria of:
 - (i) $0 < Z' < 0.5$ for a “double assay”
 - (ii) $0.5 \leq Z' < 1$ for an “excellent assay”
- Some consider Z' values below 0.5 to be weak or marginal.
- However, the NIH, and Eli Lilly, recommend using a Z' value of 0.4 as the cut-off for acceptance.
- Clearly, the interpretation of Z' relative to any cut-off would be facilitated by the addition of confidence bounds, particularly when the data are limited.

Method of Moments Estimator

- Consider a random sample of size n from a $N(\mu_u, \sigma_u^2)$ population of upper controls, and an independent random sample of size n from a $N(\mu_l, \sigma_l^2)$ population of lower controls.
- Let \bar{x}_u , \bar{x}_l and s_u and s_l be the corresponding sample means and standard deviations. Then Z' may be estimated using the observed moments as:

$$\begin{aligned}\hat{Z}' &= 1 - \frac{3(s_u + s_l)}{\bar{x}_u - \bar{x}_l} \\ &= 1 - 3W_n\end{aligned}$$

Approximation for the Standard Deviation Terms

Miller (Comm. Stat. - Theory & Methods, 1991) showed that the sample standard deviation of a normal population, such as that of the upper controls, can be expressed as,

$$s_u = \sigma_u + m^{-1/2} \sqrt{0.5} \sigma_u Y_u + O_p(m^{-1})$$

where Y_u is a standard normal variate and $m = n - 1$. By analogy, and because s_u and s_l are independent, we also have

$$s_l = \sigma_l + m^{-1/2} \sqrt{0.5} \sigma_l Y_l + O_p(m^{-1})$$

Approximation for the Numerator of W_n

Therefore, we can write

$$\begin{aligned} s_u + s_l &= \sigma_u + \sigma_l + m^{-1/2} \sqrt{0.5} \sigma_u Y_u \\ &+ m^{-1/2} \sqrt{0.5} \sigma_l Y_l + O_p(m^{-1}) \end{aligned}$$

Approximation for the Denominator of W_n

Now, if we apply a multivariate Taylor series expansion to $(\bar{x}_u - \bar{x}_l)^{-1}$, we get

$$\begin{aligned} \frac{1}{\bar{x}_u - \bar{x}_l} &= \frac{1}{\mu_u - \mu_l} - \frac{(\bar{x}_u - \mu_u)}{(\mu_u - \mu_l)^2} \\ &+ \frac{(\bar{x}_l - \mu_l)}{(\mu_u - \mu_l)^2} + O_p(n^{-1}) \end{aligned}$$

Expression for W_n

W_n of \hat{Z}' can be expressed as

$$\begin{aligned} \frac{s_u + s_l}{\bar{x}_u - \bar{x}_l} &= \left(\frac{\sigma_u + \sigma_l}{\mu_u - \mu_l} \right) - n^{-1/2} (\sigma_u Y_u + \sigma_l Y_l) \frac{(\sigma_u + \sigma_l)}{(\mu_u - \mu_l)^2} \\ &+ (\sigma_u Y_u + \sigma_l Y_l) \frac{m^{-1/2} \sqrt{0.5}}{(\mu_u - \mu_l)} + O_p(n^{-1}) \end{aligned}$$

where Y_u and Y_l are independent standard normal variables.

Distribution of W_n

- The second and third terms on the right hand side are functions of constants and standard normal variates.

- Therefore, the second term,

$n^{-1/2}(\sigma_u Y_u - \sigma_l Y_l) \frac{(\sigma_u + \sigma_l)}{(\mu_u - \mu_l)^2}$ is distributed as

$$N \left(0, n^{-1}(\sigma_u^2 + \sigma_l^2) \frac{(\sigma_u + \sigma_l)^2}{(\mu_u - \mu_l)^4} \right)$$

- and the third term, $(\sigma_u Y_u + \sigma_l Y_l) \frac{m^{-1/2} \sqrt{0.5}}{(\mu_u - \mu_l)}$ is

distributed as $N \left(0, m^{-1} 0.5 \frac{(\sigma_u^2 + \sigma_l^2)}{(\mu_u - \mu_l)^2} \right)$

Asymptotic Distribution of W_n

The asymptotic distribution of W_n can now be easily derived:

$$W_n \xrightarrow{d} N \left(\left(\frac{\sigma_u + \sigma_l}{\mu_u - \mu_l} \right), \frac{(\sigma_u^2 + \sigma_l^2)}{(\mu_u - \mu_l)^2} \left[n^{-1} \frac{(\sigma_u + \sigma_l)^2}{(\mu_u - \mu_l)^2} + m^{-1} 0.5 \right] \right)$$

where, $W_n = \frac{s_u + s_l}{\bar{x}_u - \bar{x}_l}$

Confidence Interval for Z'

Therefore, an approximate $100(1 - \alpha)\%$ confidence interval for the Z' factor based on the above argument is

$$CI_n : \left(\hat{Z}' - 3Z_{\alpha/2}V_n, \hat{Z}' + 3Z_{\alpha/2}V_n \right)$$

where

$$V_n = \sqrt{\frac{(s_u^2 + s_l^2)}{(\bar{x}_u - \bar{x}_l)^2} \left[n^{-1} \frac{(s_u + s_l)^2}{(\bar{x}_u - \bar{x}_l)^2} + m^{-1}0.5 \right]}$$

Confidence Interval for Z' : Unequal Samples

Following similar steps, it is trivial to show that for unequal sample sizes the confidence interval is of the form:

$$CI_{n_1, n_2} : \left(\hat{Z}' - 3Z_{\alpha/2}V_{n_1, n_2}, \hat{Z}' + 3Z_{\alpha/2}V_{n_1, n_2} \right)$$

where n_1 and n_2 are the sizes of random samples from $N(\mu_u, \sigma_u^2)$ and $N(\mu_l, \sigma_l^2)$, respectively. If \hat{Z}' is defined as before, but with n replaced by n_1 and n_2 where appropriate, then

$$V_{n_1, n_2} = \sqrt{\frac{(s_u + s_l)^2}{(\bar{x}_u - \bar{x}_l)^4} \left(\frac{s_u^2}{n_1} + \frac{s_l^2}{n_2} \right) + \frac{0.5}{(\bar{x}_u - \bar{x}_l)^2} \left(\frac{s_u^2}{m_1} + \frac{s_l^2}{m_2} \right)}$$

where, $m_1 = n_1 - 1$ and $m_2 = n_2 - 1$.

Simulation Studies

- Two simulation studies were conducted to evaluate the proposed confidence interval. The simulations examined the width and the coverage probability of the confidence interval for the 95% confidence intervals.
- The first study was conducted under the conditions assumed in the proof, using normal, independently distributed data.
- The second study relaxed the independence assumption, and allowed for correlations between the observations.

Simulation Study Designs

- Both simulation studies were designed to reflect the structure of assays conducted on the 384-well microtiter plates that are common in high throughput screening.
- The 384-well plate has 24 rows and 16 columns, of which, typically, 32-wells are used for the totals, and 32-wells are used for the blanks.
- In an assay validation or development contexts, it is possible that the entire plate would be split between upper and lower controls.
- Therefore, we looked at the performance of the proposed confidence interval over a range of sample sizes, ranging from 16 to 192 wells per control.

Simulation Study with Independent Samples

- Independent random samples, of equal sizes, were drawn from two normal populations.
- The parameters of the populations were chosen such that the true value of Z' ranged between 0.05 to 0.95. Ten thousand simulations were run for each setting of Z' .

Simulation Study Results for 16 and 32 Wells: *Independent Samples*

Wells	Z'	Bias in \hat{Z}'	Width	Coverage Probability
16	0.05	0.01321	0.6010	0.91
	0.25	0.01198	0.4192	0.92
	0.50	0.00862	0.2762	0.92
	0.75	0.00453	0.1285	0.93
	0.95	0.00092	0.0255	0.93
32	0.05	0.00603	0.4210	0.94
	0.25	0.00539	0.2936	0.94
	0.50	0.00386	0.1936	0.94
	0.75	0.00203	0.0900	0.94
	0.95	0.00041	0.0179	0.94

Simulation performed at the 0.05 level of significance

Simulation Study Results for 64 and 128 Wells: *Independent Samples*

Wells	Z'	Bias in \hat{Z}'	Width	Coverage Probability
64	0.05	0.00202	0.2965	0.95
	0.25	0.00200	0.2067	0.95
	0.50	0.00150	0.1363	0.95
	0.75	0.00080	0.0633	0.95
	0.95	0.00016	0.0126	0.95
128	0.05	0.00072	0.2091	0.95
	0.25	0.00061	0.1458	0.95
	0.50	0.00042	0.0962	0.95
	0.75	0.00020	0.0447	0.95
	0.95	0.00004	0.0089	0.95

performed at the 0.05 level of significance

Summary: Independent Samples Simulation Study

- The coverage of the interval is accurate at sample sizes of 64, 128 and 192 well per control article.
- At a sample size of 32, the coverage of 0.94 is only slightly below the expected value of 0.95.
- At a sample size of 16, the coverage drops to somewhere between 0.91 to 0.93.
- Bias is generally small, and as expected, decreases with sample size. Also, bias decreases as Z' increases. This is likely due to the bounded nature of Z' .
- The width of the confidence interval decreases as the sample size increases, and the width also decreases as Z' increases.

Simulation Study with Dependent Samples

- In real life the 384 wells are in very close proximity, and are filled by robots that often work sequentially along the plate.
- Some degree of spatial correlation could exist in data generated under these conditions. In our experience this has been true.
- However, for well conducted assays, the amount of spatial correlation is relatively small.

Simulation Study with Dependent Samples

- For example, we analyzed the data from 88 plates.
- The spatial correlation between the wells was modeled as an anisotropic power function of the following form:

$$\sigma^2 \rho_c^{d_{ij_c}} \rho_r^{d_{ij_r}},$$

where i and j are observations from particular wells, and the d_{ij_c} and d_{ij_r} are the number of columns, or rows, between wells i and j . The ρ terms are what is referred to, in geostatistics, as “ranges”.

- The correlation between wells i and j is given by the product: $\rho_c^{d_{ij_c}} \rho_r^{d_{ij_r}}$.

Simulation Study with Dependent Samples

- In our data set, we found the maximum observed range for the columns (ρ_c) was 0.16, and the maximum observed range for the rows (ρ_r) was 0.18.
- These ranges give rise to minor correlations, that will die off in the space of a couple of columns or a couple of rows.
- To simplify the second simulation study, we used an isotropic spatial correlation structure: $\sigma^2 \rho^{d_{ij}}$. Here d_{ij} is the Euclidean distance between wells i and j , and the range term ρ now dies out uniformly in all directions. (When this model was fit to our data set, the maximum observed range was 0.13.)

Simulation Study Results for 16 Wells: *Dependent Samples*

Number of wells = 16

Z'	$\rho = 0.2$			$\rho = 0.4$		
	Bias	Width	Prob	Bias	Width	Prob
0.05	0.02277	0.5947	0.89	0.04700	0.5781	0.83
0.25	0.02067	0.4141	0.90	0.04092	0.4022	0.84
0.50	0.01460	0.2729	0.90	0.02849	0.2651	0.84
0.75	0.00770	0.1269	0.90	0.01480	0.1233	0.85
0.95	0.00156	0.0252	0.90	0.00299	0.0245	0.85

performed at the 0.05 level of significance

Simulation Study Results for 32 Wells: *Dependent Samples*

Number of wells = 32

Z'	$\rho = 0.2$			$\rho = 0.4$		
	Bias	Width	Prob	Bias	Width	Prob
0.05	0.01736	0.4156	0.90	0.04180	0.4039	0.80
0.25	0.01536	0.2895	0.90	0.03589	0.2811	0.80
0.50	0.01075	0.1909	0.90	0.02485	0.1854	0.81
0.75	0.00560	0.0887	0.91	0.01282	0.0862	0.81
0.95	0.00113	0.0176	0.91	0.00258	0.0171	0.81

performed at the 0.05 level of significance

Simulation Study Results for 64 Wells: *Dependent Samples*

Number of wells = 64

Z'	$\rho = 0.2$			$\rho = 0.4$		
	Bias	Width	Prob	Bias	Width	Prob
0.05	0.01039	0.2937	0.89	0.02987	0.2872	0.76
0.25	0.00934	0.2046	0.90	0.02576	0.1999	0.77
0.50	0.00658	0.1349	0.91	0.01789	0.1318	0.78
0.75	0.00348	0.0626	0.91	0.00928	0.0612	0.79
0.95	0.00071	0.0124	0.91	0.00187	0.0122	0.80

performed at the 0.05 level of significance

Summary: Dependent Samples Simulation Study

- When the range was assumed to be 0.2, the coverage probability of the confidence interval dropped to roughly 90%.
- Compared to the previous simulation study, there is an increase in the bias of \hat{Z}' , and some indication that the width of the intervals might be slightly smaller.
- For the range of 0.4, we can see that the coverage has dropped to roughly 80%. Also, the bias in \hat{Z}' has increased even further, and the width of the intervals show signs of another slight decrease.
- This indicates that the interval estimator should not be used at this higher level of spatial correlation.

Conclusion

- The simple derivation presented provides a confidence interval for both the Z' and Z factors introduced by Zhang et al (1999).
- When the conditions of the proof are met, our confidence interval works well; even at moderate sample sizes.
- When the data exhibit the modest spatial correlations that are typical in high throughput screening (range ≤ 0.20), the coverage of the interval drops from 95% to 90%.
- This small decrease in coverage does not exclude the use of the interval in applied situations.

Numerical Example

The calculations needed to compute the proposed confidence interval are simple, and amenable to spreadsheet programming.

- Consider a case where 32 wells were used for each set of controls.
- The sample mean and standard deviation for the upper control samples were 3000 and 150, while for the lower control samples the mean and standard deviations were 1000 and 50.
- By plugging 3000 for \bar{x}_u , 150 for s_u , 1000 for \bar{x}_l , and 50 for s_l , the resulting estimate of the Z' factor turns out to be exactly 0.7 and the 95% confidence interval turns out to be (0.64, 0.76).