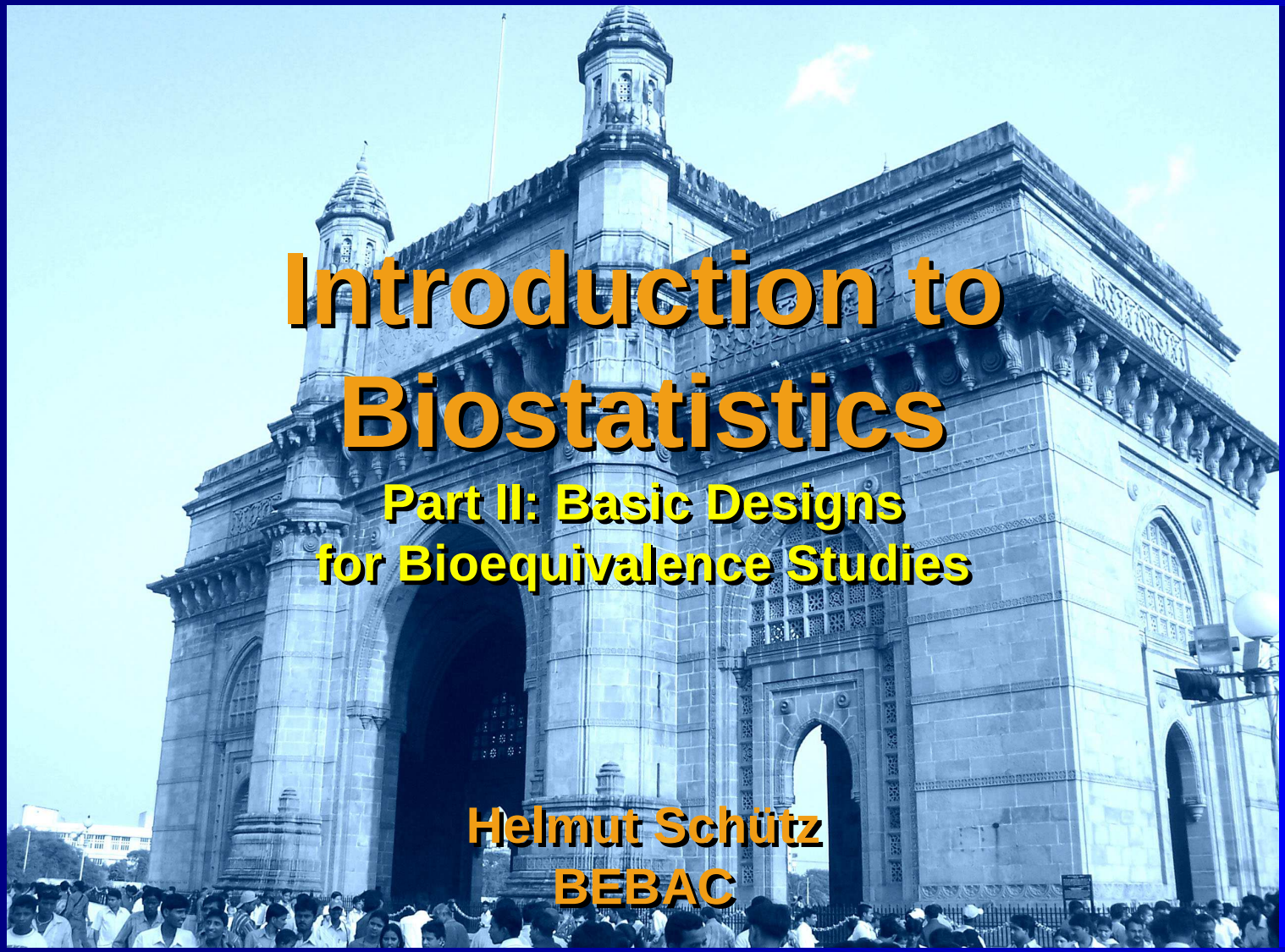


# Introduction to Biostatistics

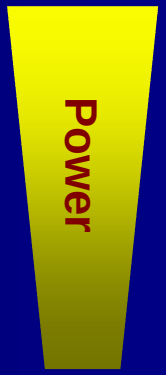
## Part II: Basic Designs for Bioequivalence Studies

Helmut Schütz
BEBAC

Wikimedia Commons • 2007 Sujit Kumar • Creative Commons Attribution-ShareAlike 3.0 Unported

# Designs

- The more ‘sophisticated’ a design is, the more information (in terms of variances) we may obtain.
  - Hierarchy of designs:




Full replicate (TRTR | RTRT)

Partial replicate (TRR | RTR | RRT)

Standard 2×2 cross-over (RT | TR)

Parallel (R | T)

# Designs

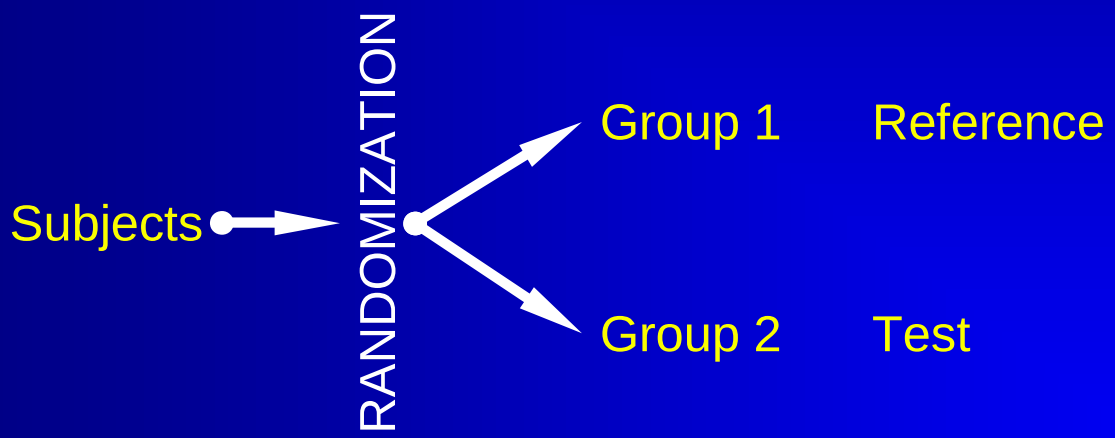

- Parallel Groups (patients, long half-life drugs)
- Cross-over (generally healthy subjects)
  - Standard 2×2×2
  - Higher Order Designs (more than two formulations)
    - Latin Squares
    - Variance Balanced Designs (Williams' Designs)
    - Incomplete Block Designs
  - Replicate designs

# Parallel design
## (independent groups)

- Two-group parallel design
  - Advantages
    - Clinical part – *sometimes* – faster than X-over.
    - Straigthforward statistical analysis.
    - Drugs with long half life.
    - Potentially toxic drugs or effect and/or AEs unacceptable in healthy subjects.
    - Studies in patients, where the condition of the disease irreversibly changes.
  - Disadvantages
    - Lower statistical power than X-over (*rule of thumb:* sample size should at least be doubled).
    - Phenotyping mandatory for drugs showing polymorphism.

# Parallel design

- Two-Group Parallel Design



# Parallel design

- One group is treated with the test formulation and another group with reference.
- Quite common that the dataset is imbalanced, *i.e.*, $n_1 \neq n_2$.
- Guidelines against the assumption of equal variance. Not implemented in PK soft-ware (Phoenix/WNL, Kinetica)!

| Subj. | Group 1 (T) | Group 2 (R) |
|---|---|---|
| 1-13 | 100 | 110 |
| 2-14 | 103 | 113 |
| 3-15 | 80 | 96 |
| 4-16 | 110 | 90 |
| 5-17 | 78 | 111 |
| 6-18 | 87 | 68 |
| 7-19 | 116 | 111 |
| 8-20 | 99 | 93 |
| 9-21 | 122 | 93 |
| 10-22 | 82 | 82 |
| 11-23 | 68 | 96 |
| 12-24 | NA | 137 |
| n | 11 | 12 |
| mean | 95 | 100 |
| $s^2$ | 298 | 314 |
| s | 17.3 | 17.7 |

# Parallel design

- Pooled variance

$$s_0^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{10 \cdot 298 + 11 \cdot 314}{10 + 11 - 2} = 306.4$$

- Pooled standard deviation

$$s_0 = \sqrt{s_0^2} = \sqrt{306.4} = 17.50$$

- 90% Confidence interval

$$CI = \left|\overline{x}_1 - \overline{x}_2\right| \pm t_{2\alpha, n_1+n_2-2} s_0 \sqrt{\frac{n_1 + n_2}{n_1 n_2}} =$$

$$= 5 \pm 1.721 \cdot 17.50 \cdot 0.4174 = [-7.6, +17.6]$$

# Parallel design

- But we want a ratio, not a difference!
  Now we have only –7.6 ≤ [T-R = -5] ≤ +17.6…
- Maybe we can use (R–7.6)/R and (R+17.6)/R to get a CI of 92.4% – 117.6%?
- No. Let’s repeat the analysis with logtransformed data.

# Parallel design

| Subj. | Group 1 (T) | ln (T) | Group 2 (R) | ln (R) |
|---|---|---|---|---|
| 1-13 | 100 | 4.605 | 110 | 4.700 |
| 2-14 | 103 | 4.635 | 113 | 4.727 |
| 3-15 | 80 | 4.382 | 96 | 4.564 |
| 4-16 | 110 | 4.700 | 90 | 4.500 |
| 5-17 | 78 | 4.357 | 111 | 4.710 |
| 6-18 | 87 | 4.466 | 68 | 4.220 |
| 7-19 | 116 | 4.754 | 111 | 4.710 |
| 8-20 | 99 | 4.595 | 93 | 4.533 |
| 9-21 | 122 | 4.804 | 93 | 4.533 |
| 10-22 | 82 | 4.407 | 82 | 4.407 |
| 11-23 | 68 | 4.220 | 96 | 4.564 |
| 12-24 | NA | NA | 137 | 4.920 |
| n | 11 | 11 | 12 | 12 |
| mean | 95 | 4.539 | 100 | 4.591 |
| $s^2$ | 298 | 0.03418 | 314 | 0.03231 |
| s | 17.3 | 0.1849 | 17.7 | 0.1798 |

$$s_0^2 = \frac{10 \cdot 0.03418 + 11 \cdot 0.03231}{10 + 11 - 2} =$$

$$= 0.03320$$

$$s_0 = \sqrt{s_0^2} = \sqrt{0.03320} = 0.1812$$

$$CI_{\ln} = 0.05203 \pm 1.721 \cdot 0.1822 \cdot 0.4174 =$$

$$= [-0.1829, +0.07886]$$

$$CI = e^{[-0.1829, +0.07886]} = [83.28\%, 108.20\%]$$

# Parallel design

- Not finished yet…
- Analysis still assumed equal variances (against GLs)!
- Degrees of freedom for the *t*-value have to be modified, *e.g.*, by the Welch-Satterthwaite approximation.

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1+1)} + \frac{s_2^4}{n_2^2(n_2+1)}}$$

# Parallel design

- Instead of the simple $\nu = n_1+n_2-2 = 21$, we get

$$\nu = \frac{\left(\frac{0.03418}{11} + \frac{0.03231}{12}\right)^2}{\frac{0.001169}{121 \cdot 12} + \frac{0.001044}{144 \cdot 13}} = 20.705$$

- Maybe it's time to leave M$-Excel.
- Easy to calculate in R.

# Parallel design

```
T <- c(100,103,80,110,78,87,116,99,
       122,82,68)
R <- c(110,113,96,90,111,68,111,93,
       93,82,96,137)
par.equal1 <- t.test(log(R), log(T),
  alternative="two.sided", mu=0,
  paired=FALSE, var.equal=TRUE,
  conf.level=0.90)
par.equal1
 Two Sample t-test

data:  log(T) and log(R)
t = 0.684, df = 21, p-value = 0.5015
alternative hypothesis: true
difference in means is not equal to 0
90 percent confidence interval:
 -0.1829099  0.0788571
sample estimates:
mean of x mean of y
 4.538544  4.590570
round(100*exp(par.equal1$conf.int),
digits=2)
  83.28 108.20
```

liberal!

```
T <- c(100,103,80,110,78,87,116,99,
       122,82,68)
R <- c(110,113,96,90,111,68,111,93,
       93,82,96,137)
par.equal0 <- t.test(log(R), log(T),
alternative="two.sided", mu=0,
paired=FALSE, var.equal=FALSE,
conf.level=0.90)
par.equal0
  Welch Two Sample t-test

data:  log(T) and log(R)
t = 0.6831, df = 20.705, p-value = 0.5021
alternative hypothesis: true difference
in means is not equal to 0
90 percent confidence interval:
 -0.18316379  0.07911102
sample estimates:
mean of x mean of y
 4.538544  4.590570
round(100*exp(par.equal0$conf.int),
digits=2)
 83.26 108.23
```

# Parallel design

- There was just a minor difference (83.28% – 108.20% *vs.* 83.26% – 108.23%), but there was also only little imbalance in the dataset ($n_1$ 11, $n_2$ 12) and the variances were quite similiar ($s_1^2$ 0.03418, $s_2^2$ 0.03231).
- If a dataset is more imbalanced and the variances are ‘truely’ different, the outcome may be *substantially* different. Generally the simple *t*-test is liberal, *i.e.*, the patients’ risk is increased!

# Parallel design

- One million simulated BE studies
  - Lognormal distribution
  - $Mean_{Test}$ 95, $Mean_{Reference}$ 100 (target ratio 95%)
  - $CV\%_{Test}$ 25%, $CV\%_{Reference}$ 40% (‘bad’ reference or inhomogenous groups)
  - $n_{Test}$ 24, $n_{Reference}$ 20
  - If width of CI (*t*-test) < CI (Welch-test) the outcome was considered ‘liberal’
  - Result: *t*-test for homogenous variances was liberal in 97.62% of cases…

# Parallel design

```
set.seed(1234567) # Use this line only to reproduce a run
sims    <- 1E6    # Number of simulations (1 mio simulations will take a couple of minutes)
nT      <- 24     # Subjects in test group
nR      <- 20     # Subjects in reference group
MeanT   <- 95     # Mean test (original scale)
MeanR   <- 100    # Mean reference (original scale)
CVT     <- 0.25   # CV test 25%
CVR     <- 0.40   # CV (bad) reference 40%
MeanlogT<- log(MeanT)-0.5*log(1+CVT^2) # Centered means log scale
MeanlogR<- log(MeanR)-0.5*log(1+CVR^2)
SDlogT  <- sqrt(log(1+CVT^2))          # Standard dev. log scale
SDlogR  <- sqrt(log(1+CVR^2))
Conserv <- 0      # Counters
Liberal <- 0
for (iter in 1:sims){
  PKT     <- rlnorm(n=nT, mean=MeanlogT, sd=SDlogT) # simulated T
  PKR     <- rlnorm(n=nR, mean=MeanlogR, sd=SDlogR) # simulated R
  TtestRes<- t.test(log(PKR), log(PKT), var.equal=TRUE,  conf.level=0.90)
  WelchRes<- t.test(log(PKR), log(PKT), var.equal=FALSE, conf.level=0.90)
  WidthT  <- abs(TtestRes$conf.int[1] - TtestRes$conf.int[2])
  WidthW  <- abs(WelchRes$conf.int[1] - WelchRes$conf.int[2])
  if (WidthT<WidthW){
    Liberal <- Liberal + 1
    }else{
    Conserv <- Conserv + 1
  }
}
result  <- paste(paste("t-test compared to Welch-test\n"),
           paste("Conservative =", 100*Conserv/sims, "%\n"),
           paste("Liberal =", 100*Liberal/sims, "%\n"),
           paste("Number of simulations =",sims,"\n"))
cat(result)
```

# Paired design
## (dependent groups)

- Every subject is treated both with test and reference.
- Generally more powerful than parallel design, because every subject acts as their own reference.
- CI is based on within- (aka *intra*-) subject variance rather than on between- (aka *inter*-) subject variance.

| Subj. | Test | Ref. | $s^2_{within}$ |
|---|---|---|---|
| 1 | 100 | 110 | 50 |
| 2 | 103 | 113 | 50 |
| 3 | 80 | 96 | 128 |
| 4 | 110 | 90 | 200 |
| 5 | 78 | 111 | 545 |
| 6 | 87 | 68 | 181 |
| 7 | 116 | 111 | 13 |
| 8 | 99 | 93 | 18 |
| 9 | 122 | 93 | 421 |
| 10 | 82 | 82 | 0 |
| 11 | 68 | 96 | 392 |
| 12 | 95 | 137 | 882 |
| n | 12 | 12 | 12 |
| mean | 95 | 100 | 240 |
| $s^2_{between}$ | 271 | 314 | |
| $s_{between}$ | 16.4 | 17.7 | |

# Paired design

| Subj. | ln (Test) | ln (Ref.) | Δ (T–R) | (Δ-mean)² |
|---|---|---|---|---|
| 1 | 4.605 | 4.700 | -0.095 | 0.00199 |
| 2 | 4.635 | 4.727 | -0.093 | 0.00176 |
| 3 | 4.382 | 4.564 | -0.182 | 0.01731 |
| 4 | 4.700 | 4.500 | +0.201 | 0.06321 |
| 5 | 4.357 | 4.710 | -0.353 | 0.09125 |
| 6 | 4.466 | 4.220 | +0.246 | 0.08830 |
| 7 | 4.754 | 4.710 | +0.044 | 0.00899 |
| 8 | 4.595 | 4.533 | +0.063 | 0.01283 |
| 9 | 4.804 | 4.533 | +0.271 | 0.10379 |
| 10 | 4.407 | 4.407 | ±0.000 | 0.00258 |
| 11 | 4.220 | 4.564 | -0.345 | 0.08649 |
| 12 | 4.554 | 4.920 | -0.366 | 0.09945 |
| n | 12 | 12 | Σ -0.609 | Σ 0.57794 |
| mean | 4.540 | 4.591 | -0.0507 | |
| $s^2_{between}$ | 0.03110 | 0.03231 | 0.0525 | $s^2_{within}$ |
| $s_{between}$ | 0.1763 | 0.1798 | 0.2292 | $s_{within}$ |

$$\bar{\Delta} = \frac{1}{n}\sum_{i=1}^{i=n}(T_i - R_i) = -\frac{0.609}{12} = -0.05075$$

$$s^2_\Delta = \frac{1}{n-1}\sum_{i=1}^{i=n}(T_i - R_i - \bar{\Delta})^2 = \frac{0.57794}{11} = 0.05254$$

$$s_\Delta = \sqrt{s^2_\Delta} = \sqrt{0.05254} = 0.2292$$

$$CI_{\ln} = \bar{\Delta} \pm t_{2\alpha,n-1} s_\Delta \sqrt{\frac{1}{n}} =$$

$$= -0.05075 \pm 1.796 \cdot 0.2292 \sqrt{\frac{1}{12}} =$$

$$= [-0.16958, +0.06808]$$

$$CI = e^{[-0.16958,+0.06808]} = [84.40\%, 107.05\%]$$

Parallel:
83.28%,108.20%

# Paired *vs.* parallel design

- Only small difference (84.40% – 107.50% *vs.* parallel 83.28% – 108.20%) since based on simulated data not accounting for different CVs (*intra vs. inter*-subject).
- Let's have a look at real data; subsets of the MPH dataset of 405 subjects.
  - 48 subjects parallel: 95.86% [75.89% – 121.10%]
  - First 12 subjects paired: 100.82% [94.91% – 107.09%]
  - Second 12 subjects paired: 91.15% [86.81% – 95.71%]
  - Width of CI of the parallel design is only ~¼ of the paired! Reason: $CV_{intra}$ ~7%, $CV_{inter}$ ~28%.

# R code

```
#Example MPH 20mg MR AUCinf
T <- c(28.39,49.42,36.78,33.36,34.81,24.29,
       28.61,45.54,59.49,28.23, 25.71,42.30,
       62.14,19.69,42.36,97.43,48.57,75.97,
       67.93,79.22,61.68,90.80,60.64,89.91)
R <- c(35.44,39.86,32.75,33.40,34.97,24.65,
       31.77,45.44,65.29,27.87,24.26,37.01,
       63.94,20.65,43.03,115.63,57.40,69.02,
       73.98,91.47,79.65,92.86,70.46,101.40)

#Parallel log-scale (n=48)
par <- t.test(log(T), log(R),
        alternative="two.sided", mu=0,
        paired=FALSE, var.equal=FALSE,
        conf.level=0.90)
result <- paste(paste(
         " Back transformed (raw data scale)\n",
         "Point estimate:",
         round(100*exp(par$estimate[1]-
           par$estimate[2]),
           digits=2),"%\n"),
         round(100*exp(par$estimate[1]-
           par$estimate[2]),
           digits=2),"%\n"),
         paste("90 % confidence interval:"),
         paste(round(100*exp(par$conf.int[1]),
           digits=2), "% to"),
         paste(round(100*exp(par$conf.int[2]),
           digits=2),"%\n"))
par
cat(result)
```

```
#Paired first 12 subjects (using first dataset)
T1     <- T[1:12]; R1 <- R[1:12]
pair1  <- t.test(log(T1), log(R1),alternative="two.sided",
            mu=0, paired=TRUE, conf.level=0.90)
result <- paste(paste" Back transformed (raw data scale)\n",
               "Point estimate:,
               round(100*exp(pair1$estimate),
               digits=2),"%\n"),
          paste("90 % confidence interval:"),
          paste(round(100*exp(pair1$conf.int[1]),
            digits=2), "% to"),
          paste(round(100*exp(pair1$conf.int[2]),
            digits=2),"%\n"))
pair1
cat(result)

#Paired second 12 subjects (using first dataset)
T2     <- T[13:24]; R2 <- R[13:24]
pair2  <- t.test(log(T2), log(R2),alternative="two.sided",
            mu=0, paired=TRUE, conf.level=0.90)
result <- paste(paste" Back transformed (raw data scale)\n",
               "Point estimate:,
               round(100*exp(pair2$estimate),
               digits=2),"%\n"),
          paste("90 % confidence interval:"),
          paste(round(100*exp(pair2$conf.int[1]),
            digits=2), "% to"),
          paste(round(100*exp(pair2$conf.int[2]),
            digits=2),"%\n"))
pair2
cat(result)
```

# R's results

```
        Welch Two Sample t-test

data:  log(T) and log(R)
t = -0.3036, df = 45.69, p-value = 0.7628
alternative hypothesis: true difference in
means is not equal to 0
90 percent confidence interval:
 -0.2759187  0.1914053
sample estimates:
mean of x mean of y
 3.840090  3.882346

Back transformed (raw data scale)
Point estimate: 95.86 %
90 % confidence interval: 75.89 % to 121.1 %
```

```
        Paired t-test

data:  log(T1) and log(R1)
t = 0.2418, df = 11, p-value = 0.8133
alternative hypothesis: true difference in means
is not equal to 0
90 percent confidence interval:
 -0.05227222  0.06854199
sample estimates:
mean of the differences
           0.008134884

Back transformed (raw data scale)
Point estimate: 100.82 %
90 % confidence interval: 94.91 % to 107.09 %
```

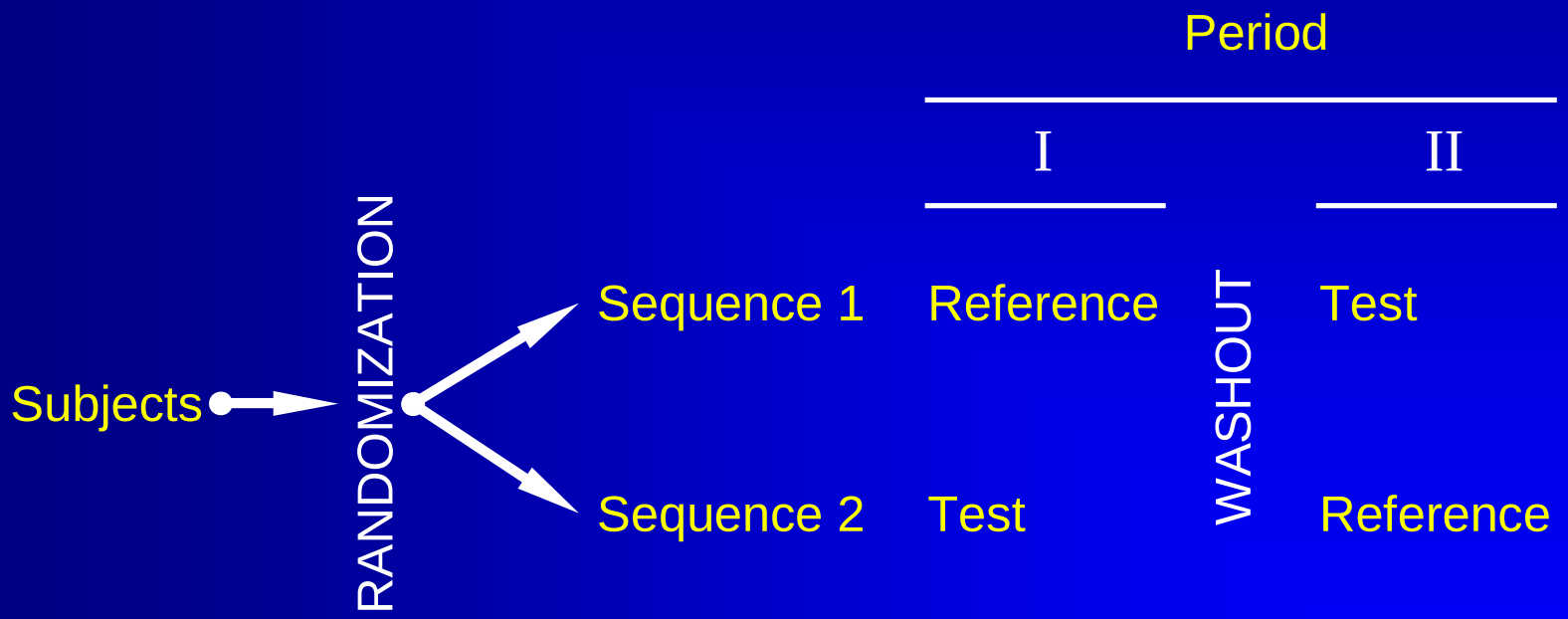
```
        Paired t-test

data:  log(T2) and log(R2)
t = -3.4076, df = 11, p-value = 0.00585
alternative hypothesis: true difference in means
is not equal to 0
90 percent confidence interval:
 -0.14147665 -0.04381995
sample estimates:
mean of the differences
             -0.0926483

Back transformed (raw data scale)
Point estimate: 91.15 %
90 % confidence interval: 86.81 % to 95.71 %
```

# Cross-over designs

- Standard 2×2×2 Design



# Cross-over designs (cont'd)

- Every subject is treated both with test and refe-rence.
- Subjects are randomized into two groups; one is receiving the formulations in the order RT and the other one in the order TR. These two orders are called sequences.
- Whilst in a paired design we must rely on the assumption that no external influences affect the periods, a cross-over design will account for that.

# Cross-over design: Model

Multiplicative Model (X-over without carryover)

$$X_{ijk} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ik} \cdot e_{ijk}$$

$X_{ijk}$: *ln*-transformed response of $j$-th subject ($j=1,\ldots,n_i$) in $i$-th sequence ($i=1,2$) and $k$-th period ($k=1,2$), $\mu$: global mean, $\mu_l$: expected formulation means ($l=1,2$: $\mu_1=\mu_{test}$, $\mu_2=\mu_{ref.}$), $\pi_k$: fixed period effects, $\Phi_l$: fixed formulation effects ($l=1,2$: $\Phi_1=\Phi_{test}$, $\Phi_2=\Phi_{ref.}$)

# Cross-over design:
## Assumptions

Multiplicative Model (X-over without carryover)

$$X_{ijk} = \mu \cdot \pi_k \cdot \Phi_l \cdot s_{ik} \cdot e_{ijk}$$

- All $ln\{s_{ik}\}$ and $ln\{e_{ijk}\}$ are independently and normally distributed about unity with variances $\sigma^2_s$ and $\sigma^2_e$.
  - ➔ This assumption may not hold true for all formulations; if the reference formulation shows higher variability than the test formulation, a *'good'* test will be penalized for the *'bad'* reference.
- All observations made on different subjects are independent.
  - ➔ This assumption should not be a problem, unless you plan to include twins or triplets in your study…

# Cross-over designs (cont'd)

- Standard 2×2×2 design
  - Advantages
    - Globally applied standard protocol for bioequivalence, PK interaction, food studies
    - Straigthforward statistical analysis
  - Disadvantages
    - Not suitable for drugs with long half life (→ parallel groups)
    - Not optimal for studies in patients with instable diseases (→ parallel groups)
    - Not optimal for HVDs/HVDPs (→ Replicate Designs)

# Cross-over design: Evaluation

- Mainly by ANOVA and LMEM (linear mixed effects modeling). Results are identical for balanced datasets, and differ only slightly for imbalanced ones.
- Avoid M$-Excel! Almost impossible to validate; tricky for imbalanced datasets – a nightmare for higher-order X-overs. Replicates impossible.
- Suitable software: SAS, Phoenix/WinNonlin, Kinetica and EquivTest/PK (both only 2×2 X-over), S+, Package *bear* for R (freeware).

# Cross-over design: Example

| subject | T | R |
|---|---|---|
| 1 | 28.39 | 35.44 |
| 2 | 39.86 | 49.42 |
| 3 | 32.75 | 36.78 |
| 4 | 33.36 | 33.40 |
| 5 | 34.97 | 34.81 |
| 6 | 24.29 | 24.65 |
| 7 | 28.61 | 31.77 |
| 8 | 45.44 | 45.54 |
| 9 | 59.49 | 65.29 |
| 10 | 27.87 | 28.23 |
| 11 | 24.26 | 25.71 |
| 12 | 42.30 | 37.01 |

| | sequence RT | | | sequence TR | |
|---|---|---|---|---|---|
| subject | P I | P II | subject | P I | P II |
| 2 | 39.86 | 49.42 | 1 | 28.39 | 35.44 |
| 3 | 32.75 | 36.78 | 4 | 33.36 | 33.40 |
| 5 | 34.97 | 34.81 | 6 | 24.29 | 24.65 |
| 8 | 45.44 | 45.54 | 7 | 28.61 | 31.77 |
| 10 | 27.87 | 28.23 | 9 | 59.49 | 65.29 |
| 11 | 24.26 | 25.71 | 12 | 42.30 | 37.01 |

Ordered by treatment sequences (RT|TR)

ANOVA on log-transformed data →

# Cross-over design: Example

| Sequence | Period 1 | | Period 2 | | Sequence mean | |
|---|---|---|---|---|---|---|
| 1 | 1R = $X_{.11}$ | 3.5103 | 1T = $X_{.21}$ | 3.5768 | $X_{..1}$ | 3.5436 |
| 2 | 2T = $X_{.12}$ | 3.5380 | 2R = $X_{.22}$ | 3.5883 | $X_{..2}$ | 3.5631 |
| Period mean | $X_{.1.}$ | 3.5241 | $X_{.2.}$ | 3.5826 | $X_{...}$ | 3.5533 |

| | | | | | | |
|---|---|---|---|---|---|---|
| RT = | $n_1$ = | 6 | | | | |
| TR = | $n_2$ = | 6 | $1/n_1+1/n_2$ | 0.3333 | | |
| ***balanced*** | n = | 12 | 1/n | 0.0833 | $n_1+n_2-2$ | 10 |

**Analysis of Variance**

| Source of variation | df | SS | MS | F | P-value | CV |
|---|---|---|---|---|---|---|
| *Inter* -subjects | | | | | | |
| Carry-over | 1 | 0.00230 | 0.00230 | 0.0144 | 0.90679 | |
| Residuals | 10 | 1.59435 | 0.15943 | 29.4312 | 4.32E-6 | 28.29% |
| *Intra* -subjects | | | | | | |
| Direct drug | 1 | 0.00040 | 0.00040 | 0.0733 | 0.79210 | |
| Period | 1 | 0.02050 | 0.02050 | 3.7844 | 0.08036 | |
| Residuals | 10 | 0.05417 | 0.00542 | | | 7.37% |
| Total | 23 | 1.67172 | | | | |

$\delta_{ML}$ **1.0082** *MLE (maximum likelihood estimator) of Delta-ML*

$X_R$ 3.5493 *LS (least squares mean for the reference formulation)* $\exp(X_R)$ 34.79

$X_T$ 3.5574 *LS (least squares mean for the test formulation)* $\exp(X_T)$ 35.07

# Cross-over design: Example

**Classical (Shortest) Confidence Interval**

± x rule: 20 [ 100 - x; 1 / (100 - x) ]

| | | | | | |
|---|---|---|---|---|---|
| $\theta_L$ | -0.2231 | $\theta_U$ | +0.2231 | $\alpha$ 0.0500 | p=1-2·$\alpha$ 0.9000 |
| $\delta_L$ | **80%** | $\delta_U$ | **125%** | $t_{2\cdot\alpha,df}$ | 1.8125 |
| $L_1$ | -0.0463 | $U_1$ | 0.0626 | *difference within Theta-L AND Theta-U; bioequivalent* | |
| $L_2$ | **95.47%** | $U_2$ | **106.46%** | ***difference within Delta-L AND Delta-U; bioequivalent*** | |

| | | |
|---|---|---|
| $\delta_{ML}$ | **100.82%** | *MLE; maximum likelihood estimator* |
| $\delta_{MVUE}$ | 100.77% | *MVUE; minimum variance unbiased estimator* |
| $\delta_{RM}$ | 100.98% | *RM; ratio of formulation means* |
| $\delta_{MIR}$ | 101.44% | *MIR; mean of individual subject ratios* |

# Cross-over design: Example

- Calculation of 90% CI (2-way cross-over)
  - Sample size ($n$) 12, Point Estimate ($PE$) 100.82%, Residual Mean Squares Error ($MSE$) from ANOVA ($ln$-transformed values) 0.005417, $t_{\alpha,n-2}$ 1.8125
    - Standard Error ($SE_\Delta$) of the mean difference

$$SE_\Delta = \sqrt{MSE}\sqrt{\frac{2}{n}} = \sqrt{0.005417}\sqrt{\frac{2}{12}} = 0.030047$$

    - Confidence Interval

$$CL_L = e^{\ln PE - t_{2\alpha,df}\cdot SE_\Delta} = e^{0.0081349 - 1.8125\times 0.030047} = 95.47\%$$

$$CL_H = e^{\ln PE + t_{2\alpha,df}\cdot SE_\Delta} = e^{0.0081349 + 1.8125\times 0.030047} = 106.46\%$$

# R code / result

```
#Cross-over 12 subjects
T1     <- c(28.39,33.36,24.29,28.61,59.49,42.30)
T2     <- c(49.42,36.78,34.81,45.54,28.23,25.71)
R1     <- c(39.86,32.75,34.97,45.44,27.87,24.26)
R2     <- c(35.44,33.40,24.65,31.77,65.29,37.01)
RT     <- log(R1) - log(T2)
TR     <- log(R2) - log(T1)
n1     <- length(RT)
mRT    <- mean(RT)
vRT    <- var(RT)
n2     <- length(TR)
mTR    <- mean(TR)
vTR    <- var(TR)
mD     <- mean(log(c(T1,T2))) - mean(log(c(R1,R2)))
MSE    <- (((n1-1)*vRT + (n2-1)*vTR)/(n1+n2-2))/2
alpha  <- 0.05
lo     <- mD - qt(1-alpha,n1+n2-2)*sqrt(MSE)*
            sqrt((1/(2*n1) + 1/(2*n2)))
hi     <- mD + qt(1-alpha,n1+n2-2)*sqrt(MSE)*
            sqrt((1/(2*n1) + 1/(2*n2)))
result <- paste(
          paste(" Back transformed (raw data scale)\n",
                "Point estimate:“,
                round(100*exp(mD), digits=2),"%\n"),
          paste("90 % confidence interval:"),
          paste(round(100*exp(lo), digits=2), "% to"),
          paste(round(100*exp(hi), digits=2),"%\n",
          paste("CVintra:",round(100*sqrt(exp(MSE)-1),
            digits=2),"%\n")))
cat(result)
```

```
Back transformed (raw data scale)
Point estimate: 100.82 %
90 % confidence interval: 95.47 % to 106.46 %
CVintra: 7.37 %
```

# Comparison of designs

- Further reduction in variability, because the influence of periods is accounted for.
  - Paired design: 100.82% [94.91% – 107.09%]
  - Cross-over design: 100.82% [95.47% – 106.46%]
  - Point estimates are the same, only variability caused by period- and/or sequence-effects is removed.

Setup | Results | Verification

Output Data

- Filtered Cells W...
- Filtered Rows W...
- Result Worksheet

| | Design | Ratio | CL90lo | CL90hi | Diff_SE | CI_width |
|---|---|---|---|---|---|---|
| 1 | Parallel | 95.86 | 75.89 | 121.1 | 0.139176 | 45.21 |
| 2 | Paired | 100.82 | 94.91 | 107.1 | 0.033636 | 12.19 |
| 3 | Xover | 100.82 | 95.47 | 106.46 | 0.030048 | 10.99 |

# Comparison of designs

- Most Important in an ANOVA table is the residual mean error (CI, $CV_{intra}$ for future studies).
  - Carry-over (aka sequence effects) can not be handled! Must be excluded by design (long enough washout).
  - Period effects are accounted for (significant *p*-values are not important). Example: all values in PII ×100…

| Dependent | Hypothesis | DF | SS | MS | F_stat | P_value |
|---|---|---|---|---|---|---|
| Ln(AUCinf) | sequence | 1 | 0.002299563 | 0.002299563 | 0.014423232 | 0.90678513 |
| Ln(AUCinf) | sequence*subject | 10 | 1.5943472 | 0.15943472 | 29.431239 | 4.3211352E-0 |
| Ln(AUCinf) | treatment | 1 | 0.000397058 | 0.000397058 | 0.07329589 | 0.79210291 |
| Ln(AUCinf) | period | 1 | 0.020500963 | 0.020500963 | 3.784425 | 0.080364101 |
| Ln(AUCinf) | Error | 10 | 0.054171935 | 0.005417193 | | |

| CI_90_Lower | CI_90_Upper |
|---|---|
| 95.471828 | 106.46102 |

| Dependent | Hypothesis | DF | SS | MS | F_stat | P_value |
|---|---|---|---|---|---|---|
| Ln(AUCinf) | sequence | 1 | 0.002299563 | 0.002299563 | 0.014423232 | 0.90678513 |
| Ln(AUCinf) | sequence*subject | 10 | 1.5943472 | 0.15943472 | 29.431239 | 4.3211352E-0 |
| Ln(AUCinf) | treatment | 1 | 0.000397058 | 0.000397058 | 0.07329589 | 0.79210291 |
| Ln(AUCinf) | period | 1 | 130.49632 | 130.49632 | 24089.286 | 0 |
| Ln(AUCinf) | Error | 10 | 0.054171935 | 0.005417193 | | |

| CI_90_Lower | CI_90_Upper |
|---|---|
| 95.471828 | 106.46102 |

# Reading ANOVA tables

**Analysis of Variance**

| Source of variation | df | SS | MS | F | P-value | CV |
|---|---|---|---|---|---|---|
| Between subjects | | | | | | |
| Carry-over | 1 | 0.00230 | 0.002300 | 0.0144 | 0.90679 | |
| Residuals | 10 | 1.59435 | 0.159435 | 29.4312 | 4.32E-6 | 28.29% |
| Within subjects | | | | | | |
| Direct drug | 1 | 0.00040 | 0.000397 | 0.0733 | 0.79210 | |
| Period | 1 | 0.02050 | 0.020501 | 3.7844 | 0.08036 | |
| Residuals | 10 | 0.05417 | 0.005417 | | | 7.37% |
| Total | 23 | 1.67172 | | | | |

Should not be tested: Design – washout!

$$CV_{inter} = \sqrt{e^{MSE_B - \frac{MSE_W}{2}} - 1}$$

$$CV_{intra} = \sqrt{e^{MSE_W} - 1}$$

Not surprising: different subjects!

Not important: Both formulations would be affected in the same way.

Not important: Significant value would old mean that 100% is not included in the CI.

balanced: $n_1 = n_2; n = n_1 + n_2$

$$CI = e^{\ln PE \pm t_{\alpha,n-2} \cdot \sqrt{MSE_W} \sqrt{\frac{2}{n}}}$$

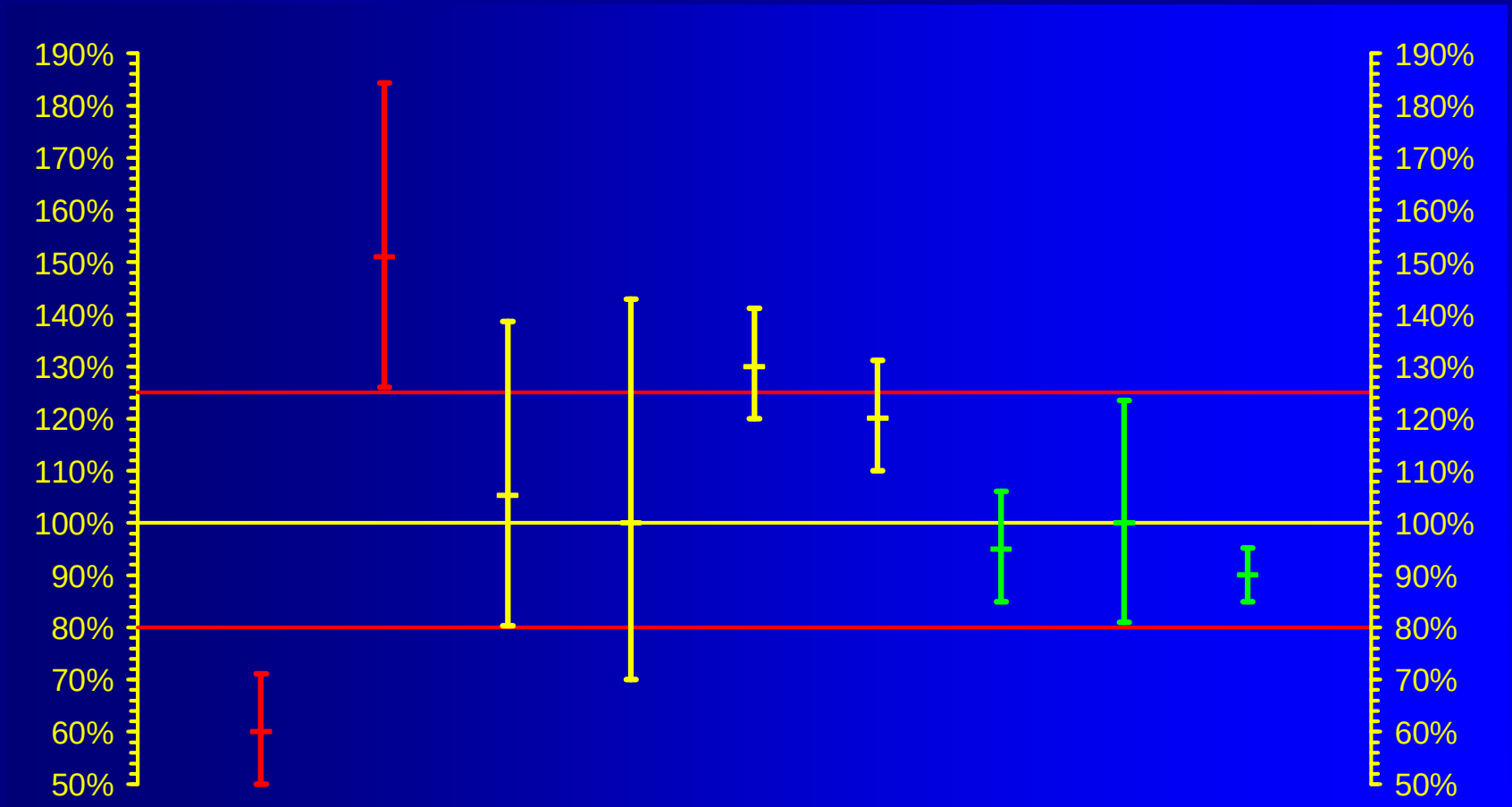
imbalanced: $n_1 \neq n_2$

$$CI = e^{\ln PE \pm t_{\alpha,n_1+n_2-2} \cdot \sqrt{MSE_W} \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}}$$

# A note on BE assessment

- The *width* of the confidence interval depends on the variability observed in the study.
- The *location* of the confidence interval depends on the observed test/reference-ratio.
- Decision rules:
  - Confidence Interval (CI) entirely outside the Acceptance Range (AR): Bioinequivalence proven.
  - CI overlaps the AR, but is not entirely within the AR: Bioequivalence not proven.
  - CI entirely within the AR: Bioequivalence proven.

# A note on BE assessment



# Cross-over designs (cont'd)

- Special case: Evaluation of $t_{max}$
  - Since $t_{max}$ is sampled from discrete values, a nonparametric method must be applied
  - Estimation of differences (linear model)
  - Wilcoxon Two-Sample Test (available in SAS 9.2 Proc NPAR1way, Phoenix/WinNonlin, EquivTest/PK, R package *coin*)
  - Since based on a discrete distribution, generally $\alpha < 0.05$ (*e.g.*, n=12: 0.0465, 24: 0.0444, 32: 0.0469, 36: 0.0485, 48: 0.0486,…)

**Hauschke D, Steinijans VW and E Diletti**
*A distribution-free procedure for the statistical analysis of bioequivalence studies*
Int J Clin Pharm Ther Toxicol 28/2, 72–78 (1990)

# Cross-over designs (cont'd)

| Sequence 1 (RT) | | | | Sequence 2 (TR) | | | |
|---|---|---|---|---|---|---|---|
| Subject | Period I | Period II | P.D. | Subject | Period I | Period II | P.D. |
| 2 | 3.0 | 1.5 | -1.5 | 1 | 2.0 | 2.0 | ±0.0 |
| 4 | 2.0 | 2.0 | ±0.0 | 3 | 2.0 | 2.0 | ±0.0 |
| 6 | 2.0 | 3.0 | +1.0 | 5 | 2.0 | 3.0 | +1.0 |
| 8 | 2.0 | 3.0 | +1.0 | 7 | 2.0 | 1.5 | -0.5 |
| 10 | 1.5 | 2.0 | +0.5 | 9 | 3.0 | 2.0 | -1.0 |
| 12 | 3.0 | 2.0 | -1.0 | 11 | 2.0 | 1.5 | -0.5 |
| 14 | 3.0 | 3.0 | ±0.0 | 13 | 3.0 | 1.5 | -1.5 |

# Cross-over designs (cont'd)

**ADDITIVE (raw data) MODEL**

**metric: $t_{max}$**

| Sequence | Period 1 | | Period 2 | |
|---|---|---|---|---|
| 1 | $R_{L1}$ = | 65 | $R_{U1}$ = | 46 |
| 2 | $R_{L2}$ = | 36 | $R_{U2}$ = | 55 |
| RT = | $n_1$ = | 7 | | |
| TR = | $n_2$ = | 7 | | |
| ***balanced*** | n = | 14 | $n_1 \cdot n_2$ | 49 |

$d_{\cdot 1}$ 0.0000 $d_{\cdot 2}$ -0.1786 *(mean period difference in sequence 1 / 2)*

$\bar{Y}_R$ 2.000 *median of the reference formulation*

$\bar{Y}_T$ 2.000 *median of the test formulation*

**Distribution-Free Confidence Interval (Moses)**

± x rule : 20

$\theta_L$ -0.429 $\theta_U$ +0.429 $\alpha$ 0.0487 $p = 1-2\cdot\alpha$ 0.9026

$\delta_L$ 80% $\delta_U$ 120%

$L_W$ -0.250 $U_W$ +0.750 ***difference outside Theta-L AND/OR Theta-U; not bioequivalent***

$\bar{\theta}$ **+0.250** ***Hodges-Lehmann estimate (median of paired differences)***

**Wilcoxon-Mann-Whitney Two One-Sided Tests Procedure (Hauschke)**

$W_L$ 37 $W_U$ 18

$W_{0.95,n1,n2}$ 38 $W_{0.05,n1,n2}$ 12 *H0(1): diff. <= Theta-L AND H0(2): diff. => Theta-U; not bioequivalent*

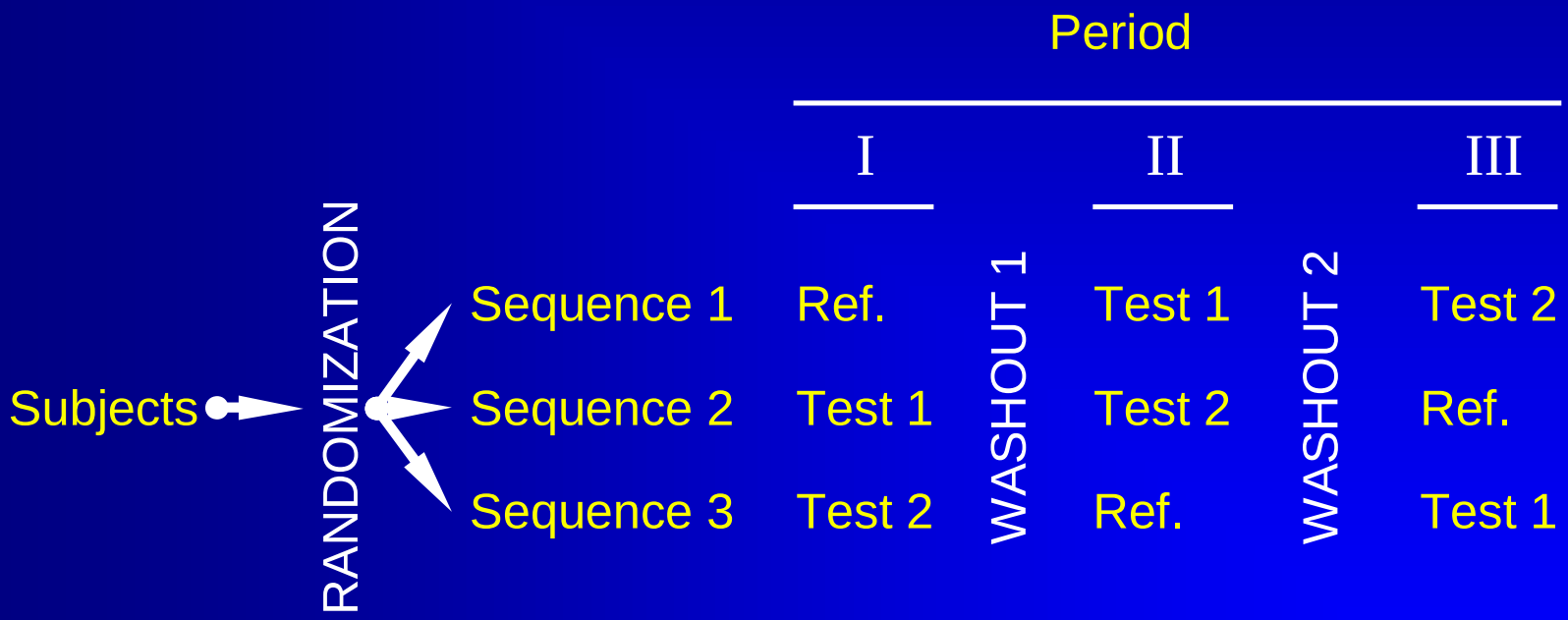
$p_1$ >0.0487 and $p_2$ >0.0487

# Cross-over designs (cont'd)

- Higher Order Designs (for more than two treatments)
  - Latin Squares
    Each subject is randomly assigned to sequences, where number of treatments = number of sequences = number of periods.
  - Variance Balanced Designs

# Cross-over designs (cont'd)

- 3×3×3 Latin Square Design



# Cross-over designs (cont'd)

- 3×3×3 Latin Square design
  - Advantages
    - Allows to choose between two candidate test formulations or comparison of one test formulation with two references.
    - Easy to adapt.
    - Number of subjects in the study is a multiplicative of three.
    - Design for establishment of Dose Proportionality.
  - Disadvantages
    - Statistical analysis more complicated (especially in the case of drop-outs and a small sample size) – not available in some pieces of software.
    - Extracted pairwise comparisons are imbalanced.
    - May need measures against multiplicity (increasing the sample size).
    - Not mentioned in any guideline.

# Cross-over designs (cont'd)

- Higher Order Designs (for more than two treatments)
  - Variance Balanced Designs (Williams' Designs)
    - For *e.g.*, three formulations there are three possible pairwise differences among formulation means (*i.e.*, form. 1 *vs.* form. 2., form 2 *vs.* form. 3, and form. 1 *vs.* form. 3).
    - It is desirable to estimate these pairwise effects with the same degree of precision (there is a common variance for each pair).
      - Each formulation occurs only once with each subject.
      - Each formulation occurs the same number of times in each period.
      - The number of subjects who receive formulation *i* in some period followed by formulation *j* in the next period is the same for all *i* # *j*.
    - Such a design for three formulations is the three-treatment six-sequence three-period Williams' Design.

# Cross-over designs (cont'd)

- Williams' Design for three treatments

| Sequence | Period | | |
|---|---|---|---|
| | I | II | III |
| 1 | R | $T_2$ | $T_1$ |
| 2 | $T_1$ | R | $T_2$ |
| 3 | $T_2$ | $T_1$ | R |
| 4 | $T_1$ | $T_2$ | R |
| 5 | $T_2$ | R | $T_1$ |
| 6 | R | $T_1$ | $T_2$ |

# Cross-over designs (cont'd)

- Williams' Design for four treatments

| Sequence | Period | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| 1 | R | $T_3$ | $T_1$ | $T_2$ |
| 2 | $T_1$ | R | $T_2$ | $T_3$ |
| 3 | $T_2$ | $T_1$ | $T_3$ | R |
| 4 | $T_3$ | $T_2$ | R | $T_1$ |

# Cross-over designs (cont'd)

- Williams' Designs
  - Advantages
    - Allows to choose between two candidate test formulations or comparison of one test formulation with two references.
    - Design for establishment of Dose Proportionality.
    - Paired comparisons (*e.g.*, for a nonparametric method) can be extracted, which are also balanced.
    - Mentioned in Brazil's (ANVISA) and EU's (EMA) guidelines.
  - Disadvantages
    - Mores sequences for an *odd* number of treatment needed than in a Latin Squares design (but equal for even number).
    - Statistical analysis more complicated (especially in the case of drop-outs) – not available in some softwares.
    - May need measures against multiplicity (increasing the sample size).

# Cross-over designs (cont'd)

- Higher Order Designs (cont'd)
  - Bonferroni-correction needed (sample size!)
    - *If more than one formulation will be marketed* (for three simultaneous comparisons without correction patients' risk increases from 5 % to 14 %).
    - *Sometimes* requested by regulators in dose proportionality.

| $k$ | $p_{\alpha=0.05}$ | $p_{\alpha=0.10}$ | $\alpha_{adj}$ | $p_{corr}$ | $\alpha_{adj}$ | $p_{corr}$ |
|---|---|---|---|---|---|---|
| 1 | 5.00% | 10.00% | 0.0500 | 5.00% | 0.100 | 10.00% |
| 2 | 9.75% | 19.00% | 0.0250 | 4.94% | 0.050 | 9.75% |
| 3 | 14.26% | 27.10% | 0.0167 | 4.92% | 0.033 | 6.67% |
| 4 | 18.55% | 34.39% | 0.0125 | 4.91% | 0.025 | 9.63% |
| 5 | 22.62% | 40.95% | 0.0100 | 4.90% | 0.020 | 9.61% |
| 6 | 26.49% | 46.86% | 0.0083 | 4.90% | 0.017 | 9.59% |

$$\alpha_{adj} = \alpha^{1/k}$$

$$p_{corr} = 1 - \left(1 - \alpha_{adj}\right)^{k}$$

# Cross-over designs (cont'd)

- Higher Order Designs (cont'd)
  - Effect of $\alpha$-adjustment on sample size (expected T/R 95%, $CV_{intra}$ 20%, power 80%)

| CV% | 2×2 $\alpha$ 0.05 | 6×3 $\alpha_{adj.}$ 0.025 | comp. 2×2 | 4×4 $\alpha_{adj.}$ 0.0167 | comp. 2×2 |
|---|---|---|---|---|---|
| 10.0 | 8 | 12 | +50% | 16 | +100% |
| 12.5 | 10 | 12 | +20% | 16 | +60% |
| 15.0 | 12 | 18 | +50% | 16 | +33% |
| 17.5 | 16 | 24 | +50% | 24 | +50% |
| 20.0 | 20 | 24 | +20% | 28 | +40% |
| 22.5 | 24 | 30 | +25% | 36 | +50% |
| 25.0 | 28 | 36 | +29% | 40 | +49% |
| 27.5 | 34 | 42 | +24% | 48 | +41% |
| 30.0 | 40 | 54 | +35% | 56 | +40% |

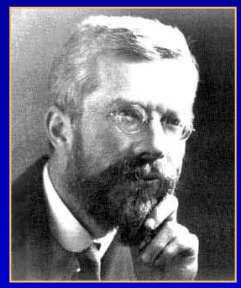
# Part II: Basic Designs for BE Studies

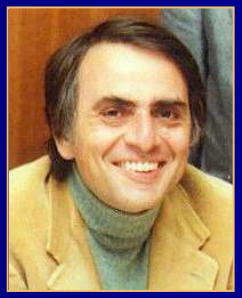
Helmut Schütz
**BEBAC**
Consultancy Services for
Bioequivalence and Bioavailability Studies
1070 Vienna, Austria
helmut.schuetz@bebac.at

# To bear in Remembrance...

To call the statistician after the experiment is done may be no more than asking him to perform a *post-mortem* examination: he may be able to say what the experiment died of.

*Ronald A. Fisher*





[The] impatience with ambiguity can be criticized in the phrase:
*absence of evidence is not evidence of absence.*

*Carl Sagan*

[…] our greatest mistake would be to forget that data is used for serious decisions in the very real world, and bad information causes suffering and death.

*Ben Goldacre*