# WAGTTBAS

## What A Great Time To Be A Statistician!

# The Duality Issues of Significance Test and Confidence Interval

Yi Tsong, FDA/CDER/OB

Based on collaborated works with Drs. Xiaoyu Dong, Meiyu Shen, Yu-Ting Weng of FDA
and
Yue-Ming Chen of U of Texas PHS

This presentation reflects the views of the authors and should not be construed to represent FDA's views or policies.

# OUTLINES

1. Introduction
2. Comparing Mean Difference of Normal Outcomes
3. Comparing Mean Difference of Binary Outcomes
4. Alternative Comparison of Normal Outcomes
5. Summary and Conclusion

# Claims often heard about p-value and confidence interval

- P-values of a significance test often misleading.
- P-values could be smaller when the non-inferiority margin is bigger.
- Clinicians or scientists are interested to measure the drug effect instead of p-values.
- Significance of drug effect can be derived from confidence interval

# I. Introduction

- The debate between reporting p-value of statistical test or confidence interval in clinical trial, a few years ago was meaningful when both CI and statistical test lead to the consistent decision to the hull hypothesis.
- However, it was never clearly emphasized on the difference on hypothesis test and general confidence interval estimation.
- Furthermore, due to the recent development of statistical testing in drug development, the duality may lead to some of the difficulties in constructing test-based confidence interval to be consistent the with significance test.
- We illustrate the problems with a few examples.

# II. Comparing Mean Difference of Normal Outcomes

Assuming $X_i \sim N(\mu_i, \sigma_i^2)$, i = T, R represent the outcome of test reference products.

Considering test of mean difference:

When the study objective is to test the mean difference

$H_0 : \mu_T - \mu_R \le -\delta$ versus $H_A : \mu_T - \mu_R > -\delta$

where $\delta$ is ≥ 0. Assuming equal sample size n and variance $\sigma^2$, the most powerful unbiased test we use is the t-test with the following statistic

$$T = \frac{\hat{\Delta}}{s\sqrt{\frac{2}{n}}}$$

with $\hat{\Delta}$ the unbiased estimate of $\mu_T - \mu_R + \delta$ and $s^2$ the estimate of common variance . T is a monotone statistic that reaches its maximum type I error rate at $\mu_T - \mu_R + \delta = 0$.

- The sampling distribution of T at $\mu_T - \mu_R + \delta = 0$ is t with degrees of freedom of freedom 2(n-1).
- We reject the null hypothesis if T > t(0.975, 2(n-1)).
- On the other hand, $\mu_T - \mu_R + \delta$ can be estimated with a (1-α)% confidence interval

$$CI = (\hat{\Delta} - s\sqrt{2}t(1-\alpha/2, 2(n-1)), \hat{\Delta} + s\sqrt{2}t(1-\alpha/2, 2(n-1)))$$

- We will also reject $H_0$ if the lower confidence level is greater than 0.
- That the lower confidence limit can be derived from the test statistic and its sampling distributions.
- In this case, decisions made with both significance test and confidence interval are consistent although they may be derived independently.

# III. Comparing Mean Difference of Binary Outcomes

Assuming $X_i \sim B(P_i)$, i = T, R represent the binary outcome of test and reference products.

Let us consider significance test, non-inferiority and equivalence tests under the setting separately.

## Superiority hypothesis

Significance test:

When comparing two proportions, the significance testing hypotheses are,

$$H_0 : P_T - P_R \le 0 \text{ versus } H_A : P_T - P_R > 0$$

Let $\Delta = P_T - P_R$, the unbiased estimate of $\Delta$ is $\hat{\Delta} = \hat{P}_T - \hat{P}_R$. The asymptotic test of the hypotheses is a score test in the form that

$$Z = \frac{\hat{P}_T - \hat{P}_R}{e(\hat{P}_T - \hat{P}_R)} \qquad \text{(III.1)}$$

$e(\cdot)$ is the standard error of estimation.

- Since the test is monotone, i.e. if $\Delta_1 \le \Delta_2$ the Z value of $\Delta_1 \le$ the Z value of $\Delta_2$, the type I error rate of Z reaches its maximum at $\Delta$ = 0.
- The sampling distribution of the statistic is derived from $Z \mid \Delta = 0$. Accordingly, the standard error is derived with restriction to $\Delta$ = 0. It leads to

$$[\, e(\hat{P}_T - \hat{P}_R) \mid \Delta = 0 \,] = \sqrt{2\overline{P}(1-\overline{P})/n} \qquad \text{(III.2)}$$

with $\overline{P} = (\overline{P}_T + \overline{P}_R)/2$. When using this significance test, we reject the null hypothesis if Z > Z(1 -α/2) asmptotically.

- When we estimate the standard error $e(\hat{P}_T - \hat{P}_R)$ without restriction to null hypothesis, we have $e(\hat{P}_T - \hat{P}_R)$ =

$$\sqrt{\frac{\overline{P}_T(1-\overline{P}_T) + \overline{P}_R(1-\overline{P}_R)}{n}} \qquad \text{(III.3)}$$

- It can be shown that $\sqrt{\frac{\overline{P}_T(1-\overline{P}_T)+\overline{P}_R(1-\overline{P}_R)}{n}} \le \sqrt{2\overline{P}(1-\overline{P})/n}$. That means restricted standard error is at least as large as the unrestricted standard error.
- The conventional confidence interval of $\Delta$ is $(\hat{\Delta}-Z(1-\alpha/2)\sqrt{\frac{\overline{P}_T(1-\overline{P}_T)+\overline{P}_R(1-\overline{P}_R)}{n}}, \hat{\Delta}+Z(1-\alpha/2)\sqrt{\frac{\overline{P}_T(1-\overline{P}_T)+\overline{P}_R(1-\overline{P}_R)}{n}})$.
- Using conventional confidence interval, one may claim superiority if $\hat{\Delta}-Z(1-\alpha/2)\sqrt{\frac{\overline{P}_T(1-\overline{P}_T)+\overline{P}_R(1-\overline{P}_R)}{n}} > 0$. It is inconsistent to the significance test constructed under null hypothesis.

- With a continuity correction, the asymptotic test based on (III.1) becomes $Z = \dfrac{\hat{P}_T - \hat{P}_R - 1/n}{e(\hat{P}_T - \hat{P}_R)}$
- The continuity correction adjusted confidence interval is then,

$$\left( (\hat{\Delta} - 1/n - Z(1-\alpha/2)\sqrt{\frac{\overline{P}_T(1-\overline{P}_T) + \overline{P}_R(1-\overline{P}_R)}{n}}\ ,\ \hat{\Delta} + 1/n + Z(1-\alpha/2)\sqrt{\frac{\overline{P}_T(1-\overline{P}_T) + \overline{P}_R(1-\overline{P}_R)}{n}} \right)$$

- It was pointed out by Farrington and Manning (1990, SIM ) that all three statistics converge to the standard normal distribution. But they argued that test statistic using (II.2) is both theoretically correct and convergent to N(0,1) faster. It is also pointed out the other two test statistics with small to moderate sample sizes, the type I error rate is not controlled.

## Non-inferiority and equivalence tests

When comparing two proportions, the non-inferiority hypotheses are,

$$H_0 : P_T - P_R \le -\delta \text{ versus } H_A : P_T - P_R > -\delta$$

where δ > 0 is a non-inferiority constant margin.

The asymptotic test of the hypotheses is a score test in the form that

$$Z = \frac{\hat{P}_T - \hat{P}_R + \delta}{e(\hat{P}_T - \hat{P}_R)} \qquad \text{(III.4)}$$

$$Z = \frac{\hat{P}_T - \hat{P}_R - \frac{1}{n} + \delta}{e(\hat{P}_T - \hat{P}_R)}$$ with continuity correction,

where $e(\cdot)$ is the standard error of estimation. The sampling distribution of the statistic is derived from $Z \mid \Delta = -\delta$. Accordingly, the standard error is derived as the maximum likelihood estimate restricted to $\Delta$ = - δ. It can be shown as (Farrington and Manning, 1990)

$$e(\hat{P}_T - \hat{P}_R) \mid -\delta = \sqrt{[\tilde{P}_T(1-\tilde{P}_T) + \tilde{P}_R(1-\tilde{P}_R)]/n}$$

where $\tilde{P}_T$ and $\tilde{P}_R$ are the maximum likelihood estimates of $P_T$ and $P_R$ restricted to $H_0$. For testing against $H_0$, $\tilde{P}_T$ and $\tilde{P}_R$ are shown to be the solutions in (δ, 1) of the following equation

$$aX^3 + bX^2 + cX + d = 0$$

with

$a = 2$

$b = -[2 + p_T + p_R + 3\delta]$

$c = \delta^2 + \delta(2p_T + 2) + p_T + p_R$

$d = -p_T\delta(1+\delta)$

where $p_T$ and $p_R$ are the sample proportions of test and reference respectively, $\tilde{P}_T = \tilde{P}_R + \delta$. Again without restriction, we have

$$e(\hat{P}_T - \hat{P}_R) = \sqrt{\frac{\overline{P}_T(1-\overline{P}_T) + \overline{P}_R(1-\overline{P}_R)}{n}}$$

This confidence decision is not different from the one for superiority test except we compare its lower limit with $-\delta$.

## Equivalence test

The equivalence test consists of two one-sided hypotheses

$$H_{01}: P_T - P_R \le -\delta \text{ versus } H_{A1}: P_T - P_R > -\delta$$

$$H_{02}: P_T - P_R \ge \delta \text{ versus } H_{A2}: P_T - P_R < \delta$$

When rejecting both null hypotheses, one shows that

$$-\delta < P_T - P_R < \delta$$

The test statistic corresponds to testing the second one-sided hypotheses is

$$Z = \frac{\hat{P}_T - \hat{P}_R - \delta}{e(\hat{P}_T - \hat{P}_R)} \qquad \text{(III.5)}$$

and

$$Z = \frac{\hat{P}_T - \hat{P}_R - \frac{1}{n} - \delta}{e(\hat{P}_T - \hat{P}_R)}$$ with continuity correction.

The standard error is derived as the maximum likelihood estimate restricted to $\Delta = \delta$. It can be derived as

$$e(\hat{P}_T - \hat{P}_R) \mid \delta = \sqrt{[\tilde{P}_T(1-\tilde{P}_T) + \tilde{P}_R(1-\tilde{P}_R)]/n}$$

where $\tilde{P}_T$ and $\tilde{P}_R$ are the maximum likelihood estimates of $P_T$ and $P_R$ restricted to $H_{02}$. For testing against $H_0$, $\tilde{P}_T$ and $\tilde{P}_R$ are shown to be the solutions in $(\delta, 1)$ of the following equation

$$aX^3 + bX^2 + cX + d = 0$$

with

$$a = 2,\ b = -[2 + p_T + p_R - 3\delta]$$

$$c = \delta^2 - \delta(2p_T + 2) + p_T + p_R,\ d = p_T\delta(1 - \delta)$$

On the other hand, using the confidence interval, the decision of equivalence is derived with the unrestricted maximum likelihood estimate

$$e(\hat{P}_T - \hat{P}_R) = \sqrt{\frac{\overline{P}_T(1-\overline{P}_T) + \overline{P}_R(1-\overline{P}_R)}{n}}$$

and equivalence if the lower confidence limit > $-\delta$ and the upper limit < $\delta$.

The RMLE confidence interval can be constructed using the intersection of two one-sided intervals defined by the tests.

The inconsistency applies to any distribution(such as Bernoulli and Poisson) of which the variance is a function is linearly dependent to the mean

## IV. Alternative Comparisons of Normal Outcomes

VI.1 Considering test for exchangeability hypotheses involving both mean and variance, the duality of significance test and confidence interval decision rules may not be as consistent.

For example, for a probability hypothesis of non-inferiority such as

$$H_0 : \Pr(X_T - X_R < L) \geq 0.5(1-P) \quad \text{vs.} \quad H_a : \Pr(X_T - X_R < L) < 0.5(1-P)$$

where L is a pre-specified margin and P a pre-specified percentage.

- Under the normality assumption, $X_T - X_R \sim N(\mu_T - \mu_R, 2\sigma^2)$, Tsong and Shen (2007) and Dong and Tsong (2015) showed the one-sided tolerance interval $(L_P, \infty)$ of $X_T - X_R$ with significance level 1-α/2 and coverage percentage 0.5(1+P). One reject the null hypothesis if $L_P$ > L. It is an exact test.

However , for an equivalence hypotheses

$$H_{0L}: \Pr(X_T - X_R < L) \geq P \quad \text{vs.} \quad H_{aL}: \Pr(X_T - X_R < L) < P$$

$$H_{0U}: \Pr(X_T - X_R > U) \geq P \quad \text{vs.} \quad H_{aU}: \Pr(X_T - X_R > U) < P$$

- Corresponding to the two one-sided tests,
  - Test based confidence interval is then $(L_P, \infty) \cap (-\infty, U_P)$.

    On the other hand, if we use a regular two-sided tolerance interval with 1 – α confidence level and P coverage, we are considering a tolerance interval

    $$(\bar{X}_T - \bar{X}_R - kS, \bar{X}_T - \bar{X}_R + kS)$$

    with k determined by the sample size n, α and P for two-sided tolerance interval.
  - In this case, the regular confidence interval provides a different decision rule that significance test.

On the other hand, if we use a regular two-sided tolerance interval with 1 – α confidence level and P coverage, we are considering a tolerance interval

$$(\overline{X}_T - \overline{X}_R - kS, \overline{X}_T - \overline{X}_R + kS)$$

with k determined by the sample size n, α and P.

- One may reject the null hypothesis if $\overline{X}_T - \overline{X}_R - kS > L$.
- Either using approximation method or exact method, this interval provides no assurance that $(-\infty, \overline{X}_T - \overline{X}_R - kS)$ covers less that < 0.5(1-P) at 1 - α/2 level.
- In this case, the regular confidence interval provides a different decision rule from the significance test.

# VI.2 Asymptotic Tests for Variance-Adjusted Equivalence with Normal Endpoints (Chen, Weng, Dong & Tsong, 2015)

Test equivalence hypothesis

$$H_0 : \mu_T - \mu_R \geq c_U\sigma_R \text{ or } \mu_T - \mu_R \leq -c_L\sigma_R$$

$$H_a : -c_L\sigma_R < \mu_T - \mu_R < c_U\sigma_R$$

$$c_L > 0, c_U > 0$$

- Two one-sided hypotheses

$$H_{0L} : \mu_T - \mu_R \leq -c_L\sigma_R \text{ versus } H_{aL} : \mu_T - \mu_R > -c_L\sigma_R,$$

$$H_{0U} : \mu_T - \mu_R \geq c_U\sigma_R \text{ versus } H_{aU} : \mu_T - \mu_R < c_U\sigma_R.$$

- Unknown parameters $(\mu_T, \mu_R, \sigma_T, \sigma_T^2, \sigma_R, \sigma_R^2)$
- Methods:
  - Unconstrained maximum likelihood estimates
  - Unconstrained uniformly minimum variance unbiased estimates (Ahn and Fessler, 2003)
  - Constrained maximum likelihood estimates (Farrington and Manning, 1990; Ng, Gu, and Tang, 2007; Stucke and Kieser, 2013)

- Under $H_{0L}$, if $\mu_T - \mu_R = -c_L\sigma_R$, the log-likelihood function

$$\log L \propto -\frac{n_T}{2}\log\sigma_T^2 - \frac{n_R}{2}\log\left(\frac{\mu_R-\mu_T}{c_L}\right)^2 - \sum_{i=1}^{n_T}\frac{(x_{Ti}-\mu_T)^2}{2\sigma_T^2} - \sum_{i=1}^{n_R}\frac{(x_{Ri}-\mu_R)^2}{2\left(\frac{\mu_R-\mu_T}{c_L}\right)^2}.$$

- Under $H_{0U}$, if $\mu_T - \mu_R = c_U\sigma_R$, the log-likelihood function

$$\log L \propto -\frac{n_T}{2}\log\sigma_T^2 - \frac{n_R}{2}\log\left(\frac{\mu_T-\mu_R}{c_U}\right)^2 - \sum_{i=1}^{n_T}\frac{(x_{Ti}-\mu_T)^2}{2\sigma_T^2} - \sum_{i=1}^{n_R}\frac{(x_{Ri}-\mu_R)^2}{2\left(\frac{\mu_T-\mu_R}{c_U}\right)^2}.$$

- Statistical inference for $\mu_T - \mu_R + c_L\sigma_R$
  - Based on $z_L = \hat{\mu}_T - \hat{\mu}_R + c_L\hat{\sigma}_R$
  - Variance estimates of $z_L$

$$s_L^2 = \frac{\dot{\sigma}_T^2}{n_T} + (1 + c_L^2 V_n)\frac{\dot{\sigma}_R^2}{n_R}$$

  - estimates $\dot{\sigma}_k^2$ including MLE, UMVUE, constrained MLE
  - Test statistic $T_L = z_L / s_L$
  - P-value = $1 - \Phi(t_L)$
  - Reject $H_{0L}$ if p-value < $\alpha$

Similar test procedure for $\mu_T - \mu_R - c_U\sigma_R$

# Type I error rate comparison based on simulation

- Set $c_L$=1.5, $c_U$=1.5, effect size=1.5,
- Equal and unequal sample size
- $\mu_R = 0, \sigma_R = 1, \mu_T = \mu_R + [\text{ effect size }] \times \sigma_R$
- $\sigma_T^2 = (0.25, 0.5, 1, 2, 4, 10)$
- Generate $X_{ki} \overset{\text{iid}}{\sim} Normal(\mu_k, \sigma_k^2), k = T, R, i = 1, \ldots, n_k$
- Repeat $m = 10^6$ times for each parameter configuration
- Significance level $\alpha = 0.05$ for each one-sided test

| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 10 | 10 | 0.2500 | 1.5634 | 1.1434 | 3.3911 |
| 2 | 10 | 10 | 0.5000 | 1.8772 | 1.4200 | 3.5167 |
| 3 | 10 | 10 | 1.0000 | 2.4358 | 1.9043 | 3.6755 |
| 4 | 10 | 10 | 2.0000 | 3.1884 | 2.5666 | 3.8109 |
| 5 | 10 | 10 | 4.0000 | 3.5344 | 2.7767 | 3.3862 |
| 6 | 10 | 10 | 10.0000 | 1.5139 | 1.0427 | 1.1197 |

| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 15 | 15 | 0.2500 | 1.8586 | 1.5418 | 3.6057 |
| 2 | 15 | 15 | 0.5000 | 2.1534 | 1.8055 | 3.6950 |
| 3 | 15 | 15 | 1.0000 | 2.6554 | 2.2629 | 3.8400 |
| 4 | 15 | 15 | 2.0000 | 3.3563 | 2.9216 | 4.0088 |
| 5 | 15 | 15 | 4.0000 | 4.0773 | 3.5949 | 4.0738 |
| 6 | 15 | 15 | 10.0000 | 2.8801 | 2.3732 | 2.2857 |

| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 25 | 25 | 0.2500 | 2.3358 | 2.1101 | 3.8937 |
| 2 | 25 | 25 | 0.5000 | 2.5907 | 2.3547 | 3.9606 |
| 3 | 25 | 25 | 1.0000 | 3.0090 | 2.7551 | 4.0532 |
| 4 | 25 | 25 | 2.0000 | 3.5830 | 3.2968 | 4.1977 |
| 5 | 25 | 25 | 4.0000 | 4.2107 | 3.9069 | 4.3411 |
| 6 | 25 | 25 | 10.0000 | 4.5404 | 4.1994 | 4.1075 |

| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 100 | 100 | 0.2500 | 3.4055 | 3.3337 | 4.3907 |
| 2 | 100 | 100 | 0.5000 | 3.5567 | 3.4821 | 4.4232 |
| 3 | 100 | 100 | 1.0000 | 3.7998 | 3.7244 | 4.4597 |
| 4 | 100 | 100 | 2.0000 | 4.1201 | 4.0448 | 4.5135 |
| 5 | 100 | 100 | 4.0000 | 4.4194 | 4.3410 | 4.6005 |
| 6 | 100 | 100 | 10.0000 | 4.7479 | 4.6670 | 4.6957 |

| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 1000 | 1000 | 0.2500 | 4.4395 | 4.4319 | 4.7989 |
| 2 | 1000 | 1000 | 0.5000 | 4.4986 | 4.4915 | 4.8107 |
| 3 | 1000 | 1000 | 1.0000 | 4.6008 | 4.5928 | 4.8305 |
| 4 | 1000 | 1000 | 2.0000 | 4.6990 | 4.6906 | 4.8586 |
| 5 | 1000 | 1000 | 4.0000 | 4.7996 | 4.7917 | 4.8776 |
| 6 | 1000 | 1000 | 10.0000 | 4.9014 | 4.8915 | 4.9188 |

Table 2: Unequal sample size $n_T \neq n_R$

| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 10 | 6 | 0.2500 | 1.0884 | 0.5979 | 2.9793 |
| 2 | 10 | 6 | 0.5000 | 1.2813 | 0.7332 | 3.0633 |
| 3 | 10 | 6 | 1.0000 | 1.6264 | 0.9831 | 3.1986 |
| 4 | 10 | 6 | 2.0000 | 2.0346 | 1.2641 | 3.2080 |
| 5 | 10 | 6 | 4.0000 | 1.9861 | 1.1382 | 2.6000 |
| 6 | 10 | 6 | 10.0000 | 0.7152 | 0.3245 | 0.8214 |

| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 10 | 25 | 0.2500 | 2.7925 | 2.4808 | 4.0718 |
| 2 | 10 | 25 | 0.5000 | 3.3870 | 2.9970 | 4.2147 |
| 3 | 10 | 25 | 1.0000 | 4.2041 | 3.7003 | 4.3829 |
| 4 | 10 | 25 | 2.0000 | 5.1182 | 4.4954 | 4.5277 |
| 5 | 10 | 25 | 4.0000 | 5.6136 | 4.8507 | 4.2419 |
| 6 | 10 | 25 | 10.0000 | 2.6361 | 2.0399 | 1.4392 |

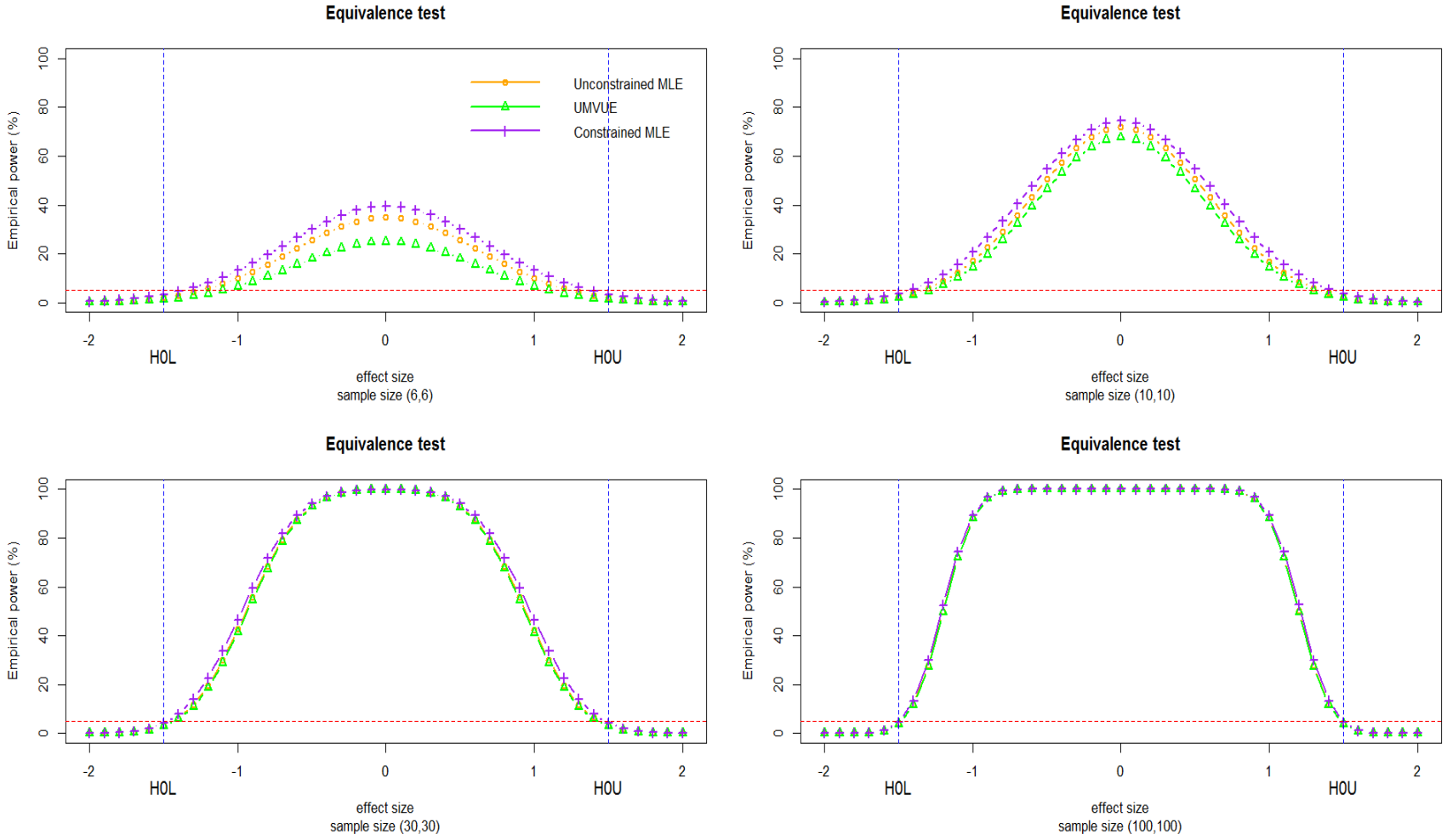
| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 25 | 10 | 0.2500 | 1.3578 | 0.9751 | 3.3213 |
| 2 | 25 | 10 | 0.5000 | 1.4874 | 1.0878 | 3.3755 |
| 3 | 25 | 10 | 1.0000 | 1.7132 | 1.3069 | 3.4486 |
| 4 | 25 | 10 | 2.0000 | 2.1263 | 1.6869 | 3.5434 |
| 5 | 25 | 10 | 4.0000 | 2.6979 | 2.2362 | 3.6815 |
| 6 | 25 | 10 | 10.0000 | 2.8482 | 2.3788 | 3.1598 |

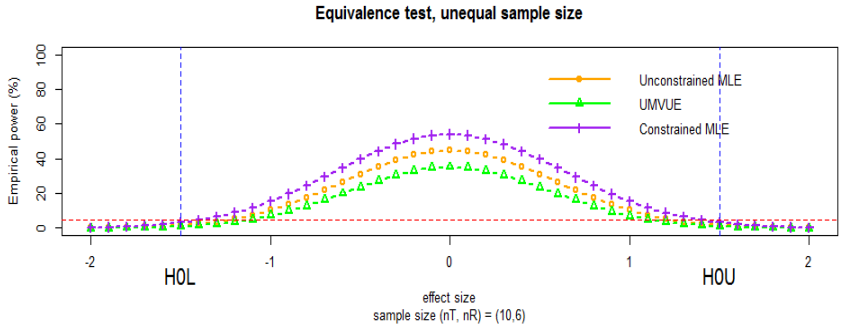
| | nT | nR | sigma_T_sq | T1_typeIerr | T2_typeIerr | T3_typeIerr |
|---|---|---|---|---|---|---|
| 1 | 100 | 10 | 0.2500 | 1.2302 | 0.8783 | 3.2443 |
| 2 | 100 | 10 | 0.5000 | 1.2681 | 0.9004 | 3.2525 |
| 3 | 100 | 10 | 1.0000 | 1.3279 | 0.9587 | 3.2727 |
| 4 | 100 | 10 | 2.0000 | 1.4505 | 1.0698 | 3.3052 |
| 5 | 100 | 10 | 4.0000 | 1.6735 | 1.2826 | 3.3711 |
| 6 | 100 | 10 | 10.0000 | 2.1957 | 1.8040 | 3.5028 |

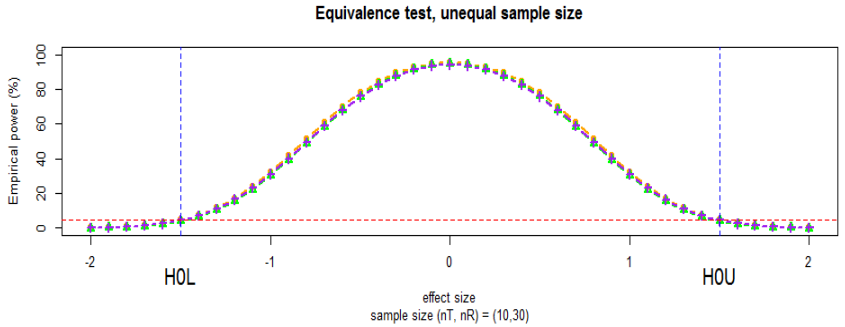
# Power comparisons based on simulation

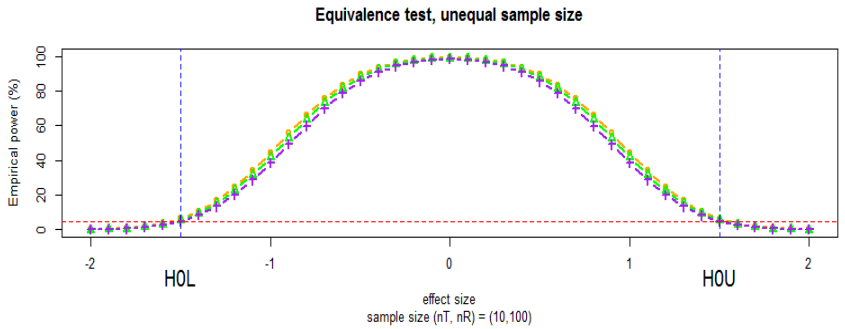
- Set $c_L$=1.5 and $c_U$=1.5
- Equal and unequal sample size
- $\mu_R = 0, \sigma_T = 1, \sigma_R = 1, \mu_T = \mu_R + \text{effect size} \times \sigma_R$
- Effect size = -2.0(0.1)2.0
- Generate $X_{ki} \overset{\text{iid}}{\sim} Normal(\mu_k, \sigma_k^2), k = T, R, i = 1, \ldots, n_k$
- Repeat $m = 10^6$ times for each parameter configuration
- Significance level $\alpha = 0.05$ for each one-sided test

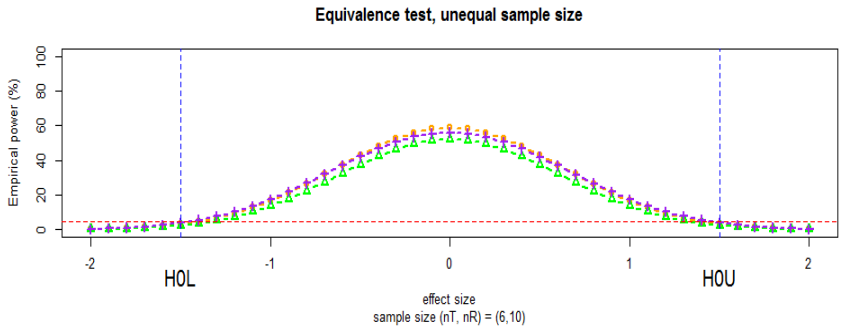
# Power comparisons based on simulation

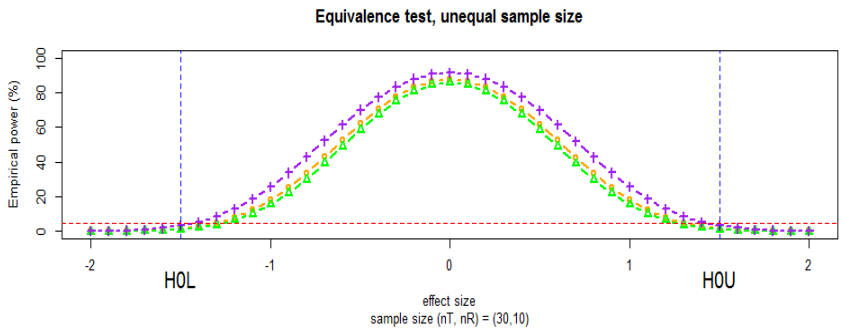


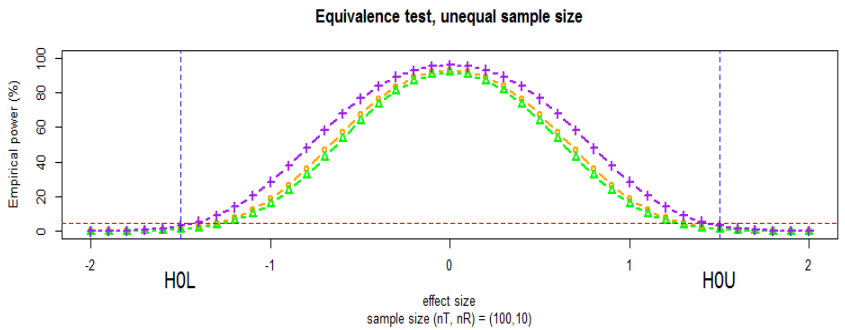

Equivalence test, unequal sample size
Empirical power (%)
Unconstrained MLE
UMVUE
Constrained MLE
H0L
H0U
effect size
sample size (nT, nR) = (10,6)


Equivalence test, unequal sample size
Empirical power (%)
H0L
H0U
effect size
sample size (nT, nR) = (10,30)


Equivalence test, unequal sample size
Empirical power (%)
H0L
H0U
effect size
sample size (nT, nR) = (10,100)


Equivalence test, unequal sample size
Empirical power (%)
H0L
H0U
effect size
sample size (nT, nR) = (6,10)

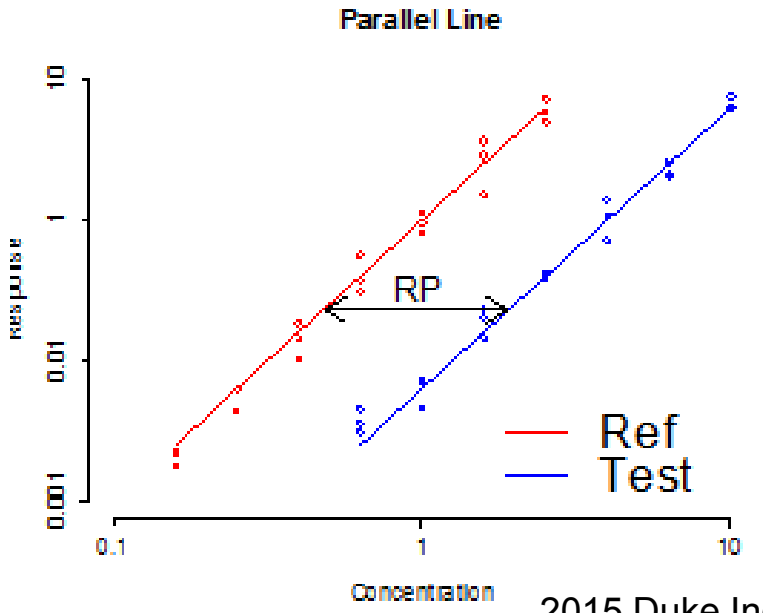

Equivalence test, unequal sample size
Empirical power (%)
H0L
H0U
effect size
sample size (nT, nR) = (30,10)


Equivalence test, unequal sample size
Empirical power (%)
H0L
H0U
effect size
sample size (nT, nR) = (100,10)

# VI.3 Statistical Methods for Parallelism Test of Bioassays

(Shao, Dong, Torigoe & Tsong, 2015)

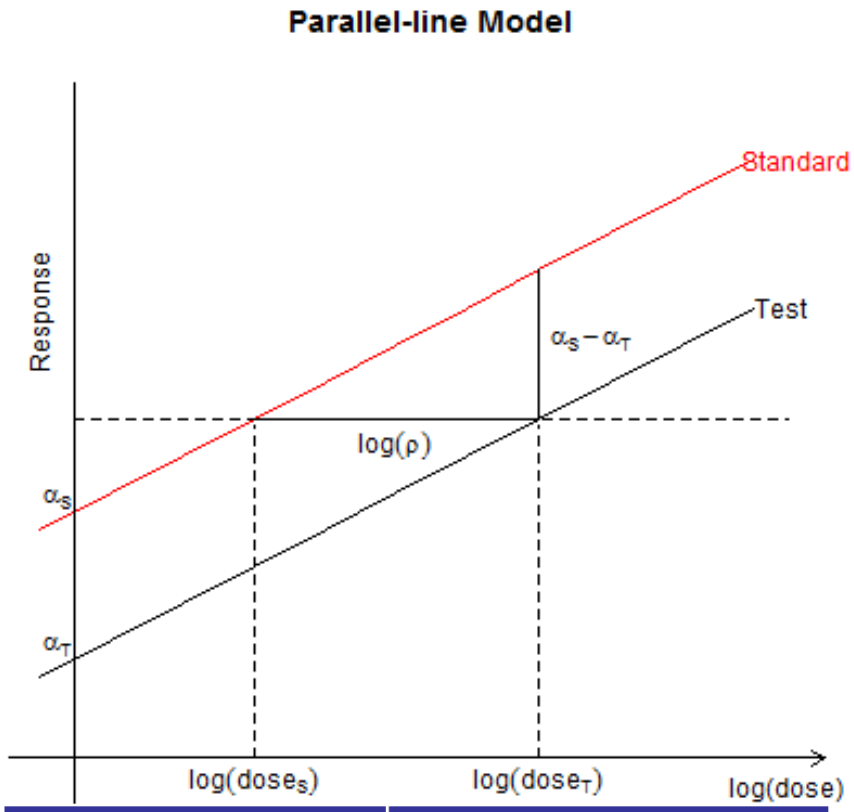
- **Bioassays** are experiment to measure biological activity (potency) of a drug as a function of concentration/dose;
- **Relative Potency** : ratio of the conc. of the test product that produces the same biological response as one unit of the conc. of the reference product




$$\rho = \frac{\text{Concentration (Test)}}{\text{Concentration (Ref)}}$$

$\rho < 1$: Test drug produces higher response (lower conc. can produce the same response as the ref.);

# Parallel-Line Model




**Model:**

$$Y_S = \alpha_S + \beta \log(dose_S) + \varepsilon$$
$$Y_T = \alpha_T + \beta \log(dose_T) + \varepsilon$$

- Independent
- Normality
- Homogenous variances of residuals

| Relative potency | $antilog \frac{\alpha_S - \alpha_T}{\beta}$ |
|---|---|

# Parallelism Test:

## Equivalence Test based on Slope Ratio

- Hypothesis: test if the ratio of slopes is close to 1.

$$H_0: \beta_T / \beta_S \le \lambda_L \text{ or } \beta_T / \beta_S \ge \lambda_U$$

$$H_a: \lambda_L < \beta_T / \beta_S < \lambda_U$$

Linearized hypothesis

$$H_0: \beta_T \le \lambda_L \beta_S \quad \text{or} \quad \beta_T \ge \lambda_U \beta_S$$

$$H_a: \lambda_L \beta_S < \beta_T < \lambda_U \beta_S$$

Use Wald test with restricted and unrestricted standard error

# Equivalence Test based on Slope Ratio: Fieller's Method

- Confidence interval for ratio of two means:
  - Widely used in bioassay analysis (USP 1034)
  - Essentially assures a **conditional coverage**
- Linearize the parameter: $\theta = \beta_T / \beta_S, \beta_T - \theta\beta_S \sim N(0, \sigma_T^2 + \theta^2\sigma_S^2)$

$$P\left(|\widehat{\beta_T} - \theta\widehat{\beta_S}|/\sqrt{SE(\widehat{\beta_T})^2 + \theta^2 SE(\widehat{\beta_S})^2} \le t_{1-\alpha,df}\right) = 1 - 2\alpha$$

- Obtain $(1 - 2\alpha)\%$ interval by solving for $\theta$

$$\frac{\frac{\widehat{\beta_T}}{\widehat{\beta_S}} \pm \frac{t_{1-\alpha,df}}{\widehat{\beta_S}}\sqrt{(1-g)SE(\widehat{\beta_T})^2 + \left(\frac{\widehat{\beta_T}}{\widehat{\beta_S}}\right)^2 SE(\widehat{\beta_S})^2}}{1-g}, \text{ where } g = \frac{t_{1-\alpha,df}^2 SE(\widehat{\beta_S})^2}{\widehat{\beta_S}^2} < 1$$

- Decision Rule: reject $H_0$ if $\lambda_L < \text{CIL} < \text{CIU} < \lambda_U$

# Simulation Studies

- Output of simulation studies:

| | |
|---|---|
| Type I Error Rate | Pr (Conclude Parallel \| Not Parallel) |
| Coverage | Pr (CI Covers the True Value of Par.) |

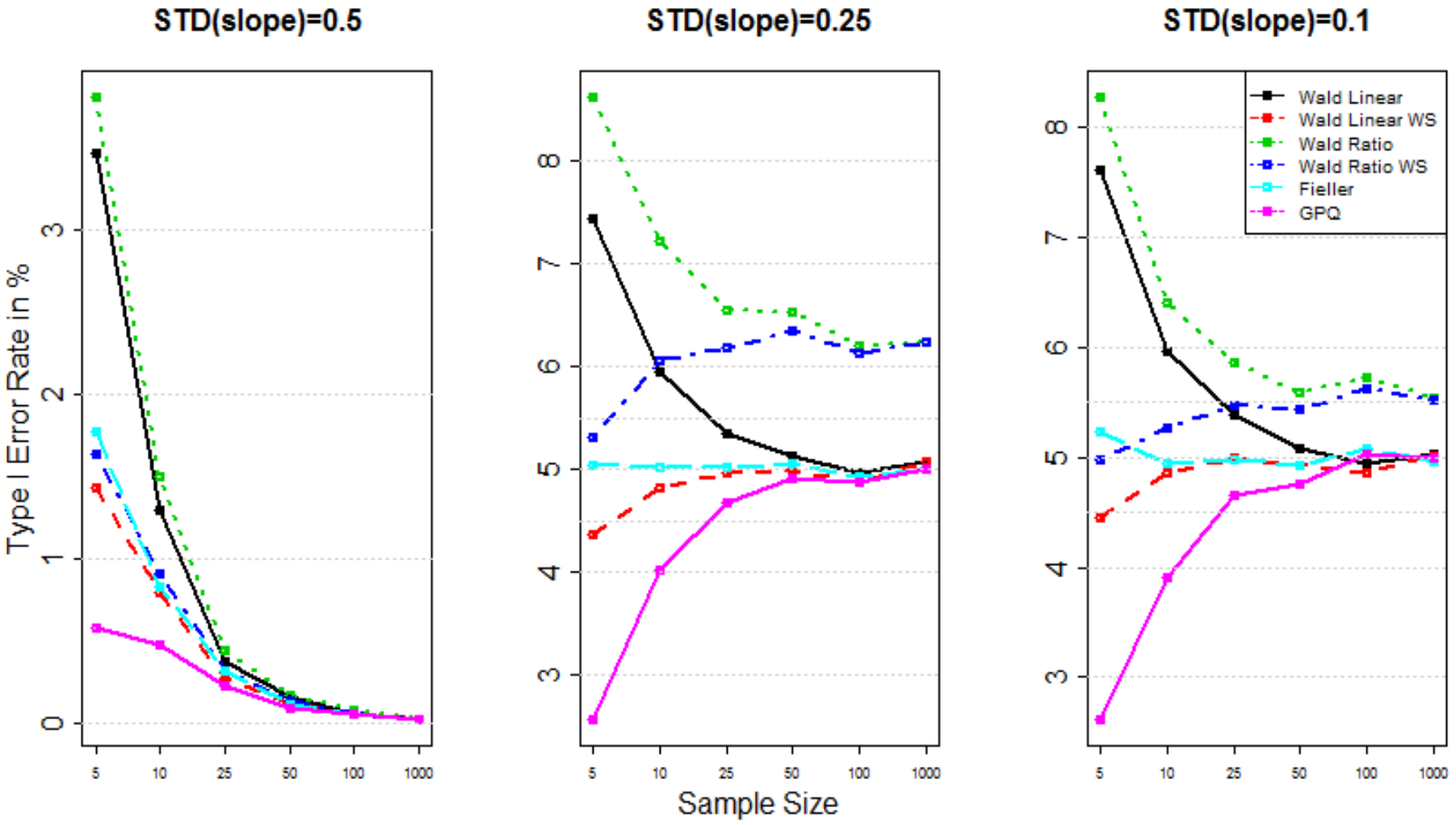
- Questions to be answered:
  - Can those methods control the type I error rate (≤ 5%)?
  - What is the coverage of CI-based approaches (close to 90%)?
  - What are the impact of sample size and variance on the Type I error rate and coverage?
  - Is the decision rule of Test Stat consistent with CI?

# Consider

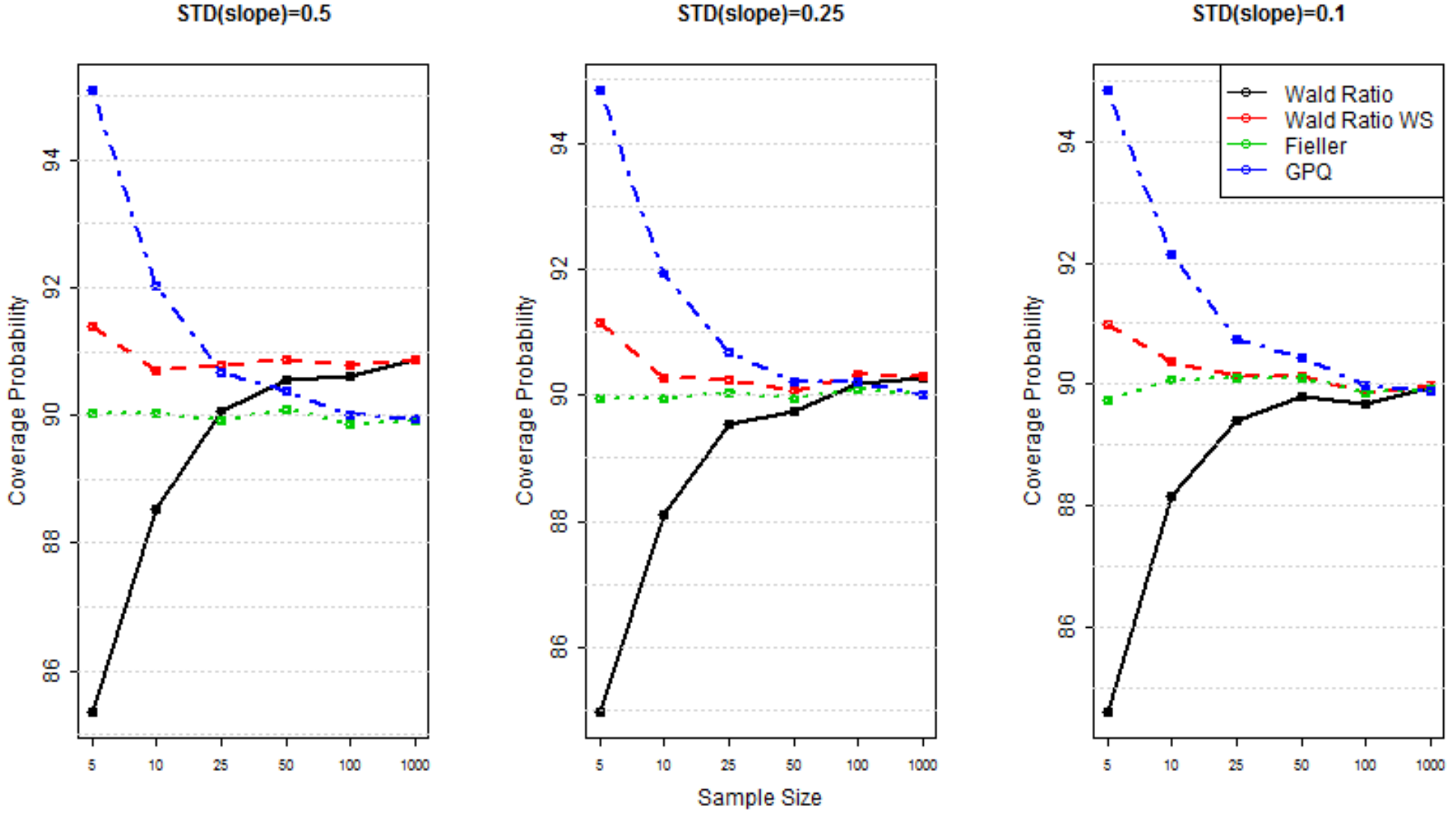
$$H_a : 0.80 < \beta_T / \beta_S < 1.25$$

- Generate Data under $H_{0:}\ \beta_T / \beta_S = 1.25$
  - $\beta_{S,j} \sim N(\beta_S, \sigma_S)$ , $\beta_{T,j} \sim N(\beta_T, \sigma_T)$
  - $\beta_S = 4, \beta_T = 5$
- $N_T = N_T = N = 5, 10, 25, 50, 100, 1000$
- $\sigma_S = \sigma_T = 0.1,\ 0.25,\ 0.5$
- Number of Simulation Steps: $10^5$

# Simulation Studies: Type I Error Rate (5%)



# Simulation Studies: Coverage (Target = 90%)



# Simulation Studies:

## Type I Error Rate vs. Coverage

- When the coverage of confidence interval can reach (1-2α) 100%, say 90%, the type I error rate may not reach α (5%):
  - The definition of coverage is consistent with decision rule of the significant test, not the equivalence test;

$H_0: \beta_T / \beta_S \neq \lambda_L$ $\qquad$ $H_0: \lambda_L < \beta_T / \beta_S < \lambda_U$

| Fieller | $\beta_T/\beta_S$ | $\beta_T$ | $\beta_S$ | SD.β | **Coverage** (%) | RR_low (%) | RR_Up (%) | **RR_Tost** (%) |
|---|---|---|---|---|---|---|---|---|
| N = 5 | 1.25 | 5 | 4 | 0.5 | **90.04** | 79.71 | 5 | **1.77** |
| N = 50 | 1.25 | 5 | 4 | 0.5 | **90.09** | 87.4 | 4.97 | **0.12** |
| N = 1000 | 1.25 | 5 | 4 | 0.5 | **89.93** | 87.84 | 4.97 | **0.02** |

- Fieller's Method:
  - provides a reliable inference for the ratio of slopes
  - controls the type I error rate;
  - However, this method is solvable only when both slopes are significant.
- A confidence interval with a 90% two-sided coverage may not assure a type error rate of 5% for equivalence test.

# Conclusion and recommendation

In many situations the standard error estimated under null hypothesis is different from the one estimated without restriction. It leads to the failing of consistency in decision making using a significance test and a traditional confidence interval in various situations.

Therefore, in various situations, estimation using regular confidence interval should be done after significance testing in order to maintain the consistency of decision making.

# Thanks for your time!

May I answer any question?