Heterogeneity in multiregional studies

Joachim Röhmel
Bremen

Reasons for regional differences can be manyfold

- Genetic sensitivity
- Culture
- Dose regimen
- Application scheme
- Disease epidemiology
- Disease definition
- Economic standing
- Health care system
- Medical practice
- Regulatory environment
- Quality of trial conduct
- Availability of concomitant medicines
- Evaluation of outcomes (in particular in composite endpoints)
- Insufficient standardisation and validation of scores (East Europe)
- Patient compliance

European Heart Journal (2013) **34**, 1846–1852
doi:10.1093/eurheartj/eht071

SPECIAL ART

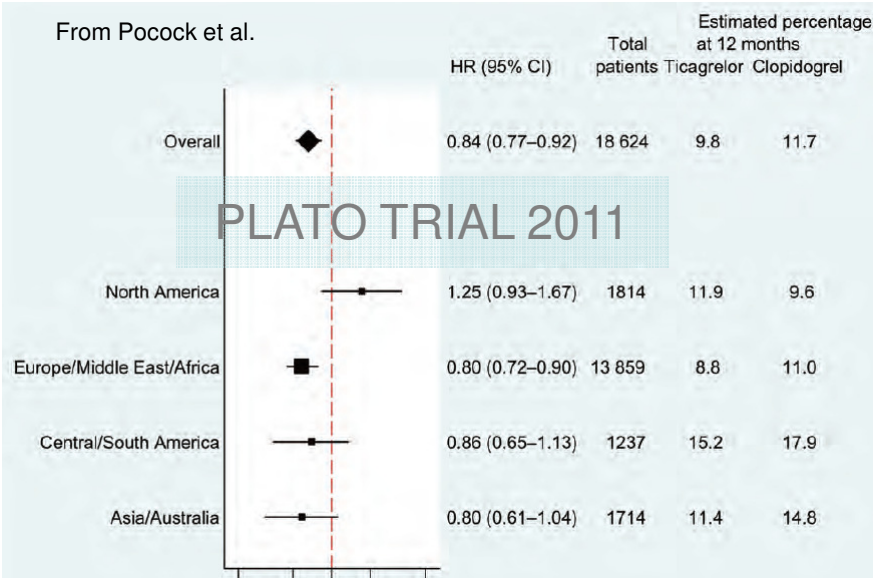
International differences in treatment effect: do they really exist and why?†

**Stuart Pocock[1]*, Gonzalo Calvo[2], Jaume Marrugat[3], Krishna Prasad[4], Luigi Tav
Lars Wallentin[6], Faiez Zannad[7], and Angeles Alonso Garcia[8]**

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK; [2]Hospital Clinic of Barcelona, Barcelo
[3]Research Group on Cardiovascular Epidemiology and Genetics, IMIM, Barcelona, Spain; [4]Medicines and Healthcare Products Regulatory Agency and Guy's and St Thon
London, UK; [5]Maria Cecilia Hospital, GVM Care & Research, Cotignola, Italy; [6]Uppsala Clinical Research Centre, Uppsala University, Uppsala, Sweden; [7]INSERM, Centr
Universitaire, Nancy, France; and [8]Hospital Universitario Puerta de Hierro, Madrid, Spain

Received 1 September 2012; revised 14 January 2013; accepted 8 February 2013; online publish-ahead-of-print 7 March 2013

From Pocock et al.



Estimated treatment effects by geographic region for the primary endpoint (CV death, MI, or stroke) of the PLATO trial (hazard ratios with 95% CIs, interaction P-value <0.05).

Conclusions of the FDA statistical review (Sep 2010)

- From the additional analyses, we continue to be troubled by the qualitative interaction between the region (US versus non-US) and treatment.
- In our view, neither play of chance nor concurrent use of ASA provides a satisfactory explanation for the US versus non-US disparity observed in this trial.
- Even though multiple factors have been screened for potential causes, the question remains unsolved.

Conclusions of the FDA statistical review (Sep 2010)

- The disparity can still be caused by the difference in standard medical practice between US and the rest of the world, which is hard to quantify and has not been quantified.
- We ought to seek further data to either confirm or dismiss this disturbing finding.
- Without the data, we would recommend that this drug not be approved.
- Another study should be required if this drug is to be approved for use in US.

Pockock's conclusions

- In the PLATO trial, the between-region comparison was one of 32 pre-planned subgroup analyses, and hence purely by chance one could expect one or two such analyses to have interaction P≤0.05.
- Furthermore, post hoc emphasis on the most striking subgroup finding (geography, in this case) means that even if the finding is not entirely due to chance, the observed data are prone to exaggerate any true disparities (between regions).
- Alternatively, one can assess all 43 countries separately, and the global interaction test for heterogeneity among the 43 hazard ratios yields P =0.95.

FDA ***APPLICATION NUMBER:*****022560Orig1s000**

- The study center effect was statistically significant in the main effect ANCOVA model. This indicates potential heterogeneity of efficacy responses across the 6 centers.
- …
- The mean percent change from baseline BMD in lumbar spine was ranging from
  - 2.5% in the US/Canada (139 subjects),
  - 3.1% in Hungary (90 subjects),
  - 3.2% in Argentina (222 subjects),
  - 3.2% in France and Belgium (64 subjects),
  - 3.8% in Poland (147 subjects), and
  - 3.9% in Estonia (140 subjects) .
- *Results of subgroups analyses are not powered to draw any meaningful statistical conclusion, mainly due to small number of subjects in subgroups.*

Regulatory consequences

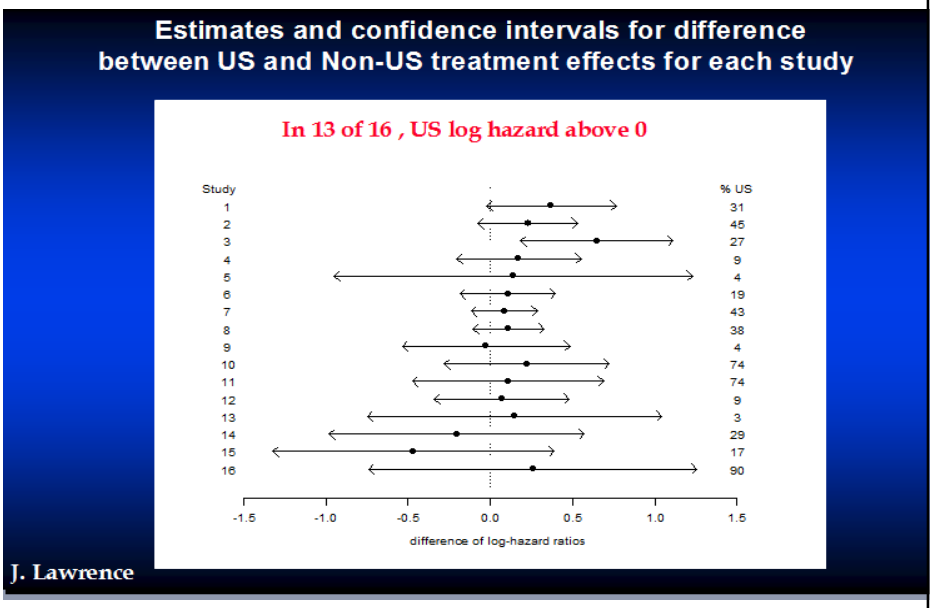
- Non approvals
  - 4 of 22 not approved because of regional heterogeneity
  - 9 of 22 approvable but more information needed - regional heterogeneity
- Need another study
- Labeling limitations or information - Merit

R.T.O'Neill, May 28, 2009

Study Undertaken by FDA statisticians to evaluate possibility of systematic regional differences

- Major cardiovascular outcome studies evaluated over the last 10 years
- Overall study result statistically positive, ie. demonstrated overall effect
- Region never pre-specified as a factor to be evaluated statistically
- 16 independent studies

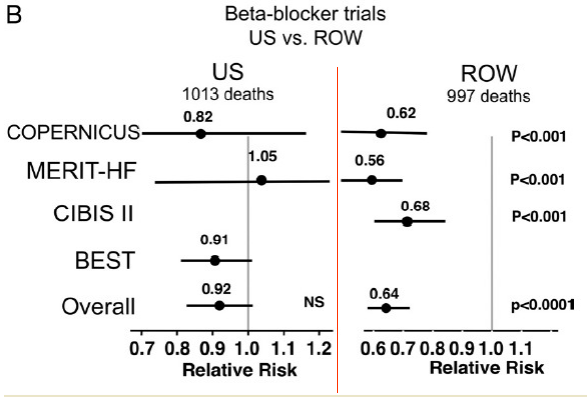
R.T.O'Neill, May 28, 2009



R.T.O'Neill, May 28, 2009

Table I International variations in mortality by treatment group in the MERIT-HF trial

	Metoprolol	Placebo	Hazard ratio (95% CI)
USA vs. the rest (*post hoc* interaction test $P = 0.003$)			
n	1990	2001	
Deaths	145	217	$P = 0.00009$
USA	51	49	1.05 (0.71, 1.56)
Other countries	94	168	0.55 (0.43, 0.70)



Social Court of the Berlin-Brandenburg
Reference number: L 1 KR 140/11 KL Dec 6, 2011

- Company complains against Escitalopram being merged with all others SSRIs, which means low reimbursement
  - Company wins first stage battle in court
- Health Insurance replies (actually based on IQWiG arguments) :
  - The results of the Yevtushenko study (2007) (conducted solely in Russia) lie extraordinarily above the estimates of the other studies. Comparability is therefore critical.
  - Furthermore, the applicability of study results may not be given in the context of German patient care. Generally, it is necessary to take stronger regard to cultural aspects in depression.

How do we define region?
How do we define consistency?

Issues and Questions (2)

- How do we define 'region'? Should there be a regulatory standard agreed to cover all trials ?
- What should be the allocation of N across regions / countries? How do we determine This?
- What is meant by 'consistency' ? How do we define this? How do we assess it? What is the value and role of routine homogeneity testing of regional results? And graphical methods?
- Should a random effects analysis be the standard in MRCTs? What are the consequences if so?

K. J. Caroll, AstraZeneca, 2011

What constitutes a region?

- America
  - North
  - Latin
  - South
- Europe
  - North
  - East
  - South
- Asia
  - China
  - India
  - Japan
  - South-East

- By Country?
- Significance of Interaction often disappears when 3 or more regions are included (Caroll, 2011)

Consistency Consideration - Design

Japan MHLW (2007): Meet the following "consistency" criterion

$$M1: \quad \hat{\delta}_1 \geq \pi \hat{\delta}_{all} \quad , \ \pi \geq 0.5$$

$$M2: \quad \hat{\delta}_i > 0 \quad , \quad \forall i = 1,\ldots,K$$

Have substantial implications on sample size distribution to the regions

J.Hung 2011 PSI-DIA Journal Club 18

Common criteria (Quan et al. DIJ, 2010)

1. Achieving in each region a proportion of the observed overall effect
2. Observing in each region an effect above a certain threshold
3. Tighten 1. by subsituting the lower limit of CIs instead of the observed values
4. Absense of statistical significance in interaction tests, usually at significance levels >>0.05
5. Lack of clinically significant differences from the overall

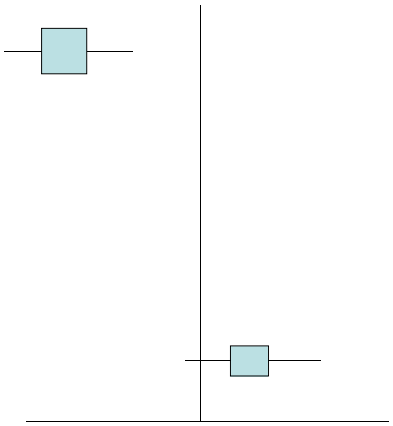
Probability of effect reversal may increase as number of regions increases or sample size allocation is more unbalanced

The smaller the sample size fraction for a region, the larger the probability of showing an effect reversal in this region will be

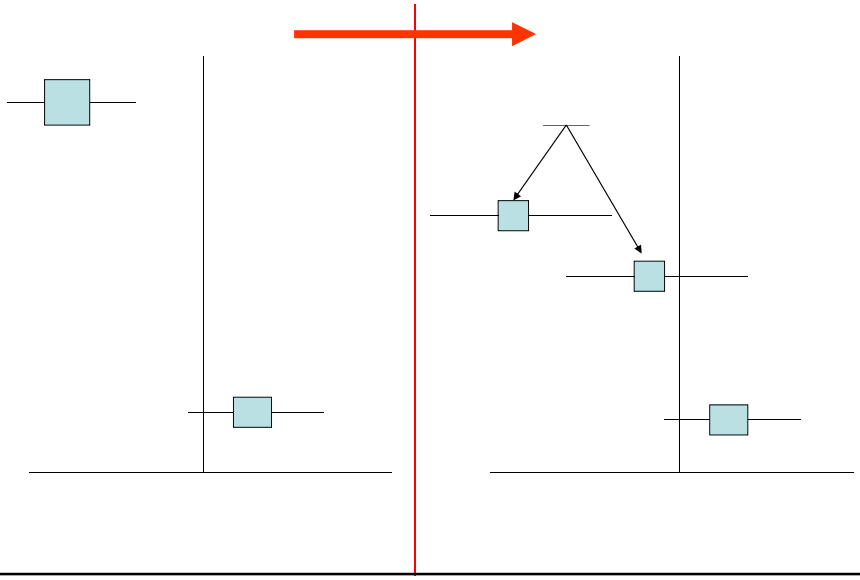
More sensible to strive for equal sample size allocation

J.Hung 2011 BBS Conference 18

Strong Interaction



Strong Interaction may look less impressive when splitting one category into two



Hung (2010)

- If, in truth, $\tau^2 > 0$, then Type I error will be inflated

$$1-\Phi\left(z_\alpha\left(1+V\tau^2\sum_{i=1}^{r}f_i^2\right)^{-0.5}\right)$$

- V should be increased to

$$\tilde{V}=\left(\frac{\theta^2}{(z_\alpha+z}\right) \quad \text{so that} \quad \frac{V}{\tilde{V}}=1-\frac{\tau^2}{\theta^2}(z_\alpha+z_\beta)^2\sum_{i=1}^{r}f_i^2$$

$$\text{and power} = \Phi\left(-z_\alpha+(z_\alpha+z_\beta)\left(1+V\tau^2\sum_{i=1}^{r}f_i^2\right)^{-0.5}\right)$$

- For $V=\frac{N}{\sigma^2}$ sample size inflation is:

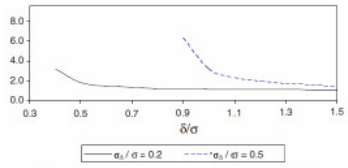


Figure 1. Sample size ratio N/N_0 versus (δ/σ).

Digging fast and deep is possible

Qualitative interaction trees:a tool to identify qualitativetreatment–subgroup interactions

Elise Dusseldorp;Iven Van Mechelen

- Modeling Algorithm (STIMA) [16, 17], Int… on Trees [18, 19], Virtual Twins [20], and Subgroup I… ation Based on Differential Effect Search (SIDES
- The goal of STIMA and Int… Trees is to partition the total group of patients … groups that differ as much as possible in relati… ent effectiveness; this implies that the two met… k for subgroups involved in an as large as possible treat…ent–subgroup interaction. The other two methods, Virtual Twins and SIDES, start by considering one of the two treatment alternatives as the reference treatment and the other as the alternative treatment; subsequently, the methods aim at identifyingspecific subgroups of patients in which the alternative treatment outperforms as much as possible the reference treatment,

Suitable for sleepless nights

An alternative conditional (permutation) approach to interaction

Full population

	Treat1	Treat2
Events	a_1	a_2
~Events	m_1-a_1	m_2-a_2
Totals	m_1	m_2

subgroup1

	Treat1	Treat2
Events	b_{11}	b_{12}
~Events	B_{11}	B_{12}
Totals	n_{11}	n_{12}

...

subgroup k

	Treat1	Treat2
Events	b_{k1}	b_{k2}
~Events	B_{k1}	B_{k2}
Totals	n_{k1}	n_{k2}

$$0 \le \begin{matrix}(b_{11},\ldots,b_{k1}) \\ (b_{12},\ldots,b_{k2})\end{matrix} \le \begin{matrix}(n_{11},\ldots,n_{k1}) \\ (n_{12},\ldots,n_{k2})\end{matrix}$$

conditional on

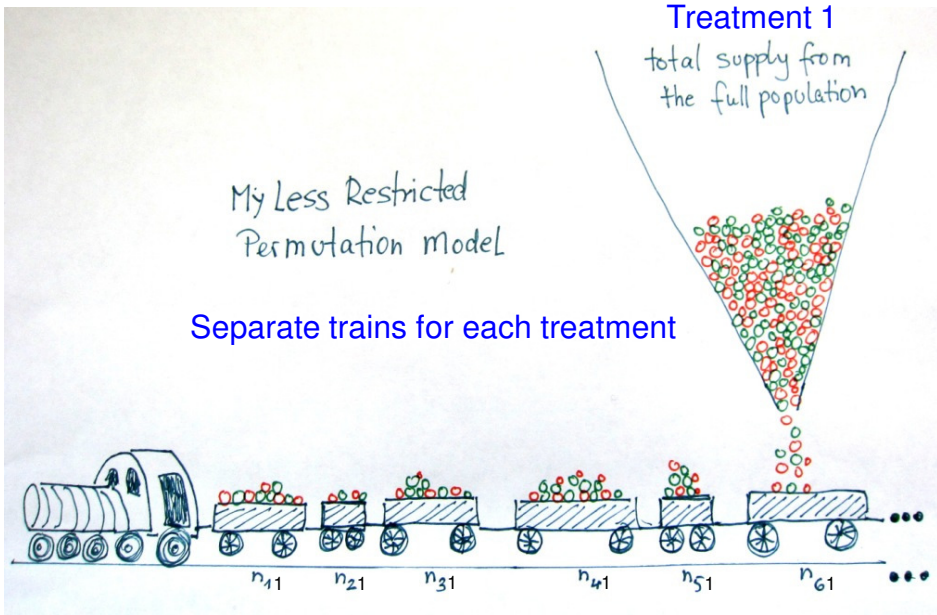
$$a_1 = \sum_1^k b_{i1}$$

$$a_2 = \sum_1^k b_{i2}$$

Comparison
"Zelen test" vs. "less restricted permuations"

	Treat1	Treat2
Events	b_{k1}	b_{k2}
~Events	B_{k1}	B_{k2}
Totals	n_{k1}	n_{k2}

	ZT/BD	LRP
Column totals per each stratum constant	√	√
Total no of Events in Treat 1 (sum over all strata) constant	√	√
Total no of Events in Treat 2 (sum over all strata) constant	√	√
Row totals per stratum constant	√	---



The LRP distribution

$$\text{prob}(b_{11},\ldots,b_{k1} \mid n_{11},\ldots,n_{k1};a_1) \cong \frac{1}{\prod_i^k (b_{i1}!)\prod_i^k ((n_{i1}-b_{i1})!)}$$

$$\text{prob}(b_{12},\ldots,b_{k2} \mid n_{12},\ldots,n_{k2};a_2) \cong \frac{1}{\prod_i^k (b_{i2}!)\prod_i^k ((n_{i2}-b_{i2})!)}$$

For independent samples:

$d(\mathbf{b}_1,\mathbf{b}_2) = \text{prob}(\mathbf{b}_1,\mathbf{b}_2|\mathbf{n}_1,\mathbf{n}_2,a_1,a_2)$

$= \text{prob}(\mathbf{b}_1|\mathbf{n}_1,a_1)\cdot\text{prob}(\mathbf{b}_2|\mathbf{n}_2,a_2)$

After knowing the permutation distribution

We need to develop measures of discepency

Starting with two subgroups

Full population

	Treat1	Treat2
Events	a_1	a_2
~Events	m_1-a_1	m_2-a_2
Totals	m_1	m_2

US

	Treat1	Treat2
Events	b_{11}	b_{12}
~Events	n_{11}-b_{11}	n_{12}-b_{12}
Totals	n_{11}	n_{12}

ROW

	Treat1	Treat2
Events	b_{21}	b_{22}
~Events	…	…
Totals	n_{21}	n_{22}

Full population

	Treat1	Treat2
Events	a_1	a_2
~Events	m_1-a_1	m_2-a_2
Totals	m_1	m_2

US

	Treat1	Treat2
Events	b_{11}	b_{12}
~Events	n_{11}-b_{11}	n_{12}-b_{12}
Totals	n_{11}	n_{12}

ROW

	Treat1	Treat2
Events	b_{21}	b_{22}
~Events	…	…
Totals	n_{21}	n_{22}

Conditional approach: a_1,a_2, n_{11}, n_{12}, n_{21}, n_{22} are given

Full population

	Treat1	Treat2
Events	a_1	a_2
~Events	m_1-a_1	m_2-a_2
Totals	m_1	m_2

US

	Treat1	Treat2
Events	b_{11}	b_{12}
~Events	n_{11}-b_{11}	n_{12}-b_{12}
Totals	n_{11}	n_{12}

ROW

	Treat1	Treat2
Events	b_{21}	b_{22}
~Events	…	…
Totals	n_{21}	n_{22}

Conditional approach: a_1,a_2, n_{11}, n_{12}, n_{21}, n_{22} are given

Permutation approach: b_{11},b_{12} are random.
(b_{21}= a_1-b_{11} ,b_{22}=a_2-b_{12})

Measures of effect in the full population

- Difference $\Delta = (a_1 \div m_1) - (a_2 \div m_2)$
- Relative risk $RR = (a_1 \div m_1) \div (a_2 \div m_2)$
- Odds ratio $OR = (a_1 \div (m_1-a_1)) \div (a_2 \div (m_2-a_2))$

Measures of discrepency

- Difference — Difference of Differences
  - CI for interaction (Newcombe 1998)
  - k=2: $\Delta_{discepency} = \Delta_{subgroup\ 1} - \Delta_{subgroup\ 2}$
  - k>2: $\Delta_{discepency} = \Sigma_{ij} (\Delta_{subgroup\ i} - \Delta_{subgroup\ j})^2$
- Relative risk — Ratio of relative risks
  - $RR_{discepency} = \log [RR_{subgroup\ 1}/RR_{subgroup\ 2}]$
  - $RR_{discepency} = \Sigma_{ij} (\log [RR_{subgroup\ i} /RR_{subgroup\ j}])^2$
- Odds ratio — Zelen's test for homogeneity of odds ratios?
  - Breslow-Day test?

P-values (for differences)

- 1-sided: $p_1(b1,b2) = \sum_{\Delta(\eta1,\eta2)\le\Delta(b1,b2)} d(\eta1,\eta2)$

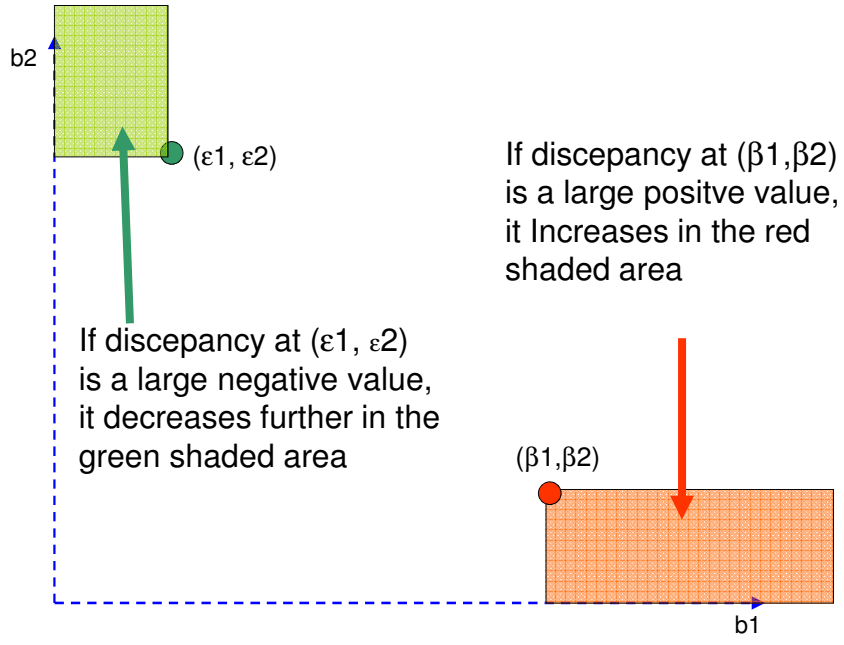
$$p_2(b1,b2) = \sum_{\Delta(\eta1,\eta2)\ge\Delta(b1,b2)} d(\eta1,\eta2)$$

- 2-sided

$$p(b1,b2)=2\cdot\min\{p_1(b1,b2),p_2(b1,b2)\}$$

Logic Order on the permutation space

- $\Delta(b1,b2) > \min\{\Delta(b1-1,b2), \Delta(b1,b2+1)\}$
- $RR(b1,b2) > \min\{RR(b1-1,b2), RR(b1,b2+1)\}$



Confidence intervals

- The lower 100(1-α)% CI limit is a unique δ such that

$$\alpha/2 \geq \sum_{\Delta(\eta 1,\eta 2)\leq\delta} d(\eta 1,\eta 2)$$

and δ is largest value with this property

More details for three or more regions

♑ ❄♒♓♦ ♓♦ ♋■ ♓♦♦◆♏ ♐□❐ ♦♒♏ ■♏⊠♦ ●♏♦♦□■ ❄♒♋■& ♘□◆ ❖♏❐♘ ❍◆♍♒ ♐□❐ ♘□◆❐ ♋♦♦♏■♦♓□■

- (I propose to study this in the next lesson)

- Thank you very much for your attention