Implementation details of the power calculations via simulations for scaled ABE in R-package "PowerTOST"

Version 0.07 (Jan 2015)

Author: Detlew Labes

EMA method (`power.scABEL()`)

Method description in a cook book manner:

- Evaluate all data (log-transformed) via an ANOVA equal to the classical cross-over design with treatment, period, sequence and subject within sequence.
  Get the point estimate (pe) for T-R and the mse from that ANOVA.
  The 90% confidence interval is obtained from pe and mse according to
  $$[lL, uL] = pe \pm t_{(1-alpha),df} * \sqrt{mse * b_{k(ni)} * \sum \frac{1}{n_i}}$$
  The term under the square root is s_d^2, the variance of the pe. $b_{k(ni)}$ is the design constant in terms of n_i = number of subjects in the sequence groups.
  The term $b_{k(ni)} * \sum \frac{1}{n_i}$ is named C2.
- Evaluate the data (log-transformed) for the reference only via an ANOVA with period, sequence and subject within sequence. The mse of that evaluation is s_{wR}^2 (within-subject variance for the reference). It has df(RR) degrees of freedom associated.
- If $CV_{wR} = sqrt(\exp(s_{wR}^2) - 1)$ is greater 0.3 calculate the widened acceptance limits (in the log domain) according to
  $$[lABEL, uABEL] = \pm 0.760 * s_{wR}$$
  If $CV_{wR} \leq 0.3$ use $[-\log(1.25, \log(1.25)]$.
  If $CV_{wR} > 0.5$ use the acceptance limits for $CV_{wR} = 0.5$ (cap on widening).

  0.760 is the regulatory constant set by the EMA, derived from $\log(1.25) / s_{wR} = 0.7601283$ at $s_{wR} = 0.2935604$, the value of the error standard deviation for $CV_{wR} = 0.3$.
- Decide BE if the 90% confidence interval is contained in the scaled (widened) acceptance limits.

The covered replicate crossover designs have the following characteristics (N=Σn_i):

Design	df	$b_{k(ni)}$	b_k	df (RR)	E(mse)
2x3x3 (partial replicate)	2*N - 3	1/6	1.5	N-2	$(\sigma_{wT}^2 + 2 * \sigma_{wR}^2)/3$
2x2x4 (full replicate)	3*N - 4	1/4	1	N-2	$(\sigma_{wT}^2 + \sigma_{wR}^2)/2$
2x2x3 (TRT|RTR)	2*N - 3	3/6	1.5	N/2 -1	$(\sigma_{wT}^2 + \sigma_{wR}^2)/2$
unbalanced				$n_2 - 1$	$w_1(2 * \sigma_{wT}^2 + \sigma_{wR}^2)/3$ $+ w_2(\sigma_{wT}^2 + 2 * \sigma_{wR}^2)/3$

b_k is the design constant assuming $n_i = N/seqs$.
n_1=n(TRT), n_2=n(RTR) and $w_i = n_i / (n_1 + n_2)$.
E(mse) is the expectation of the mean squared error from a model without subject by treatment interaction composed from the intra-subject variabilities of Test and Reference, respectively.

Simulation implementation

Instead of simulating subject data[1] and performing the above described evaluation it would be much more efficient according to a suggestion of Zheng et al.[2] to simulate the needed statistics for the BE decision methods via their associated distributions. This gives a boost in respect to the run times of the simulations from some hours to fraction of minutes.

A first attempt (implemented in PowerTOST V1.1-00, V1.1-02)

- pe is normal distributed with mean=log(GMR) and sd=sqrt($E(mse) * C2$)
  GMR is the true (assumed) ratio for the population.
- s_d^2*df/(E(mse)*C2) is chi-squared distributed and simulated via
  $s_d^2 = E(mse) * C2 * rchi(nsims, df)/df$
- $s_{wR}^2 * df_{RR}/\sigma_{wR}^2$ is chi-squared distributed and simulated via
  $s_{wR}^2 = \sigma_{wR}^2 * rchi(nsims, df_{RR})/df_{RR}$

With the so simulated statistics the above described method for the BE decision is performed. The cases of BE=TRUE will counted and pBE = count(BE=TRUE)/nsims is calculated.

The above described simulation attempt proved as too naïve.
The agreement of the power so calculated with values obtained via the 'classical' way of simulating subject data was not very satisfactory, especially for the partial replicate (2x3x3) crossover design. This holds true regardless of assuming equal variabilities for Test and Reference or not. See Appendix.

The conclusion could only be that the simulation of mse and s_{wR}^2 via *independent* chi-square distributions is not appropriate. One consequence of this attempt is that studies are simulated in which s_{wT}^2 if calculated via the relations given in the Table above becomes negative.

To avoid this it was simulated (V1.1-03 ff) as following:

- $s_{wR}^2 * df_{RR}/\sigma_{wR}^2$ is chi-squared distributed and simulated via
  $s_{wR}^2 = \sigma_{wR}^2 * rchi(nsims, df_{RR})/df_{RR}$
- $s_{wT}^2 * df_{TT}/\sigma_{wT}^2$ is chi-squared distributed and simulated via
  $s_{wT}^2 = \sigma_{wT}^2 * rchi(nsims, df_{TT})/df_{TT}$
- mse is calculated from the constituents s_{wR}^2 and s_{wT}^2 according to the relations given in the Table above and from that $s_d^2 = mse * C2$.

This approach however has the flaw that we are not able to give the df_{TT} in case of the 2x3x3 design. It was choosen equal to df_{RR}. So this approach is totally empirical for the 2x3x3 design and only justified by the better numeric agreement of the power values compared to those obatined via subject data simulations. It has further the flaw that within the EMA approach indeed negative variance components are imaginable, analogous to negativ inter-subject variances for the 2x2x2 crossover design.

A closer look at the results of V1.1-03 ff showed that the approach via simulations of the mse constituents overcorrects the power values for the 2x2x4 design. The new introduced 2x2x3 design showed the same behavior. Therefore it was decided to use the mean from both approaches from V1.1-07 on. Indeed this empirical attempt gave the most satisfactory agreement to the power results via subject data simulations. See Appendix.

<u>Open questions, understanding problems</u>:

1. Is there a better way to handle the simulations of mse and s_{wR}^2 via *dependent* chi-square distributions?

2. The E(mse) for the 2x3x3 (partial replicate) design was decided from subject data simulations. How can we derive this via theoretically arguments?

3. Is working with different variabilities within the EMA method reasonable at all? Or does the model used only allow equal variabilities?

   An indication for that is the observation that the EMA method and the FDA method via intra-subject contrasts (ISC, see next paragraph) lead to different expected standard errors of the mean T-R :

   EMA: $sqrt(\frac{1}{3}(\sigma_{wT}^2 + 2 * \sigma_{wR}^2) * \frac{1}{6} * \sum \frac{1}{n_i}) = sqrt((\sigma_{wT}^2/2 + \sigma_{wR}^2) * \frac{1}{9} * \sum \frac{1}{n_i}$

   FDA: $sqrt((\sigma_{wT}^2 + \sigma_{wR}^2/2) * \frac{1}{9} * \sum \frac{1}{n_i})$

   These formulas give only identical results if $\sigma_{wT}^2 = \sigma_{wR}^2$ is assumed.

   A second indication is the poor agreement of the power values compared to those obtained via subject data simulations for the "2x3x3" design in case of $\sigma_{wT}^2 \neq \sigma_{wR}^2$.
   In case of $CV_{wT} < CV_{wR}$ the power values via subject data simulations are markedly lower (more conservative), in case of $CV_{wT} > CV_{wR}$ they are markedly higher (more liberal) compared to the simulations of the key statistics *pe*, mse and σ_{wR}^2.

   To explore into this the question arose ***"How performs the EMA recommended evaluation of the replicate designs ("Use the same ANOVA model as for the classical 2x2x2 crossover") for deciding pure ABE in terms of type I error (alpha), especially if the homoscedasticity assumption is not true?"*** The following table summarizes the simulated alpha values via subject data:

ABE decision, design 2x3x3 (partial replicate) 1E6 sims

CVwT	CVwR	pooled CV	n	'alpha'	power.TOST	power.scABEL (details=T)
0.3	0.3	0.3	12	0.0440	0.0445	0.0459
			24	*0.0505*	0.0500	0.0505
			36	0.0500	0.0500	0.0502
0.4	0.4	0.4	12	0.0164	0.0164	0.0186
			24	0.0483	0.0482	0.0488
			36	0.0500	0.0500	0.0502
0.5	0.5	0.5	12	0.0027	0.0028	0.0038
			24	0.0323	0.0324	0.0330
			36	0.0484	0.0484	0.0487
0.3	0.4	0.3690	12	0.0202	0.0251	0.0294
			24	***0.0361***	0.0495	0.0507
			36	***0.0363***	0.0500	0.0507
0.3	0.5	0.4407	12	0.0068	0.0084	0.0140
			24	***0.0267***	0.0443	0.0459
			36	***0.0285***	0.0498	0.0508
0.4	0.3	0.3359	12	0.0434	0.0354	0.0362
			24	**0.0659**	0.0499	0.0500
			36	**0.0663**	0.0500	0.0500
0.5	0.3	0.3754	12	0.0319	0.0232	0.0241
			24	**0.0757**	0.0493	0.0495
			36	**0.0783**	0.0500	0.0500

'alpha' from subject data sims with CV's as given and GMR=1.25.
power.TOST results calculated with pooled CV
(= mse2CV((CV2mse(CVwT) + 2* CV2mse(CVwR))/3)).

Wow! While performing almost as expected if $\sigma_{wT}^2 = \sigma_{wR}^2$ the empirical alpha values are much too conservative in case of $CV_{wT} < CV_{wR}$. In case of $CV_{wT} > CV_{wR}$ they are too liberal up to an alpha inflation!

This observation resembles well known results for one-way or two-way ANOVA, showing that the usual F-test for testing effects are no longer valid if the assumption of equal variances is violated.

The only explanation could be that the distributional assumptions ("mse is chi-squared distributed") no longer holds if the homoscedasticity is not true. As far as I know there is no way out here since there is no solution to the question of the mse distribution within the crossover ANOVA for the case of heteroscedasticity, beside to use mixed model software. Moreover the EMA forced us to use this fixed effects ANOVA without allowing mixed models evaluation.

Thus we had to stick with subject data simulations with the burden of very long simulation run-times if we wish to calculate empirical power for the EMA method within a 2x3x3 design in case of heteroscedasticity, i.e. $\sigma_{wT}^2 \neq \sigma_{wR}^2$.

FDA method (power.RSABE())

Method description in a cook book manner:

- Calculate the intra-subject contrasts **T-R** (of the log-transformed PK metrics) and analyze them via an ANOVA(1) with sequence as soley effect. The intercept of this ANOVA gives the point estimator (pe) of μ_T - μ_R.
  The std error associated with the pe is
  $$s_d = sqrt(mse_1 * \frac{1}{seqs^2} * \sum \frac{1}{n_i})$$
  The associated degrees of freedom are df=N-seqs. The term $\frac{1}{seqs^2} * \sum \frac{1}{n_i}$ is named C3. In case of equal number of subjects in sequence groups n_i=N/seqs the term C3 reduces to 1/N.
- Calculate the intra-subject contrasts **R-R** (of the log-transformed PK metrics) and analyze them via an ANOVA(2) with sequence as solely effect. The intra-subject variance for the reference is $s_{wR}^2 = mse_2/2$. The associated degrees of freedom are also df(RR)=N-seqs.
- In case of the full replicate design (2x2x4) the previous step can be repeated for **T-T** to obtain s_{wT}^2. But this value isn't used further down. It's only a nice to have.
- If s_{wR} > 0.2935604 calculate the linearized reference scaled ABE criterion
  $$crit = pe^2 - s_d^2 - theta^2 * s_{wR}^2$$
  where theta = log(1.25)/0.25 = 0.8925742 is the regulatory constant set by the FDA.
  Calculate a 95% upper confidence interval of this criterion via Howe[3] approximation according to
  $$E_m = pe^2 - s_d^2$$
  $$C_m = (abs(pe) + t_{(1-alpha),df} * s_d)^2$$
  $$E_s = theta^2 * s_{wR}^2$$
  $$C_s = \mathrm{E}_s * df_{RR}/Chi_{(1-alpha),dfRR}$$
  $$bound = E_m - E_s + sqrt((C_m - E_m)^2 + (C_s - E_s)^2)$$
  Decide BE if the upper bound is lower than zero.
- If s_{wR} ≤ 0.2935604 (CV_{wR}≤ 0.3) then perform ABE evaluation, i.e. calculate 90% confidence intervals and decide BE if these are contained in the acceptance range $[-\log(1.25, \log(1.25)]$. The FDA demands to use the Proc MIXED code[4,5] for this evaluation, regardless of the design.

Simulation implementation

Instead of simulating via subject data we are simulating the needed statistics via their associated distributions:

- pe is normal distributed with mean=log(GMR) and sd = sqrt($E(mse_1) * C3$)
  GMR is the true (assumed) ratio for the population.
- s_d^2*df/(E(mse_1)*C3) is chi-squared distributed and simulated via
  $s_d^2 = E(mse_1) * C3 * rchi(nsims, df)/df$
- $s_{wR}^2 * df_{RR}/\sigma_{wR}^2$ is chi-squared distributed and simulated via
  $s_{wR}^2 = \sigma_{wR}^2 * rchi(nsims, df_{RR})/df_{RR}$

With the so simulated statistics the above described method for the BE decision is performed. The cases of BE=TRUE are counted and pBE = count(BE=TRUE)/nsims is calculated.

The expectation of the mse1 are taken from the literature about IBE as:

Design	E(mse$_1$)
2x3x3 (partial replicate) [6,7]	$\sigma_{wT}^2 + \sigma_{wR}^2/2$
2x2x4 (full replicate) [8, 9]	$(\sigma_{wT}^2 + \sigma_{wR}^2)/2$
2x2x3 (TRT|RTR)	$(\sigma_{wT}^2 + \sigma_{wR}^2)/2$
unbalanced	$w_1(\sigma_{wT}^2 + \sigma_{wR}^2/2) + w_2(\sigma_{wT}^2/2 + \sigma_{wR}^2)$

$w_1 = df_{RR}\,/(df_{TT} + df_{RR}),\ w_2 = df_{TT}\,/(df_{TT} + df_{RR})$

<u>Open questions, understanding problems</u>:

1. The ABE evaluation (90% CI's) in case of $s_{wR} \leq 0.2935604$ ($CV_{wR} \leq 0.3$) is done via the results from the ANOVA(1), i.e. we calculate the 90%CI with pe and s_d from that step. How does this affect the results? How could we test this?
   If there is a considerable effect, how can we then simulate the ABE decision?

2. The "unknown x", i.e. the term $-s_d^2$ in E_m (taken from the SAS code of the progesterone guidance[5]): Where did it came from? Have the two Laszlo's used it in their simulations? Their earlier papers do not contain this term.
   Mueller-Cohrs[10] notes that pe^2 is only approximately unbiased for (μT-μR)^2 and a user in the BEBA forum (http://forum.bebac.at/mix_entry.php?id=5943) gave the hint that it may be the bias correction which is done by subtraction of s_d^2.

The design 2x2x3 (TRT|RTR) has the peculiarity that the intrasubject contrasts for T versus R have different variances in the two sequence groups. See Chow & Liu, Chapter 9.3.4.[11]
Pooling of the sequence groups therefore has the danger that in case of heteroscedasticity the BE decision via an ANOVA with sequence as effect may be too conservative or too liberal.

To explore this some spare subject data sims of the ABE decision (90% CI's in the usual acceptance range 0.8 – 1.25) are performed. The results are shown in the next Table:

Design 2x2x3 (TRT|RTR), 1E6 sims

CVwT	CVwR	n1	n2	EMA ANOVA	FDA ISC
0.5	0.2	18	18	0.0509	0.0496*
		21	15	0.0509	**0.0559***
		15	21	0.0507	***0.0430****
		24	24	0.0505	
0.5	0.3	18	18	0.0503	0.0502
		19	17		0.0516
		21	15	0.0502	**0.0541**
		15	21	0.0503	***0.0459***
0.2	0.5	18	18	0.0509	0.0500*
		21	15	0.0511	***0.0438****
		15	21	0.0506	**0.0570***
		24	24	0.0505	
0.3	0.5	18	18	0.0505	0.0504
		19	17		0.0490
		21	15	0.0505	***0.0469***
		15	21	0.0503	**0.0542**

n1 = n(TRT), n2 = n(RTR)
* 1E5 sims only

There is no hint of an alpha-inflation or too conservative alpha values if one uses the EMA recommended evaluation (same ANOVA as for the classical 2x2x2 crossover), regardless of grade of heteroscedasticity and unbalancedness analyzed.

The evaluation via intra-subject contrasts also does not show noticeable deviations from the nominal level 0.05 as long as the design is balanced or only slightly unbalanced. Alpha-inflation or too conservative type I error values are only observed if the sequence groups are strongly unbalanced. Thus this design is much more 'friendly' in respect to heteroscedasticity than the 2x3x3 design in the EMA evaluation.

An appreciable effect on the power values of the RSABE method is not observed unless the design is very unbalanced. See Appendix.

FDA method for NTID's (`power.NTIDFDA()`)

Method description[12] in a cook book manner, design 2x2x4:

- Calculate the intra-subject contrasts **T-R** (of the log-transformed PK metrics) and analyze them via an ANOVA(1) with sequence as soley effect. The intercept of this ANOVA gives the point estimator (pe) of $\mu_T - \mu_R$.
  The std error associated with the pe is
  $$s_d = sqrt(mse_1 * \frac{1}{4} * \sum \frac{1}{n_i})$$
  The associated degrees of freedom are df = N-2. The term $\frac{1}{4} * \sum \frac{1}{n_i}$ is named C3. In case of equal number of subjects in sequence groups n_i=N/2 the term C3 reduces to 1/N.
- Calculate the intra-subject contrasts **R-R** (of the log-transformed PK metrics) and analyze them via an ANOVA(2) with sequence as soley effect. The intra-subject variance for the reference is $s_{wR}^2 = mse_2/2$. The associated degrees of freedom are dfRR = N-2.
- Repeated the previous step for **T-T** to obtain s_{wT}^2. This variance has the associated degrees of freedom dfTT = N-2 (= dfRR).
- Calculate the linearized reference scaled ABE criterion according to
  $$crit = pe^2 - s_d^2 - theta^2 * s_{wR}^2$$
  where theta = -log(0.9)/0.1 = 1.053605157.
  Calculate a 95% upper confidence interval of this criterion via Howe[3] approximation according to
  $$E_m = pe^2 - s_d^2$$
  $$C_m = (abs(pe) + t_{(1-alpha),df} * s_d)^2$$
  $$E_s = theta^2 * s_{wR}^2$$
  $$C_s = \mathrm{E}_s * df_{RR}/Chi_{(1-alpha),dfRR}$$
  $$bound = E_m - E_s + sqrt((C_m - E_m)^2 + (C_s - E_s)^2)$$
- Decide BE if the upper confidence limit of the linearized reference scaled ABE criterion is ≤0 and if the conventional ABE test (90% CI of T versus R within ABE acceptance range) shows BE. The latter is operational identical to placing a cap at CV_{wR}>=0.2142 ($s_{wR} = \log(1.25)/theta = 0.2117905$) on the widening of the implied acceptance limits.
- Additionally the ratio of s_{wT}/s_{wR} should be ≤2.5. This is tested by calculating an upper confidence interval of this ratio via
  $$UL = \frac{s_{wT}/s_{wR}}{\sqrt{F_{1-alpha/2,dfTT,dfRR}}} \leq 2.5$$
  where $F_{1-alpha/2,dfTT,dfRR}$ is the value of the F-distribution with ν_1=dfTT and ν_2=dfRR degrees of freedom that has probability 1-alpha/2 to its **right** (see [12]).
  In R: `Fval=qf(1-alpha/2,dfTT,dfRR,lower.tail=FALSE)`.
  Alpha is set =0.1.

Simulation implementation

Instead of simulating via subject data we are simulating the needed statistics via their associated distributions:

- pe is normal distributed with mean=log(GMR) and sd=sqrt($E(mse_1) * C3$)
  GMR is the true (assumed) ratio for the population.
- $s_d^2 * df/(E(mse_1) * C3)$ is chi-squared distributed and simulated via
  $s_d^2 = E(mse_1) * C3 * rchi(nsims, df)/df$
- $s_{wR}^2 * df_{RR}/\sigma_{wR}^2$ is chi-squared distributed and simulated via
  $s_{wR}^2 = \sigma_{wR}^2 * rchi(nsims, df_{RR})/df_{RR}$
- $s_{wT}^2 * df_{TT}/\sigma_{wT}^2$ is chi-squared distributed and simulated via
  $s_{wT}^2 = \sigma_{wT}^2 * rchi(nsims, df_{TT})/df_{TT}$

E(mse1) is taken as $\sigma_D^2 + (\sigma_{wT}^2 + \sigma_{wR}^2)/2$. See above and reference[7]. The subject by formulation interaction term σ_D^2 is assumed to be zero. It is only present for an eventually enhancement in future. The σ_{xY}^2 are the population values for the respective variances.

With the so simulated statistics the above described method for the BE decision is performed. The cases of BE=TRUE are counted (implies BE(ABE) =TRUE, BE(scABE)=TRUE and ratio $s_{wT}/s_{wR} \leq 2.5$). From the counts pBE = count(BE=TRUE)/nsims is calculated as 'empirical' power.

Open questions, understanding problems:

- The ABE evaluation (90% CI's) is done via the results from the ANOVA(1), i.e. we calculate the 90%CI with pe and s_d from that step. The FDA on the other hand recommends to do that by the Proc MIXED code for SAS[12]. This is scarcely implementable in R.
  How does this affect the results? How could we test this?
  If there is a considerable effect, how can we then simulate the ABE decision?

Appendix: Results of simulations via subject data

EMA method, GMR=0.95, 5E+5 subject sims if not otherwise given
power.scABEL() with nsims=1E6

CVwT	CVwR	n	sims	pBE	power.scABEL			
					V1.1-02c	Diff.	V1.1-07	Diff.
Design 2x3x3								
0.2	0.2	12		0.7522	0.7517	0.0005	0.7517	0.0005
		24		0.9617	0.9618	-0.0001	0.9615	0.0002
0.3	0.3	12		0.4066	0.3958	**0.0108**	0.4112	-0.0046
		24		0.7790	0.7714	0.0076	0.7818	-0.0028
		48		0.9632	0.9698	-0.0066	0.9636	-0.0004
0.4	0.4	12		0.2899	0.2984	-0.0085	0.2896	0.0003
		12	1E6	0.2895	0.2984	-0.0089	0.2896	-0.0001
		24		0.7390	0.7233	**0.0157**	0.7456	-0.0066
		24	1E6	0.7398	0.7233	**0.0165**	0.7456	-0.0058
		48		0.9597	0.9543	0.0054	0.9612	-0.0015
0.5	0.5	12		0.1954	0.2202	**-0.0248**	0.1907	0.0047
		12	1E6	0.1952	0.2202	**-0.0250**	0.1907	0.0045
		24		0.7048	0.6953	0.0095	0.7091	-0.0043
		48		0.9620	0.9579	0.0041	0.9630	-0.0010
0.3	0.5	12		0.3762	0.3486	**0.0276**	0.3497	**0.0265**
		24		0.8623	0.8042	**0.0581**	0.8215	**0.0408**
		48		0.9934	0.9810	**0.0134**	0.9853	0.0081
0.5	0.3	12		0.1452	0.1625	**-0.0173**	0.1489	-0.0037
		24		0.5179	0.5585	**-0.0406**	0.5652	**-0.0473**
		48		0.8284	0.8657	**-0.0373**	0.8690	**-0.0406**
Design 2x2x4								
0.2	0.2	12		0.9024	0.9017	0.0007	0.9031	-0.0007
		24		0.9948	0.9949	-0.0001	0.9950	-0.0002
0.3	0.3	12		0.6553	0.6447	**0.0106**	0.6538	0.0015
		12	1E6	0.6552	0.6447	**0.0105**	0.6538	0.0014
		24		0.9120	0.9080	0.0040	0.9115	0.0005
		36		0.9771	0.9756	0.0015	0.9770	0.0001
0.4090	0.4090	12	1E6	0.5482	0.5334	**0.0148**	0.5446	0.0036
		24	1E6	0.8878	0.8780	0.0098	0.8852	0.0026
		36		0.9695	0.9657	0.0038	0.9687	0.0008
0.5	0.5	12		0.4705	0.4662	0.0043	0.4676	0.0029
		24		0.8787	0.8718	0.0069	0.8771	0.0016
		36		0.9710	0.9677	0.0033	0.9706	0.0004
0.3	0.5	12		0.6955	0.6770	**0.0185**	0.6920	0.0035
		24		0.9603	0.9526	0.0077	0.9590	0.0013
		36		0.9931	0.9916	0.0015	0.9929	0.0003
0.5	0.3	12		0.3011	0.2971	0.0040	0.2964	0.0047
		24		0.6974	0.6949	0.0025	0.6976	-0.0002
		36		0.8586	0.8561	0.0025	0.8576	0.0010

Red: abs(diff)>0.002; Red, bold: abs(diff)>0.01

Agreement not perfect but – except the calculations with CVwT ≠CVwR for design "2x3x3" – to some degree satisfactory for me.

EMA method
Design 2x2x3 (TRT|RTR) (new)
GMR=0.95, 5E+5 subject data sims

CVwT	CVwR	n1	n2	pBE	power.scABEL()	diff
0.3	0.3	12	12	0.7884	0.7891	-0.0007
		18	18	0.9144	0.9138	0.0006
		21	15	0.9080	0.9084	-0.0004
0.4	0.4	12	12	0.7137	0.7146	-0.0009
		18	18	0.8803	0.8802	0.0001
		21	15	0.8659	0.8665	-0.0006
0.5	0.5	12	12	0.6574	0.6579	-0.0005
		18	18	0.8669	0.8672	-0.0003
		21	15	0.8468	0.8476	-0.0008
		15	21	0.8660	0.8649	0.0011
0.5	0.3	12	12	0.4857	0.4867	-0.0010
		18	18	0.7043	0.7039	0.0004
		21	15	0.6815	0.6817	-0.0002
		15	21	0.6987	0.6981	0.0006

Agreement totally satisfactory for me.

FDA method, GMR=0.95, 1E5 subject data sims if not otherwise given

CVwT	CVwR	n	sims	pBE	power.RSABE	Diff
Design 2x3x3						
0.2	0.2	12		0.7106	0.7108	-0.0002
		24		0.9560	0.9561	-0.0001
0.3	0.3	12		0.4123	0.4132	-0.0009
		24		0.7980	0.7990	-0.0010
		48		0.9700	0.9691	0.0009
0.40898	0.40898	12		0.3808	0.3801	0.0006
		24		0.8089	0.8104	-0.0016
		48		0.9831	0.9827	0.0004
0.5	0.5	12		0.3795	0.3779	0.0017
		24		0.8132	0.8153	-0.0020
		48		0.9763	0.9765	-0.0003
0.3	0.5	12		0.6296	0.6289	0.0006
		24		0.9406	0.9416	-0.0009
		48		0.9962	0.9961	0.0001
Design 2x2x4						
0.2	0.2	12		0.8737	0.8744	0.0007
		24		0.9931	0.9933	-0.0002
0.3	0.3	12		0.6374	0.6321	0.0054
		12	1E6	0.6355	0.6348	0.0007
		24		0.9172	0.9165	0.0006
		48		0.9948	0.9948	0.0000
0.40898	0.40898	12		0.5933	0.5913	0.0020
		24		0.9234	0.9231	0.0003
		48		0.9968	0.9971	-0.0003
0.5	0.5	12		0.5912	0.5903	0.0009
		24		0.9238	0.9235	0.0003
		48		0.9935	0.9938	-0.0002
0.3	0.5	12		0.7491	0.7483	0.0008
		24		0.9709	0.9710	-0.0002
		48		0.9986	0.9990	-0.0004
0.5	0.3	12		0.3263	0.3264	-0.0002
		24		0.7264	0.7244	0.0020
		48		0.9457	0.9444	0.0014

Red: abs(diff)>0.002

Agreement totally satisfactory for me.

FDA method, Design 2x2x3 (TRT|RTR) (new)

GMR=0.95, 1E5 subject data sims

CVwT	CVwR	n1	n2	pBE	power.RSABE	diff
0.3	0.3	12	12	0.7967	0.7955	0.0012
		18	18	0.9204	0.9208	-0.0004
		21	15	0.9133	0.9140	-0.0007
0.4	0.4	12	12	0.7626	0.7612	0.0014
		18	18	0.9139	0.9133	0.0006
		21	15	0.8971	0.8975	-0.0004
0.5	0.5	12	12	0.7604	0.7588	0.0016
		18	18	0.9153	0.9145	0.0008
		21	15	0.8969	0.8972	-0.0003
0.5	0.3	12	12	0.5178	0.5167	0.0011
		18	18	0.7362	0.7355	0.0007
		21	15	0.7278	0.7342	-0.0064
0.3	0.5	12	12	0.8725	0.8719	0.0006
		18	18	0.9659	0.9653	0.0006
		21	15	0.9572	0.9526	0.0046

GMR=1.25, 1E6 subject data sims

CVwT	CVwR	n1	n2	pBE	power.RSABE	diff
0.5	0.2	18	18		0.0501	
		21	15		0.0503	
		15	21		0.0499	
0.5	0.3	18	18	0.1127	0.1124	0.0003
		19	17	0.1151	0.1135	0.0016
		21	15	0.1045	0.1138	-0.0093
		15	21	0.1184	0.1094	0.0090
0.2	0.5	18	18		0.4473	
		21	15		0.4318	
		15	21		0.4590	
0.3	0.5	18	18	0.4418	0.4417	0.0001
		19	17	0.4366	0.4367	-0.0001
		21	15	0.4232	0.4246	-0.0014
		15	21	0.4527	0.4518	0.0009

Agreement totally satisfactory for me, except the red marked cases of heavy in-balance.

FDA method for NTID's, design 2x2x4, 1E+5 sims if not otherwise given

GMR=0.95

CVwT	CVwR	n	sims	pBE	power.NTIDFDA	Diff
GMR=0.95						
0.05	0.05	12		0.0564	0.0583	-0.0019
		24		0.0644	0.0633	0.0011
0.075	0.075	12		0.2492	0.2505	-0.0014
		24		0.4266	0.4283	-0.0017
		48		0.6738	0.6707	0.0030
0.1	0.1	12		0.4037	0.4029	0.0008
		24		0.6871	0.6865	0.0006
		48		0.9222	0.9198	0.0023
0.125	0.125	12		0.4982	0.4968	0.0014
		24		0.8134	0.8123	0.0012
		48		0.9774	0.9752	0.0021
0.15	0.15	12		0.5597	0.5568	0.0029
		24		0.8762	0.8749	0.0013
		48		0.9914	0.9911	0.0003
0.175	0.175	12		0.5954	0.5939	0.0014
		24		0.9090	0.9094	-0.0004
		36		0.9804	0.9800	0.0004
0.2	0.2	12		0.6115	0.6103	0.0012
		24		0.9288	0.9291	-0.0003
		36		0.9874	0.9871	0.0003
0.3	0.3	12		0.4364	0.4357	0.0007
		24		0.8610	0.8607	0.0003
		36		0.9638	0.9627	0.0011
0.125	0.175	12		0.7231	0.7209	0.0022
		24		0.9526	0.9518	0.0008
		36		0.9926	0.9925	0.0001
0.175	0.125	12		0.2861	0.2872	-0.0011
		24		0.6338	0.6359	-0.0021
		36		0.8302	0.8306	-0.0004

Red: abs(diff)>0.002

Agreement satisfactory for me.

References

[1] Laszlo Tothfalusi and Laszlo Endrenyi
"Sample Sizes for Designing Bioequivalence Studies for Highly Variable Drugs"
J. Pharm. Pharmaceut. Sci. (www.cspsCanada.org) 15(1) 73 - 84, 2011
http://ejournals.library.ualberta.ca/index.php/JPPS/article/download/11612/9489

[2] Zheng C., Wang J. and Zhao L.
"Testing bioequivalence for multiple formulations with power and sample size calculations"
Pharmaceut. Statist. 2012, 11 334-341

[3] Howe W.G.
"Approximate confidence limits on the mean of X+Y where X and Y are two tabled independent random variables "
J. of the American Statistical Association 1974, 69(347): 789-794

[4] FDA Draft guidance "Statistical Approaches to Establishing Bioequivalence"
CDER January 2001, Appendix E
http://www.fda.gov/downloads/Drugs/Guidances/ucm070244.pdf

[5] FDA "Draft Guidance on Progesterone"
Recommended Apr 2010; Revised Feb 2011
http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM209294.pdf

[6] McNally R.J.
"Tests for Individual and Population Bioequivalence Using 3-Period Crossover Designs"
http://www.stat.colostate.edu/statresearch/stattechreports/Technical%20Reports/2002/02-7%20McNally.pdf

[7] Shein-Chung Chow, Jun Shao, and Hansheng Wang
"Individual Bioequivalence testing under 2×3 designs"
Statist. Med. 2002; 21:629–648
http://hansheng.gsm.pku.edu.cn/pdf/2002/IBE.pdf

[8] McNally R.J., Iyer H., Mathew T.
"Tests for Individual and Population Bioequivalence Based on Generalized p-Values"
Statist. Med. 2003 Jan 15;22(1):31-53.
http://www.stat.colostate.edu/statresearch/stattechreports/Technical%20Reports/2001/01-11%20McNally%20Iyer%20Mathew.pdf

[9] Hansheng Wang and Shein-Chung Chow
"On statistical power for average Bioequivalence testing under Replicated crossover designs"
J. of Biopharmaceutical Statistics Vol. 12, No. 3, pp. 295–309, 2002
http://hansheng.gsm.pku.edu.cn/pdf/2002/power.pdf

[10] Mueller-Cohrs J.
"Analysis of a three-period two-treatment pharmacokinetic study to assess scaled average bioequivalence "
PHUSE 2009, Paper SP04
www.phusewiki.org/docs/2009%20PAPERS/SP04.pdf

[11] Chow, S.-C. and Liu, J.-P.
"Design and Analysis of Bioavailability and Bioequivalence Studies"
Third edition, CRC/Chapman & Hall, Boca-Raton 2009

[12] FDA "Draft Guidance on Warfarine Sodium"
Recommended Dec 2012
http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201283.pdf