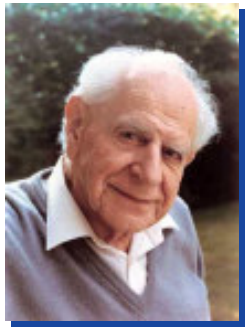
# Sample Size Estimation

Helmut Schütz

# To bear in Remembrance...

**Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve.**



Karl R. Popper

**Even though it's *applied* science we're dealin' with, it still is – *science*!**

Leslie Z. Benet

# Assumptions

## All models rely on assumptions.

- Bioequivalence as a surrogate for therapeutic equivalance.
  - Studies in healthy volunteers in order to minimize variability (*i.e.*, lower sample sizes than in patients).
  - Current emphasis on *in vivo* release ('human dissolution apparatus').
- Concentrations in the sample matrix reflect concentrations at the target receptor site.
  - In the strict sense only valid in steady state.
  - *In vivo* similarity in healthy volunteers can be extrapolated to the patient population(s).
- $f = \mu_T / \mu_R$ assumes that
  - $D_T = D_R$ *and*
  - inter-occasion clearances are constant.

# Assumptions

## All models rely on assumptions.

- Log-transformation allows for additive effects required in ANOVA.
- No carry-over effect in the model of crossover studies.
  - Cannot be statistically adjusted.
  - Has to be avoided *by design* (suitable washout).
  - Shown to be a statistical artifact in meta-studies.
  - Exception: Endogenous compounds (biosimilars!)
- Between- and within-subject errors are independently and normally distributed about unity with variances $\sigma^2_s$ and $\sigma^2_e$.
  - If the reference formulation shows higher variability than the test, the 'good' test will be penalized for the 'bad' reference.
- All observations made on different subjects are independent.
  - No monocygotic twins or triplets in the study!

# Error(s)

**All *formal* decisions are subjected to two 'Types' of Error.**

- **$\alpha$: Probability of Type I Error (aka Risk Type I)**
- **$\beta$: Probability of Type II Error (aka Risk Type II)**

**Example from the justice system – which presumes that the defendant is *not guilty*:**

| Verdict | Defendant *innocent* | Defendant *guilty* |
|---|---|---|
| Presumption of innocence *rejected* (*guilty*) | wrong | correct |
| Presumption of innocence *accepted* (*not guilty*) | correct | wrong |

# Hypotheses

## In statistical terminology

- Null hypothesis ($H_0$): innocent
- Alternative hypothesis ($H_a$ aka $H_1$): guilty

| Decision | Null hypothesis *true* | Null hypothesis *false* |
|---|---|---|
| $H_0$ rejected | Type I Error | Correct (accept $H_a$) |
| Failed to reject $H_0$ | Correct (accept $H_0$) | Type II Error |

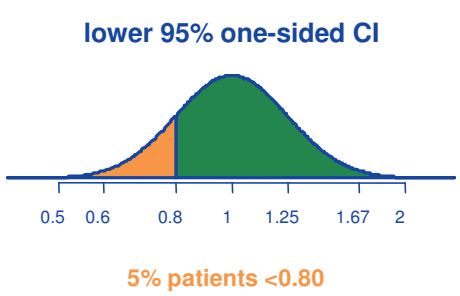
## In BE the Null hypothesis is bioinequivalence ($\mu_T \neq \mu_R$)!

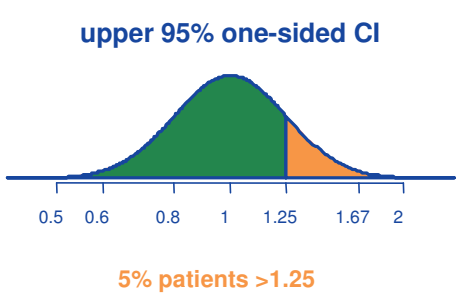
| Decision | Null hypothesis *true* | Null hypothesis *false* |
|---|---|---|
| $H_0$ rejected | Patient's risk ($\alpha$) | Correct (BE) |
| Failed to reject $H_0$ | Correct (not BE) | Producer's risk ($\beta$) |

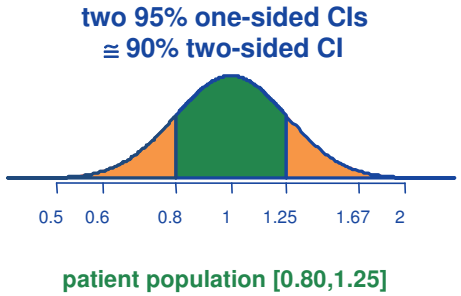
# Type I Error

## $\alpha$: Patient's risk to be treated with an inequivalent formulation ($H_0$ *falsely* rejected)

- BA of the test compared to reference in a *particular* patient is considered to be risky *either* below 0.80 *or* above 1.25.
  - If we keep the risk of *particular* patients at $\alpha$ 0.05 (5%), the risk of the entire *population* of patients (where BA <0.80 *and* >1.25) is $2\alpha$ (10%) – expressed as a confidence interval: $100(1 - 2\alpha) = 90\%$.
  - However, since in a patient BA cannot be <0.80 *and* >1.25 *at the same time*, the patient's risk from a 90% CI is still 5%!



# Type II Error

β: **Producer's risk to get no approval of an equivalent formulation (*$H_0$ falsely* not rejected)**

- **Fixed in study planning to 0.1 – ≤0.2 (10 – ≤20%), where power = 1 – β = ≥80 – 90%.**

  **If all assumptions in sample size estimations turn out to be correct and power was set to 80%,**

  **one out of five studies will fail just by chance!**


- ***A posteriori* (*post hoc*) power is irrelevant!**

  **Either a study has demonstrated bioequivalence or not.**

# Review of Guidelines

## Minimum Sample Size.

- 12 WHO, EU, CAN, NZ, AUS, AR, MZ, ASEAN States, RSA, Russia ('Red Book'), EAEU, Ukraine
- 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.'*
- 18 Russia (2008)
- 20 RSA (MR formulations)
- 24 Saudia Arabia (12 to 24 if statistically justifiable)
- 24 Brazil; USA (replicate designs intended for RSABE)
- 24 EU (RTR|TRT replicate designs intended for ABEL)
- 'Sufficient number' Japan
- 'Adequate' India

# Review of Guidelines

## Maximum Sample Size.

- Generally *not* specified (decided by IEC/IRB and/or local Authorities).
- ICH E9, Section 3.5 states:

  *'The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed.'*

# Power *vs.* Sample Size

**It is not possible to *directly* obtain the required sample size.**

- **The required sample size depends on**
  - **the acceptance range (AR) for bioequivalence;**
  - **the error variance ($s^2$) associated with the PK metrics as estimated from**
    - **published data,**
    - **a pilot study, or**
    - **previous studies;**
  - **the fixed significance level ($\alpha$);**
  - **the expected deviation ($\Delta$) from the reference product and;**
  - **the desired power ($1 - \beta$).**
- **Three values are *known and fixed* (AR, $\alpha$, $1 - \beta$), one is an *assumption* ($\Delta$), and one an *estimate* ($s^2$). Hence, the correct term is 'sample size *estimation*'.**

# Power *vs.* Sample Size

## Only power is accessible.

- The sample size is searched in an iterative procedure until at least the desired power is obtained.
  Example: $\alpha$ 0.05, target power 80% ($\beta$ 0.2), expected *GMR* 0.95, $CV_{intra}$ 20% $\rightarrow$ minimum sample size 19 (power 81.3%), rounded *up* to the next even number in a 2×2×2 study (power 83.5%).

| *n* | power (%) |
|---|---|
| 16 | 73.5 |
| 17 | 76.4 |
| 18 | 79.1 |
| 19 | 81.3 |
| 20 | 83.5 |

  - Exact methods for ABE in parallel, crossover, and replicate designs available.
  - Simulations suggested for Group-Sequential and Two-Stage Designs.
  - Simulations mandatory for reference-scaling methods.

# Power *vs.* Sample Size

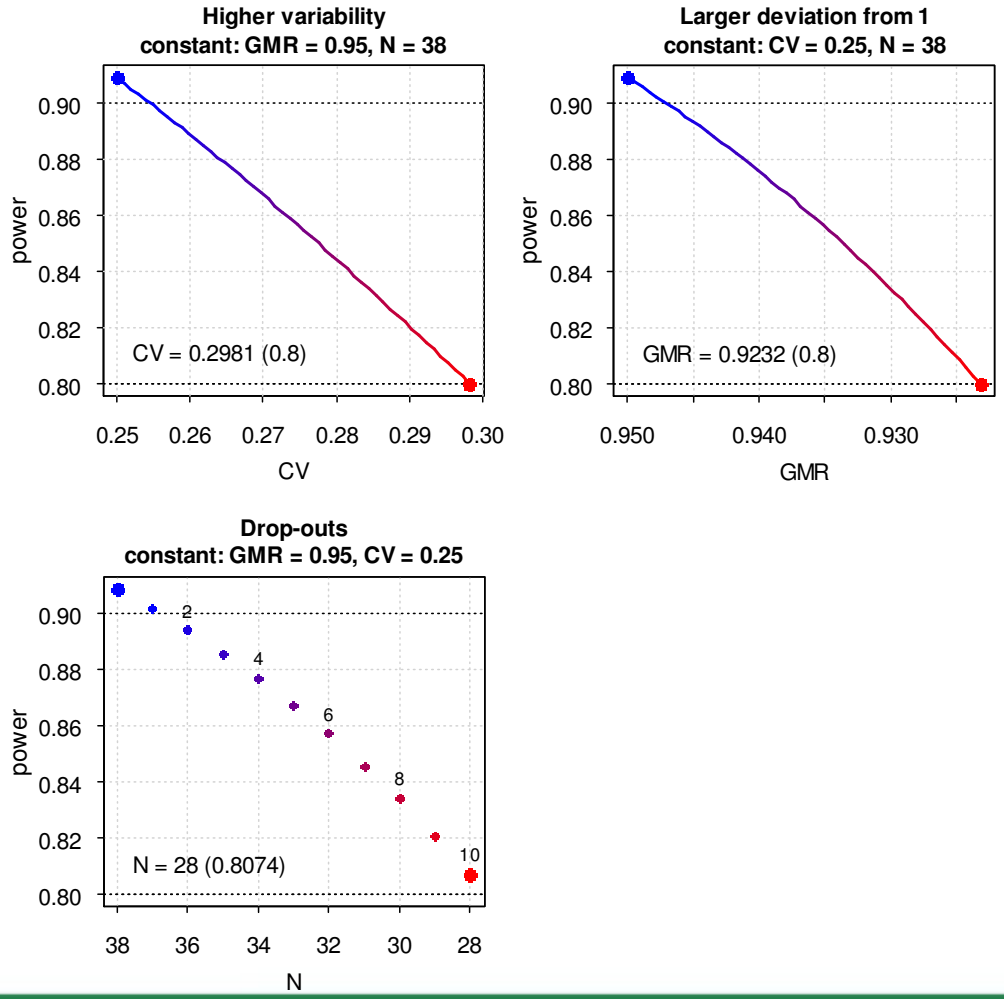
## How many subjects are 'enough'?

- Most guidelines recommend 80 – 90% power.
  - If a study is planned for ≤70% power, problems with the ethics committee are possible (ICH E9).
  - If a study is planned for >90% power (especially with low variability drugs), additional problems with regulators are possible ('forced bioequivalence').
  - Some subjects ('alternates') may be added to the estimated sample size according to the expected drop-out rate – especially for studies with more than two periods or multiple-dose studies.
- According to ICH E9 a sensitivity analysis is mandatory to explore the impact on power if values deviate from assumptions.

# Power Analysis

## Example 2×2×2, ABE

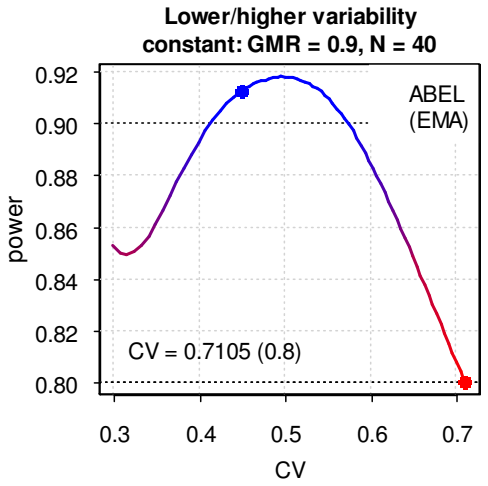
- Assumed *GMR* 0.95, $CV_w$ 0.25, desired power 0.9, min. acceptable power 0.8.
  - Sample size 38 (power 0.909)
  - $CV_w$ can increase to 0.298 (rel. +19%)
  - *GMR* can decrease to 0.923 (rel. –2.8%)
  - 10 drop-outs acceptable (rel. –26%)
  - Most critical is the *GMR*!

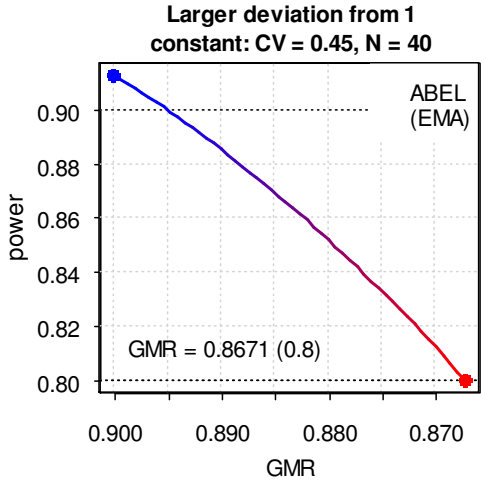


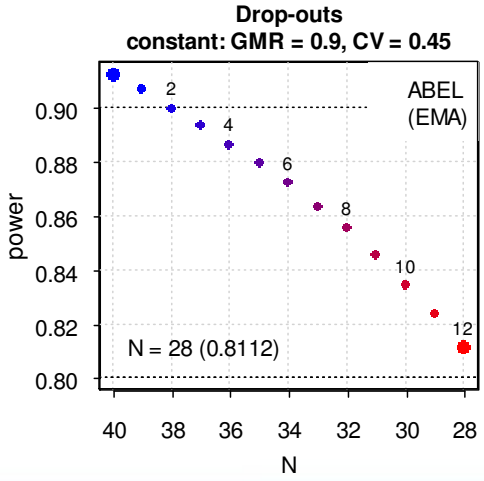
# Power Analysis

## Example 2×2×4, ABEL

- Assumed *GMR* 0.90, $CV_{wR}$ 0.45, desired power 0.9, min. acceptable power 0.8.
  - Sample size 40 (power 0.912)
  - $CV_w$ can increase to 0.711 (rel. +58%)
  - *GMR* can decrease to 0.867 (rel. –3.7%)
  - 12 drop-outs acceptable (rel. –30%)
  - Most critical is the *GMR*!



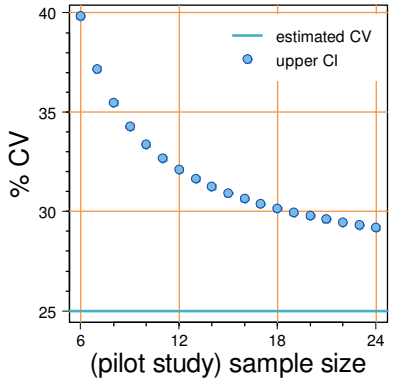



# Dealing with Uncertainty

## Nothing is 'carved in stone'.

- Never assume perfectly matching products.
  - Generally a $\Delta$ of not better than 5% should be assumed (0.9500 – 1.0526).
  - For HVD(P)s do not assume a $\Delta$ of <10% (0.9000 – 1.1111).
- Do not use the *CV* but one of its confidence limits.
  - Suggested $\alpha$ 0.2 (here: the producer's risk).
  - For ABE the upper CL.
  - For reference-scaling the lower CL.
- Better alternatives.
  - Group-Sequential Designs
    Fixed total sample size, interim analysis for early stopping.
  - (Adaptive) Sequential Two-Stage Designs
    Fixed stage 1 sample size, re-estimation of the total sample size in the interim analysis.



# Excursion

## Type I Error.

- In BE the Null Hypothesis ($H_0$) is *inequivalence*.
  - TIE = Probability of falsely rejecting $H_0$ (*i.e.*, accepting $H_a$ and claiming BE).
  - Can be calculated for the nominal significance level ($\alpha$) assuming a *GMR* ($\theta_0$) at one of the limits of the acceptance range [$\theta_1$ , $\theta_2$].
    - Example: 2×2×2 crossover, *CV* 20%, *n* 20, $\alpha$ 0.05, $\theta_0$ = [$\theta_1$ 0.80 or $\theta_2$ 1.25].

```
library(PowerTOST)
AR <- c(1-0.20, 1/(1-0.20)) # common acceptance range: 0.80-1.25
power.TOST(CV=0.20, n=20, alpha=0.05, theta0=AR[1])
[1] 0.0499999
power.TOST(CV=0.20, n=20, alpha=0.05, theta0=AR[2])
[1] 0.0499999
```

    - TOST is not a uniformly most powerful (UMP) test.

```
power.TOST(CV=0.20, n=12, alpha=0.05, theta0=AR[2])
[1] 0.04976374
```

    - However, the TIE never exceeds the nominal level.

```
power.TOST(CV=0.20, n=72, alpha=0.05, theta0=AR[2])
[1] 0.05
```

Labes D, Schütz H, Lang B. *PowerTOST: Power and Sample size based on Two One-Sided t-Tests (TOST) for (Bio)Equivalence Studies.* R package version 1.4-2. 2016. https://cran.r-project.org/package=PowerTOST

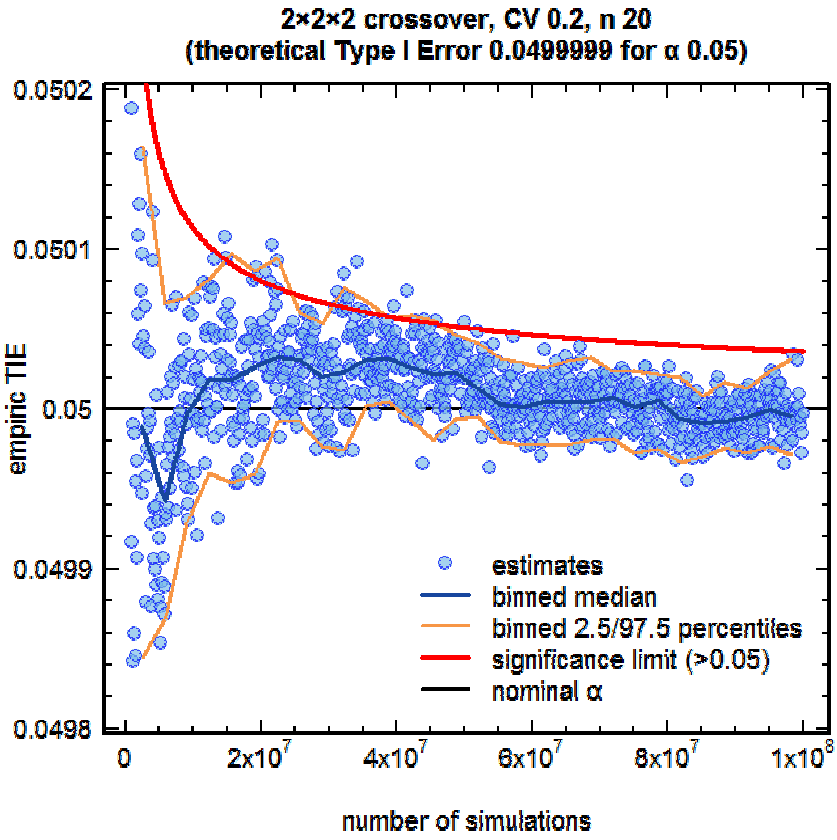
# Excursion

## Type I Error.

- Alternatively perform simulations to obtain an *empiric* Type I Error.

```
power.TOST.sim(CV=0.20, n=20, alpha=0.05, theta0=AR[2],
                    nsims=1e8)
[1] 0.04999703
```

- In other settings (*i.e.*, frameworks like Two-Stage Designs or reference-scaled ABE) analytical solutions for power – and therefore, the TIE – are not possible:
  Simulations are required.

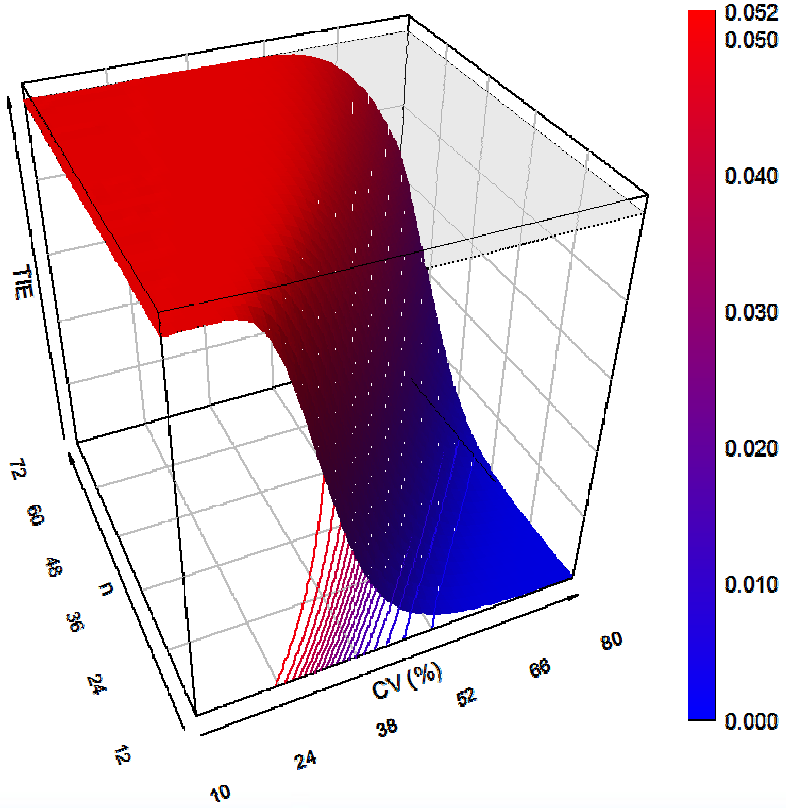


# Excursion

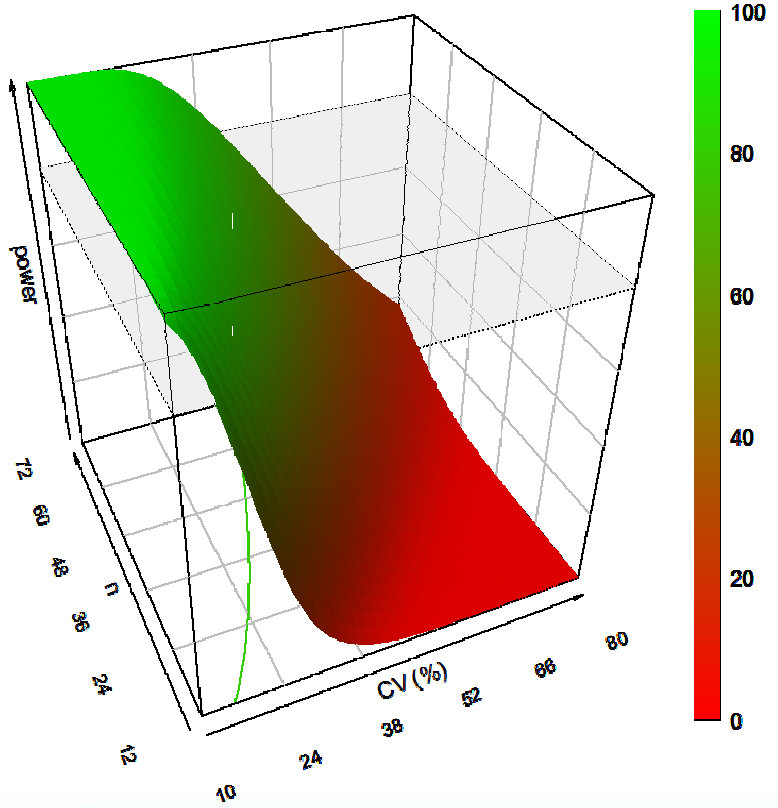
## Type I Error and power.

- Fixed sample 2×2×2 design (α 0.05). *GMR* 0.95, *CV* 10 – 80%, *n* 12 –72

TIE




power



# R Package PowerTOST

## Examples

- **Install the package from CRAN if necessary and attach it.**

```
if (!("PowerTOST" %in% installed.packages()[, "Package"])) {
  install.packages("PowerTOST")
}
library(PowerTOST)
```

- **ABE**
  - **2×2×2 crossover, $CV_{intra}$ 25%, $\theta_0$ 0.95, targetpower 90%.**

```
sampleN.TOST(CV=0.25, theta0=0.95, targetpower=0.9,
             print=FALSE)[["Sample size"]]
[1] 38
```

  - **2×2×2 crossover, $CV_{intra}$ 10%, NTID (AR 90.00–111.11%), $\theta_0$ 0.95.**

```
sampleN.TOST(CV=0.10, theta0=0.95, theta1=0.9,
             print=FALSE)[["Sample size"]]
[1] 44
```

  - **Parallel design, $CV_{total}$ 40%, $\theta_0$ 0.95.**

```
sampleN.TOST(CV=0.20, theta0=0.95, design="parallel",
             print=FALSE)[["Sample size"]]
[1] 130
```

# R Package PowerTOST

- **ABEL (reference-scaling according to the EMA)**
  - **4-period full replicate, $CV_{wR}$ 35%, $\theta_0$ 0.90.**

```
sampleN.scABEL(CV=0.35, theta0=0.90, design="2x2x4", details=TRUE)
++++++++++++ scaled (widened) ABEL ++++++++++++
             Sample size estimation
   (simulation based on ANOVA evaluation)
--------------------------------------------
Study design:  2x2x4 (full replicate)

alpha  = 0.05, target power = 0.8
CVw(T) = 0.35; CVw(R) = 0.35
True ratio = 0.9
ABE limits / PE constraint = 0.8 ... 1.25
EMA regulatory settings
- CVswitch             = 0.3
- cap on scABEL if CVw(R) > 0.5
- regulatory constant = 0.76
- pe constraint applied

Sample size search
 n    power
30   0.7702
32   0.7929
34   0.8118
```

# R Package PowerTOST

- **ABEL (reference-scaling according to the EMA, iteratively adjusted α to preserve the consumer risk at ≤0.05: Labes and Schütz 2016)**
  - **4-period full replicate, $CV_{wR}$ 35%, $\theta_0$ 0.90.**

```
sampleN.scABEL.ad(CV=0.35, theta0=0.90, design="2x2x4", details=TRUE)
++++++++++++ scaled (widened) ABEL ++++++++++++
             Sample size estimation
       for iteratively adjusted alpha'
-------------------------------------------
Study design: 2x2x4 (RTRT|TRTR)

Expected CVwR 0.35
Nominal alpha      : 0.05
True ratio         : 0.9000
Target power       : 0.8
Regulatory settings: EMA (ABEL)
Switching CVwR     : 0.3
Regulatory constant: 0.76
Expanded limits    : 0.7723...1.2948
Upper scaling cap  : CVwR > 0.5
PE constraints     : 0.8000 ... 1.2500
n  34, nomin. alpha: 0.05000 (power 0.8118), TIE: 0.0656
n  34,   adj. alpha: 0.03630 (power 0.7728)
n  38,   adj. alpha: 0.03610 (power 0.8100), TIE: 0.05000
```

# Sample Size Estimation

## Thank You!

## *Open Questions?*


**Helmut Schütz**

**BEBAC**

Consultancy Services for
Bioequivalence and Bioavailability Studies
1070 Vienna, Austria

helmut.schuetz@bebac.at