

Oct 21 webinar

Introduction

- Alex Dmitrienko (Eli Lilly and Company), Chair of Distance Training, Biopharmaceutical Section of ASA

Stratified Analyses: Tips for Improving Power

- Devan Mehrotra (Merck Research Laboratories)

Handouts

- Can be downloaded from BioPharmNet's web site at <http://www.biopharmnet.com/doc/doc03002-05.html>

Dr. Devan V. Mehrotra

Senior Director

- Biostatistics Department of Merck Research Laboratories

Publications

- Papers and presentations on a wide variety of biopharmaceutical topics

Professional services

- Associate Editor for the American Statistician and the Biometrical Journal
- Adjunct Scholar at the University of Pennsylvania
- Past President of the ASA Philadelphia Chapter
- Secretary of the ASA Biopharmaceutical Section
- Fellow of the American Statistical Association

Stratified Analyses - Tips for Improving Power

Devan V. Mehrotra
Merck Research Laboratories
e-mail: devan_mehrotra@merck.com

ASA/Biopharmaceutical Section Webinar
October 21, 2008

Outline

- Why stratify?
- **Part I:** stratified analyses with **binary data**
 - > Mantel-Haenszel and related tests
 - > Other competing tests
 - > Simulation results
 - > Conclusions
- **Part II:** stratified analyses with **ranked data**
 - > van Elteren test
 - > Other competing tests
 - > Simulation results
 - > Conclusions

Why Stratify?

Motivating Example

Percentage of "responders" to treatment A and B

	Treatment A (Test)	Treatment B (Control)	A - B	
	48%	49%	-1%	
covariate +	95% (38/40)	80% (48/60)	+15%	p = .034*
covariate -	17% (10/60)	3% (1/40)	+14%	p = .027*
"Pooled"	48% (48/100)	49% (49/100)	-1%	

* 2-tailed p-value from z-test; "pooled" = ignoring covariate (unstratified)

Failure to stratify on prognostic covariate(s) can yield **misleading** and/or **inefficient** analyses.

3

Part I

Stratified Analyses with Binary Data

4

Stratified Trials with Binary Data

- Test (A) vs. Control (B), number of strata = s
Binary response (responder/non-responder)
- p_{ij} = true (population) proportion for strat i , trt j
 $\delta_i = p_{iA} - p_{iB}$ = true difference for strat i
 f_i = true (population) relative frequency for strat i
 $\delta = \sum_i f_i \delta_i$ = true overall difference
- \hat{p}_{ij} = observed proportion for strat i , trt j
 n_{ij} = observed number of subjects in strat i , trt j
 w_i = weight assigned to stratum i , $\hat{\delta}_i = \hat{p}_{iA} - \hat{p}_{iB}$.

5

Hypothesis Testing: General Framework

Superiority or Non-Inferiority Trials

$H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$ (superiority trial)

$H_0 : \delta \leq -\delta_0$ vs. $H_1 : \delta > -\delta_0$ (non-inferiority trial)

$$Z_w = \frac{\left| \sum_i w_i \hat{\delta}_i \right| - cc}{\sqrt{\sum_i a_i w_i^2 V(\hat{\delta}_i)}} \quad (\text{for superiority trial})$$

$$Z_w = \frac{\left(\sum_i w_i \hat{\delta}_i + \delta_0 \right) - cc}{\sqrt{\sum_i a_i w_i^2 V(\hat{\delta}_i)}} \quad (\text{for non-inferiority trial})$$

a_i = finite sample term, cc = continuity correction

IMPORTANT : What to use for $V(\hat{\delta}_i)$, w_i , a_i , and cc ?

6

Mantel & Haenszel Test (1959)

Superiority Trials

$$Z_{MH}^2 = \frac{(\sum_i w_i \hat{\delta}_i - cc)^2}{\sum_{i=1}^s a_i w_i^2 V(\hat{\delta}_i)}$$

$$V(\hat{\delta}_i) = \left(\frac{1}{n_{iA}} + \frac{1}{n_{iB}} \right) \bar{p}_i (1 - \bar{p}_i), \text{ where } \bar{p}_i = \frac{n_{iA} \hat{p}_{iA} + n_{iB} \hat{p}_{iB}}{n_{iA} + n_{iB}}$$

$$w_i^{CMH} = \frac{(n_{iA} n_{iB}) / (n_{iA} + n_{iB})}{\sum_i (n_{iA} n_{iB}) / (n_{iA} + n_{iB})}$$

$$a_i = (n_{iA} + n_{iB}) / (n_{iA} + n_{iB} - 1)$$

$$cc = 0.5 \left(\sum_i \frac{n_{iA} n_{iB}}{n_{iA} + n_{iB}} \right)^{-1}$$

Note: MH test is **optimal**
if and only if $\frac{p_{iA} / (1 - p_{iA})}{p_{iB} / (1 - p_{iB})}$
is constant across strata.

7

Choice of Variance

- **Null** variance [Miettinen & Nurminen, 1985] (MN)

$$V(\hat{\delta}_i) = \frac{\tilde{p}_{iA}(1 - \tilde{p}_{iA})}{n_{iA}} + \frac{\tilde{p}_{iB}(1 - \tilde{p}_{iB})}{n_{iB}} \equiv \tilde{V}_i$$

\tilde{p}_{ij} = m.l.e. of p_{ij} under the restriction $p_{iA} - p_{iB} = -\delta_0$

Note: MH test uses the null variance.

- **Observed (OBS)** variance

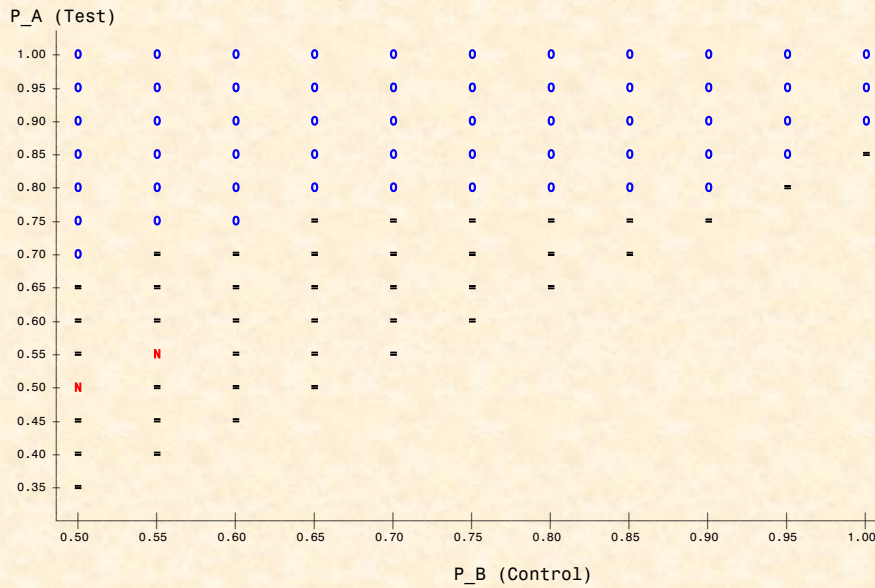
$$V(\hat{\delta}_i) = \frac{\hat{p}_{iA}(1 - \hat{p}_{iA})}{n_{iA}} + \frac{\hat{p}_{iB}(1 - \hat{p}_{iB})}{n_{iB}} \equiv \hat{V}_i$$

- Note: With 1:1 randomization, \hat{V}_i is always $< \tilde{V}_i$ for superiority trials, and often (but not always) so for non-inferiority trials.

8

(p_A, p_B) pairs where **Null** or **Observed** Variance is "Better"
 Non-Inferiority Margin = 15%

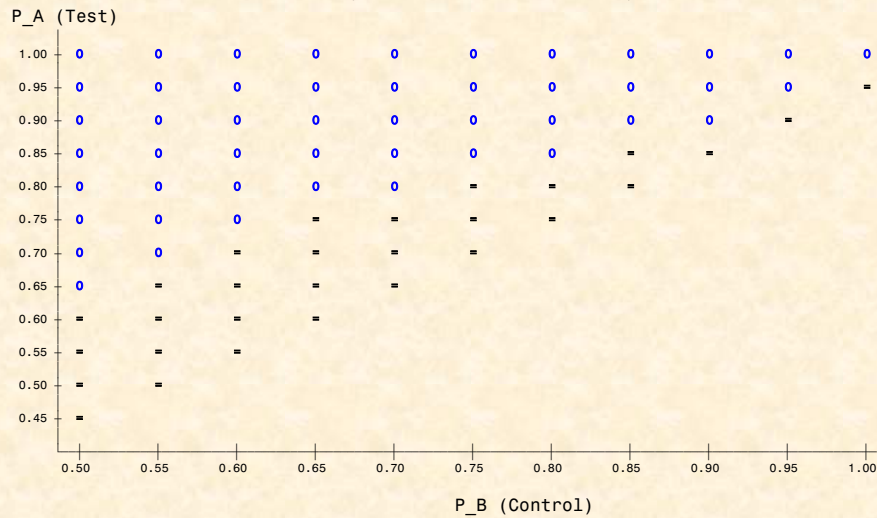
VR = VARIANCE RATIO (NULL/OBSERVED)
 VR < 0.98 (N), 0.98 < VR < 1.02 (=), VR > 1.02 (O)



(Above, P_T minus $P_C \geq -0.15$; 1:1 RANDOMIZATION)

(p_A, p_B) pairs where **Null** or **Observed** Variance is "Better"
 Non-Inferiority Margin = 5%

VR = VARIANCE RATIO (NULL/OBSERVED)
 VR < 0.98 (N), 0.98 < VR < 1.02 (=), VR > 1.02 (O)



(Above, P_T minus $P_C \geq -0.05$; 1:1 RANDOMIZATION)

Choice of Weights

- **Cochran-Mantel-Haenszel (CMH)** weights

$$w_i^{CMH} = \frac{n_{iA}n_{iB} / (n_{iA} + n_{iB})}{\sum_i n_{iA}n_{iB} / (n_{iA} + n_{iB})} \quad (= \hat{f}_i \text{ if } n_{iA} : n_{iB} \text{ is constant})$$

>> Estimator of δ is ~ unbiased.

- **Minimum Risk (MR)** weights [Mehrotra & Railkar, 2000]

Formula for two strata trial:

$$w_1^{MR} = \frac{\hat{V}_1^{-1} + \hat{f}_1(\hat{\delta}_1 - \hat{\delta}_2)^2 \hat{V}_1^{-1} \hat{V}_2^{-1}}{\hat{V}_1^{-1} + \hat{V}_2^{-1} + (\hat{\delta}_1 - \hat{\delta}_2)^2 \hat{V}_1^{-1} \hat{V}_2^{-1}}, \quad w_2^{MR} = 1 - w_1^{MR}$$

For general formula (> two strata), see Mehrotra & Railkar (2000)

>> Estimator of δ has smallest mean squared error.

>> If $\hat{\delta}_i \approx \text{constant}$, $w_i^{MR} \approx \frac{\hat{V}_i^{-1}}{\sum_i \hat{V}_i^{-1}}$ (optimal weights!)

11

Choice of Finite Sample Term (FST)

- With **CMH** weights:

$$a_i = (n_{iA} + n_{iB}) / (n_{iA} + n_{iB} - 1)$$

Mantel & Haenszel (1959), Miettinen & Nurminen (1985).

- With **MR** weights:

$$a_i = 1$$

Mehrotra & Railkar (2000).

12

Choice of Continuity Correction

- With **CMH** weights:

$$cc = 0.5 \left(\sum_i \frac{n_{iA} n_{iB}}{n_{iA} + n_{iB}} \right)^{-1} \text{ is used by the MH test.}$$

$cc = 0$ is less conservative, used by the MN test.

- With **MR** weights:

$$cc = \frac{3}{16} \left(\sum_i \frac{n_{iA} n_{iB}}{n_{iA} + n_{iB}} \right)^{-1} \text{ is used.}$$

Mehrotra & Railkar (2000)

13

Summary of Competing Methods Stratified Binary Data

Method	Variance	Weights	Finite sample term?	Continuity correction?
MH	Null	CMH	Yes	Yes
MN	Null	CMH	Yes	No
MR _{null}	Null	MR	No	Yes
MR _{obs}	Observed	MR	No	Yes

- MH = Mantel & Haenszel (1959) test
- MN = Miettinen & Nurminen (1985) test
- MR = minimum risk test of Mehrotra & Railkar (2000)
- Superiority trials: all four methods can be used
- Non-inferiority trials: all except MH can be used

Is there a "best" method?

14

Illustrative Example #1 Test for Superiority

$$H_0 : \delta = 0 \text{ vs. } H_1 : \delta \neq 0$$

Strat <i>i</i>	Test (A) \hat{P}_{iA}	Control (B) \hat{P}_{iB}	A - B $\hat{\delta}_i$	A:B OR_i	Null $\sqrt{\hat{V}_i}$	Obs $\sqrt{\hat{V}_i}$
1	.556 (30/54)	.463 (25/54)	.093	1.450	.0962	.0958
2	.917 (33/36)	.722 (26/36)	.194	4.231	.0907	.0877

Method	Weights		$\sum_i w_i \hat{\delta}_i$	$\sqrt{\sum_i a_i w_i^2 V(\hat{\delta}_i)}$	cc	2-tailed p-value
	w1	w2				
MH	.60	.40	.133	.069	.0111	.075
MN	.60	.40	.133	.069	0	.052
MR_null	.51	.49	.142	.066	.0042	.037*
MR_obs	.51	.49	.142	.065	.0042	.034*

$$a_1 = 1.0094, a_2 = 1.0141$$

* establishes superiority at 2-tailed $\alpha = .05$

15

Illustrative Example # 2 Test for Non-Inferiority

$$H_0 : \delta \leq -0.10 \text{ vs. } H_1 : \delta > -0.10 \quad (\delta_0 = 0.10)$$

Stratum <i>i</i>	Test (A) \hat{P}_{iA}	Control (B) \hat{P}_{iB}	A - B $\hat{\delta}_i$	Null $\sqrt{\hat{V}_i}$	Observed $\sqrt{\hat{V}_i}$
1	.699 (107/153)	.732 (112/153)	-.033	.0514	.0515
2	.889 (64/72)	.903 (65/72)	-.014	.0535	.0509

Method	Weights		$\sum_i w_i \hat{\delta}_i$	$\sqrt{\sum_i a_i w_i^2 V(\hat{\delta}_i)}$	cc	1-tailed p-value
	w ₁	w ₂				
MN	.68	.68	-.027	.039	0	.030
MR_null	.51	.49	-.023	.037	.0017	.022*
MR_obs	.51	.49	-.023	.036	.0017	.019*

$$a_1 = 1.0033, a_2 = 1.0070$$

* establishes non-inferiority at 1-tailed $\alpha = .025$

16

Simulation Study Test for Superiority

- 1:1 randomization, N subjects per treatment group
- $f_1 = 0.5, f_2 = 0.5$ and $n_{iA} = n_{iB} \sim B(N, f_i)$
- Response proportions simulated:

Case		B (Control)	A (Test)	Comments
I	<i>Overall</i>	60%	$60\% + \delta_I$	<ul style="list-style-type: none"> • Null: $\delta_I = 0\%$, Alt: $\delta_I = 20\%$ • $N = 82$ for $\sim 80\%$ power based on unstratified test*
	Stratum 1	$p_{1B} = 50\%$	$p_{1A} = 50\% + \delta_{1, I}$	
	Stratum 2	$p_{2B} = 70\%$	$p_{2A} = 70\% + \delta_{2, I}$	
II	<i>Overall</i>	75%	$75\% + \delta_{II}$	<ul style="list-style-type: none"> • Null: $\delta_{II} = 0\%$, Alt: $\delta_{II} = 10\%$ • $N = 250$ for $\sim 80\%$ power based on unstratified test*
	Stratum 1	$p_{1B} = 65\%$	$p_{1A} = 65\% + \delta_{1, II}$	
	Stratum 2	$p_{2B} = 85\%$	$p_{2A} = 85\% + \delta_{2, II}$	
III	<i>Overall</i>	90%	$90\% + \delta_{III}$	<ul style="list-style-type: none"> • Null: $\delta_{III} = 0\%$, Alt: $\delta_{III} = 5\%$ • $N = 435$ for $\sim 80\%$ power based on unstratified test*
	Stratum 1	$p_{1B} = 86\%$	$p_{1A} = 86\% + \delta_{1, III}$	
	Stratum 2	$p_{2B} = 94\%$	$p_{2A} = 95\% + \delta_{2, III}$	

* Farrington and Manning (1990)

- $\delta_{case} = f_1 \delta_{1, case} + f_2 \delta_{2, case}$
- $\delta_{1, case}$ and $\delta_{2, case}$ were chosen such that $g(p_{1A}) - g(p_{1B}) = g(p_{2A}) - g(p_{2B})$, i.e., no T x S interaction for given scale $g(\cdot)$

Simulation Results Test for Superiority

Type 1 Error rate (target $\alpha = 5\%$)

Case	δ	N	No TxS scale: $g(p)$	Method			
				MH	MN	MR _{null}	MR _{obs}
I	0	82	All	3.4	4.9	4.6	4.8
II	0	250	All	3.9	5.1	5.0	5.0
III	0	500	All	3.7	4.9	4.9	4.9

100,000 simulations; 5% + 2 std. errors = 5.14%

MH = Mantel & Haenszel test (uses CMH weights, cc, finite sample term, null variances)

MN = Miettinen & Nurminen test (uses CMH weights, finite sample term, null variances)

MR_{null} = Mehrotra & Railkar test (uses MR weights, null variances)

MR_{obs} = Mehrotra & Railkar test (uses MR weights, observed variances)

Simulation Results

Test for Superiority

Case I: $N=82/\text{group}$, $p_B=60\%$, $\delta=20\%$

Power and Relative Efficiency (RE)

No TxS scale: $g(p)$	Power (%)				RE (%)
	MH	MN	MR _{null}	MR _{obs}	
logit(p)	77	82	81	82	100
arcsine (p)	77	82	81	82	100
p	78	83	83	84	102
$p^{1/2}$	78	83	84	85	103
$p^{1/3}$	78	83	84	85	103
log(p)	79	83	85	86	103
1/p	80	84	88	89	104

100,000 simulations

$$RE = 100 \times \text{MSE}(\text{CMH wts}) / \text{MSE}(\text{MR wts})$$

MSE = mean squared error

19

Simulation Results

Test for Superiority

Case II: $N=250/\text{group}$, $p_B=75\%$, $\delta=10\%$

Power and Relative Efficiency (RE)

No TxS scale: $g(p)$	Power (%)				RE (%)
	MH	MN	MR _{null}	MR _{obs}	
logit(p)	79	82	81	82	102
arcsine (p)	79	82	83	83	104
p	80	83	87	87	109
$p^{1/2}$	80	83	88	88	110
$p^{1/3}$	80	83	88	88	111
log(p)	80	83	89	89	111
1/p	81	84	90	90	109

100,000 simulations

$$RE = 100 \times \text{MSE}(\text{CMH wts}) / \text{MSE}(\text{MR wts})$$

MSE = mean squared error

20

Simulation Results

Test for Superiority

Case III: N=435/group, $p_B=90\%$, $\delta=5\%$

Power and Relative Efficiency (RE)

No TxS scale: $g(p)$	Power (%)				RE (%)
	MH	MN	MR _{null}	MR _{obs}	
logit(p)	78	81	80	81	102
arcsine (p)	78	82	83	83	106
p	78	82	89	89	116
$p^{1/2}$	78	82	89	90	117
$p^{1/3}$	78	82	90	90	117
log(p)	78	82	90	90	118
1/p	78	82	91	91	119

100,000 simulations

RE = 100 × MSE(CMH wts)/MSE(MR wts)

MSE = mean squared error

21

Simulation Results

Test for Non-inferiority

N = sample size per treatment group

$n_{1A} = n_{1B} \sim B(f_1 = .50, N), n_{2A} = n_{2B} = N - n_{1A}$

$p_{1A} = .70, p_{2A} = .90, p_{iA} - p_{iB} = -\delta_0 \forall i \Rightarrow \delta = -\delta_0$ (H_0 true)

$-\delta_0$	N	Type I Error Rate (target $\alpha = 2.5\%$)		
		MN	MR _{null}	MR _{obs}
-.20	74	2.5	2.2	2.3
-.15	130	2.4	2.3	2.5
-.10	285	2.4	2.4	2.5
-.05	1130	2.5	2.5	2.5

Results based on 100,000 simulations.

22

Simulation Results: Power Test for Non-inferiority

N = sample size per treatment group

$$n_{1A} = n_{1B} \sim B(f_1 = .50, N), n_{2A} = n_{2B} = N - n_{1A}$$

$$p_{1A} = .70, p_{2B} = .90, p_{iA} - p_{iB} = 0 \forall i \Rightarrow \delta = 0 (H_1 \text{ true})$$

$-\delta_0$	N	Power (%)			\$\$\$ saved* C vs. A
		MN [A]	MR _{null} [B]	MR _{obs} [C]	
-.20	74	87	88	90	\$80K
-.15	130	87	89	90	\$150K
-.10	285	86	89	90	\$330K
-.05	1130	86	90	90	\$1.42M

* Based on \uparrow in N required to achieve 90% power with method A, and assuming \$5,000 per subject; 100,000 simulations.

23

Summary (Part I)

For stratified trials with binary responses:

- The Mantel-Haenszel and Miettinen-Nurminen tests uses **CMH** weights with **null** variances. They have optimal power properties if and only if the odds ratio is constant across strata.
- Using **minimum risk (MR)** weights with either **null** or **observed** variances can provide notable power gains over the MH and MN tests in some cases.
- No method is uniformly the best ☹
Use **simulations** to check the type I error rate and power of competing methods when planning a new superiority or non-inferiority trial.

Note: user-friendly simulation SAS macro available from author.

24

Part II

Stratified Analyses with Ranked Data

Joint work with Xiaomin Lu (U of Florida) and Xiaoming Li (Merck)

25

Example # 1

Hypothetical viral loads of HIV+ subjects (\log_{10} copies/ml)

Stratum	Placebo	Vaccine
Females	3.90, 3.96 Median=3.93	1.40, 2.80, 2.90 Median=2.80
Males	3.50, 3.50, 3.56, 3.59, 3.69, 3.85, 4.06, 4.36, 4.36, 4.43, 4.68, 4.69, 4.70, 4.85, 5.06, 5.50 Median=4.36	1.79, 2.32, 2.54, 3.42, 3.59, 3.89, 4.64, 5.23, 5.32 Median=3.59

Compared to placebo, the VLs for vaccine appear to be "shifted" to the left. **Is the overall shift statistically significant using a stratified rank-based test?**

26

Example # 1 (continued)

Stratified rank-based analysis (via popular SAS modules)

- **PROC FREQ;**
TABLES gender * trt * vload/CMH **SCORES=RANK;**
RUN;
- **PROC FREQ;**
TABLES gender * trt * vload/CMH **SCORES=MODRIDIT;**
RUN;
- **PROC TWOSAMPL;** [Note: Part of PROC StatXact module]
WI/AS;
PO trt;
RE vload;
ST gender;
RUN;

27

Example # 1 (continued)

- 2-tailed p-values using the three "methods":

PROC FREQ ^{RANK}	PROC FREQ ^{MODRIDIT}	PROC TWOSAMPL
p = .1506	p = .0642	p = .0436

- Why are the p-values so different?
PROC FREQ
 - > Ranks based on pooled sample within each stratum ("stratum-specific" ranks)
 - > SCORES = RANK → equal stratum weights
 - SCORES = MODRIDIT → unequal stratum weights**PROC TWOSAMPL:** Ranks based on overall pooled sample, ignoring strata ("**stratum-invariant**" ranks), with equal stratum weights.

28

Technical Details

Stratified Rank-Based Tests

- Y_{ijk} = response for stratum i , treatment j , subject k
($i = 1, \dots, s$; $j = 1, 2$; $k = 1, \dots, n_{ij}$)
- Assumptions
 - $Y_{i1k} \sim \text{i.i.d } F(y + \beta_i)$ [Group 1]
 - $Y_{i2k} \sim \text{i.i.d } F(y + \beta_i - \delta_i)$ [Group 2]
 - $\beta_i \in R$ is the fixed effect of stratum i
 - δ_i is the treatment effect ("shift") in stratum i
 - No T x S interaction $\Rightarrow \delta_i = \delta \forall i$ (constant shift)
- $H_0 : \delta_i = 0 \forall i$ vs. $H_1 : \delta_i \neq 0$ for at least one i

29

Technical Details (continued)

- Let R_{ijk} = rank of Y_{ijk} (stratum-specific OR stratum-invariant)

$$\bullet Z_{obs} = \frac{\sum_i w_i [S_i - E(S_i | H_0)]}{\sqrt{\sum_i w_i^2 V(S_i | H_0)}}, \text{ p-value} = 2 \times P(Z > |Z_{obs}|)$$

$$S_i = \sum_{k=1}^{n_{i1}} R_{i1k}, w_i = \text{weight for stratum } i$$

$$E(S_i | H_0) = \frac{n_{i1}}{n_{i1} + n_{i2}} \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} R_{ijk}$$

$$V(S_i | H_0) = \frac{n_{i1}n_{i2}}{(n_{i1} + n_{i2})(n_{i1} + n_{i2} - 1)} \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} \left(R_{ijk} - \frac{E(S_i | H_0)}{n_{i1}} \right)^2$$

30

Technical Details (continued)

Three Popular Rank-Based Tests

Test	Stratum weights	Comments
T_{EQ}	$w_i = 1$	<ul style="list-style-type: none">• PROC FREQ SCORES = <u>RANK</u>• Stratum-specific ranks
T_{vE}	$w_i = 1/(n_i + 1)$	<ul style="list-style-type: none">• PROC FREQ SCORES = <u>MODRIDIT</u>• Stratum-specific ranks• van Elteren test (1960)
T_{EQ}^*	$w_i = 1$	<ul style="list-style-type: none">• PROC TWOSAMPL• Stratum-invariant ranks

Note: $n_i = n_{i1} + n_{i2}$

31

Technical Details (continued)

- If there is no true treatment by stratum ($T \times S$) interaction ($\delta_i = \delta$ for all i), the **van Elteren** test is optimal among all the stratified tests that use **stratum-specific ranks**, i.e., $w_i = 1/(n_i + 1)$ are optimal weights.
- However, if **interaction** exists, the van Elteren test can suffer from a **power loss** because it does not use optimal weights!
- In general, is there an asymptotically optimal test (with **stratum-specific ranks**) that allows for a $T \times S$ interaction?

YES ... we derived it: T_{opt}

32

Technical Details (continued)

- Weights for T_{opt} :

$$w_{i,opt} = \frac{\lambda_i - 0.5}{n_i + 1}, \text{ with } \lambda_i = P(Y_{i1j} > Y_{i2k})$$

Since λ_i is unknown, we can use a test using estimated optimal weights (\hat{T}_{opt}).

λ_i is estimated as $\hat{\lambda}_i = \sum_{j,k} I(Y_{i1j} > Y_{i2k}) / (n_{i1}n_{i2})$

- Note: there are other published methods for doing stratified rank-based analyses, of which we studied two (next two slides).

33

Technical Details (continued)

"BPS" test (T_{BPS}) [Brunner, Puri and Sun, JASA, 1995]

Define overall treatment effect as:

$$\eta_s = \sum_{i=1}^s (\hat{p}_i - \frac{1}{2})^2 = \sum_{i=1}^s n_{i1}^{-2} (\bar{R}_{i2.} - \frac{n_{i1} + n_{i2} + 1}{2})^2, \text{ where}$$

$$\hat{p}_i = \frac{1}{n_{i1}} (\bar{R}_{i2.} - \frac{n_{i2} + 1}{2}), \bar{R}_{i2.} = n_{i2}^{-1} \sum_{j=1}^{n_{i2}} R_{i2j} \quad (R_{ij} \text{ is a stratum-specific rank})$$

Under $H_0 : \eta_s = 0$ (equivalent to $H_0 : \delta_i = 0 \forall i$)

$$D = \sum_{i=1}^s n_{i1}^{-2} \sigma_i^{-2} (\bar{R}_{i2.} - \frac{n_{i1} + n_{i2} + 1}{2})^2 \sim \chi_s^2$$

$$\hat{\sigma}_i^2 = \frac{1}{n_{i1}n_{i2}} \left(\sum_{t=1}^2 \frac{n_{i1} + n_{i2} - n_{it} - 1}{n_{it}(n_{i1} + n_{i2} - n_{it})} \times \sum_{j=1}^{n_{it}} [R_{ij} - R_{ij}^{(t)} - \bar{R}_{it.} + \frac{n_{it} + 1}{2}]^2 + \frac{1}{4} \right),$$

where $R_{ij}^{(t)}$ is the rank within the stratum by treatment cell

Technical Details (continued)

Aligned rank test (T_{align}) [Hodges and Lehmann, Annals of Stat, 1962]

Step 1: Calculate $Y_{ijk}^{align} = Y_{ijk} - b_i$, where b_i is the Hodges-Lehmann estimate of the stratum "location" (median of all pairwise means of the observations in stratum i)

Step 2: Perform unstratified Wilcoxon rank sum test using Y_{ijk}^{align}

Note: If there is no true $T \times S$ interaction, T_{align} should be theoretically at least as powerful as T_{vE}

35

Technical Details (continued)

Proposed (new) adaptive test [Mehrotra, Lu and Li; submitted]

- Recall that $w_{i,opt} = \frac{\lambda_i - 0.5}{n_i + 1}$, with $\lambda_i = P(Y_{i1j} > Y_{i2k})$
- Let $p_{T \times S}$ = p-value for $T \times S$ interaction, and $r_S(\hat{\lambda} - 0.5, n)$ = Spearman's rank correlation between the (estimated) treatment effect and corresponding stratum size.
- **Adaptive test** (for 1-tailed alternative; see paper for 2-tailed alternative):
If $p_{T \times S} < 0.01$ and $r_S(\hat{\lambda} - 0.5, n) > 0$, then $T_{adap} = T_{EQ}$
else $T_{adap} = T_{align}$
- We studied two ways to calculate $p_{T \times S} \Rightarrow$ they led to adaptive tests T_{adap1} and T_{adap2} , respectively.

36

Technical Details (continued)

- TxS interaction test of Öhrvik [1999]:

Let Z_{ijk} = rank of Y_{ijk}^{align} (**stratum-invariant rank**)

$$\text{Test statistic: } Q_{\text{int}} = \frac{12}{N(N+1)} \sum_{i=1}^s \sum_{j=1}^2 n_{ij} \left(Z_{ij.} - \frac{N+1}{2} \right)^2$$

$$\text{where } N = \sum_{i=1}^s \sum_{j=1}^2 n_{ij} \text{ and } Z_{ij.} = \sum_{k=1}^{n_{ij}} Z_{ijk}$$

$Q_{\text{int}} \sim \chi_{s-1}^2$ under the hypothesis of no TxS interaction

$$\text{p-value: } p_{\text{TxS}} = P(\chi_{s-1}^2 > Q_{\text{int}})$$

37

Technical Details (continued)

- TxS interaction test of Brunner *et al.* [1995]:

$$\text{Test statistic: } Q_B = \sum_{i=1}^s \frac{1}{\sigma_i^2} \left(\hat{p}_i - \frac{1}{\sum_{j=1}^s (1/\sigma_j^2)} \sum_{j=1}^s \frac{\hat{p}_j}{\sigma_j^2} \right)^2,$$

where σ_i^2 and \hat{p}_i are as described for the **BPS test**

$Q_B \sim \chi_{K-1}^2$ under the hypothesis of no TxS interaction

$$\text{p-value: } p_{\text{TxS}} = P(\chi_{K-1}^2 > Q_B)$$

38

Technical Details (continued)

Estimate and $100(1-\alpha)\%$ CI for δ Obtained by Inverting the Given Test

- Let $\tilde{Y}_{ijk}(c) = Y_{ijk}$ if $j = 1$
 $= Y_{ijk} + c$ if $j = 2$

Let $p(c) = 1$ -tailed p-value for test applied to $\tilde{Y}_{ijk}(c)$

- Point estimate ($\hat{\delta}$) $\Rightarrow c$ for which $p(c) = .50$

Lower limit (δ_L) $\Rightarrow c$ for which $p(c) = \frac{\alpha}{2}$

Upper limit (δ_U) $\Rightarrow c$ for which $p(c) = 1 - \frac{\alpha}{2}$

Obtained via a **numerical search**.

39

Example #1 Revisited 2-tailed p-values

T_{vE}	.064
T_{EQ}	.151
T_{EQ}^*	.044*
\hat{T}_{opt}	.099
T_{BPS}	.025*
T_{align}	.065
T_{adap1}	.065
T_{adap2}	.065

* statistically significant at 2-tailed $\alpha = .05$

40

Example # 1 (continued)
Estimates and 95% CIs for δ (selected methods)

Stratum		Placebo (P)	Vaccine (V)	P - V
Females	Median n	3.93 2	2.80 3	1.13
Males	Median n	4.36 16	3.59 9	0.77

Method:	T_{vE}	T_{EQ}	T_{EQ}^*	T_{align}
p-value	.064	.151	.044*	.065
Estimate	1.00	.80	0.94	.84
95% CI	(-.04, 1.61)	(-0.28, 1.61)	(.01, 1.71)	(-.09, 1.53)

41

Example #2

Immune Responses in a Real Vaccine Clinical Trial

Stratum	Group 1	Group 2
1	15, 23, 29, 30, 58, 64, 79, 81, 88, 129, 189, 234, 410 Median = 71.5	25, 34, 66, 129, 135, 155, 321, 379, 389 Median = 135
2	6, 11, 14, 19, 30, 34, 35, 39, 49, 60, 94, 123, 136, 139, 144, 148, 155, 189, 294, 376, 843 Median = 94	13, 24, 38, 103, 111, 119, 139, 144, 185, 195, 228, 270, 458 Median = 139
3	24, 24, 24, 25, 31, 54, 55, 65, 119, 124, 184, 265, 413 Median = 55	23, 38, 45, 50, 78, 85, 98, 141, 230, 235, 408, 935 Median = 91.5
4	5, 29, 65, 69, 84, 169, 225 Median = 69	50, 58, 125, 486 Median = 91.5

Responses for Group 2 appear to be "shifted" to the right of those in Group 1 for each stratum. **Is the overall shift statistically significant using a stratified rank-based test?**

42

Example #2 (continued)

2-tailed p-values

T_{vE}	.054
T_{EQ}	.063
T_{EQ}^*	.031*
\hat{T}_{opt}	.412
T_{BPS}	.377
T_{align}	.035*
T_{adap1}	.035*
T_{adap2}	.035*

* statistically significant at 2-tailed $\alpha = .05$

43

Simulation Study

- 2 treatments, 1:1 randomization per stratum
- Number of strata = 2, 4, 6, 8
- Stratum size (n_i): $10 \cdot i$ for stratum i
- Different choices of δ_i :
 - **constant** for each stratum (no TxS interaction)
 - **positively** or **negatively** associated with stratum size (TxS interaction, ~ 40% power to detect it)
- Four different **distributions** for Y:
 - Normal
 - Log Normal
 - Mixture of Normals: $0.9N(m,v) + 0.1N(m^*,v^*)$
 - t_3

44

Simulation Results
 Type I Error Rate (nominal $\alpha = 5\%$)
 Distribution = **normal**

Test	No. of strata			
	2	4	6	8
T_{vE}	4.6	4.4	4.7	5.0
T_{EQ}	4.1	4.9	4.8	5.0
T_{EQ}^*	4.8	4.7	4.7	5.0
\hat{T}_{opt}	4.4	3.8	4.8	4.2
T_{BPS}	11.0	11.3	12.1	11.7
T_{align}	5.3	5.1	5.0	5.2
T_{adap1}	5.4	5.1	5.1	5.1
T_{adap2}	5.4	5.2	5.1	5.2

Note: 5.0% + 2 std. errors = 5.62% (5000 simulations)

45

Simulation Results
 Type I Error Rate (nominal $\alpha = 5\%$)
 Distribution = **lognormal**

Test	No. of strata			
	2	4	6	8
T_{vE}	4.6	4.4	4.7	5.0
T_{EQ}	4.1	4.9	4.8	5.0
T_{EQ}^*	4.9	4.8	5.1	5.0
\hat{T}_{opt}	4.4	3.8	4.8	4.2
T_{BPS}	11.0	11.3	12.1	11.7
T_{align}	5.0	5.1	5.1	5.1
T_{adap1}	5.1	5.1	5.1	5.1
T_{adap2}	5.1	5.3	5.1	5.1

Note: 5.0% + 2 std. errors = 5.62% (5000 simulations)

46

Simulation Results
 Type I Error Rate (nominal $\alpha = 5\%$)
 Distribution = mixture of normals

Test	No. of strata			
	2	4	6	8
T_{vE}	4.7	4.8	4.8	4.8
T_{EQ}	4.3	4.8	4.7	5.1
T_{EQ}^*	4.5	4.9	5.0	4.8
\hat{T}_{opt}	4.9	4.1	4.3	4.0
T_{BPS}	11.1	11.2	11.8	11.1
T_{align}	5.3	5.2	5.1	5.2
T_{adap1}	5.4	5.2	5.1	5.2
T_{adap2}	5.4	5.2	5.2	5.2

Note: 5.0% + 2 std. errors = 5.62% (5000 simulations)

47

Simulation Results
 Type I Error Rate (nominal $\alpha = 5\%$)
 Distribution = t_3

Test	No. of strata			
	2	4	6	8
T_{vE}	4.4	4.6	4.4	4.8
T_{EQ}	3.9	4.7	4.4	5.0
T_{EQ}^*	4.4	4.7	4.6	4.7
\hat{T}_{opt}	4.3	3.6	4.5	4.7
T_{BPS}	11.4	11.2	11.9	12.0
T_{align}	4.9	5.0	4.8	4.8
T_{adap1}	4.9	5.0	4.8	4.8
T_{adap2}	5.0	5.1	4.8	4.8

Note: 5.0% + 2 std. errors = 5.62% (5000 simulations)

48

Simulation Results
Power (%)
No T x S interaction

Test	Normal				Lognormal			
	No. of strata				No. of strata			
	2	4	6	8	2	4	6	8
T_{vE}	81	81	82	80	81	80	81	81
T_{EQ}	77	77	77	76	77	76	77	77
T_{EQ}^*	82	82	81	79	82	79	79	77
\hat{T}_{opt}	69	58	52	46	69	58	51	48
T_{align}	84	83	83	82	84	82	83	83
T_{adap1}	84	83	83	82	84	82	83	83
T_{adap2}	84	83	83	82	84	82	83	83

Power: $T_{vE} < (T_{align}, T_{adap1}, T_{adap2})$

49

Simulation Results
Power (%)
No T x S interaction

Test	Mix. of normals				t_3			
	No. of strata				No. of strata			
	2	4	6	8	2	4	6	8
T_{vE}	81	81	80	82	80	81	81	81
T_{EQ}	77	78	77	78	77	77	77	76
T_{EQ}^*	81	79	76	76	81	78	74	71
\hat{T}_{opt}	70	58	51	49	69	58	51	47
T_{align}	83	83	82	83	81	82	82	82
T_{adap1}	83	83	82	83	82	82	82	82
T_{adap2}	83	83	82	83	82	82	82	82

Power: $T_{vE} < (T_{align}, T_{adap1}, T_{adap2})$

50

Summary (Part II)

For stratified rank-based analyses:

- No single method is uniformly the best ☹
- **Recommendation:** use either of the two proposed **adaptive tests** (T_{adap1} or T_{adap2}) or the **aligned rank test** (T_{align}); all three were more powerful than the van Elteren test (T_{vE}) in every case studied, whether or not there was a T x S interaction.
- **It is time to retire the popular van Elteren test!**

55

References

Part I

- Mantel N and Haenszel W (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. Journal of the National Cancer Institute, 22, 719-748.
- Mehrotra DV and Railkar R (2000). Minimum Risk Weights for Comparing Treatments in Stratified Binomial Trials. Statistics in Medicine, 19, 811-825.
- Mehrotra DV (2001). Stratification Issues with Binary Endpoints. Drug Information Journal, 35, 1343-1350.
- Miettinen O and Nurminen M (1985). Comparative analysis of two rates. Statistics in Medicine, 4, 213-226.
- Wang W, Mehrotra DV, Chan ISF and Heyse JF (2006). Non-Inferiority /Equivalence Trials in Vaccine Development. Journal of Biopharmaceutical Statistics, 16, 429-441.

Part II

- Brunner E, Puri ML, and Sun S (1995). Nonparametric Methods for Stratified Two-Sample Designs with Application to Multiclinic Trials. Journal of American Statistical Association, 90, 1004-1014.
- Hodges JL and Lehman EC (1962). Rank Methods for Combination of Independent Experiments in the Analysis of Variance. Annals of Mathematical Statistics, 33, 482-497.
- Mehrotra DV, Lu X and Li X. Rank-Based Analyses of Stratified Experiments: Retire the van Elteren Test? Submitted.
- Öhrvik J (1999). Aligned Ranks: A Method of Gaining Efficiency in Rank Tests. <http://www.stat.fi/isi99/proceedings/arkisto/varasto/hrvi0423.pdf>
- van Elteren PH (1960). On the Combination of Independent Two Sample Tests of Wilcoxon. Bulletin of the Institute of International Statistics, 37, 351-361.