Statistical Tests of Agreement Based on Non-Standard Data

Elizabeth Stanwyck
Bimal Sinha
Department of Mathematics and Statistics
University of Maryland, Baltimore County
Barry Nussbaum
Office of Environmental Information
U.S. Environmental Protection Agency

Proving equivalence is increasingly important

- Testing is expensive & time consuming
- Newer methods and procedures are being developed
- Common goal: *assess agreement between two methods of measurement*

Applications to EPA problems

- Demonstrating equivalence between primary and secondary methods for measuring formaldehyde emissions from composite wood products
  - Large chamber test is expensive (single measurement)
  - Small chamber test is easier and less costly (multiple measurements)

- Prediction of Dioxin-Furan Congener (TEQ) toxicity in fresh-water fish based on fatty acid methyl ester (FAME) profiles
  - Equivalence between KVL and NERL labs for FAME
  - Equivalence between KVL & ECL labs for TEQ

Common methods for assessing agreement

- Hypothesis testing of the correlation coefficient
- Regression analysis
- Paired t-tests
- Least-squares analysis for intercept and slope
- Within-subject coefficient of variation

Mean, variance, covariance approach

- Some current tests are based only on the mean and standard deviation of the differences:

$$d_i = x_i - \bar{y}_i, i = 1, \ldots, n$$

- *Does not guarantee equivalence!!*

$$[(10, 22), (15, 12), (18, 10), (25, 17), (17, 25), (22, 18), (12, 15)]$$

$$\bar{d} = 0; s_x^2 = s_y^2 = 28; r_{xy} = -0.1012$$

- *Even high correlation, by itself, does not guarantee agreement!*

$$[(10, 15), (15, 25), (18, 25), (20, 26), (25, 30), (30, 36)]$$

$$r_{xy} = 0.965; \bar{d} = -6.5; s_x^2 = 50.67, s_y = 47.77$$

Assessing agreement

- Likelihood ratio test for combined hypothesis:

$$H_0 : \mu_x = \mu_y, \sigma_x = \sigma_y, \rho \geq \rho_0$$

(Yimprayoon et al., 2006)

- Interval hypothesis test

$$H_0 : |\mu_x - \mu_y| < \delta_1, \delta_2 < |\frac{\sigma_x}{\sigma_y}| < \delta_3, \rho \geq \rho_0$$

  - Extremely difficult and complicated test

- **Equivalence is not the same as equality!**

Nonstandard data problem

- Inference usually based on paired data *X* and *Y* (bivariate normal assumption)
  - Yinprayoon, Tiensuwan, and Sinha, 2006

- Generalize the LRT approach for **nonstandard** data

$$\left[\left(x_i, y_{i1}, \ldots, y_{i,m_i}\right), i = 1, \ldots, n\right]$$

  - Balanced case: $m_1 = \ldots = m_n = m$
  - Unbalanced case: $m_1 \neq \ldots \neq m_n$

Restricted dataset

$$[(x_i, \bar{y}_i), i = 1, \ldots, n]$$

- Likelihood function is based on marginal likelihood of *X* and conditional likelihood of *Y*

$$x_i \sim N\left[\mu_x, \sigma_x^2\right]$$

$$\bar{y}_i | x_i \sim N\left[\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x_i - \mu_x), \frac{\sigma_y^2(1-\rho^2)}{m_i}\right]$$

Likelihood function

$$L(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho | data) \sim (\sigma_x \sigma_y)^{-n} (1-\rho^2)^{-n/2} \times$$

$$exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu_x)^2}{\sigma_x^2} - \frac{1}{2\sigma_y^2(1-\rho^2)}\sum_{i=1}^{n} m_i(\bar{y}_i - \mu_y - \rho\frac{\sigma_y}{\sigma_x}(x_i-\mu_x))^2\right]$$

$$A = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad C = \sum_{i=1}^{n} m_i(x_i - \bar{\bar{x}})^2$$

$$D = \sum_{i=1}^{n} m_i(\bar{y}_i - \bar{\bar{y}})^2, \qquad E = \sum_{i=1}^{n} m_i(x_i - \bar{\bar{x}})(\bar{y}_i - \bar{\bar{y}})$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad \bar{\bar{y}} = \frac{\sum m_i \bar{y}_i}{M}, \qquad \bar{\bar{x}} = \frac{\sum m_i x_i}{M}, \qquad M = \sum m_i$$

Unrestricted maximization

- Maximum likelihood estimates

$$\hat{\mu}_x = \bar{x}, \qquad \hat{\mu}_y = \bar{\bar{y}} + \frac{E}{C}(\bar{x} - \bar{\bar{x}})$$

$$\hat{\sigma}_x^2 = \frac{A}{n}, \quad \hat{\sigma}_y^2 = \frac{1}{n}[D + M\frac{AE^2}{nC^2} - \frac{E^2}{C}], \quad \hat{\rho}^2 = \frac{E^2\hat{\sigma}_x^2}{C^2\hat{\sigma}_y^2}$$

- Maximized likelihood

$$\left[\frac{C}{A(DC - E^2)}\right]^{n/2}$$

Restricted maximization

- Maximum likelihood estimates

$$\hat{\mu}_\rho = \frac{n\bar{x}(1+\rho) + M(\bar{\bar{y}} - \rho\bar{\bar{x}})}{M(1-\rho) + n(1+\rho)}$$

$$2n\hat{\sigma}^2_\rho = Q_1(\rho) = A + \frac{D + C\rho^2 - 2E\rho}{1-\rho^2} + \frac{nM\left[\bar{\bar{y}} - \bar{x} + \rho\left(\bar{x} - \bar{\bar{x}}\right)\right]^2}{(1-\rho)\left[M\left(1-\rho\right) + n\left(1-\rho\right)\right]}$$

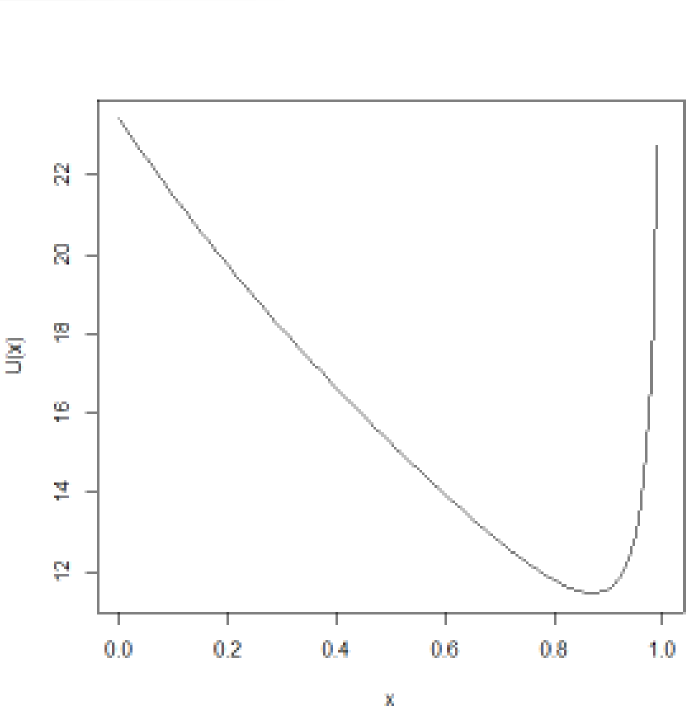
- Likelihood function, maximized wrt μ and σ^2

$$L_1\left(\rho \middle| \, \text{data}\right) \sim \left[\left(1-\rho^2\right)^{\frac{1}{2}} \times Q_1\left(\rho\right)\right]^{-n}$$

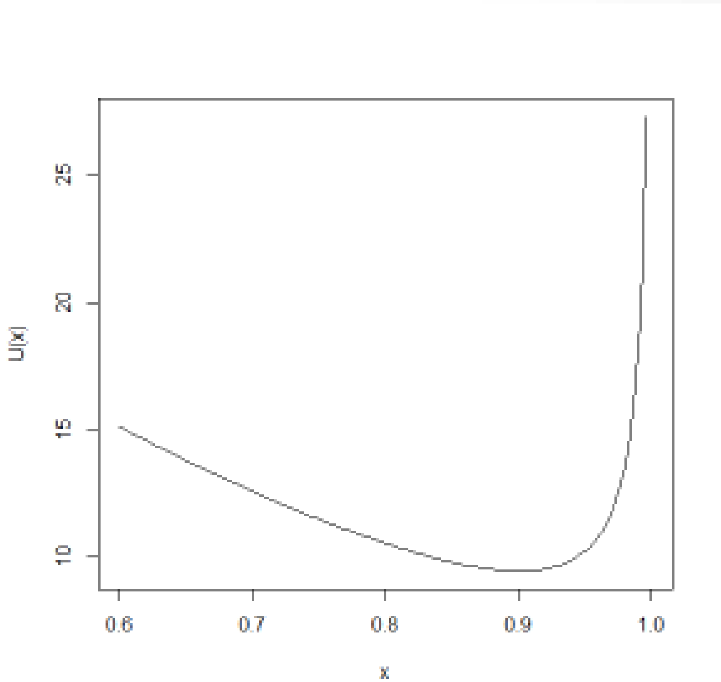
- To maximize the likelihood, minimize wrt ρ

$$U_1\left(\rho\right) = \left[\left(1-\rho^2\right)^{\frac{1}{2}} \times Q_1\left(\rho\right)\right]$$

Images of U_1



$\rho = 0.9, \rho_0 = 0.9, n = 15, m = 1$



$\rho = 0.9, \rho_0 = 0.9, n = 15, m = 3$

Likelihood ratio test statistic

- Test statistic

$$\lambda = \frac{\sup_{H_0} L(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy} | \text{ data })}{\sup_{\text{unrestricted}} L(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy} | \text{data})}$$

- Reject H0 for large values of T_1

$$T_1 = \left[\min_{\rho \geq \rho_0} U_1(\rho)\right] \times \left[\frac{C}{A(DC - E^2)}\right]^{\frac{1}{2}}$$

- Select cutoff d_1 so that

$$\alpha = P\left[T_1 > d_1 | H_0 : \mu_x = \mu_y, \sigma_x = \sigma_y, \rho = \rho_0\right]$$

Remarks

- T_1 is location and scale invariant
- Composite null hypothesis: determine the cutoff value d1 under $\rho = \rho_0$ and verify size is less than or equal to alpha for $\rho > \rho_0$
- Simulations: different correlation, means, variances, and combinations thereof to get an idea of power

Unrestricted dataset

$$\left[x_i, \left(y_{i1}, \ldots, y_{im_i}\right), i = 1, \ldots, n\right]$$

- Likelihood function:

$$L(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho | data) \sim (\sigma_x)^{-n}[\sigma_y^2(1-\rho^2)]^{-M/2} \times$$

$$exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu_x)^2}{\sigma_x^2} - \frac{1}{2\sigma_y^2(1-\rho^2)}\{\sum_{i=1}^{n} m_i(\bar{y}_i - \mu_y - \rho\frac{\sigma_y}{\sigma_x}(x_i-\mu_x))^2 + W_y\}\right]$$

$$W_y = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2 : \text{ additional term}$$

Unrestricted maximization

- Maximum likelihood estimates

$$\hat{\mu}_x = \bar{x}, \quad \hat{\mu}_y = \bar{\bar{y}} + \frac{E}{C}(\bar{x} - \bar{\bar{x}}), \quad \hat{\sigma}_x^2 = \frac{A}{n}$$

$$\hat{\sigma}_y^2 = \frac{1}{M}[W_y + D + \frac{MAE^2}{nC^2} - \frac{E^2}{C}], \quad \hat{\rho} = \frac{E\hat{\sigma}_x}{C\hat{\sigma}_y}$$

- Maximized likelihood

$$\frac{1}{A^{\frac{n}{2}} \times [D - \frac{E^2}{C} + W_y]^{\frac{M}{2}}}$$

Restricted maximization

- Maximum likelihood estimates

$$\hat{\mu}_\rho = \frac{n\bar{x}(1+\rho) + M(\bar{\bar{y}} - \rho\bar{\bar{x}})}{M(1-\rho) + n(1+\rho)} \qquad \hat{\sigma}^2_\rho = \frac{1}{n+M} Q_2(\rho)$$

$$Q_2(\rho) = A + \frac{D + C\rho^2 - 2E\rho + W_y}{1-\rho^2} + \frac{nM[\bar{\bar{y}} - \bar{x} + \rho(\bar{x} - \bar{\bar{x}})]^2}{(1-\rho)[M(1-\rho) + n(1+\rho)]}$$

- Likelihood maximized wrt μ and σ²

$$L_2(\rho|\,\mathrm{data}) \sim \left[\left(1-\rho^2\right)^{\frac{M}{2}} \times Q_2(\rho)^{\frac{n+M}{2}}\right]^{-1}$$

- To maximize likelihood, minimize

$$U_2(\rho) = \left[\left(1-\rho^2\right) \times Q_2(\rho)^{1+\frac{n}{M}}\right]$$

Likelihood ratio test statistic

- Test statistic

$$\lambda = \frac{\sup_{H_0} L(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy} | \text{ data })}{\sup_{\text{unrestricted}} L(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy} | \text{data})}$$

- Reject H0 for large values of T_2

$$T_2 = \frac{1}{A} \times \left[\frac{\min_{\rho \geq \rho_0} U_2(\rho)}{D - \frac{E^2}{C} + W_y} \right]^{\frac{M}{n}}$$

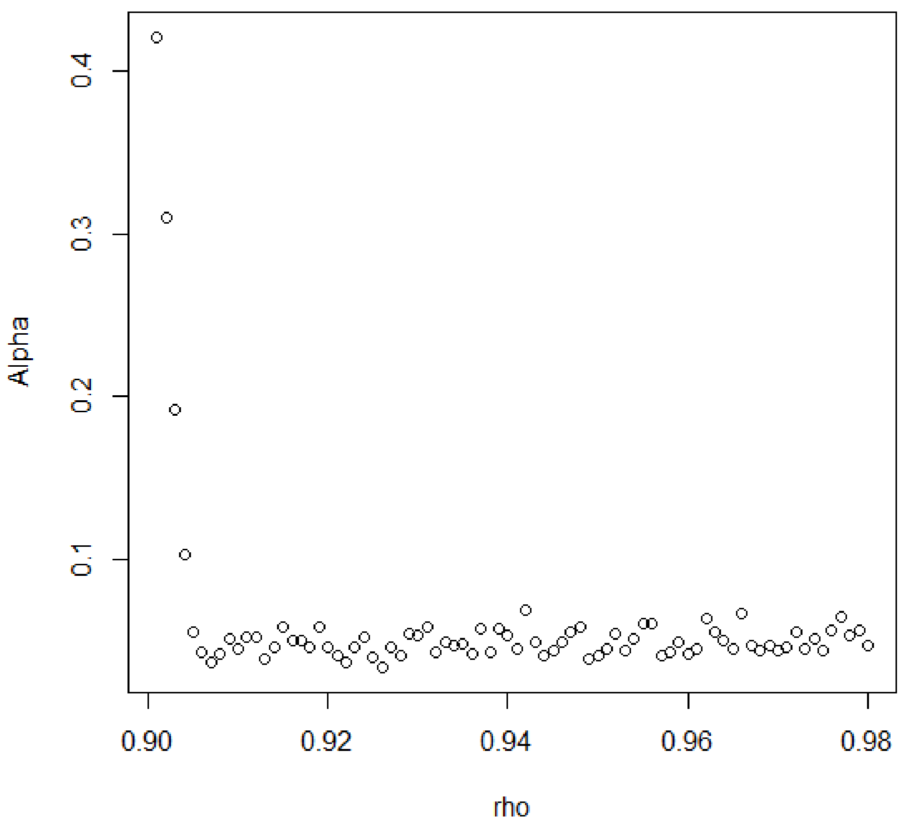
- Select cutoff d_2 so that

$$\alpha = P\left[T_2 > d_2 | H_0 : \mu_x = \mu_y, \sigma_x = \sigma_y, \rho = \rho_0\right]$$

Restricted dataset Simulations: Type I Error rates

ρ	ρ_0	n	m	α
0.92	0.9	5	1	0.0439
0.92	0.9	10	1	0.0396
0.92	0.9	15	1	0.0371
0.92	0.9	5	3	0.0452
0.92	0.9	10	3	0.0409
0.92	0.9	15	3	0.0335
0.95	0.9	5	1	0.033
0.95	0.9	10	1	0.0299
0.95	0.9	15	1	0.0274
0.95	0.9	5	3	0.0374
0.95	0.9	10	3	0.0305
0.95	0.9	15	3	0.0237
0.99	0.9	5	1	0.0299
0.99	0.9	10	1	0.0254
0.99	0.9	15	1	0.0253
0.99	0.9	5	3	0.0309
0.99	0.9	10	3	0.0277
0.99	0.9	15	3	0.0266

Type I Error rates



Simulations: Power

ρ	ρ_0	μ_y	σ_y^2	n	m	α	$1-\beta$
0.5	0.9	0	1	5	1	0.05	0.2458
0.5	0.9	0	1	10	1	0.05	0.642
0.5	0.9	0	1	15	1	0.05	0.8527
0.5	0.9	0	1	5	3	0.05	0.4265
0.5	0.9	0	1	10	3	0.05	0.8875
0.5	0.9	0	1	15	3	0.05	0.9723

ρ	ρ_0	μ_y	σ_y^2	n	m	α	$1-\beta$
0.9	0.9	1	1	5	1	0.05	0.8815
0.9	0.9	1	1	10	1	0.05	0.9999
0.9	0.9	1	1	15	1	0.05	1
0.9	0.9	1	1	5	3	0.05	0.9996
0.9	0.9	1	1	10	3	0.05	1
0.9	0.9	1	1	15	3	0.05	1

Simulations: Power

ρ	ρ_0	μ_y	σ_y^2	n	m	α	$1-\beta$
0.9	0.9	0	4	5	1	0.05	0.5481
0.9	0.9	0	4	10	1	0.05	0.961
0.9	0.9	0	4	15	1	0.05	0.9984
0.9	0.9	0	4	5	3	0.05	0.9096
0.9	0.9	0	4	10	3	0.05	0.9996
0.9	0.9	0	4	15	3	0.05	1

ρ	ρ_0	μ_y	σ_y^2	n	m	α	$1-\beta$
0.9	0.9	1	4	5	1	0.05	0.8197
0.9	0.9	1	4	10	1	0.05	0.9976
0.9	0.9	1	4	15	1	0.05	1
0.9	0.9	1	4	5	3	0.05	0.9885
0.9	0.9	1	4	10	3	0.05	1
0.9	0.9	1	4	15	3	0.05	1

Simulations: Power

ρ	ρ_0	μ_y	σ_y^2	n	m	α	$1-\beta$
0.5	0.9	1	1	5	1	0.05	0.6795
0.5	0.9	1	1	10	1	0.05	0.9836
0.5	0.9	1	1	15	1	0.05	0.9988
0.5	0.9	1	1	5	3	0.05	0.9515
0.5	0.9	1	1	10	3	0.05	1
0.5	0.9	1	1	15	3	0.05	1

ρ	ρ_0	μ_y	σ_y^2	n	m	α	$1-\beta$
0.5	0.9	0	4	5	1	0.05	0.5043
0.5	0.9	0	4	10	1	0.05	0.9442
0.5	0.9	0	4	15	1	0.05	0.9955
0.5	0.9	0	4	5	3	0.05	0.5077
0.5	0.9	0	4	10	3	0.05	0.9486
0.5	0.9	0	4	15	3	0.05	0.9888

Simulations

ρ	ρ_0	μ_y	σ_y^2	n	m	α	$1-\beta$
0.5	0.9	1	4	5	1	0.05	0.6653
0.5	0.9	1	4	10	1	0.05	0.9862
0.5	0.9	1	4	15	1	0.05	0.9995
0.5	0.9	1	4	5	3	0.05	0.8536
0.5	0.9	1	4	10	3	0.05	0.9978
0.5	0.9	1	4	15	3	0.05	0.9998

- Test is most powerful when means are different
- Least powerful when only variances are different

Tests based on combinations of P-values

- Consider the composite hypothesis test

$$H_{01} : \mu_x = \mu_y; H_{02} : \sigma_x^2 = \sigma_y^2; H_{03} : \rho \geq \rho_0$$

$$\text{versus}$$

$$H_{11} : \mu_x \neq \mu_y; H_{12} : \sigma_x^2 \neq \sigma_y^2; H_{13} : \rho < \rho_0$$

- We consider three separate tests for H_{01}, H_{02}, and H_{03}, and combine the resulting P-values to derive an overall test.

Testing H01

- Paired t-test:

$$x_i - \bar{y}_i = d_i \sim N\left[\mu_x - \mu_y, (\sigma_x - \rho\sigma_y)^2 + \frac{\sigma_y^2\left(1-\rho^2\right)}{m_i}\right]$$

  - Assumption: $m_1 = \cdots = m_n = m$

- Reject the null for large values of $|t_d| = \left|\frac{\sqrt{n}\bar{d}}{s_d}\right|$

$$d_i = x_i - \bar{y}_i, \bar{d} = \frac{\sum_{i=1}^n d_i}{n}, s_d^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}$$

- P-value $p_1 = Pr\left(|t_{n-1}| > |t_d|\right)$

Testing H02

- Modified Pittman-Morgan

$$u_i = x_i + \bar{y}_i \left(\frac{m_i}{1 + (m_i - 1)\rho_0^2} \right)^{\frac{1}{2}}, \qquad v_i = x_i - \bar{y}_i \left(\frac{m_i}{1 + (m_i - 1)\rho_0^2} \right)^{\frac{1}{2}}$$

$$H_{02} \equiv H_{02}^* : \rho_{uv} = 0 \qquad t_{uv} = \frac{r_{uv}(n-2)^{\frac{1}{2}}}{(1 - r_{uv}^2)^{\frac{1}{2}}}$$

- P-value $p_2 = Pr\left(|t_{n-2}| > |t_{uv}|\right)$

Testing H03

- assume $m_1 = \cdots = m_n = m$

$$\rho_{x\bar{y}} = \left(\frac{m\rho^2}{1 + (m-1)\rho^2} \right) = \rho^*$$

$$z^* = \frac{1}{2} \ln \frac{1 + r^*}{1 - r^*}; \zeta^* = \frac{1}{2} \ln \frac{1 + \rho_0^*}{1 - \rho_0^*} \text{ with } \rho_0^* = \left(\frac{m\rho_0^2}{1 + (m-1)\rho_0^2} \right)$$

- P-value $p_3 = Pr\left(N(0,1) < z^*(n-3)^{\frac{1}{2}} \right)$

Tests based on P-values

1. Tippett's test:

$$\text{Reject } H_0 \text{ when } min(p_1, p_2, p_3) < c_1$$

2. Fisher's test:

$$\text{Reject } H_0 \text{ when } \; -2\left[\ln p_1 + \ln p_2 + \ln p_3\right] > c_2$$

3. Stouffer's test:

$$\text{Reject } H_0 \text{ when } \; \left[\Phi^{-1}(p_1) + \Phi^{-1}(p_2) + \Phi^{-1}(p_3)\right] < c_3$$

Tests based on P-values

Simulations: Type I Error rates

ρ	ρ_0	n	m	Tippett	Fisher	Stouffer
0.92	0.9	5	1	0.0498	0.0481	0.0358
0.92	0.9	10	1	0.0468	0.0439	0.0327
0.92	0.9	15	1	0.0409	0.0343	0.0248
0.92	0.9	5	3	0.0484	0.0448	0.0349
0.92	0.9	10	3	0.0416	0.0354	0.0271
0.92	0.9	15	3	0.0412	0.0402	0.0271
0.95	0.9	5	1	0.0457	0.0402	0.0183
0.95	0.9	10	1	0.0388	0.0314	0.0092
0.95	0.9	15	1	0.039	0.025	0.0053
0.95	0.9	5	3	0.0474	0.0442	0.0172
0.95	0.9	10	3	0.0473	0.0421	0.0116
0.95	0.9	15	3	0.0551	0.0427	0.0088
0.99	0.9	5	1	0.0399	0.0309	0.0007
0.99	0.9	10	1	0.0386	0.0262	0
0.99	0.9	15	1	0.0388	0.023	0
0.99	0.9	5	3	0.1112	0.1067	0.0018
0.99	0.9	10	3	0.3148	0.2344	0.0001
0.99	0.9	15	3	0.5378	0.4211	0

Simulations: Power

ρ	ρ_0	μ_y	σ_y^2	n	m	Tippett	Fisher	Stouffer
0.5	0.9	0	1	5	1	0.2151	0.2762	0.3224
0.5	0.9	0	1	10	1	0.6453	0.6981	0.5593
0.5	0.9	0	1	15	1	0.8661	0.8714	0.6836
0.5	0.9	0	1	5	3	0.2984	0.3835	0.4372
0.5	0.9	0	1	10	3	0.8323	0.8956	0.7832
0.5	0.9	0	1	15	3	0.9764	0.9898	0.9391

ρ	ρ_0	μ_y	σ_y^2	n	m	Tippett	Fisher	Stouffer
0.9	0.9	1	1	5	1	0.8507	0.8843	0.6941
0.9	0.9	1	1	10	1	0.9998	0.9998	0.9243
0.9	0.9	1	1	15	1	1	1	0.9796
0.9	0.9	1	1	5	3	0.9981	0.9984	0.8461
0.9	0.9	1	1	10	3	1	1	0.9781
0.9	0.9	1	1	15	3	1	1	0.9987

Simulations: Power

ρ	ρ_0	μ_y	σ_y^2	n	m	Tippett	Fisher	Stouffer
0.9	0.9	0	4	5	1	0.403	0.4249	0.3596
0.9	0.9	0	4	10	1	0.9154	0.9615	0.754
0.9	0.9	0	4	15	1	0.994	0.9984	0.9189
0.9	0.9	0	4	5	3	0.6942	0.7543	0.5457
0.9	0.9	0	4	10	3	0.9971	0.9994	0.916
0.9	0.9	0	4	15	3	1	1	0.9925

ρ	ρ_0	μ_y	σ_y^2	n	m	Tippett	Fisher	Stouffer
0.9	0.9	1	4	5	1	0.5252	0.7668	0.7903
0.9	0.9	1	4	10	1	0.9759	0.9979	0.9904
0.9	0.9	1	4	15	1	0.9991	0.9999	0.9993
0.9	0.9	1	4	5	3	0.823	0.9734	0.9505
0.9	0.9	1	4	10	3	1	1	0.9994
0.9	0.9	1	4	15	3	1	1	1

Simulations: Power

ρ	ρ_0	μ_y	σ_y^2	n	m	Tippett	Fisher	Stouffer
0.5	0.9	1	1	5	1	0.4099	0.6822	0.7232
0.5	0.9	1	1	10	1	0.9163	0.9835	0.9622
0.5	0.9	1	1	15	1	0.9957	0.9997	0.9963
0.5	0.9	1	1	5	3	0.6486	0.9415	0.9381
0.5	0.9	1	1	10	3	0.995	0.9999	0.9993
0.5	0.9	1	1	15	3	1	1	1

ρ	ρ_0	μ_y	σ_y^2	n	m	Tippett	Fisher	Stouffer
0.5	0.9	0	4	5	1	0.3051	0.5042	0.5782
0.5	0.9	0	4	10	1	0.8448	0.9602	0.9209
0.5	0.9	0	4	15	1	0.9982	0.9969	0.9846
0.5	0.9	0	4	5	3	0.3223	0.458	0.5203
0.5	0.9	0	4	10	3	0.8789	0.9489	0.8779
0.5	0.9	0	4	15	3	0.9886	0.9962	0.9783

Simulations: Power

ρ	ρ_0	μ_y	σ_y^2	n	m	Tippett	Fisher	Stouffer
0.5	0.9	1	4	5	1	0.3575	0.6788	0.7638
0.5	0.9	1	4	10	1	0.8987	0.9887	0.979
0.5	0.9	1	4	15	1	0.9929	0.9995	0.9984
0.5	0.9	1	4	5	3	0.5109	0.852	0.8831
0.5	0.9	1	4	10	3	0.9796	0.9987	0.9964
0.5	0.9	1	4	15	3	0.9999	1	0.9998

- Stouffer’s test has the lowest Type I Error rates (of all tests, including LRT)

- LRT and Fisher’s tests have similar power
  - Fisher’s test has the highest power of the combined P-value tests in almost every case
  - Stouffer’s has a higher power in some small sample size (n=5) cases

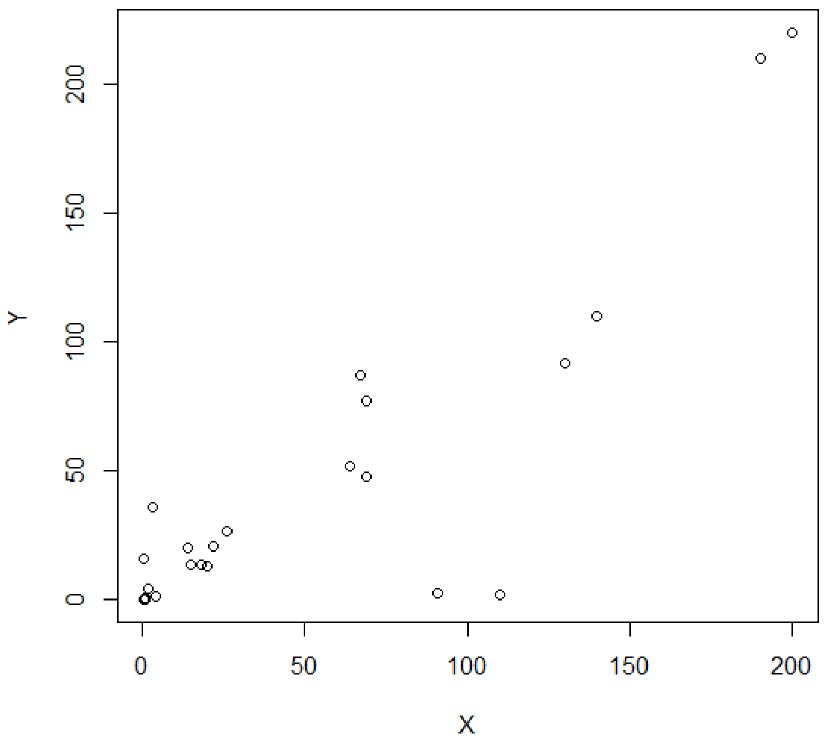
Applications

- Application to EPA data: measuring concentrations of pollutants in groundwater
  - Conventional purging methods i.e. low-flow sampling methods
    - A pump slowly collects groundwater so that the sample is not contaminated by water at different levels
  - New HydraSleeve method
    - A tube is lowered into the well and left there long enough for sediment etc. to settle, then water is collected as the tube is pulled upwards
- Focus: specific pollutants

Results

- TCE

$$H_0 : \mu_x = \mu_y, \sigma_x = \sigma_y, \rho \geq 0.9$$



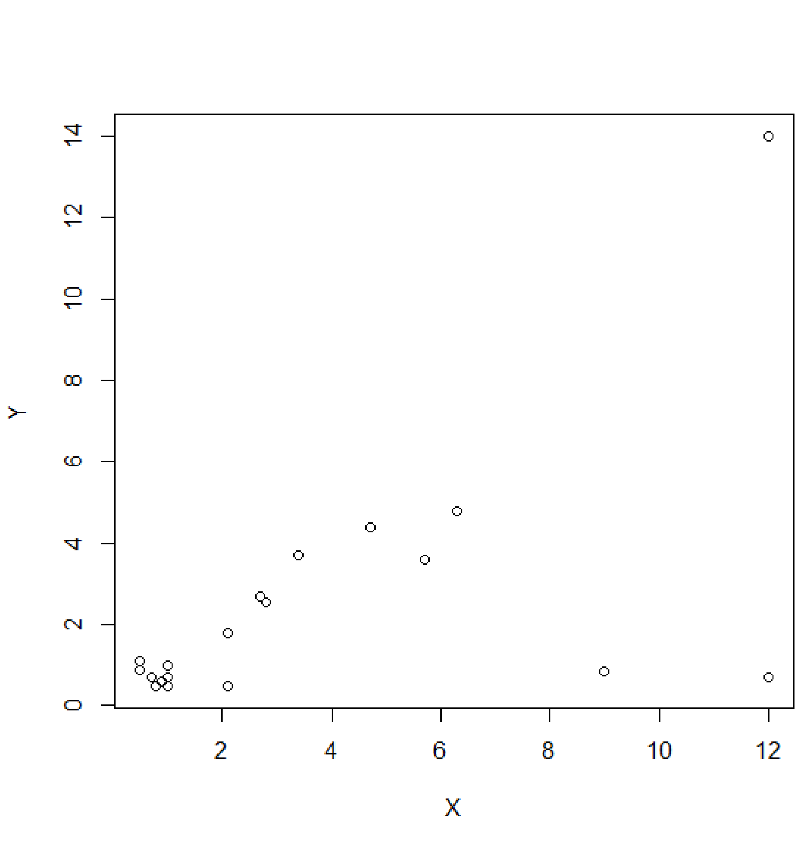
Test	Cutoff	Test Statistic	Conclusion
LRT	2.37547	2.206056	Do not reject
Tippett	0.01803122	0.2217555	Do not reject
Fisher	11.74769	5.849823	Do not reject
Souffer	-2.473122	0.4399887	Do not reject

$$n = 23$$

Results

- DCA

$$H_0 : \mu_x = \mu_y, \sigma_x = \sigma_y, \rho \geq 0.9$$



Test	Cutoff	Test Statistic	Conclusion
LRT	2.462177	3.641468	Reject
Tippett	0.01858661	0.0007817254	Reject
Fisher	11.65932	20.72726	Reject
Souffer	-2.418705	-4.703667	Reject

$$n = 19$$

Strong resemblance to bioequivalence testing

- In an equivalence trial, the aim is to show that two treatments are not too different in characteristics
- **Not too different** is defined in a clinical manner
- Called **bioequivalence testing**
- Nature of the data for bioequivalence testing
  - Same patients
  - Washout period
  - Crossover designs

Bioequivalence testing

- Often data are collected from healthy volunteers

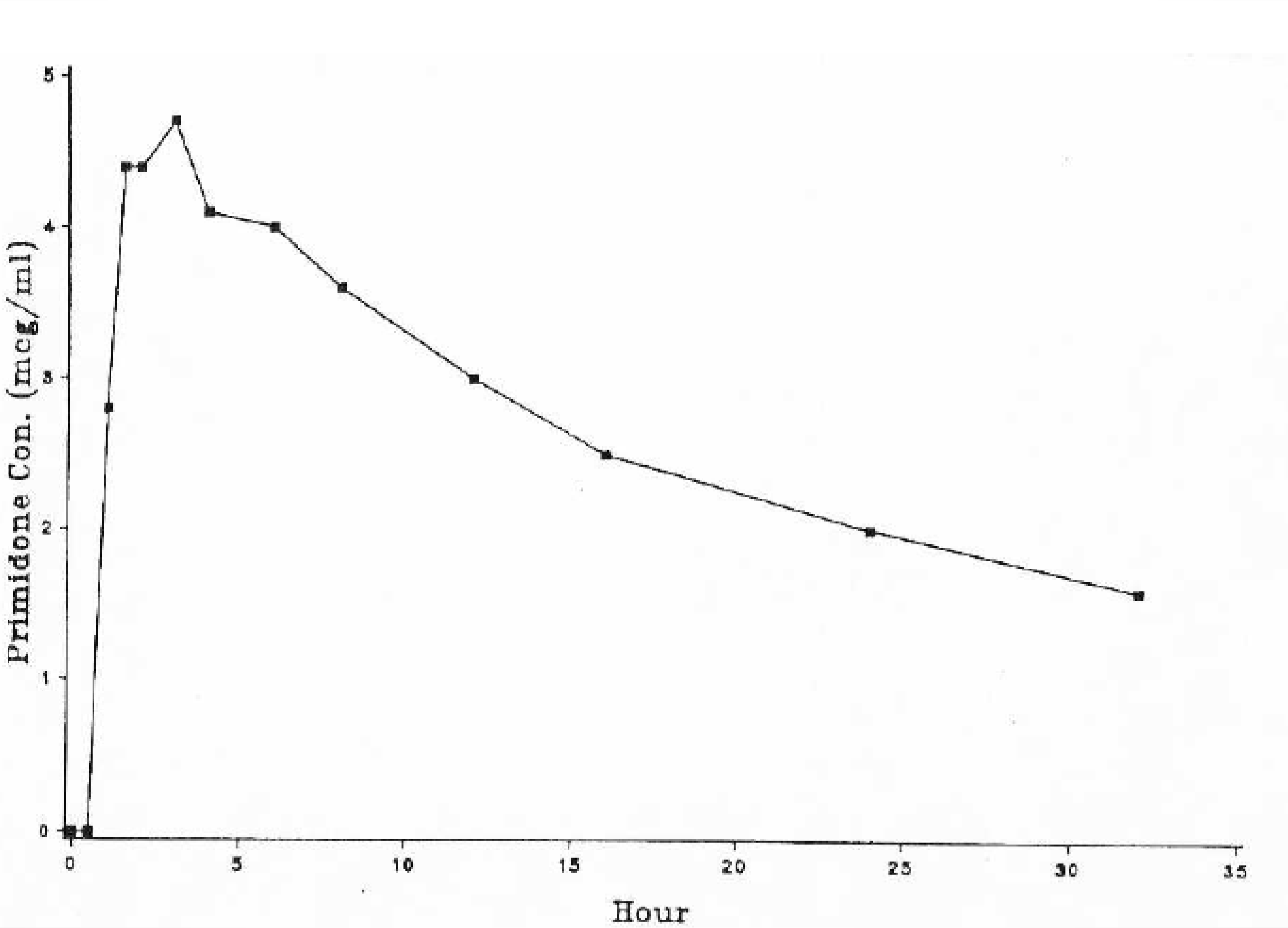
- If two drug products perform the same in healthy volunteers, the assumption is made that they will perform the same in patients with the disease

- Data obtained on three patient characteristics
  - Area under the curve (AUC)
  - Maximum blood concentration C_{max}
  - Time to reach the maximum concentration T_{max}

Bioequivalence testing

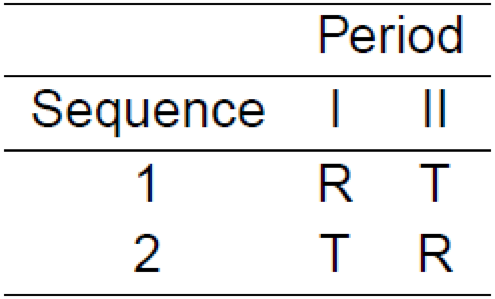
- Two drug products are *bioequivalent* if they have similar rate and extent of absorption into the blood.
- Two drug products are *therapeutically equivalent* if they provide similar therapeutic effects.
- **Fundamental bioequivalence assumption**: If two drug products are bioequivalent, they are also therapeutically equivalent

Data for bioequivalence testing



Experimental designs

- Reference drug (R)
- Test drug (T)
- Each subject receives both R and T, separated by a washout period
- Crossover designs are used
- A two sequence–two period crossover design:



	Period	
Sequence	I	II
1	R	T
2	T	R

Average bioequivalence

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

- The response is usually AUC, after log-transformation (could be C_{max} or T_{max}).

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

Average bioequivalence

- μ_T and μ_R are considered equivalent if $|\mu_T - \mu_R| < \ln(1.25)$.
- Hypothesis 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 H0 is rejected after a statistical test based on the log-transformed AUC data.

A canonical form

- Under an appropriate model for the log-transformed data, a canonical form is

$$D \sim N\left(\mu_T - \mu_R, c^2\sigma^2\right) \qquad \nu\frac{S^2}{\sigma^2} \sim \chi^2_\nu$$

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

- Rewrite as

$$H_{01} : \mu_T - \mu_R \leq -\ln(1.25) \text{ vs. } H_{11} : \mu_T - \mu_R > -\ln(1.25)$$

$$H_{02} : \mu_T - \mu_R \geq \ln(1.25) \text{ vs. } H_{12} : \mu_T - \mu_R < \ln(1.25)$$

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

Assessing bioequivalence

- Carry out t-tests: conclude average bioequivalence at significance level a if

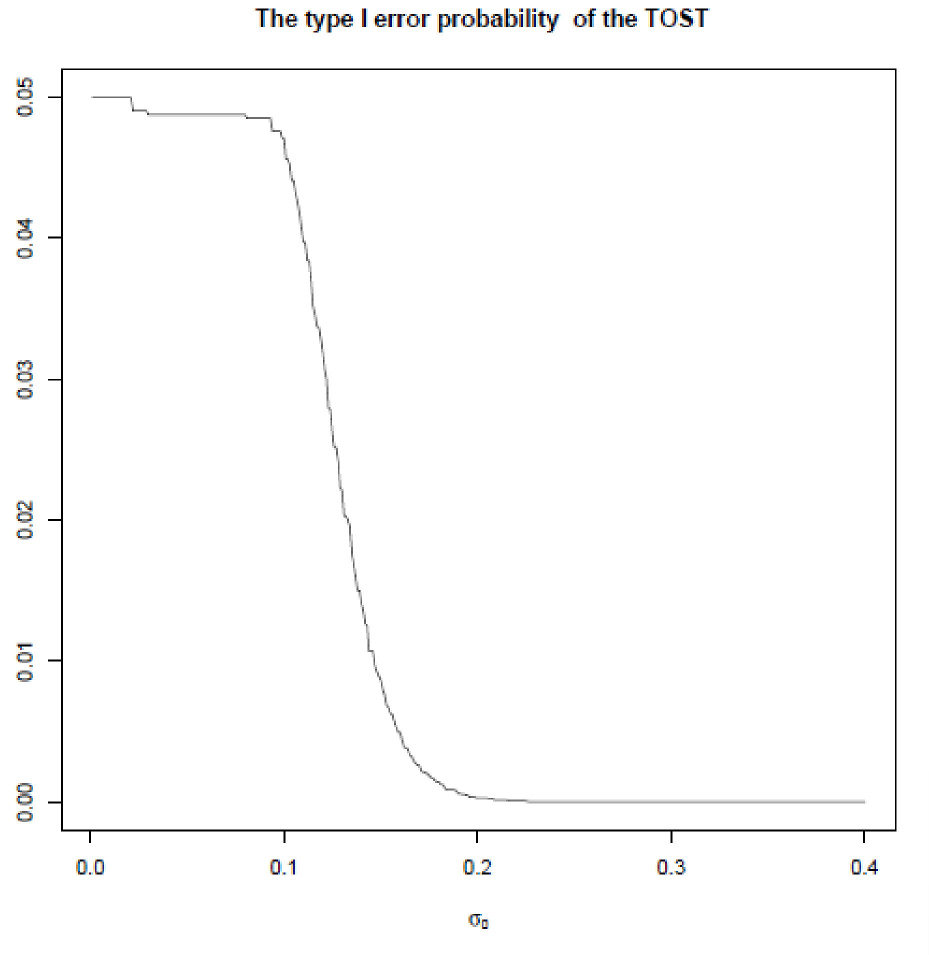
$$\frac{D+\ln(1.25)}{cS} > t_{\nu}(\alpha) \text{ and } \frac{D-\ln(1.25)}{cS} < -t_{\nu}(\alpha)$$

- Equivalently, if $\dfrac{|D|-\ln(1.25)}{cS} < -t_{\nu}(\alpha)$

- Two one-sided t-test (TOST)
  - Schuirmann (1981), *Biometrics*
  - Schuirmann (1987), *Journal of Pharmacokinetics and Biopharmaceutics*

- Main drawback: <u>not</u> scale invariant
  - Performance depends on unknown σ

Type I Error rate: TOST



Improvements on 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*

- Improvement in power at values of σ that are unlikely.

Criterion for equivalence

$$X : \text{ measurements made by the standard device (SD)}$$
$$Y : \text{ measurements made by the alternative device (AD)}$$

- If the probability that Y/X is around 1 is large, conclude that the standard device and the alternative device are equivalent.
- Let $\theta = P\left(1 - \delta \leq \frac{Y}{X} \leq 1 + \delta\right)$

  for small δ.
- If θ is large, conclude that the standard device and the alternative device are equivalent.

Criterion for equivalence

- A usual choice is δ= 0.25

$$\theta = P\left(0.75 \leq \frac{Y}{X} \leq 1.25\right)$$

- Use the data to test

$$H_0 : \theta \leq 0.90 \text{ versus } H_1 : \theta \geq 0.90$$

- Accept equivalence if H_0 is rejected, i.e., if θ ≥ 0.90 is concluded.

References

Casella, George; Roger L. Berger. *Statistical Inference.* 2nd Edition. : Duxbury Press, California 2001.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Lin, L.I.K. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268.

Lin, L.I.K. (2000). Total deviation index for measuring individual agreement with application in laboratory performance and bioequivalence. *Statistics in Medicine*, 19, 255-270.

Lin, L.I.K., Hedayat, A.S., Sinha, Bikas, and Yang, M. (2002). Statistical methods in assessing agreement: models, issues, and tools. *Journal of American Statistical Association*, 97,257-270.

Yinprayoon, P., Tiensuwan, M. and Sinha, Bimal (2006). Some statistical aspects of assessing agreement: theory and applications. *Festschrift for Tarmo Pukkila on his 60th Birthday*. Edited by Liski, Isotalo, Niemela, Puntanen, Styan. Department of Mathematics, Statistics, and Philosophy, University of Tampere, 327-346.

Bioequivalence references

Anderson ,S. and Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics: Theory and Methods* 12: 2663 – 92

Berger, R.L.; Hsu, J.C. (1996), Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* **11**(4): 283-319 (with discussion).

Brown, L.D., Hwang, J. and Munk, A. (1997), An unbiased test for the bioequivalence problem. *Annals of Statistics* **25**, 2345-2367

Cao, L. and Mathew, T. (2008), A Simple Numerical Approach Towards Improving the Two One-Sided Test for Average Bioequivalence. *Biometrical Journal*, **50**: 205–211

Munk, A. (1993),An Improvement on Commonly Used Tests in bioequivalence Assessment. *Biometrics*, **49**(4): 1225-1230

Munk, A.; Hwang, J.T.; Brown, L. (2000), Testing Average Equivalence – Finding a Compromise between Theory and Practice. *Biometrical Journal*, **42**(5): 531-552

Schuirmann, D. L. (1981), On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics* 37(617): 137.

Schuirmann, D.L. (1987), A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 15(6): 657-680.