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

Malcolm Hudson
NHMRC Clinical Trials Centre,*
University of Sydney

malcolm@ctc.usyd.edu.au

July, 2009

*Thanks to Ayse Bilgin, Petra Graham and Martin Stockler for comments. Initial analysis conducted at Macquarie University.

Testing treatment effects in TTO data

Patient Preference Data

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

- In cancer studies, the choices of patients may involve trading off discomfort and inconvenience of treatment 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
- Outcome is the minimum benefit necessary to make treatment worthwhile
- May be heavily influenced by the patient's experience of treatment

Time tradeoff outcomes

TTO = the minimum survival benefit necessary to make the discomfort and inconvenience of chemotherapy worthwhile.
After a full course of chemo, questions of the form:

> Imagine you knew that:
> - without chemotherapy life expectancy is 5 years, and;
> - with chemotherapy, life expectancy is 5 years and *6 months*.
>
> In other words, having chemotherapy would increase life expectancy by *6 months*.
> Based on your own experiences of chemotherapy, which would you prefer?

- The 6 month extra survival benefit is varied from 1 day to 20 years. (Patient can nominate that no survival gain would suffice).
- The responses to a structed sequence of questions provide a single survival benefit balance point, the patient's TTO.

Survival time trade-offs

TTOs

- T: extra survival 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. [1]
- Statistical Analysis perspectives
  - ☐ grouped data with underlying continuous distribution?
  - ☐ ordinal discrete (esp. survival categories, e.g, 'low-realistic')?
  - ☐ mixture distribution?
    - both non-traders ($T = 0$, discrete) and continuous ($T > 0$) outcomes

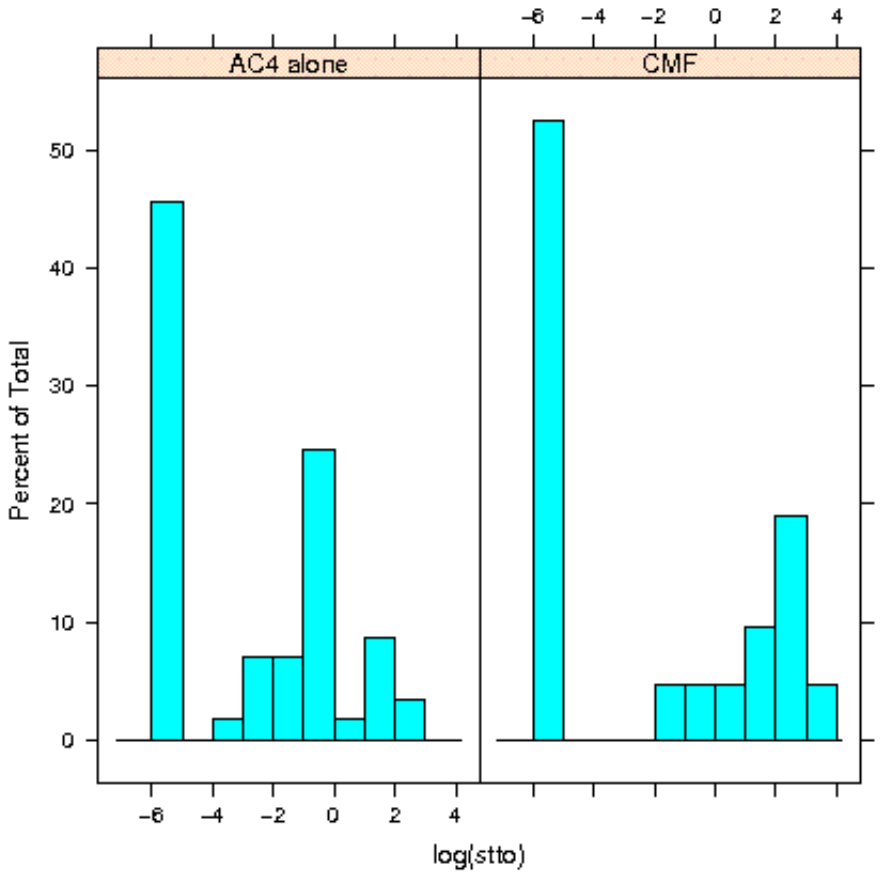
[1] Duric et al, Annals of Onc, 2005

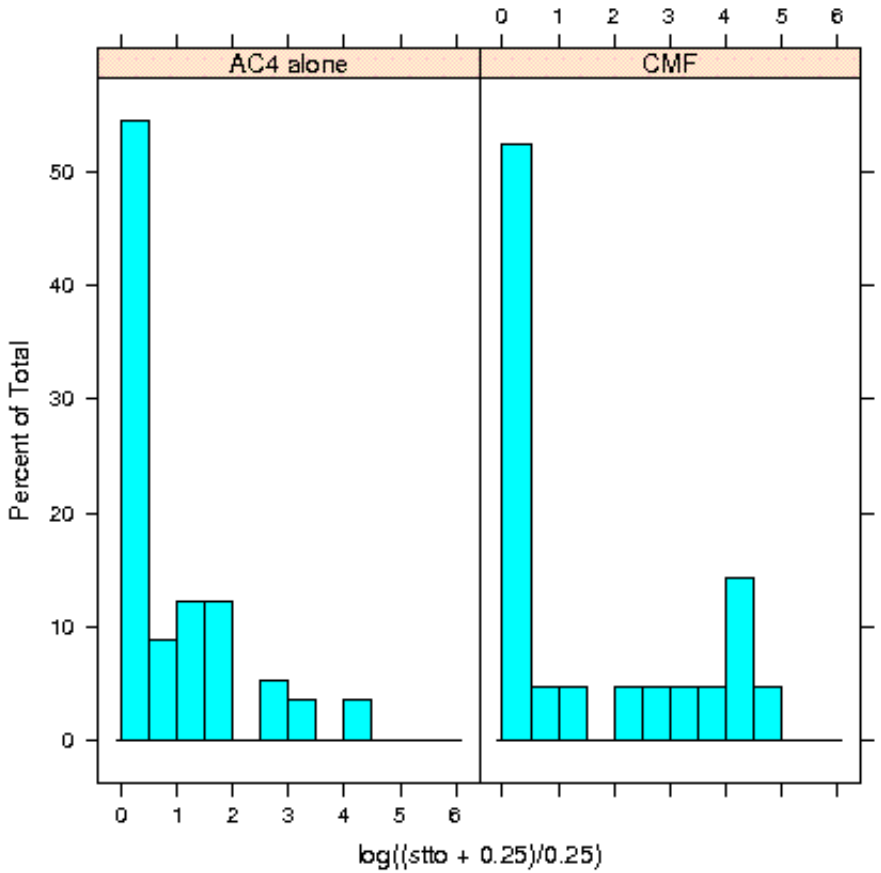
Continuous TTO inference

- 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 (grouped data)
    - observed outcomes are discrete (1 day, 1 month, 3 mths, . . .
    - pre-assign 'scores'
      - t-test, score TTO levels using log
      - rank tests scores are the o.s. under a distribution
      - break ties?

Comparisons by scores of two treatments

AC4 (n=57) vs CMF (n=21)

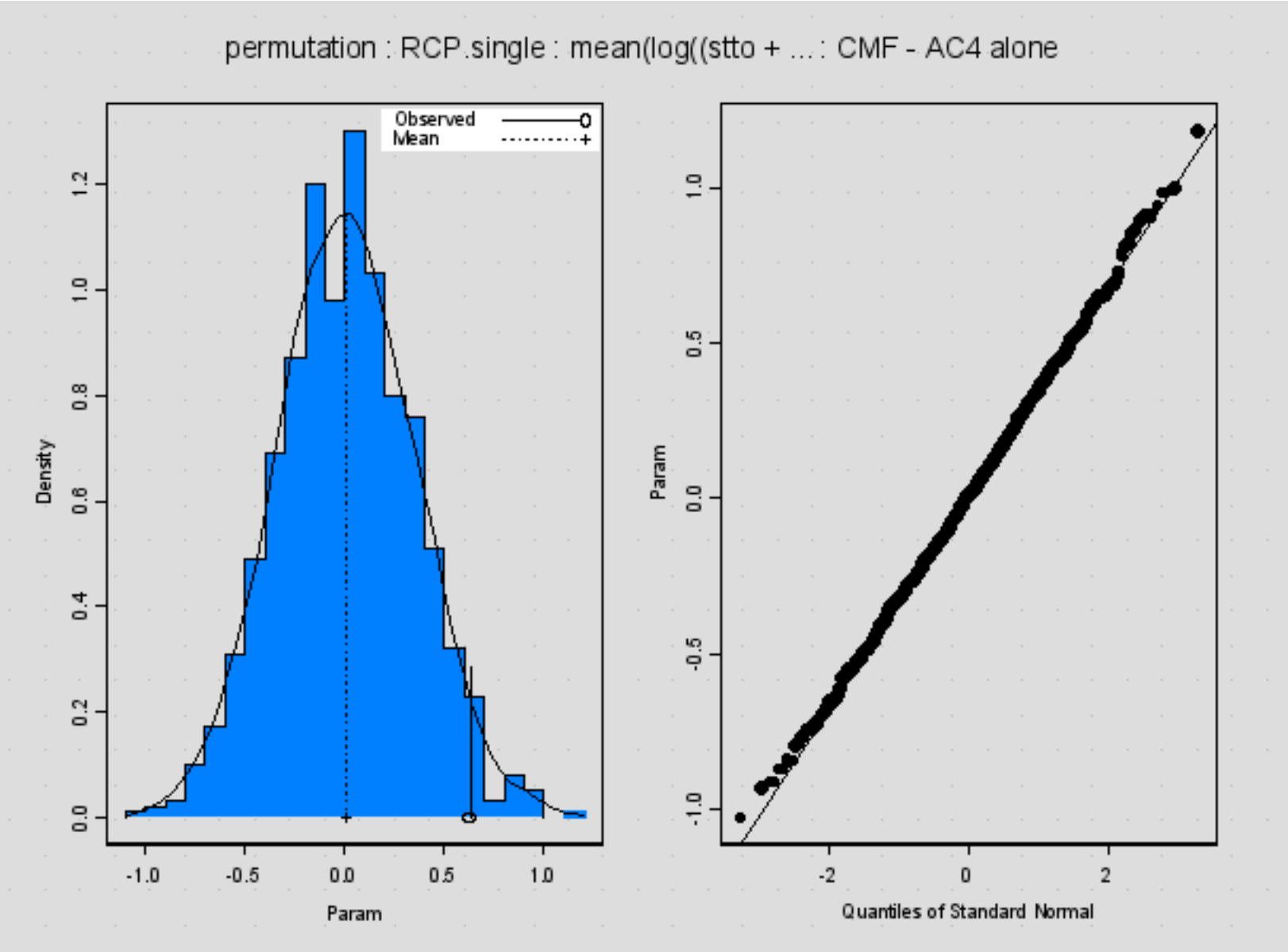




Displayed by scoring: (a) logT; (b) log[(T+0.25)/0.25].

Effect of ties on P-values

statistic = mean difference in log((STTO+0.25)/0.25)



'log' analysis


```
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
```

Inconsistent findings

- Ad hoc analysis by t-test of log(1+TTO/0.25) suggests a treatment difference.
- Wilcoxon-Mann-Whitney, Normal scores tests are the standard analysis of TTOs (NS: P=0.57);
- Ad hoc analysis by jittering to break grouping ties.
- Logrank test (as in survival analysis) is an alternative;
- Logrank test P-value?
- Effect of ties in Cox PH models?

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

Size & power calculations

- Exact conditional size (under the null hypothesis),
- exact conditional power (under an alternative);
- estimated by crude Monte-Carlo method (generation of 10,000 permuted data sets);
- test rejection cutoff for nominal significance level, using an asymptotic (normal) distribution of finite population sampling without replacement;
- See Hilton & Mehta, Biometrics 1993, Rabee et al., SMMR, 2003

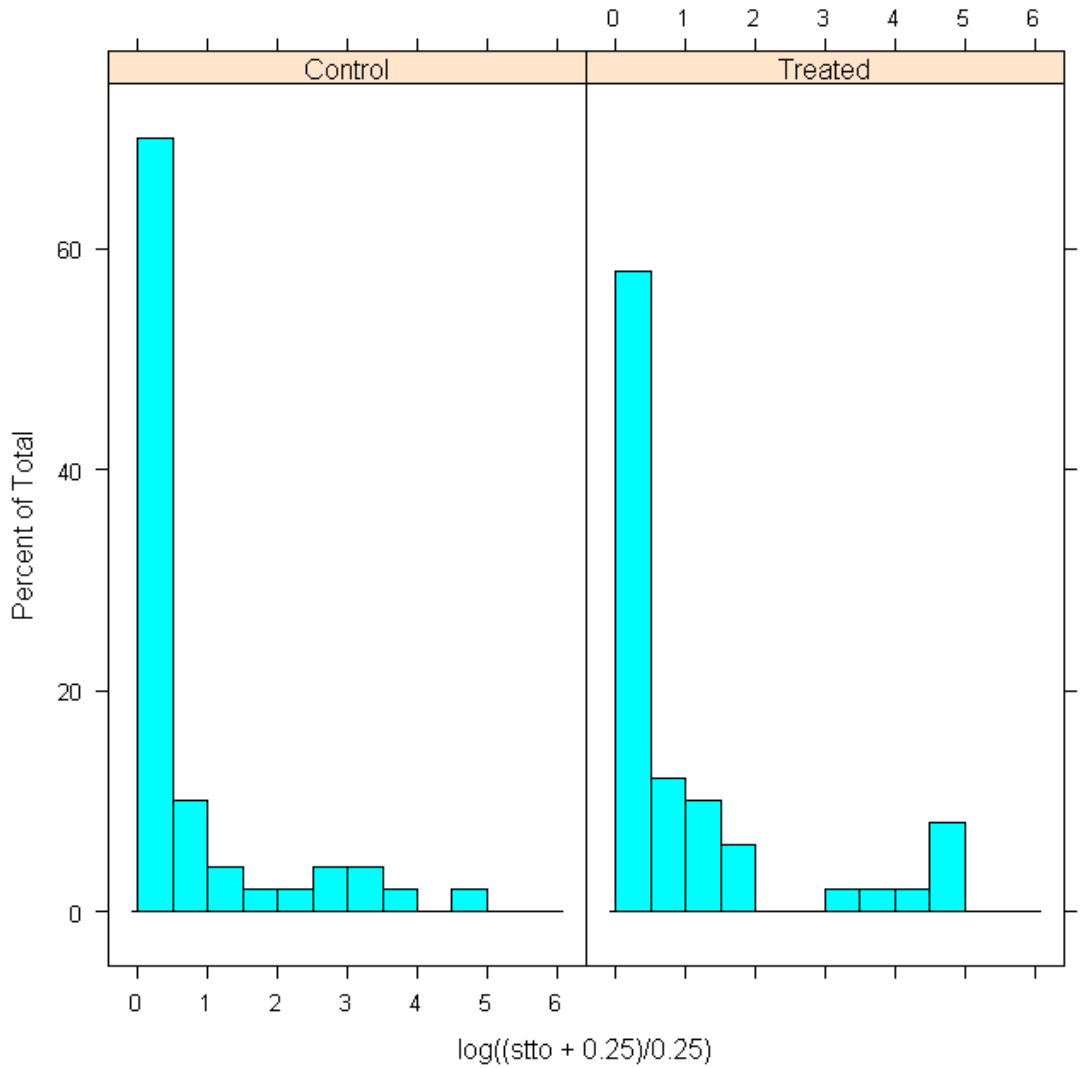
Simulated data: NULL effect

Null effect: type 1 error rates

Table 1: Rejection rates under the null versus nominal significance

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

Location shift (log) alternative

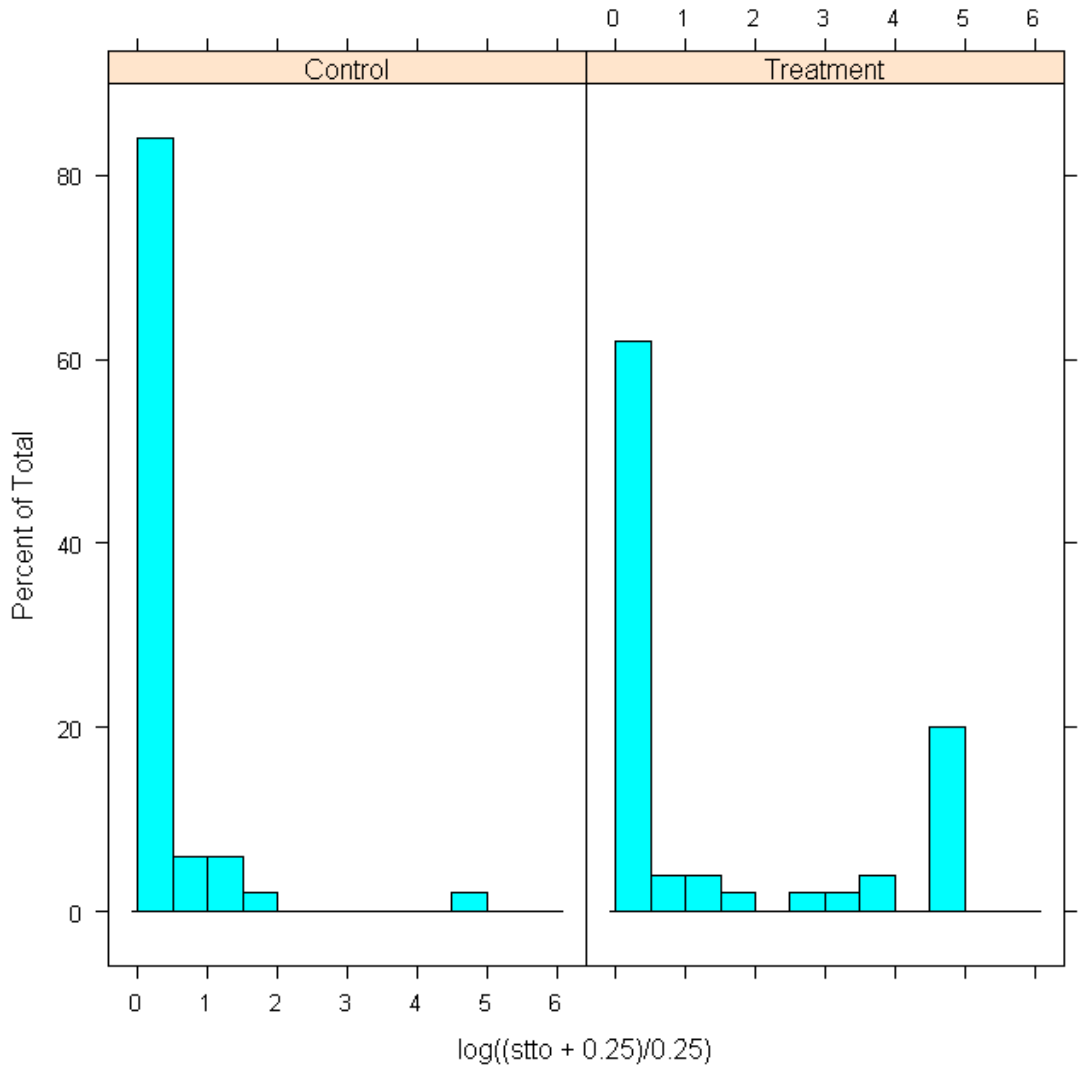


Effect: location shift

Table 1: Power: SHIFT alternative

Equal sample sizes	Effect: SHIFT		0.5*SD	
N=100	Rejection rate*			
Test	%	%	%	%
alpha	0.1	1	5	10
	%	%	%	%
Wilcoxon RS	14	36	62	73
Normal scores	14	37	63	74
(unconditional)	15	38	63	74
Logrank (exponential scores)	13	32	57	68
t-test (permutation)	6	25	50	63
(unconditional)	7	25	50	63
*N=10000 replicated data sets				

Polarisation



Effect: polarisation

Table 2: Power: POLAR alternative

Equal sample sizes	Effect: POLARISE		2.0*SD	
N=100	Rejection rate*			
Test	%	%	%	%
alpha	0.1	1	5	10
	%	%	%	%
Wilcoxon RS	0.6	5	15	24
Normal scores	1.2	8	22	32
(unconditional)	2	8	22	32
Logrank (exponential scores)	5	21	43	57
t-test (permutation)	6	25	50	63
(unconditional)	10	36	63	75
*N=10000 replicated data sets				

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
- Poor power in heterogeneous groups, relative to permutation t-test and logrank test
- To do: mixture model analysis
- log rank tests for TTO and PTO data!
- agrees with ad hoc analysis: $\log(1 + T/0.25)$.

R-notes

coin: Conditional Inference

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	Pearson's χ^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	Spearman's test
wilcoxsign_test	Wilcoxon-Signed-Rank test

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.

Power Calculation code snippets

```
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
```

Power Calculation code snippets

```
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)
}
```

Power Calculation code snippets

```
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

References

- Hajek, Sidak, Sen, Theory of rank tests, Wiley, 1998.
- Kolassa, SIM, 1995
- Lesaffre, SIM, 1993
- Simes and Coates, JNCI Monographs, 2001
- Leung, 2007 (BCA)
- Hilton and Mehta, Biometrics, 1993
- Varice, Weil, Exact non-null distributions of rank statistics, Communications in Statistics, 2001

Power Calculation code snippets

```
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)))
```