Sample Size Calculations
...or the Myth of Power

Helmut Schütz
BEBAC
Sample Size (Limits)

- Minimum
  - 12: WHO, EU, CAN, NZ, AUS, AR, MZ, ASEAN States, RSA
  - 12: USA ‘A pilot study that documents BE can be appropriate, provided its design and execution are suitable and a sufficient number of subjects (e.g., 12) have completed the study.’
  - 20: RSA (MR formulations)
  - 24: Saudia Arabia (12 to 24 if statistically justifiable)
  - 24: Brazil
  - Sufficient number: JPN
Sample Size (Limits)

- **Maximum**
  - **NZ:** ‘If the calculated number of subjects appears to be higher than is ethically justifiable, it may be necessary to accept a statistical power which is less than desirable. Normally it is not practical to use more than about 40 subjects in a bioavailability study.’
  - **All others:** Not specified (judged by IEC/IRB or local Authorities). ICH E9, Section 3.5 applies: ‘The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed.’
Power & Sample Size

• Reminder

- Generally power is set to at least 80% ($\beta$, error type II: producers’s risk to get no approval for a bioequivalent formulation; power = $1 - \beta$).

  1 out of 5 studies will fail just by chance!

- If you plan for power of less than 70%, problems with the ethics committee are likely (ICH E9).

- If you plan for power of more than 90% (especially with low variability drugs), problems with the regulator are possible (‘forced bioequivalence’).

- Add subjects (‘alternates’) according to the expected drop-out rate – especially for studies with more than two periods or multiple-dose studies.
US FDA, Canada TPD

- Statistical Approaches to Establishing Bioequivalence (2001)
  - Based on maximum difference of 5%.
  - Sample size based on 80% – 90% power.

- Draft GL (2010)
  - Consider potency differences.
  - Sample size based on 80% – 90% power.
  - *Do not* interpolate linear between CVs (as stated in the GL)!
EU

• EMEA NfG on BA/BE (2001)
  ▪ Detailed information (data sources, significance level, expected deviation, desired power).

• EMA GL on BE (2010)
  ▪ Batches must not differ more than 5%.
  ▪ The number of subjects to be included in the study should be based on an appropriate sample size calculation.

Cookbook?
Hierarchy of Designs

- The more ‘sophisticated’ a design is, the more information can be extracted.

  - Hierarchy of designs:
    - Full replicate (TRTR | RTRT)
    - Partial replicate (TRR | RTR | RRT)
    - Standard 2×2 cross-over (RT | RT)
    - Parallel (R | T)

  - Variances which can be estimated:
    - Parallel: total variance (between + within)
    - 2×2 Xover: + between, within subjects
    - Partial replicate: + within subjects (reference)
    - Full replicate: + within subjects (reference, test)
Coefficient(s) of Variation

From any design one gets variances of lower design levels also.

- Total CV% from a 2×2 cross-over used in planning a parallel design study:
  - Intra-subject CV% (within)
  - Inter-subject CV% (between)
  - Total CV% (pooled)

\[
CV_{\text{intra}} \% = 100 \cdot \sqrt{e^{MSE_w} - 1}
\]

\[
CV_{\text{inter}} \% = 100 \cdot \sqrt{\frac{MSE_B - MSE_w}{MSE_w}} - 1
\]

\[
CV_{\text{total}} \% = 100 \cdot \sqrt{\frac{MSE_B + MSE_w}{MSE_w} - 1}
\]
**Coefficient(s) of Variation**

- CVs of *higher* design levels not available.
  - If only mean ± SD of reference is available…
    - Avoid ‘rule of thumb’ $\text{CV}_{\text{intra}} = 60\%$ of $\text{CV}_{\text{total}}$
    - Don’t plan a cross-over based on $\text{CV}_{\text{total}}$
    - Examples (cross-over studies)

<table>
<thead>
<tr>
<th>drug, formulation</th>
<th>design</th>
<th>n</th>
<th>metric</th>
<th>$\text{CV}_{\text{intra}}$</th>
<th>$\text{CV}_{\text{inter}}$</th>
<th>$\text{CV}_{\text{total}}$</th>
<th>% intra/total</th>
</tr>
</thead>
<tbody>
<tr>
<td>methylphenidate MR</td>
<td>SD</td>
<td>12</td>
<td>$\text{AUC}_t$</td>
<td>7.00</td>
<td>19.1</td>
<td>20.4</td>
<td>34.3</td>
</tr>
<tr>
<td>paroxetine MR</td>
<td>MD</td>
<td>32</td>
<td>$\text{AUC}_\tau$</td>
<td>25.2</td>
<td>55.1</td>
<td>62.1</td>
<td>40.6</td>
</tr>
<tr>
<td>lansoprazole DR</td>
<td>SD</td>
<td>47</td>
<td>$C_{\text{max}}$</td>
<td>47.0</td>
<td>25.1</td>
<td>54.6</td>
<td>86.0</td>
</tr>
</tbody>
</table>

- Pilot study unavoidable, unless
- Two-stage sequential design is used
Hints

- Literature search for CV%
  - Preferably other BE studies (the bigger, the better!)
  - PK interaction studies (Cave: Mainly in steady state! Generally lower CV than after SD).
  - Food studies (CV higher/lower than fasted!)
  - If $CV_{\text{intra}}$ not given (quite often), a little algebra helps. All you need is the 90% geometric confidence interval and the sample size.
Algebra...

- **Calculation of CV\textsubscript{intra} from CI**
  - Point estimate ($PE$) from the Confidence Limits
    \[ PE = \sqrt{CL_{lo} \cdot CL_{hi}} \]
  - Estimate the number of subjects / sequence (example 2×2 cross-over)
    - If total sample size ($N$) is an even number, *assume* (!)
      \[ n_1 = n_2 = \frac{1}{2}N \]
    - If $N$ is an odd number, *assume* (!)
      \[ n_1 = \frac{1}{2}N + \frac{1}{2}, \quad n_2 = \frac{1}{2}N - \frac{1}{2} \quad (not \; n_1 = n_2 = \frac{1}{2}N!) \]
  - Difference between one $CL$ and the $PE$ in log-scale; use the $CL$ which is given with more significant digits
    \[ \Delta_{CL} = \ln PE - \ln CL_{lo} \quad or \quad \Delta_{CL} = \ln CL_{hi} - \ln PE \]
Algebra...

- Calculation of $CV_{\text{intra}}$ from CI (cont’d)
  - Calculate the Mean Square Error ($MSE$)

$$MSE = 2 \left( \frac{\Delta_{CL}}{\sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right) \cdot t_{1-2\alpha,n_1+n_2-2}}} \right)^2$$

- $CV_{\text{intra}}$ from $MSE$ as usual

$$CV_{\text{intra}} \% = 100 \cdot \sqrt{e^{MSE} - 1}$$
Calculation of $CV_{\text{intra}}$ from CI (cont’d)

Example: 90% CI $[0.91 - 1.15]$, N 21 ($n_1 = 11$, $n_2 = 10$)

\[
PE = \sqrt{0.91 \cdot 1.15} = 1.023
\]

\[
\Delta_{CL} = \ln 1.15 - \ln 1.023 = 0.11702
\]

\[
MSE = 2 \left( \frac{0.11702}{\sqrt{\left( \frac{1}{11} + \frac{1}{10} \right) \times 1.729}} \right)^2 = 0.04798
\]

\[
CV_{\text{intra}} \% = 100 \times \sqrt{e^{0.04798} - 1} = 22.2\%
\]
**Algebra...**

- **Proof: CI from calculated values**
  - **Example:** 90% CI \([0.91 - 1.15]\), N 21 (n_1 = 11, n_2 = 10)

\[
\begin{align*}
\ln PE &= \ln \sqrt{CL_{lo} \cdot CL_{hi}} = \ln \sqrt{0.91 \times 1.15} = 0.02274 \\
SE_\Delta &= \sqrt{\frac{2 \cdot MSE}{N}} = \sqrt{\frac{2 \times 0.04798}{21}} = 0.067598 \\
CI &= e^{\ln PE \pm t \cdot SE_\Delta} = e^{0.02274 \pm 1.729\times0.067598} \\
CI_{lo} &= e^{0.02274 - 1.729\times0.067598} = 0.91 \\
CI_{hi} &= e^{0.02274 + 1.729\times0.067598} = 1.15
\end{align*}
\]

\(\checkmark\)
Sensitivity to Imbalance

- If the study was more imbalanced than assumed, the estimated CV is conservative.
  - Example: 90% CI [0.89 – 1.15], N 24 (n₁ = 16, n₂ = 8, but not reported as such); CV 24.74% in the study.

<table>
<thead>
<tr>
<th>n₁</th>
<th>n₂</th>
<th>CV%</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>12</td>
<td>26.29</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>26.20</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>25.91</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>25.43</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>24.74</td>
</tr>
</tbody>
</table>
No Algebra...

- Implemented in R-package *PowerTOST*, function *CVfromCI* (not only 2×2 cross-over, but also parallel groups, higher order cross-overs, replicate designs). Previous example:

```r
require(PowerTost)
CVfromCI(lower=0.91, upper=1.15, n=21, design = "2x2", alpha = 0.05)
[1] 0.2219886
```
Sample Size Calculations …or the Myth of Power

Literature data

Doxicycline (37 studies from Blume/Mutschler, Bioäquivalenz: Qualitätsbewertung wirkstoffgleicher Fertigarzneimittel, GOVI-Verlag, Frankfurt am Main/Eschborn, 1989-1996)
Pooling of CV%

- Intra-subject CV from different studies can be pooled (LA Gould 1995, Patterson and Jones 2006)
  - In the parametric model of log-transformed data, additivitiy of variances (not of CVs!) apply.
  - Do not use the arithmetic mean (or the geometric mean either) of CVs.
  - Before pooling variances must be weighted according to the studies’ sample size – larger studies are more influential than smaller ones.
Pooling of CV%

- Intra-subject CV from different studies
  - Calculate the variance from CV
    \[ \sigma_w^2 = \ln(CV_{\text{intra}}^2 + 1) \]
  - Calculate the total variance weighted by df
    \[ \sum \sigma_w^2 df \]
  - Calculate the pooled CV from total variance
    \[ CV = \sqrt{e^{\sum \sigma_w^2 df / \sum df} - 1} \]
  - Optionally calculate an upper \((1-\alpha)\) % confidence limit on the pooled CV (recommended \(\alpha = 0.25\))
    \[ CL_{CV} = \sqrt{e^{\sum \sigma_w^2 df / \chi^2_{\alpha, \sum df}} - 1} \]
## Pooling of CV%

### Example 1: $n_1 = n_2$;

$CV_{\text{Study}1} < CV_{\text{Study}2}$

<table>
<thead>
<tr>
<th>studies</th>
<th>N</th>
<th>df (total)</th>
<th>$\alpha$</th>
<th>$1-\alpha$</th>
<th>total</th>
<th>$CV_{\text{pooled}}$</th>
<th>$CV_{\text{mean}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>24</td>
<td>20</td>
<td>0.25</td>
<td>0.75</td>
<td>1.2540</td>
<td>0.254</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\chi^2(\alpha, df)$</td>
<td>15.452</td>
<td></td>
<td></td>
<td>0.291</td>
<td>+14.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$CV_{\text{intra}}$</th>
<th>n</th>
<th>seq.</th>
<th>df (mj)</th>
<th>$\sigma_W$</th>
<th>$\sigma^2_W$</th>
<th>$\sigma^2_W \times df$</th>
<th>$CV_{\text{intra/pooled}}$</th>
<th>$&gt;CL_{\text{upper}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.200</td>
<td>12</td>
<td>2</td>
<td>10</td>
<td>0.198</td>
<td>0.0392</td>
<td>0.3922</td>
<td>78.6%</td>
<td>no</td>
</tr>
<tr>
<td>0.300</td>
<td>12</td>
<td>2</td>
<td>10</td>
<td>0.294</td>
<td>0.0862</td>
<td>0.8618</td>
<td>117.9%</td>
<td>yes</td>
</tr>
</tbody>
</table>
Pooling of CV%

Example 2: $n_1 < n_2$;

$CV_{\text{Study1}} < CV_{\text{Study2}}$

<table>
<thead>
<tr>
<th>studies</th>
<th>N</th>
<th>df (total)</th>
<th>$\alpha$</th>
<th>$1-\alpha$</th>
<th>total</th>
<th>$CV_{\text{pooled}}$</th>
<th>$CV_{\text{mean}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>36</td>
<td>32</td>
<td>0.25</td>
<td>0.75</td>
<td>2.2881</td>
<td>0.272</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\chi^2(\alpha, df)$</td>
<td></td>
<td></td>
<td></td>
<td>26.304</td>
<td>0.301</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$CV_{\text{intra}}$</th>
<th>n</th>
<th>df (mj)</th>
<th>$\sigma_W$</th>
<th>$\sigma^2_W$</th>
<th>$\sigma^2_W \times df$</th>
<th>$CV_{\text{intra/pooled}}$</th>
<th>$&gt;CL_{\text{upper}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.200</td>
<td>12</td>
<td>10</td>
<td>0.198</td>
<td>0.0392</td>
<td>0.3922</td>
<td>73.5%</td>
<td>no</td>
</tr>
<tr>
<td>0.300</td>
<td>24</td>
<td>22</td>
<td>0.294</td>
<td>0.0862</td>
<td>1.8959</td>
<td>110.2%</td>
<td>no</td>
</tr>
</tbody>
</table>
### Pooling of CV%

**Example 3:** \( n_1 > n_2; \)

\[ CV_{\text{Study1}} < CV_{\text{Study2}} \]

<table>
<thead>
<tr>
<th>studies</th>
<th>N</th>
<th>df (total)</th>
<th>( \alpha )</th>
<th>( 1 - \alpha )</th>
<th>total</th>
<th>( CV_{\text{pooled}} )</th>
<th>( CV_{\text{mean}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>36</td>
<td>32</td>
<td>0.25</td>
<td>0.75</td>
<td>1.7246</td>
<td>0.235</td>
<td>0.245</td>
</tr>
</tbody>
</table>

\[ \chi^2(\alpha, df) = 26.304 \]

\[ \text{CV}_{\text{intra}} \times \text{df}_{\text{pooled}} > \text{CL}_{\text{upper}} \]

<table>
<thead>
<tr>
<th>( CV_{\text{intra}} )</th>
<th>n</th>
<th>seq.</th>
<th>df (mj)</th>
<th>( \sigma_W )</th>
<th>( \sigma_{W}^2 )</th>
<th>( \sigma_{W}^2 \times df )</th>
<th>( CV_{\text{intra}} / \text{pooled} )</th>
<th>&gt;CL_{\text{upper}}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.200</td>
<td>24</td>
<td>2</td>
<td>22</td>
<td>0.198</td>
<td>0.0392</td>
<td>0.8629</td>
<td>85.0%</td>
<td>no</td>
</tr>
<tr>
<td>0.300</td>
<td>12</td>
<td>2</td>
<td>10</td>
<td>0.294</td>
<td>0.0862</td>
<td>0.8618</td>
<td>127.5%</td>
<td>yes</td>
</tr>
</tbody>
</table>
Pooling of CV%

- R package *PowerTost* function *CVpooled*, data of last example.

```r
require(PowerTOST)
CVs <- (
  PKmetric | CV   | n  | design | source
  AUC     | 0.20 | 24 | 2x2    | study 1
  AUC     | 0.30 | 12 | 2x2    | study 2
)
txtcon <- textConnection(CVs)
CVdata <- read.table(txtcon, header=TRUE, sep=" |",
                    strip.white=TRUE, as.is=TRUE)
close(txtcon)
CVsAUC <- subset(CVdata, PKmetric=="AUC")
print(CVpooled(CVsAUC, alpha=0.25), digits=3, verbose=TRUE)

Pooled CV = 0.235 with 32 degrees of freedom
Upper 75% confidence limit of CV = 0.260
```
Pooling of CV%

- Or you may combine pooling with an estimated sample size based on uncertain CVs (we will see later what that means).

*R* package *PowerTost* function *expsampleN.TOST*, data of last example. CVs and degrees of freedom must be given as vectors:

\[ CV = c(0.2, 0.3), \text{dfCV} = c(22, 10) \]
Pooling of CV%

require(PowerTOST)
expsampleN.TOST(alpha=0.05, targetpower=0.8, theta1=0.8, theta2=1.25, theta0=0.95, CV=c(0.2,0.3), dfCV=c(22,10), alpha2=0.05, design="2x2", print=TRUE, details=TRUE)

+++++++++ Equivalence test - TOST ++++++++ Sample size est. with uncertain CV
---------------------------------------
Study design: 2x2 crossover
Design characteristics:
df = n-2, design const. = 2, step = 2

log-transformed data (multiplicative model)

alpha = 0.05, target power = 0.8
BE margins = 0.8 ... 1.25
Null (true) ratio = 0.95
Variability data
   CV df
   0.2 22
   0.3 10
CV(pooled) = 0.2353158 with 32 df
one-sided upper CL = 0.2995364 (level = 95%)

Sample size search
   n  exp. power
   24  0.766585
   26  0.800334
\( \alpha \)- vs. \( \beta \)-Error

- \( \alpha \)-Error: **Patient’s risk** to be treated with a bioinequivalent formulation.
  - Although \( \alpha \) is generally set to 0.05, sometimes <0.05 (e.g., NTDIs in Brazil, multiplicity, interim analyses).

- \( \beta \)-Error: **Producer’s risk** to get no approval for a bioequivalent formulation.
  - Generally set in study planning to \( \leq 0.2 \), where power = \( 1 - \beta \geq 80\% \).
  - There is no \textit{a posteriori} (aka \textit{post hoc}) power!
  - Either a study has demonstrated BE or not. Phoenix’/WinNonlin’s output is statistical nonsense!
Power Curves

Power to show BE with 12 – 36 subjects for $\text{CV}_{\text{intra}} = 20\%$

- $n$: 24 → 16:
  - Power: 0.896 → 0.735
- $\mu_T/\mu_R$: 1.05 → 1.10:
  - Power: 0.903 → 0.700
Power vs. Sample Size

- It is not possible to \textit{directly} calculate the required sample size.
- Power is calculated instead, and the lowest sample size which fulfills the minimum target power is used.

Example: $\alpha 0.05$, target power 80\% ($\beta 0.2$), $T/R$ 0.95, $CV_{\text{intra}}$ 20\% $\rightarrow$ minimum sample size 19 (power 81\%), rounded \textit{up} to the next even number in a 2$\times$2 study (power 83\%).

<table>
<thead>
<tr>
<th>$n$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>73.54%</td>
</tr>
<tr>
<td>17</td>
<td>76.51%</td>
</tr>
<tr>
<td>18</td>
<td>79.12%</td>
</tr>
<tr>
<td>19</td>
<td>81.43%</td>
</tr>
<tr>
<td>20</td>
<td>83.47%</td>
</tr>
</tbody>
</table>
Power vs. Sample Size

2×2 cross-over, T/R 0.95, 80%–125%, target power 80%

- sample size
- power
- power for n=12
Tools

- Sample Size Tables (Phillips, Diletti, Hauschke, Chow, Julious, ...)
- Approximations (Diletti, Chow, Julious, ...)
- General purpose (SAS, R, S+, StaTable, ...)
- Specialized Software (nQuery Advisor, PASS, FARTSSIE, StudySize, ...)
- Exact method (Owen – implemented in R-package *PowerTOST*)

* Thanks to Detlew Labes!
Background

- Reminder: Sample Size is not directly obtained; only power
- Solution given by DB Owen (1965) as a difference of two bivariate noncentral t-distributions
  - Definite integrals cannot be solved in closed form
  - ‘Exact’ methods rely on numerical methods (currently the most advanced is AS 243 of RV Lenth; implemented in R, FARTSSIE, EFG). nQuery uses an earlier version (AS 184).
Background

- Power calculations...
  - ‘Brute force’ methods (also called ‘resampling’ or ‘Monte Carlo’) converge asymptotically to the true power; need a good random number generator (e.g., Mersenne Twister) and may be time-consuming
  - ‘Asymptotic’ methods use large sample approximations
    - Approximations provide algorithms which should converge to the desired power based on the $t$-distribution
## Comparison

<table>
<thead>
<tr>
<th>original values</th>
<th>Method</th>
<th>Algorithm</th>
<th>CV%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.75</td>
</tr>
<tr>
<td>PowerTOST 0.8-2 (2011)</td>
<td>exact Owen's Q</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Patterson &amp; Jones (2006)</td>
<td>noncentr. t AS 243</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Diletti et al. (1991)</td>
<td>noncentr. t Owen's Q</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>nQuery Advisor 7 (2007)</td>
<td>noncentr. t AS 184</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>FARTSSIE 1.6 (2008)</td>
<td>noncentr. t AS 243</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>EFG 2.01 (2009)</td>
<td>noncentr. t AS 243</td>
<td>brute force ElMaestro</td>
<td>4</td>
</tr>
<tr>
<td>StudySize 2.0.1 (2006)</td>
<td>central t</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Hauschke et al. (1992)</td>
<td>approx. t</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Chow &amp; Wang (2001)</td>
<td>approx. t</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Kieser &amp; Hauschke (1999)</td>
<td>approx. t</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

### CV% Table

<table>
<thead>
<tr>
<th>original values</th>
<th>Method</th>
<th>Algorithm</th>
<th>CV%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>22.5</td>
</tr>
<tr>
<td>PowerTOST 0.8-2 (2011)</td>
<td>exact Owen's Q</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Patterson &amp; Jones (2006)</td>
<td>noncentr. t AS 243</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Diletti et al. (1991)</td>
<td>noncentr. t Owen's Q</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>nQuery Advisor 7 (2007)</td>
<td>noncentr. t AS 184</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>FARTSSIE 1.6 (2008)</td>
<td>noncentr. t AS 243</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>EFG 2.01 (2009)</td>
<td>noncentr. t AS 243</td>
<td>brute force ElMaestro</td>
<td>23</td>
</tr>
<tr>
<td>StudySize 2.0.1 (2006)</td>
<td>central t</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Hauschke et al. (1992)</td>
<td>approx. t</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Chow &amp; Wang (2001)</td>
<td>approx. t</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Kieser &amp; Hauschke (1999)</td>
<td>approx. t</td>
<td></td>
<td>NA</td>
</tr>
</tbody>
</table>
Sample Size Calculations ...or the Myth of Power

Approximations

Hauschke et al. (1992)

Patient’s risk $\alpha$ 0.05, Power 80% (Producer’s risk $\beta$ 0.2), AR $[0.80 - 1.25]$, CV 0.2 (20%), T/R 0.95
1. $\Delta = \ln(0.8) - \ln(1.25) = -0.1719$
2. Start with e.g. n=8/sequence
   1. $df = n \cdot 2 - 1 = 8 \times 2 - 1 = 14$
   2. $t_{\alpha,df} = 1.7613$
   3. $t_{\beta,df} = 0.8681$
   4. new $n = \left( \frac{(t_{\alpha,df} + t_{\beta,df})^2 \cdot (CV/\Delta)}{1.7613+0.8681} \right)^2 = 9.3580$
3. Continue with n=9.3580/sequence (N=18.716 → 19)
   1. $df = 16.716; \text{roundup to the next integer} 17$
   2. $t_{\alpha,df} = 1.7396$
   3. $t_{\beta,df} = 0.8633$
   4. new $n = \left( \frac{(t_{\alpha,df} + t_{\beta,df})^2 \cdot (CV/\Delta)}{1.7396+0.8633} \right)^2 = 9.1711$
4. Continue with n=9.1711/sequence (N=18.3422 → 19)
   1. $df = 17.342; \text{roundup to the next integer} 18$
   2. $t_{\alpha,df} = 1.7341$
   3. $t_{\beta,df} = 0.8620$
   4. new $n = \left( \frac{(t_{\alpha,df} + t_{\beta,df})^2 \cdot (CV/\Delta)}{1.7341+0.8620} \right)^2 = 9.1233$
5. Convergence reached (N=18.2466 → 19):
   Use 10 subjects/sequence (20 total)

S-C Chow and H Wang (2001)

Patient’s risk $\alpha$ 0.05, Power 80% (Producer’s risk $\beta$ 0.2), AR $[0.80 - 1.25]$, CV 0.2 (20%), T/R 0.95
1. $\Delta = \ln(T/R) - \ln(1.25) = 0.1719$
2. Start with e.g. n=8/sequence
   1. $df_{\alpha} = \text{roundup}(2 \cdot n - 2) - 2 = (2 \times 8 - 2) = 26$
   2. $df_{\beta} = \text{roundup}(4 \cdot n - 2) = 4 \times 8 - 2 = 30$
   3. $t_{\alpha,df} = 1.7056$
   4. $t_{\beta/2,df} = 0.8538$
   5. new $n = \beta^2 \cdot \left( \frac{(t_{\alpha,df} + t_{\beta/2,df})}{\Delta} \right)^2 = 0.2^2 \times (1.7056+0.8538)^2 / 0.1719^2 = 8.8723$
3. Continue with n=8.8723/sequence (N=17.7446 → 18)
   1. $df_{\alpha} = \text{roundup}(2 \cdot n - 2) - 2 = (2 \times 8.8723 - 2) = 30$
   2. $df_{\beta} = \text{roundup}(4 \cdot n - 2) = 4 \times 8.8723 - 2 = 34$
   3. $t_{\alpha,df} = 1.6973$
   4. $t_{\beta/2,df} = 0.8523$
   5. new $n = \beta^2 \cdot \left( \frac{(t_{\alpha,df} + t_{\beta/2,df})}{\Delta} \right)^2 = 0.2^2 \times (1.6973+0.8523)^2 / 0.1719^2 = 8.8045$
4. Convergence reached (N=17.6090 → 18):
   Use 9 subjects/sequence (18 total)

<table>
<thead>
<tr>
<th>sample size</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>power %</td>
<td>79.124</td>
<td>81.428</td>
<td>83.468</td>
</tr>
</tbody>
</table>

Pharma Edge
Approximations obsolete

- Exact sample size tables still useful in checking the plausibility of software’s results
- Approximations based on noncentral $t$ (FARTSSIE17)

Approximations obsolete

http://individual.utoronto.ca/ddubins/FARTSSIE17.xls
or R / S+ →

- Exact method (Owen) in R-package PowerTOST

```
alpha <- 0.05  # alpha
CV    <- 0.30  # intra-subject CV
theta1<- 0.80  # lower acceptance limit
theta2<- 1/theta1 # upper acceptance limit
ratio <- 0.95  # expected ratio T/R
PwrNeed<- 0.80  # minimum power
Limit  <- 10000 # Upper Limit for Search
n     <- 4     # start value of sample size search
s     <- sqrt(2)*sqrt(log(CV^2+1))
repeat{
t     <- qt(1-alpha,n-2)
nc1   <- sqrt(n)*(log(ratio)-log(theta1))/s
nc2   <- sqrt(n)*(log(ratio)-log(theta2))/s
prob1 <- pt(+t,n-2,nc1); prob2 <- pt(-t,n-2,nc2)
power <- prob2-prob1
n     <- n+2  # increment sample size
if(power >= PwrNeed | (n-2) >= Limit) break
}
if(total == Limit){
cat("Search stopped at Limit",Limit,
    " obtained Power",power*100,"\%\n")
} else
cat("Sample Size",Total,"(Power",power*100,"\%\n")
```
Sensitivity Analysis

- ICH E9 (1998)
  - Section 3.5 Sample Size, paragraph 3
    - The method by which the sample size is calculated should be given in the protocol […]. The basis of these estimates should also be given.
    - It is important to investigate the sensitivity of the sample size estimate to a variety of deviations from these assumptions and this may be facilitated by providing a range of sample sizes appropriate for a reasonable range of deviations from assumptions.
    - In confirmatory trials, assumptions should normally be based on published data or on the results of earlier trials.
Sensitivity Analysis

- Example
  nQuery Advisor: \[ \sigma_w = \sqrt{\ln(CV_{\text{intra}}^2 + 1)}; \sqrt{\ln(0.2^2 + 1)} = 0.198042 \]

### Table: Sensitivity Analysis

<table>
<thead>
<tr>
<th></th>
<th>90% power</th>
<th>25% CV</th>
<th>4 drop outs</th>
<th>25% CV + d.o.</th>
<th>PE 90%</th>
<th>worst case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test significance levels, ( \alpha ) (one-sided)</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
</tr>
<tr>
<td>Lower equivalence limit for ( \mu_1 / \mu_2 ), ( A_1 )</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
<td>0.800</td>
</tr>
<tr>
<td>Upper equivalence limit for ( \mu_1 / \mu_2 ), ( A_u )</td>
<td>1.250</td>
<td>1.250</td>
<td>1.250</td>
<td>1.250</td>
<td>1.250</td>
<td>1.250</td>
</tr>
<tr>
<td>Expected ratio, ( \mu_1 / \mu_2 )</td>
<td>0.950</td>
<td>0.950</td>
<td>0.950</td>
<td>0.950</td>
<td>0.900</td>
<td>0.900</td>
</tr>
<tr>
<td>Crossover ANOVA, sqrt(MSE) (ln scale)</td>
<td>0.198042</td>
<td>0.246221</td>
<td>0.198042</td>
<td>0.246221</td>
<td>0.198042</td>
<td>0.246221</td>
</tr>
<tr>
<td>SD differences, ( \sigma_4 ) (ln scale)</td>
<td>0.280074</td>
<td>0.348209</td>
<td>0.260074</td>
<td>0.348209</td>
<td>0.280074</td>
<td>0.348209</td>
</tr>
<tr>
<td>Power (%)</td>
<td>90.00</td>
<td>77.60</td>
<td>86.88</td>
<td>69.53</td>
<td>66.94</td>
<td>45.09</td>
</tr>
<tr>
<td>n per sequence group</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td>11</td>
</tr>
</tbody>
</table>

- 20% CV: \( n=26 \)
- 25% CV: power 90% → 78%
- 20% CV, 4 drop outs: power 90% → 67%
- 25% CV, 4 drop outs: power 90% → 70%
Sensitivity Analysis

**Example**

*PowerTOST*, function *sampleN.TOST*

```r
require(PowerTost)
sampleN.TOST(alpha = 0.05, targetpower = 0.9, logscale = TRUE,
             theta1 = 0.8, theta2 = 1.25, theta0 = 0.95, CV = 0.2,
             design = "2x2", exact = TRUE, print = TRUE)
```

```
+++++++++++ Equivalence test - TOST +++++++++++
Sample size estimation

Study design: 2x2 crossover
log-transformed data (multiplicative model)
alpha = 0.05, target power = 0.9
BE margins = 0.8 ... 1.25
Null (true) ratio = 0.95, CV = 0.2
Sample size
n power
26 0.917633
```
Sensitivity Analysis

To calculate Power for a given sample size, use function `power.TOST`

```r
require(PowerTost)
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
    theta0=0.95, CV=0.25, n=26, design="2x2", exact=TRUE)
[1] 0.7760553
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
    theta0=0.95, CV=0.20, n=22, design="2x2", exact=TRUE)
[1] 0.8688866
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
    theta0=0.95, CV=0.25, n=22, design="2x2", exact=TRUE)
[1] 0.6953401
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
    theta0=0.90, CV=0.20, n=26, design="2x2", exact=TRUE)
[1] 0.6694514
power.TOST(alpha=0.05, logscale=TRUE, theta1=0.8, theta2=1.25,
    theta0=0.90, CV=0.25, n=22, design="2x2", exact=TRUE)
[1] 0.4509864
```
Sensitivity Analysis

- Must be done before the study (*a priori*)
- The Myth of retrospective (*a posteriori*)

  - High values do not further support the claim of already demonstrated bioequivalence.
  - Low values do not invalidate a bioequivalent formulation.

Further reader:

RV Lenth (2000)
JM Hoenig and DM Heisey (2001)
P Bacchetti (2010)
Data from Pilot Studies

- Estimated CVs have a high degree of uncertainty (in the pivotal study it is more likely that you will be able to reproduce the PE, than the CV)
  - The smaller the size of the pilot, the more uncertain the outcome.
  - The more formulations you have tested, lesser degrees of freedom will result in worse estimates.
  - Remember: CV is an *estimate* – *not carved in stone!*
**Pilot Studies: Sample Size**

- Small pilot studies (sample size < 12)
  - Are useful in checking the sampling schedule and the appropriateness of the analytical method, but are not suitable for the purpose of sample size planning!
  - Sample sizes (T/R 0.95, power ≥ 80%) based on a n=10 pilot study

```
require(PowerTOST)
expsampleN.TOST(alpha=0.05, targetpower=0.80, theta1=0.80, theta2=1.25, theta0=0.95, CV=0.40, dfCV=24-2, alpha2=0.05, design="2x2")
```

<table>
<thead>
<tr>
<th>CV%</th>
<th>CV</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fixed</td>
<td>uncertain</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>25</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>30</td>
<td>40</td>
<td>52</td>
</tr>
<tr>
<td>35</td>
<td>52</td>
<td>68</td>
</tr>
<tr>
<td>40</td>
<td>66</td>
<td>86</td>
</tr>
</tbody>
</table>

If pilot n=24: n=72, ratio 1.091
Moderate sized pilot studies (sample size ~12–24) lead to more consistent results (both CV and PE).

- If you stated a procedure in your protocol, even BE may be claimed in the pilot study, and no further study will be necessary (US-FDA).
- If you have some previous hints of high intra-subject variability (>30%), a pilot study size of at least 24 subjects is reasonable.
- A Sequential Design may also avoid an unnecessarily large pivotal study.
Pilot Studies: Sample Size

- *Do not* use the pilot study’s CV, but calculate an upper confidence interval!
  - Gould (1995) recommends a 75% CI (*i.e.* a producer’s risk of 25%).
  - Apply Bayesian Methods (Julious and Owen 2006, Julious 2010) implemented in *R’s PowerTOST/expsampleN.TOST*.
  - Unless you are under time pressure, a Two-Stage Sequential Design will help in dealing with the uncertain estimate from the pilot study.
Sample Size Calculations

...or the Myth of Power

Helmut Schütz
BEBAC
Consultancy Services for Bioequivalence and Bioavailability Studies
1070 Vienna, Austria
helmut.schuetz@bebac.at
To bear in Remembrance...

Power. That which statisticians are always calculating but never have.

Power: That which is wielded by the priesthood of clinical trials, the statisticians, and a stick which they use to beta their colleagues.

Power Calculation – A guess masquerading as mathematics.  

Stephen Senn

You should treat as many patients as possible with the new drugs while they still have the power to heal.  

Armand Trousseau
The Myth of Power

There is simple intuition behind results like these: If my car made it to the top of the hill, then it is powerful enough to climb that hill; if it didn’t, then it obviously isn’t powerful enough. Retrospective power is an obvious answer to a rather uninteresting question. A more meaningful question is to ask whether the car is powerful enough to climb a particular hill never climbed before; or whether a different car can climb that new hill. Such questions are prospective, not retrospective.

The fact that retrospective power adds no new information is harmless in its own right. However, in typical practice, it is used to exaggerate the validity of a significant result (“not only is it significant, but the test is really powerful!”), or to make excuses for a nonsignificant one (“well, P is .38, but that’s only because the test isn’t very powerful”). The latter case is like blaming the messenger.

RV Lenth

*Two Sample-Size Practices that I don’t recommend*

http://www.math.uiowa.edu/~rlenth/Power/2badHabits.pdf
References

- Collection of links to global documents
  - http://bebac.at/Guidelines.htm
- ICH
- EMA-CPMP/CHMP/EWP
  - Points to Consider on Multiplicity Issues in Clinical Trials (2002)
  - BA/BE for HVDs/HVDPs: Concept Paper (2006)
- Questions & Answers: Positions on specific questions addressed to the EWP therapeutic subgroup on Pharmacokinetics (2010)
- US-FDA
  - Center for Drug Evaluation and Research (CDER)
    - Statistical Approaches Establishing Bioequivalence (2001)
    - Bioequivalence Recommendations for Specific Products (2007)
- Midha KK, Ormsby ED, Hubbard JW, McKay G, Hawes EM, Gavalas L, and IJ McGilveray
  - Logarithmic Transformation in Bioequivalence: Application with Two Formulations of Perphenazine
- Hauschke D, Steinijans VW, and E Diletti
  - Presentation of the intrasubject coefficient of variation for sample size planning in bioequivalence studies
- Diletti E, Hauschke D, and VW Steinijans
  - Sample size determination for bioequivalence assessment by means of confidence intervals
  - Int J Clin Pharm Ther Toxicol 29/7, 376-378 (1994)
- Hauschke D, Steinijans VW, Diletti E, and M Burke
  - Sample Size Determination for Bioequivalence Assessment Using a Multiplicative Model
  - J Pharmacokin Biopharm 20/5, 557-561 (1992)
- S-C Chow and H Wang
  - On Sample Size Calculation in Bioequivalence Trials
  - J Pharmacokin Pharmacodyn 28/2, 155-169 (2001)
- DB Owen
  - A special case of a bivariate non-central t-distribution
  - Biometrika 52, 3/4, 437-446 (1965)
References

- LA Gould
  *Group Sequential Extension of a Standard Bioequivalence Testing Procedure*
  DOI: 10.1007/BF02353786

- Tóthfalusi L, Endrenyi L, and A Garcia Arieta
  *Evaluation of Bioequivalence for Highly Variable Drugs with Scaled Average Bioequivalence*

- RV Lenth
  *Two Sample-Size Practices that I don’t recommend*
  Joint Statistical Meetings, Indianapolis (2000)
  [http://www.math.uiowa.edu/~rlenth/Power/2badHabits.pdf](http://www.math.uiowa.edu/~rlenth/Power/2badHabits.pdf)

- Hoenig JM and DM Heisey
  *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis*
  The American Statistician 55/1, 19–24 (2001)

- P Bacchetti
  *Current sample size conventions: Flaws, harms, and alternatives*
  BMC Medicine 8:17 (2010)
  [http://www.biomedcentral.com/content/pdf/1741-7015-8-17.pdf](http://www.biomedcentral.com/content/pdf/1741-7015-8-17.pdf)

- Jones B and MG Kenward
  *Design and Analysis of Cross-Over Trials*

- Patterson S and B Jones
  *Determining Sample Size*, in: *Bioequivalence and Statistics in Clinical Pharmacology*

- SA Julious
  *Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data*

- Julious SA and RJ Owen
  *Sample size calculations for clinical studies allowing for uncertainty about the variance*

- SA Julious
  *Sample Sizes for Clinical Trials*

- D Labes
  *Package ‘PowerTOST’*
  Version 0.8-2 (2011-01-10)
  [http://cran.r-project.org/web/packages/PowerTOST/PowerTOST.pdf](http://cran.r-project.org/web/packages/PowerTOST/PowerTOST.pdf)