Part II
The statistical assessment of bioequivalence

Thomas Mathew

Department of Mathematics and Statistics
University of Maryland Baltimore County (UMBC)
Baltimore, Maryland 21250

`mathew@umbc.edu`

Outline

- A canonical form for 2×2 crossover designs
- Average bioequivalence
- Testing average bioequivalence
- Other topics
  - Comparing variances
  - Highly variable drugs
- The assessment of biosimilarity
- Complex generic forms
- Other equivalence testing problems

A canonical form for 2 × 2 crossover designs

	Period	
Sequence	I	II
1	R	T
2	T	R

The washout period is usually long enough so that there is no carry-over effect from the first period to the second period.

It is often the case that there are no period and sequence effects.

We proceed under the above assumptions.

In view of the above assumptions, there is no need to distinguish between the two sequences while modeling the responses.

Data obtained on three variables:

Area under the curve (AUC)

Maximum blood concentration (C$_{\mathrm{max}}$)

Time to reach the maximum concentration (T$_{\mathrm{max}}$)

Statistical analysis of the data is done to determine if the two drugs are equivalent.

Univariate bioequivalence testing: Based on the separate modeling and analysis of the univariate responses AUC, $C_{\max}$ and $T_{\max}$.

Multivariate bioequivalence testing: Based on the joint modeling and analysis of all the three responses AUC, $C_{\max}$ and $T_{\max}$, or any two of them (typically, AUC and $C_{\max}$).

Usually log-transformed data are analyzed.

We shall discuss only univariate bioequivalence.

Suppose there are n healthy volunteers in the study

A few are assigned to the first sequence and a few to the second sequence.

y_{jT}, y_{jR}: responses to T and R for the jth subject, $j = 1, 2, ..., n$.

Usually, the response is the log-transformed AUC

Recall: We are dealing with a 2×2 crossover design.

We assume the models

$$\begin{aligned} y_{jT} &= \mu_T + \eta_{jT} + e_{jT} \\ y_{jR} &= \mu_R + \eta_{jR} + e_{jR} \end{aligned}$$

$(\eta_{jT}, \eta_{jR})'$: a bivariate random effect corresponding to the jth subject. Captures between subject variability

e_{jT}, e_{jR}: within subject random errors.

We assume

$$(\eta_{jT}, \eta_{jR})' \sim N(0, \Sigma_B), \; e_{jT} \sim N(0, \sigma^2_{WT}) \text{ and } \; e_{jR} \sim N(0, \sigma^2_{WR}),$$

and all the random variables are independent, $j = 1$, 2, ..., n.

Note: Due to lack of replication, the parameters Σ_B, σ^2_{WT} and σ^2_{WR} cannot be estimated.

Write

$$\Sigma_B = \begin{pmatrix} \sigma^2_{BT} & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma^2_{BR} \end{pmatrix}$$

From the models for y_{jT} and y_{jR}, we note that

$$d_j = y_{jT} - y_{jR} \sim N(\mu_T - \mu_R, \sigma^2),$$

where

$$\begin{aligned} \sigma^2 &= V\left(y_{jT} - y_{jR}\right) \\ &= V\left[(\eta_{jT} - \eta_{jR}) + (e_{jT} - e_{jR})\right] \\ &= \sigma^2_{BT} + \sigma^2_{BR} - 2\rho\sigma_{BT}\sigma_{BR} + \sigma^2_{WT} + \sigma^2_{WR} \end{aligned}$$

Average bioequivalence can be tested using the data

$$d_j = y_{jT} - y_{jR} \sim N(\mu_T - \mu_R, \sigma^2),$$

$j = 1,\ 2,\ ....,\ n$, where $\sigma^2 = V(d_j)$ is as defined before.

σ^2 can be estimated using the sample variance among the d_js.

$\bar{d}$, S^2: Sample mean and sample variance among the d_js.

Under a 2×2 crossover design, a canonical form for the average bioequivalence problem consists of

$$\bar{d} \sim N\left(\mu_T - \mu_R, \frac{\sigma^2}{n}\right) \quad \text{and} \quad (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

A similar canonical form can be obtained under other crossover designs.

We shall assume the canonical form to be

$$D \sim N(\mu_T - \mu_R, c^2\sigma^2) \text{ and } v\frac{S^2}{\sigma^2} \sim \chi_v^2$$

where the known constant c^2 and the df v depend on the assumed model, the design and the sample sizes.

Average bioequivalence

μ_T, μ_R: average responses among the population of patients who take the test drug, and the reference drug, respectively.

The response is usually AUC, after log-transformation (could be $\mathsf{C}_{\max}$ or $\mathsf{T}_{\max}$).

Average bioequivalence holds if μ_T and μ_R are equivalent, i.e., they are "close"

μ_T and μ_R are considered equivalent if $|\mu_T - \mu_R| < \ln(1.25)$.

After obtaining data from a crossover design, perform a statistical test to decide if $|\mu_T - \mu_R| \geq \ln(1.25)$ or if $|\mu_T - \mu_R| < \ln(1.25)$

Hypotheses to be tested

$$H_0: \ |\mu_T - \mu_R| \geq \ln(1.25) \ \text{ versus } \ H_1: \ |\mu_T - \mu_R| < \ln(1.25)$$

Conclude average bioequivalence if H_0 is rejected after a statistical test based on the log-transformed AUC data.

Can we switch H_0 and H_1?

That is, test the hypotheses

$$H_0: \ |\mu_T - \mu_R| \leq \ln(1.25) \ \text{ versus } \ H_1: \ |\mu_T - \mu_R| > \ln(1.25)$$

Now conclude average bioequivalence if H_0 is accepted after a statistical test.

In order to decide the formulation of H_0 and H_1, let's look at the interpretation of a type I error probability.

First consider the standard formulation

$$H_0 : \ |\mu_T - \mu_R| \geq \ln(1.25) \ \text{ versus } \ H_1 : \ |\mu_T - \mu_R| < \ln(1.25)$$

We conclude average bioequivalence by rejecting H_0.

Type I error: Reject H_0 when it is true

That is, conclude average bioequivalence when average bioequivalence does not hold.

Since the generic drug will be taken by patients, we don't want to market a drug that is not equivalent to the brand name drug. That is, we don't want to commit a type I error.

Thus under the standard formulation, controlling the type I error probability is very important.

Consider once again the standard formulation

$$H_0 : \; |\mu_T - \mu_R| \geq \ln(1.25) \;\; \text{versus} \;\; H_1 : \; |\mu_T - \mu_R| < \ln(1.25)$$

Type II error: Do not reject H_0 when it is not true

That is, do not conclude average bioequivalence when average bioequivalence holds.

If a type II error is committed, then a generic manufacturer may not be able to market a drug that is bioequivalent to the brand name.

Less serious than marketing a generic drug that is not bioequivalent.

Under the standard formulation, type I error becomes the more serious error, and the test is carried out to control the type I error probability.

If we switch the two hypotheses, we have

$$H_0: \ |\mu_T - \mu_R| \leq \ln(1.25) \ \text{ versus } \ H_1: \ |\mu_T - \mu_R| > \ln(1.25),$$

Now average bioequivalence is concluded when H_0 is accepted

Since the test is carried out after fixing the type I error probability, we are now unable to control the more serious error.

Thus for concluding average bioequivalence, the hypotheses are formulated as:

$$H_0 : \ |\mu_T - \mu_R| \geq \ln(1.25) \ \text{ versus } \ H_1 : \ |\mu_T - \mu_R| < \ln(1.25)$$

Conclude average bioequivalence if H_0 is rejected after a statistical test.

Testing average bioequivalence

Write $\mu_T - \mu_R = \mu$

$$D \sim N(\mu, c^2\sigma^2),\ v\frac{S^2}{\sigma^2} \sim \chi^2_v$$

$$H_0 : |\mu| \geq \ln(1.25), \quad \text{vs} \quad H_1 : |\mu| < \ln(1.25).$$

Rewrite as

$$\begin{aligned} H_{01} : \mu \leq -\ln(1.25) \text{ vs. } H_{11} : \mu > -\ln(1.25), \\ H_{02} : \mu \geq \ln(1.25) \text{ vs. } H_{12} : \mu < \ln(1.25). \end{aligned}$$

Average bioequivalence is concluded if both H_{01} and H_{02} are rejected.

First consider $H_{01}: \mu \leq -\ln(1.25)$ vs. $H_{11}: \mu > -\ln(1.25)$

Recall the canonical form $D \sim N(\mu, c^2\sigma^2),\ v\frac{S^2}{\sigma^2} \sim \chi_v^2$

t−test: Reject H_0 when $\frac{D+\ln(1.25)}{cS} > t_v(\alpha)$

Now consider $H_{02}: \mu \geq \ln(1.25)$ vs. $H_{12}: \mu < \ln(1.25)$

t−test: Reject H_0 when $\frac{D-\ln(1.25)}{cS} < -t_v(\alpha)$.

Conclude average bioequivalence at significance level α if

$$\frac{D + \ln(1.25)}{cS} > t_v(\alpha) \quad \text{and} \quad \frac{D - \ln(1.25)}{cS} < -t_v(\alpha)$$

$$\text{Equivalently, if} \quad \frac{|D| - \ln(1.25)}{cS} < -t_v(\alpha)$$

Two one-sided t-test (TOST)

Schuirmann (1981), *Biometrics*

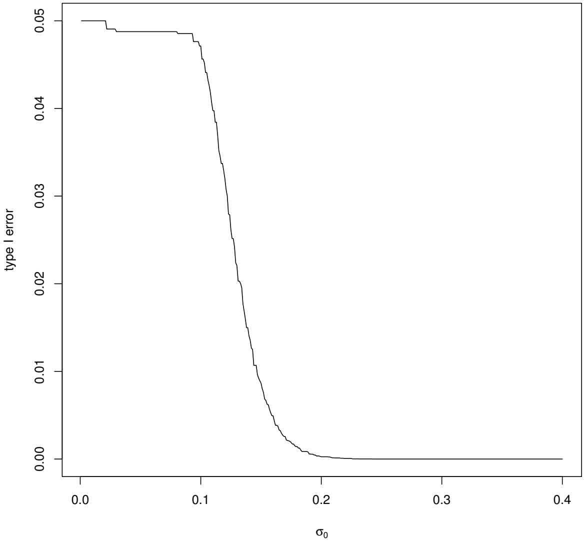
Schuirmann (1987), *Journal of Pharmacokinetics and Biopharmaceutics*.

The TOST is equivalent to computing a 90% confidence interval for $\mu = \mu_T - \mu_R$, and verifying if the interval is contained in $(-\ln(1.25),\ \ln(1.25))$.

The following can be theoretically shown:

- Even though the TOST consists of two t-tests, each carried out using an α significance level, the TOST has an overall type I error probability not exceeding α.
- The type I error probability approaches zero as σ becomes large.
- The type I error probability approaches α as σ approaches zero.

The following plot shows the behavior of the type I error probability against values of $\sigma_0 = c\sigma$ (plotted for $\alpha = 0.05$).



The type I error probability of the TOST

The TOST can be quite conservative as σ gets large

Improved tests due to:

Anderson and Hauck (1983), *Communications in Statistics*

Munk (1993), *Biometrics*

Berger and Hsu (1996), *Statistical Science*

Brown, Hwang and Munk (1997), *Annals of Statistics*

Munk, Brown and Hwang (2000), *Biometrical Journal*

Cao and Mathew (2008), *Biometrical Journal*

Very often, the σ value is not that large.

In cases where it is large, the usual average bioequivalence criterion, and the TOST are not used.

If the TOST is implemented using a significance level greater than 5%, the actual type I error probability could become close to 0.05.

What significance level should be used in order to achieve this?

The significance level to be used will depend on σ, and could also depend on the sample size.

Since σ is unknown, the actual significance level to be used has to be estimated from the data

We use the bootstrap to estimate the significance level

The process is called bootstrap calibration.

In order to explain the bootstrap calibrated TOST, recall the canonical form:

$$D \sim N(\mu_T - \mu_R, c^2\sigma^2) \text{ and } v\frac{S^2}{\sigma^2} \sim \chi^2_v$$

The rejection rule for the TOST, using a 5% significance level, is

$$\frac{|D| - \ln(1.25)}{cS} < -t_v(0.05)$$

Generate M parametric bootstrap samples (D_{i*}, S^2_{i*}) according to

$$D_{i*} \sim N(\ln(1.25), c^2S^2) \text{ and } v\frac{S^2_{i*}}{S^2} \sim \chi^2_v,$$

$i = 1, 2, ..., M$.

Compute M values of the test statistics $\frac{|D_{i*}|-\ln(1.25)}{cS_{i*}}$, $i = 1, 2, ..., M$.

Now numerically select α for which the proportion

$$\frac{|D_{i*}| - \ln(1.25)}{cS_{i*}} < -t_v(\alpha)$$

is close to 0.05. Let $\hat{\alpha}$ denote the value so selected.

Note: $\hat{\alpha}$ is actually estimated from the data.

The TOST is now carried out using the significance level $\hat{\alpha}$. The corresponding rejection rule is

$$\frac{|D| - \ln(1.25)}{cS} < -t_v(\hat{\alpha})$$

"Bootstrap calibrated TOST"

Type I error rates after bootstrap calibration: based on 10,000 simulated samples with $M = 200$ bootstrap samples ($\alpha = 0.05$).

BOOT*: Test based on bootstrap calibration

		$\sigma = 0.20$	$\sigma = 0.30$
$v = 10$	TOST	0.00238	0.00012
	BOOT*	0.04396	0.04714
$v = 22$	TOST	0.00025	0.00000
	BOOT*	0.04806	0.04890

The type I error probabilities of BOOT* are very close to the assumed 5% significance level.

It is also possible to estimate the required percentile by bootstrapping the distribution of $\frac{|D|-\ln(1.25)}{cS}$.

This gave type I error probabilities very close to those for BOOT*.

A four period crossover design

Let s = number of sequences

$s = 2$:

	Period			
Sequence	I	II	III	IV
1	T	T	R	R
2	R	R	T	T

$s = 4$:

	Period			
Sequence	I	II	III	IV
1	T	T	R	R
2	R	R	T	T
3	T	R	R	T
4	R	T	T	R

For the $s-$sequence crossover design having four periods, let n_i be the number of subjects assigned to the ith sequence.

We note that each subject receives two administrations of T and two administrations of R. We now have replication.

y_{ijTl}: the lth response corresponding to T for the jth subject in the ith sequence, $l = 1, 2; j = 1, 2, ..., n_i; i = 1, 2, ..., s, .$

Similarly define y_{ijRl}

The following model is often assumed for four period designs:

$$\begin{aligned} y_{ijTl} &= \mu_T + \gamma_{iTl} + \eta_{ijT} + e_{ijTl} \\ y_{ijRl} &= \mu_R + \gamma_{iRl} + \eta_{ijR} + e_{ijRl} \end{aligned}$$

γ_{iTl} and γ_{iRl} are fixed effects satisfying the identifiability conditions:

$$\sum_{i=1}^{s}\sum_{l=1}^{2}\gamma_{iTl} = 0 \text{ and } \sum_{i=1}^{s}\sum_{l=1}^{2}\gamma_{iRl} = 0$$

$(\eta_{ijT}, \eta_{ijR})'$, e_{ijTl} and e_{ijRl} have the same distributions as in the case of the 2×2 crossover design.

$$\text{Canonical form:} \quad D \sim N(\mu_T - \mu_R, c^2\sigma^2) \quad \text{and} \quad v\frac{S^2}{\sigma^2} \sim \chi^2_v$$

$$\text{where,} \quad c^2 = \frac{1}{s^2}\sum_{i=1}^{s}\frac{1}{n_i}, \ v = \sum_{i=1}^{s} n_i - s$$

$$\sigma^2 = \sigma^2_{BT} + \sigma^2_{BR} - 2\rho\sigma_{BT}\sigma_{BR} + \frac{1}{2}\left(\sigma^2_{WT} + \sigma^2_{WR}\right).$$

The TOST can be carried out using D and S^2

We can also derive estimates S^2_{WT} and S^2_{WR} of the within-subject variances σ^2_{WT} and σ^2_{WR}, respectively.

$$v\frac{S^2_{WT}}{\sigma^2_{WT}} \sim \chi^2_v, \quad \text{and} \quad v\frac{S^2_{WR}}{\sigma^2_{WR}} \sim \chi^2_v$$

An example

Example and data taken from Chow and Liu (2009, Table 12.3.3)

A bioequivalence study to compare a test and reference formulation of the anti-hypertensive drug Verapamil.

AUC data were obtained based on a 4 × 4 crossover design with 6 subjects each, for the sequences TRTR, RTRT, TRRT, and 5 subjects for the sequence RTTR.

Thus the total number of subjects is 23.

We have the canonical form

$$D \sim N(\mu_T - \mu_R, c^2\sigma^2) \text{ and } v\frac{S^2}{\sigma^2} \sim \chi^2_v$$

where

$$c^2 = \frac{1}{s^2}\sum_{i=1}^{s}\frac{1}{n_i} = 0.04375, \quad \text{and} \quad v = \sum_{i=1}^{s} n_i - s = 19$$

We also have the summary data: $D = -0.0196$, $cS = 0.2434$

To test the hypothesis of average bioequivalence:
H_0: $|\mu_T - \mu_R| \geq \ln(1.25)$, H_1: $|\mu_T - \mu_R| < \ln(1.25)$

Test statistic: $\frac{|D| - \ln(1.25)}{cS} = -0.8363$

$t_{19}(0.05) = 1.7291$

Since $\frac{|D| - \ln(1.25)}{cS} > -t_{19}(0.05)$, we cannot reject H_0.

Thus we cannot conclude average bioequivalence.

Applying a bootstrap calibration, we estimate the significance level to be used as $\hat{\alpha} = 0.35$ (based on 500 parametric bootstrap samples).

$t_{19}(\hat{\alpha}) = 0.3912$

Now $\frac{|D| - \ln(1.25)}{cS} < -t_{19}(0.35)$

Hence we reject H_0 and conclude average bioequivalence.

Due to the conservatism of the TOST, the conclusion maybe against average bioequivalence, when average bioequivalence holds.

The bootstrap calibration corrects this.

Other topics

- Comparing variances
- Highly variable drugs

Comparing variances

Suppose average bioequivalence holds. That is μ_T and μ_R are close.

If σ^2_{WT} and σ^2_{WR} are different, it is difficult to conclude that T and R are bioequivalent, even if average bioequivalence holds.

This is a major concern in the assessment of biosimilarity

Thus we need to consider the equivalence of the variances σ^2_{WT} and σ^2_{WR}.

We shall consider an $s \times 4$ crossover design, so that it is possible to estimate σ^2_{WT} and σ^2_{WR}.

For an $s \times 4$ crossover design, we recall that it is possible to compute sample variances S^2_{WT} and S^2_{WR} so that

$$v\frac{S^2_{WT}}{\sigma^2_{WT}} \sim \chi^2_v, \quad \text{and} \quad v\frac{S^2_{WR}}{\sigma^2_{WR}} \sim \chi^2_v$$

In order to test the equivalence of σ^2_{WT} and σ^2_{WR}, we can test if the ratio $\frac{\sigma^2_{WT}}{\sigma^2_{WR}}$ is around 1.

However, what we need is not the equivalence of σ^2_{WT} and σ^2_{WR}.

Rather, we want to make sure that the variance σ^2_{WT} for the test drug is not too large compared to the variance σ^2_{WR} for the reference drug.

The above requirement is verified by testing

$$H_0 : \sigma^2_{WT} - \sigma^2_{WR} \geq 0.02 \text{ vs } \ H_1 : \sigma^2_{WT} - \sigma^2_{WR} < 0.02.$$

If H_0 is rejected, we conclude that the within-subject variance σ^2_{WT} is not too large compared to the variance σ^2_{WR} for the reference drug.

Even if σ^2_{WT} is larger than σ^2_{WR}, we conclude that the difference is no more than 0.02.

The test procedure consists of computing an upper confidence limit for $\sigma^2_{WT} - \sigma^2_{WR}$, and rejecting H_0 when the upper confidence limit is less than 0.02.

To compute the required upper confidence limits for $\sigma^2_{WT} - \sigma^2_{WR}$, use the **modified large sample** (MLS) procedure.

Burdick, R. and Graybill, F. A. (1992). *Confidence Intervals for Variance Components*. Marcel–Dekker.

The MLS procedure

$$v\frac{S_{WT}^2}{\sigma_{WT}^2} \sim \chi_v^2, \quad \text{and} \quad v\frac{S_{WR}^2}{\sigma_{WR}^2} \sim \chi_v^2$$

Need an upper confidence limit for $\sigma_{WT}^2 - \sigma_{WR}^2$.

The usual $100(1-\alpha)\%$ upper confidence limit for σ_{WT}^2:

$$\sigma_{WT}^2 \leq v\frac{S_{WT}^2}{\chi_{v;\alpha}^2} = S_{WT}^2 + \sqrt{\left(v\frac{S_{WT}^2}{\chi_{v;\alpha}^2} - S_{WT}^2\right)^2}.$$

The usual $100(1-\alpha)\%$ lower confidence limit for σ_{WR}^2:

$$\sigma_{WR}^2 \geq S_{WR}^2 - \sqrt{\left(v\frac{S_{WR}^2}{\chi_{v;1-\alpha}^2} - S_{WR}^2\right)^2}.$$

According to the MLS procedure, an approximate $100(1-\alpha)\%$ upper confidence limit for $\sigma^2_{WT} - \sigma^2_{WR}$:

$$S^2_{WT} - S^2_{WR} + \sqrt{\left(v\frac{S^2_{WT}}{\chi^2_{v;\alpha}} - S^2_{WT}\right)^2 + \left(v\frac{S^2_{WR}}{\chi^2_{v;1-\alpha}} - S^2_{WR}\right)^2}.$$

The MLS method can be extended to obtain one-sided or two-sided confidence limits for an arbitrary linear combination of variances.

Known to be accurate for large samples.

Highly variable drugs

Highly variable drugs are defined as drugs with within-subject coefficient of variation being 30% or larger. The TOST has low power for such drugs.

A reference scaled criterion is used to assess average bioequivalence for highly variable drugs

Instead of the parameter $\mu = \mu_T - \mu_R$, consider $\frac{\mu}{\max(0.25^2, \sigma_R^2)}$

$$H_0 : \frac{\mu^2}{\max(0.25^2, \sigma_R^2)} \geq K \quad vs \quad H_1 : \frac{\mu^2}{\max(0.25^2, \sigma_R^2)} < K,$$

$$K = \left(\frac{ln(1.25)}{0.25} \right)^2.$$

σ_R^2: variance of the response to the reference drug.

Cannot be tested using a 2×2crossover design.

Will consider four-period crossover designs.

Equivalent to testing

$$H_0 : \mu^2 - K \times \max(0.25^2, \sigma_R^2) \geq 0,\ H_1 : \mu^2 - K \times \max(0.25^2, \sigma_R^2) < 0.$$

It is possible to compute a 95% upper confidence limit for $\mu^2 - K \times \max(0.25^2, \sigma_R^2)$ using the MLS idea.

Reject H_0 if the MLS upper confidence limit so obtained is less than zero.

The assessment of biosimilarity

For assessing bioequivalence in the case of chemical drugs, US FDA recommends crossover designs.

Usually, such designs are recommended for a product with a short half-life (e.g., shorter than 5 days).

For biotechnology drugs with a short half-life, FDA (2016) still recommends crossover designs for assessing biosimilarity.

For biological products with a longer half-life, parallel study designs are recommended.

If parallel designs are to be used, a two arm parallel design is a natural choice, since we are comparing two drugs, the original biotechnology drug R, and a copy T.

Here the subjects are randomized between the two arms with one arm corresponding to R, and the second arm corresponding to T.

The criterion suggested in the FDA (2016) document for the assessment of biosimilarity is the same criterion used for assessing average bioequivalence, namely the TOST.

FDA (2016). Guidance for Industry: *Clinical Pharmacology Data to Support a Demonstration of Biosimilarity to a Reference Product*. Available online.

A number of researchers have expressed serious concerns on the use of a criterion that compares only the averages, since biological products may exhibit higher variability.

Thus, at the very least, variabilities should also be compared.

However, there is no properly defined criterion for assessing biosimilarity by the regulatory agencies.

A proposal for testing biosimilarity

A natural criterion to consider is the amount of overlap between the response distributions corresponding to the test drug and the reference drug.

We can conclude bioequivalence if there is significant overlap, which can be assessed using a specified threshold.

The overlap coefficient (OVL) is a parameter that measures the amount of overlap between two probability distributions.

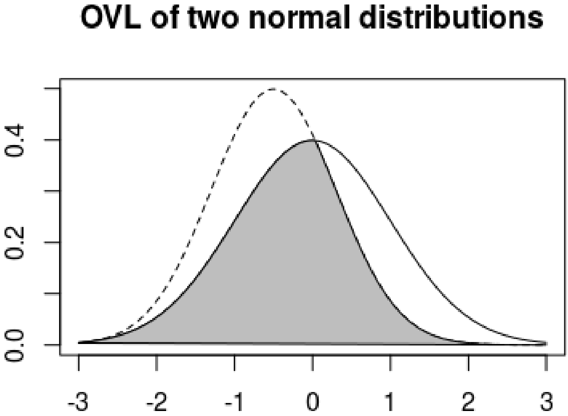
The OVL takes values in the range $[0, 1]$; OVL values close to one indicates that the two distributions are more similar.

Let $f_T(x)$ and $f_R(x)$ denote the response densities for the test drug and the reference drug, respectively. The OVL is defined as:

$$\mathrm{OVL} = \int_{-\infty}^{\infty} \min\left(f_T(x), f_R(x)\right) dx,$$

see Weitzman (1970).

The shaded area in the figure gives the OVL for two normal densities.



The OVL between $N(\mu_T, \sigma_T^2)$ and $N(\mu_R, \sigma_R^2)$

Let $\theta = \dfrac{\mu_T - \mu_R}{\sigma_R}$, $\lambda = \dfrac{\sigma_T}{\sigma_R}$.

$$\text{Define} \quad x_i = \frac{\theta \pm \sqrt{\lambda^2\theta^2 - 2\lambda^2(1-\lambda^2)\ln(\lambda)}}{1-\lambda^2}$$

(the two points at which the two densities intersect)

$$\begin{aligned}
\text{OVL} &= 1 + \Phi(x_1) - \Phi\left(\frac{x_1-\theta}{\lambda}\right) - \Phi\left(x_2\right) + \Phi\left(\frac{x_2-\theta}{\lambda}\right), \text{ for } \lambda > 1 \\
&= 1 - \Phi(x_1) + \Phi\left(\frac{x_1-\theta}{\lambda}\right) + \Phi\left(x_2\right) - \Phi\left(\frac{x_2-\theta}{\lambda}\right) \text{ for } \lambda < 1 \\
&= 2\Phi\left(\frac{-|\theta|}{2}\right), \text{ for } \lambda = 1.
\end{aligned}$$

The biosimilarity hypothesis can be formulated as:

$$H_0 : \mathsf{OVL} \leq \delta \quad vs \quad H_1 : \mathsf{OVL} > \delta, \tag{1}$$

where δ is a pre-specified threshold.

We conclude biosimilarity if H_0 is rejected.

Can be tested based on data from crossover designs or parallel designs.

Approaches that can be used: Bootstrap methods and fiducial inference

Complex generic forms

Complex due to products with complex active ingredients, formulations, routes of delivery or dosage forms, complex drug-device combinations, ...

Examples:

Locally acting drugs such as dermatological products and complex ophthalmological products

Dry powder inhalers, nasal sprays

FDA will provide product specific guidance for the approval of complex generic forms.

"FDA will strive to issue guidance for a complex product as soon as scientific recommendations are available."

Other equivalence testing problems

Rose, E. M., Mathew, T., Coss, D. A., Lohr, B. and Omland, K. E. (2018). A new statistical method to test equivalence: an application in male and female eastern bluebird song. *Animal Behavior*, 145, 77-85.

Multivariate bioequivalence testing based on five-dimensional data: five standard features of acoustic structure of the bird songs.

Dolado, J., Carmen, M. & Mark, O. (2014). Equivalence hypothesis testing in experimental software engineering. *Software Quality Journal*, 22, 215-238.

Used the TOST to re-examine the behavior of experts and novices when handling code with side effects, compared to side-effect free code.

Garber Jr., L. L., Boya, U. O &Hyatt, E. M. (2018). Hypotheses of equivalence and their testing. *Journal of Marketing Theory and Practice*, 26, 280-288.

Equivalence testing for functional data:

Fogarty, C. B. and Small, D. S. (2014). Equivalence testing for functional data with an application to comparing pulmonary function devices. *Annals of Applied Statistics*, 8, 2002-2026.

Comparison of two devices for pulmonary function testing.

Data used consists of the flow of air into and out of the lungs versus volume of air within the lungs over time, for each breath.

The authors consider testing the equivalence of means, and testing the equivalence of variances.

The necessary thresholds are defined as a function of time.

Bootstrap and Bayesian approaches.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence*, Second Edition, Chapman and Hall/CRC Press.

A reference that provides the US FDA perspective:

Yu, L. X. and Li, B. V. (2014), Editors. *FDA Bioequivalence Standards*. Springer.