Applications in medical statistics - meta-analysis, nonparametric testing, and power calculations

Malcolm Hudson*
Professor, Department of Statistics
Macquarie University

malcolm@ctc.usyd.edu.au

June, 2008

*Thank NHMRC CTC: supporting R-Workshop & ASC08 attendance
I. Meta-analysis graphics

Womens Health Initiative study of HRT

This talk: Graphic synthesis

Sources of bias in observational studies

Reducing bias

Meta analysis models and weighting

Cross-design RE models

Meta-Analysis: HRT studies up to WHI 2002

EM Algorithm

Model estimates

Findings

Malcolm Hudson
I. Meta-analysis graphics

Womens Health Initiative study of HRT

This talk: Graphic synthesis

Sources of bias in observational studies

Reducing bias

Meta analysis models and weighting

Cross-design RE models

Meta-Analysis: HRT studies up to WHI 2002

EM Algorithm

Model estimates

Findings

Malcolm Hudson

Center cases

Carcinoma in situ
Thailand 327
Philippines 319

All in situ 1462

Invasive cancer
Colombia 96
Spain 115

All invasive 211

All 1673

OR
I. Meta-analysis graphics

Malcolm Hudson¹, Victor DeGruttola², Carol Hargreaves and Val Gebski³

¹Macquarie University
²Harvard School of Public Health
³NHMRC Clinical Trials Centre
Early stopping by the SDMC raised questions

**Ethical issues**  Weigh individual risk of trial participants vs. community benefit

**Statistical interpretation of findings**  *Over-estimate* risk of the adverse treatment effect (breast cancer) that led to stopping the trial;

- Statistical estimation of odds ratios requires adjusting for multiple outcomes
- Stopping rule based on a mix of outcomes (1 primary, 7 adverse) implies limited information about each.
- Should we adjust the OR of breast cancer down?

**Specific effects**  (inducing trial-specific bias) apply to randomized trials
**Aim:** to review variation in published HRT trial results and the potential for combining risk estimates from RCTs with those of cohort and case-control studies

**Cross-design synthesis (CDS):** synthesis of evidence from multiple (trial) sources and designs (RCTs, observational) identify sources of variation in reported outcomes appropriate identification, adjustments for bias statistical model and methods evaluation in meta-analysis of 28+ HRT studies

**Issues** bias- variance compromise selection criteria for study inclusion in meta-analysis

**Scope & Limitations** uses reported summary statistics not IPD; known within trial measurement uncertainty
Sources of bias in observational studies

“Observational evidence is clearly better than opinion, but it is thoroughly unsatisfactory.” (Archibald Cochran)

Therapy is chosen to affect outcome.

**Treatment imbalances:** Confounding. *Why* did the patient get treatment?

**Time origin:** Time since study enrolment? Subject age?

**Temporal change**

In observational studies estimating HRT effect on breast cancer, necessary to allow for biases:

- earlier diagnosis, differential reporting of use,
- potential confounders: time since menopause, BMI, delay starting HRT after menopause, years of HRT
- lead to substantial underestimation of risk of breast cancer associated with the use of HRT

---

4 Collaborative Group on Hormonal Factors in Breast Cancer (HFBC) Lancet, 1997
Exclusion strategy: In a meta-analysis Peto\textsuperscript{5} excluded trials:

\textquote{“treatment assignment was not by strict randomisation”}

Sources of bias in RCTs

Lack of treatment concealment

Outcome evaluation not double blind

Study quality

In observational studies

Stratification and model adjustment for confounders

\textsuperscript{5} Stampfer, Goldhaber, Yusuf, Peto and Henneken (NEJM 307)
Meta analysis models and weighting

Single true meta effect (fixed effect) versus Inhomogeneity (random effects).

RE model (DerSimonian and Laird\(^6\))

\[ Y_j = \delta + u_j + e_j, \]

\( e_j \), measurement error in the estimated treatment effect in study \( j \), is distributed \( \text{N}(0, V_{0j}^2) \).

- \( Y_j \) is the apparent effect,
- \( \delta \) – average (meta) effect of treatment,
- \( u_j \), mean 0, variance \( \sigma_1^2 \), varies treatment effect due to specific study effects
- \( V_{0j} \) – measurement variance in the estimate of effect in study \( j \).

Weightings of trial estimates are inverse to their variance: \( V_{0j} + \sigma_1^2 \)

\(^6\)DerSimonian & Laird, 1986
**Stratified binary outcomes:** e.g. DerSimonian-Laird method with log odds-ratio estimates $Y_j$.

**Study classes:** e.g. randomised R, non-randomised NR. Postulate LME model:

\[
E(Y_j | u) = \mu + u_{j1} \sim N(0, \sigma_1^2), \quad \text{for } j \in R
\]

\[
E(Y_j | u) = \mu + \delta + u_{j1} + u_{j2} \sim N(0, \sigma_1^2 + \sigma_2^2), \quad \text{for } j \in NR.
\]

**Notes:**

- Introduces an extra source of variation in NR studies
- If $\delta = 0$, pooling class meta-estimates is legitimate.
- Not covered by DerSimonian-Laird theory.
1. Included:
   - all studies included in the HFBC (1997) meta-analysis (RCTs 0);
   - published papers since this date (n=4, RCTs 2). Total N=28 estimates.

2. Goal: meta-estimate and display

3. Outcome: HRT effect on invasive breast cancer incidence
   Odds-ratio (adjusted) comparing HRT (ever) vs HRT never.

4. Trial types case-control (hospital controls; community based controls), prospective/ cohort,
   two recent randomized clinical trials.
I. Meta-analysis graphics

Womens Health Initiative study of HRT

This talk: Graphic synthesis

Sources of bias in observational studies

Reducing bias

Meta analysis models and weighting

Cross-design RE models

Meta-Analysis: HRT studies up to WHI 2002

EM Algorithm

Model estimates

Findings

hidden
EM Algorithm

Meta-analysis graphics
I. Meta-analysis graphics
Womens Health Initiative study of HRT
This talk: Graphic synthesis
Sources of bias in observational studies
Reducing bias
Meta analysis models and weighting
Cross-design RE models
Meta-Analysis: HRT studies up to WHI 2002
EM Algorithm
Model estimates
Findings

{  # For two groups, R (n1 trials, type=1), NR (n2 trials, type != 1)
    # [snip: skip E-step and outer iteration loop]
    # M-step: update variance components d1 (sigma_1^2) and d2 (sigma_2^2)

    V <- V0 + d1 * rep(1, n)
    V[type!=1] <- V[type!=1] + d2 * rep(1, n2)
    w <- 1/V

    res <- (y-mu)
    ss1 <- sum(w^2* res^2)
    d1 <- ( d1^2 * ss1 + d1* (n - d1 * sum(w)) )/n
    d1var[itr] <- d1

    ss2 <- sum(w[type!=1]^2 * res[type!=1]^2)
    d2 <- ( d2^2 * ss2 + d2 * (n2 - d2* sum(w[type!=1])) )/n2
    d2var[itr] <- d2

    mu <- mu + sum(w * res)/sum(w0)
    means[itr] <- mu
}

EM.Searle.100 <- EM3(y, V0, nitn=100, d1=0.0001, d2=0.0004, type=Study.Type)
Model estimates

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter estimate</th>
<th>$-2l$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneous model</td>
<td>$\hat{\mu} = 0.186$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no random effects</td>
<td>$\sigma_1^2 = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\sigma_2^2 = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneity but shared mean</td>
<td>$\hat{\mu} = 0.188$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-randomised studies only,</td>
<td>$\sigma_1^2 = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{\sigma}_2^2 = 0.00684$</td>
<td></td>
<td>27.407</td>
</tr>
<tr>
<td>Heterogeneity but shared mean</td>
<td>$\hat{\mu} = 0.188$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in both RCTs and NRCTs, shared mean</td>
<td>$\sigma_1^2 = 0.00011$</td>
<td></td>
<td>27.405</td>
</tr>
<tr>
<td></td>
<td>$\hat{\sigma}_2^2 = 0.00672$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Odds ratios of risk of invasive breast cancer were generally consistent over the 28 studies, once stratified by age, parity, age at first child, years since menopause and BMI.

2. Exceptions can either be seen as ‘outlier’ trials, or as providing support for extra variation (or over-dispersion) in OR estimates among non-randomized studies (of any design class).

3. Outlier trials were indicated by a discrepancy between the naive variance \( \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \) and the correct pooled variance after stratification.

4. In either case, there is extra variation but no statistical evidence of consistent bias when studies are classified by their design class.

5. The data is generally consistent with an average log OR comparing (HRT ever use) with (HRT never use) between 0.16 and 0.22 with 95% confidence.
R-notes

Meta-analysis in R: references
Early days: S-Plus 6 on PC
Variance component estimation

Current environment
Lattice Display
Lattice Display: grouped by study type
Meta Plot via Lattice Graphic
R packages: meta and rmeta

hidden
Error Bar Plot via S-Plus Object Oriented Graphics (non R-compatible)

Data <- HRT5[, c("Y", "ID", "SE.Y")]
Data[,3] <- 1.96*Data[,3]

guiPlot( PlotType = "HorizErrorBar",
         GraphSheet="ErrorBarPlot",
         DataSetValues = Data) \#HRT5[, c("Y", "ID", "SE.Y")]

guiCreate("ReferenceLine", Name = "ErrorBarPlot$1$1",
          LineColor = "Black", LineStyle="ShortDash",
          Orientation = "Vertical", Position = 0)


guiModify( "Graph2D", Name = "ErrorBarPlot$1$",
           PanelType = "Condition",
           ConditionColumns = "TT",
           ConditionType = "Discrete")
Variance component estimation

Single population, home grown code, snippet

```
"EM1" = function(y, V0, maxitn = 1, mu = sum(y/V0)/sum(1/V0), d1 = 0.2, cc1 = cc2 = 0.001)
{
    # Searle's algorithm (8.15)
    # d1 variance component
    # input logOR (unscaled) for specified subgp as y
    # V0 measurement variances
    # Searle's convergence criterion
```

Malcolm Hudson
Mandriva Linux 2007.1
R ver 2.6
KDE 3.5.6
RKWard 0.4.6 R GUI interface (fantastic)
kile LaTeX editor
library(lattice)
yscale <- round(exp(c(-0.4, -0.2, 0, 0.2, 0.4, 0.6)),1) # 1 decimal place

# basic plot, no grouping
xyplot(ID~Y, data=hrt5s, sd=hrt5s$SE.Y,
  panel=function(x,y,subscripts,sd,...){
    panel.xyplot(x,y,...)
    larrows(x-sd[subscripts],y,
      x+sd[subscripts],y,
      angle=90,code=3,len=0.1,#lwd=1/sd[subscripts])/4,
      ...
    }
  panel.abline(v=0,lty=2)
  panel.abline(v=0.18)
},
scales=list(x=list(at=log(y尺度), labels=y尺度))
Lattice Display: grouped by study type

R-notes
Meta-analysis in R: references
Early days: S-Plus 6 on PC
Variance component estimation
Current environment
Lattice Display
Lattice Display: grouped by study type
Meta Plot via Lattice Graphic
R packages: meta and rmeta

Malcolm Hudson
```r
xyplot(ID^Y | TT, data=hrt5s, sd=hrt5s$SE.Y,
  panel=function(x, y, subscripts, sd,...) {
    panel.xyplot(x, y, ...)
    larrows(x-sd[subscripts], y,
      x+sd[subscripts], y,
      angle=90, code=3, len=0.1, #lwd=1/sd[subscripts] )/4,
      ...)
  panel.abline(v=0, lty=2)
  panel.abline(v=0.18)
},
  scales=list(x=list(at=log(yscale), labels=yscale))
)```

R packages: meta and rmeta

metabin(meta)  Meta-analysis of binary outcome data
metacont(meta) Meta-analysis of continuous outcome data
metacum(meta)  Cumulative meta-analysis
metagen(meta)  Generic inverse variance meta-analysis
metainf(meta)  Influence analysis in meta-analysis
trimfill(meta) Trim and fill method for meta-analysis
plot(meta)     meta-analysis plots

Type 'help(FOO, package = PKG)' to inspect entry 'FOO(PKG) TITLE'

rmeta:
der Simmonian and Laird RE, produces some nice graphs
Power comparisons

II. Power comparisons
Statistical analysis of survival trade-offs
Continuous TTO inference
Comparisons by scores vs. ranks
Comparison by scores
Effect of ties on P-values
'log' analysis
Simulation study goals
Nominal P-value
Location shift (log) alternative
Power
Conclusion

Malcolm Hudson
Power comparisons

Context: parametric and rank tests: grouped outcomes with zero-spike.
Survival trade-off outcomes:

- In cancer studies, preferences between treatments may depend on trading off discomfort and inconvenience for enhanced survival.
- Two forms of outcome measure:
  - time trade-off (TTO): offer extra survival time
  - probability trade-off (PTO): offer higher probability of survival
  - minimum outcome necessary to make treatment worthwhile
STOs

- $T$: survival gain required for treatment to be worthwhile
- 50-70% of women judged a 1% improvement in 5 year survival rates or a 3 month improvement in life expectancy to make either 6 cycles of CMF or 4 cycles of AC worthwhile. \(^7\)

**Analysis perspectives**

- underlying/latent continuous outcome?
- ordinal discrete (esp. survival categories, e.g, 'low-realistic')?
- mixture distribution?

- both non-traders ($T = 0$, discrete) and continuous ($T > 0$) outcomes

\(^7\) Duric et al, Annals of Onc, 2005
II. Power comparisons

Statistical analysis of survival trade-offs

Continuous TTO inference

Comparisons by scores vs. ranks

Comparison by scores

Effect of ties on P-values

'log' analysis

Simulation study goals

Nominal P-value

Location shift (log) alternative

Power

Power

Conclusion

Hidden

$T$ - time required for ACT to be worthwhile

- t-test, 'log'-transformation (ad hoc)?
- rank tests?
  - Wilcoxon-Mann-Whitney
  - Normal scores (common choice, underlying lognormal)?
  - rank tests are invariant to (monotone) transformation

- discrete distributions (binning)?
  - observed outcomes are discrete (1 day, 1 month, 3 mths, ...)
  - pre-assign 'scores'
    - t-test, score STO levels using log
    - rank tests scores are the o.s. under a distribution
Comparisons by scores vs. ranks

AC4 vs CMF

- Two sets of histograms are shown, one for AC4 alone and the other for CMF.
- The x-axis represents different values, with labels such as log(stto) and unclass(as.ordered(stto)).
- The y-axis is labeled Percent of Total, ranging from 0 to 50.

Malcolm Hudson
AC4 vs CMF

Comparison by scores

II. Power comparisons
Statistical analysis of survival trade-offs
Continuous TTO inference
Comparisons by scores vs. ranks

Comparison by scores
Effect of ties on P-values
'log' analysis
Simulation study goals
Nominal P-value
Location shift (log) alternative

Power
Power
Conclusion

Malcolm Hudson

ASC2008-R satellite – 30 / 55
Effect of ties on P-values

\[ \text{statistic} = \text{mean difference in } \log((\text{STTO} + 0.25)/0.25) \]
### Power comparisons

II. Power comparisons

Statistical analysis of survival trade-offs
Continuous TTO inference
Comparison by scores vs. ranks
Comparison by scores
Effect of ties on P-values

'log' analysis

Simulation study goals
Nominal P-value
Location shift (log) alternative

Key Points:
- **P=0.07**
- *** Permutation Test Results ***
- Number of Replications: 999
- Summary Statistics:
  - Observed Mean SE alternative p.value
  - Param 0.6302 0.006444 0.3365 two.sided 0.07
- Percentiles:
  - 2.5% 5% 95% 97.5%
  - Param -0.6539397 -0.5460052 0.5544847 0.6529697
Further analysis

II. Power comparisons

- Statistical analysis of survival trade-offs
- Continuous TTO inference
- Comparisons by scores vs. ranks
- Comparison by scores
- Effect of ties on P-values
- 'log' analysis

Simulation study goals

Nominal P-value
Location shift (log) alternative
Power

Conclusion
Simulation study goals

- Validity of P-values reported in discrete TTO data
  - based on asymptotic normality (finite sampling theory)
  - permutation distribution P-values are gold standard

- Power comparisons
  - location-shift alternatives to latent log-normal TTOs
  - alternative: multiplicative factor changes latent TTO
  - grouped in fixed intervals to form the discrete distributions

- Tests considered
  - log- scores (permutation t-test)
  - Wilcoxon (rank) test
  - Normal scores (rank) test
  - Exponential scores (Savage rank) test
## Null effect: type 1 error rates

<table>
<thead>
<tr>
<th>Equal sample sizes</th>
<th>Effect: NULL</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N=100</td>
<td>Rejection rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>alpha</td>
<td>0.1</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Wilcoxon RS</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>Normal scores</td>
<td>0.06</td>
<td>0.92</td>
<td>5.0</td>
<td>9.9</td>
<td></td>
</tr>
<tr>
<td>(unconditional)</td>
<td>0.07</td>
<td>0.90</td>
<td>5.0</td>
<td>9.9</td>
<td></td>
</tr>
<tr>
<td>Logrank</td>
<td>0.10</td>
<td>0.98</td>
<td>5.0</td>
<td>9.8</td>
<td></td>
</tr>
<tr>
<td>exponential scores</td>
<td>0.08</td>
<td>1.00</td>
<td>4.7</td>
<td>9.5</td>
<td></td>
</tr>
<tr>
<td>t-test</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>(permutation)</td>
<td>0.02</td>
<td>0.69</td>
<td>4.7</td>
<td>9.8</td>
<td></td>
</tr>
<tr>
<td>(unconditional)</td>
<td>0.02</td>
<td>0.70</td>
<td>4.6</td>
<td>9.7</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Rejection rates
II. Power comparisons

Statistical analysis of survival trade-offs
Continuous TTO inference
Comparisons by scores vs. ranks
Comparison by scores
Effect of ties on P-values
'log' analysis
Simulation study goals
Nominal P-value
Location shift (log) alternative

Power comparisons

Power

Conclusion
### Table 2: Power: SHIFT alternative

<table>
<thead>
<tr>
<th>Test</th>
<th>Alpha 0.1</th>
<th>N=100</th>
<th>N=500</th>
<th>N=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon RS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(unconditional)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logrank (exponential scores)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-test (permutation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(unconditional)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*N=10000 replicated data sets
II. Power comparisons
Statistical analysis of survival trade-offs
Continuous TTO inference
Comparisons by scores vs. ranks
Comparison by scores
Effect of ties on P-values
'log' analysis
Simulation study goals
Nominal P-value
Location shift (log) alternative
Power
Power
Conclusion
<table>
<thead>
<tr>
<th>Equal sample sizes</th>
<th>Effect: POLARISE</th>
<th>2.0*SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=100</td>
<td>Rejection rate*</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>alpha</th>
<th>%</th>
<th>%</th>
<th>%</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon RS</td>
<td>0.1</td>
<td>0.6</td>
<td>5</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Normal scores</td>
<td></td>
<td>1.2</td>
<td>8</td>
<td>22</td>
<td>32</td>
</tr>
<tr>
<td>(unconditional)</td>
<td></td>
<td>2</td>
<td>8</td>
<td>22</td>
<td>32</td>
</tr>
<tr>
<td>Logrank (exponential scores)</td>
<td></td>
<td>5</td>
<td>21</td>
<td>43</td>
<td>57</td>
</tr>
<tr>
<td>t-test (permutation)</td>
<td></td>
<td>6</td>
<td>25</td>
<td>50</td>
<td>63</td>
</tr>
<tr>
<td>(unconditional)</td>
<td></td>
<td>10</td>
<td>36</td>
<td>63</td>
<td>75</td>
</tr>
</tbody>
</table>

*N=10000 replicated data sets

Table 3: Power: POLAR alternative
II. Power comparisons

Statistical analysis of survival trade-offs
Continuous TTO inference
Comparisons by scores vs. ranks
Comparison by scores
Effect of ties on P-values
’log’ analysis
Simulation study goals
Nominal P-value
Location shift (log) alternative
Power
Power
Conclusion

- Nominal type 1 error rates (finite sample asymptotics) are reliable for STO data
- Standard method, normal scores tests, Wilcoxon share good performance under translation shift alternatives
- Very poor power in heterogeneous groups, relative to permutation t-test and logrank test
- mixture model analysis
- log rank tests for TTO and STO data!
- agrees with ad hoc analysis: \( \log(1 + T/0.25) \).
R-notes

coin: Conditional Inference
Related R bootstrap packages

hidden
Exact and asymptotic permutation distribution probabilities:

T. Horthorn **R News**, Vol 1/1, January 2001, p11

- **oneway_test**: two- and K-sample permutation test
- **wilcox_test**: Wilcoxon-Mann-Whitney rank sum test
- **normal_test**: van der Waerden normal quantile test
- **ansari_test**: Ansari-Bradley test
- **fligner_test**: Fligner-Killeen test
- **chisq_test**: Pearsons $\chi^2$ test
- **cmh_test**: Cochran-Mantel-Haenszel test
- **lbl_test**: linear-by-linear association test
- **surv_test**: two- and K-sample logrank test
- **spearman_test**: Spearmans test
- **wilcoxsing_test**: Wilcoxon-Signed-Rank test
**Related R bootstrap packages**

**boot**: This package incorporates quite a wide variety of bootstrapping tricks.

**bootstrap**: A package of relatively simple functions for bootstrapping and related techniques.

**coin**: A package for permutation tests (discussed above).

**MChtest**: This package is for Monte Carlo hypothesis tests, that is, tests using some form of resampling. This includes code for sampling rules where the number of samples taken depend on how certain the result is.

**permtest**: A package containing a function for permutation tests of microarray data.

**resper**: A package for doing restricted permutations.

**scaleboot**: This package produces approximately unbiased hypothesis tests via bootstrapping.

**simpleboot**: A package of a few functions that perform (or present) bootstraps in simple situations, such as one and two samples, and linear regression.
```r
Nscores.2 <- normal.scores(stto.2)
nscores.out2 <- t.test(Nscores.2[group==0], Nscores.2[group==1])
nscores.out2$p.value
test.NS.2 <- sum(Nscores.2[group==1])
?replicate
sum(replicate(10000,sum(Nscores.2[sample(n,n1)])) >= test.NS.2) / 10000
sum(replicate(10000,sum(Nscores.2[sample(n,n1)])) <= test.NS.2) / 10000
Nscores.2.rep <- apply(stto.2.rep, 2, normal.scores)
nscoresP <- t.test(normal.scores(stto.2)[1:50], normal.scores(stto.2)[51:100])$p.value
nscores.P.2 <- apply(Nscores.2.rep, 2, function(x){t.test(x[1:50], x[51:100])$p.value})
summary(nscores.P.2)
qqplot(nscores.P.2, unif.os)
for(alpha in c(0.001, 0.01, 0.05, 0.10)){
  print(sum(nscores.P.2 <= alpha) / 10000)
}
## more precise P
```
norm.approx <- function(obs, scores, n1, N)
{
  # approx permutation P-value from sampling without replacement mean, var
  # many scores are tied, but jittering leaves unchanged mu, V and sum of second group
  # test conditional on values observed, no continuity correction
  mu <- n1 * mean(scores)
s2 <- var(scores)
f <- n1/N
V <- n1*s2*(1-f)
z <- (obs-mu)/sqrt(V)
P1 <- pnorm(z)
P2 <- 1-P1
P <- ifelse(P1<=0.5, 2*P1, 2*P2)
list(z, P, P1, P2)
}
```
nscores.P.2b <- apply(Nscores.2.rep, 2, function(x){
    norm.approx(sum(x[51:100]), x, 50, 100)[[2]]
})
summary(nscores.P.2b)
qqplot(nscores.P.2b, unif.os)
for(alpha in c(0.001, 0.01, 0.05, 0.10)){
    print(sum(nscores.P.2b <= alpha)/10000)
}
## Exponential scores rank test
```
Appendix
Women’s Health Inititiative Study of HRT

Key features

- Second large RCT\(^8\) on Estrogen/progestin vs. placebo
- First of a pair of RCTs conducted by WHI with different HRT treatments
- Primary outcome CHD, primary adverse outcome invasive breast cancer
- Healthy post-menopausal women aged 50-79 yrs
- Population sample (direct mailing campaign)
- Multiple outcomes – CHD, colorectal cancer, hip fractures, ...
- Global index of monitored outcomes: balancing risks and benefits

Controversial

- Settled advice to women
- Trial was stopped early (5 yrs vs 8.5 yrs) by the SDMC
- Stopping rule based on mix of outcome boundaries (1 positive, 8 adverse)
- Adverse boundaries were for breast cancer and 7 other outcomes
  the latter employed 1-sided \( \alpha = 0.05/7 \) boundaries
- Compliance: treatment non-compliance 25%-30% at 5 yrs

\(^8\)WHI Investigators, JAMA 2002
LETTER

Large-Database Research

Complement to Randomized Trials?

Jose A. Sacristan, MD; Javier Soto, MD, PhD; and Ines Galende, MD

15 May 1998 | Volume 128 Issue 10 | Page 875

To the Editor: We are disappointed by the emphasis that the articles on database research in the supplement published on 15 October 1997 place on sophisticated new mathematical models to control for confounding factors and by the classic commonplace that clinical database studies are "attractive alternatives to randomized trials" [1]. Using large databases to compare therapies remains controversial [2]. By design, databases record observations made in clinical practice. Because treatment decisions are not randomly allocated, any observed therapeutic effect may be due to unrecognized factors affecting the treatment allocation rather than the treatment itself.

It is surprising that a supplement focused on the future of databases did not mention new research methods, such as cross-design synthesis [3], directed toward the generation of results with an acceptable balance between internal and external validity. Specifically, cross-design synthesis proposes the assessment, adjustment, and combination of treatment effects obtained with randomized studies and database analyses.
Searle’s random effects model: $Y = X\mu + \sum_{i=1}^{2} Z_iu_i + e$, where $u_1 = (u_{11}, \ldots, u_{1N})^T$ and $u_2 = (u_{2N_1+1}, \ldots, u_{2N})^T$, $X$ is an arbitrary design matrix for fixed effects, $Z_1$ is an $N \times N$ identity matrix and $Z_2 = \begin{bmatrix} 0^T : I \end{bmatrix}^T$ is $N \times N_2$, with $N = N_1 + N_2$.

The log-likelihood $l$ is conveniently expressed as

$$-2l = \sum_j \log(V_j) + \sum_j \frac{(y_j - \mu)^2}{V_j}$$

where $V_j$ represents the variance of the treatment summary outcome in trial $j$ according to the model. For example, in the model with two strata:

$$V_j = V_{j0} + \sigma_1^2 \quad \text{for } j = 1, \ldots, N_1$$

$$= V_{j0} + \sigma_1^2 + \sigma_2^2 \quad \text{for } j = (N_1 + 1), \ldots, N$$ \hspace{1cm} (1)
Nested models may readily be compared by difference in log-likelihoods, once variance parameters are estimated.
Differences in twice log-likelihood $-2\Delta l$ should be compared with half the tabled value for chi-square with degrees of freedom the number of extra variance parameters $^9$.

$^9$Stram, Biometrics, 1994
References

- Best Non-parametric comparisons of two histograms, Biometrics, 1994
- Buning et al. Power of generalised Wilcoxon test, Communications in Statistics
- Tanizaki Power comparisons of non-parametric tests: small-sample properties from Monte-Carlo experiments, 1997
```
simul2 <- data.frame(stto=stto.2, group)

## R graphics
library(lattice)
histogram(~log(stto) | group, data=simul2, breaks=(-7:4))
dev.set(2)
dev2bitmap("simul2plot1.png", type="png256", res=72.00000000)
histogram(~log((stto+0.25)/0.25) | group, data=simul2, breaks=(seq(0,6,by=0.5)))
```