



## **Society for Clinical Trials 31<sup>st</sup> Annual Meeting**

### **Workshop P3 Design and Interim Data Analysis of Clinical Trials**

**Sunday, May 16, 2010  
8:00 AM - 12:00 PM  
Harborview Ballroom E**



# Society for Clinical Trials

Pre-Conference Workshop Evaluation  
Baltimore, Maryland  
May 16, 2010

## WORKSHOP 3 – Design and Interim Data Analysis of Clinical Trials

1. Overall, did the subject context of this workshop meet your expectations and needs?  
Yes ( ) No ( )

If yes, in what way? If no, why not? \_\_\_\_\_

\_\_\_\_\_

2. Was the content of this workshop of value to you personally or on the Job?  
Yes ( ) No ( )
3. Was the content of the workshop: New ( ) New/Review ( ) Review ( )
4. The level and complexity of this workshop was: Too elementary ( ) Correct ( ) Too advanced ( )

**Please complete the following questions by circling the appropriate description using the rating scale listed below.**

**1 = excellent 2 = very good 3 = good 4 = fair 5 = poor**

5. Rate the extent to which this workshop:
- |   |   |   |   |   |   |
|---|---|---|---|---|---|
| a. Presented content clearly  | 1 | 2 | 3 | 4 | 5 |
| b. Allowed sufficient time for discussion and audience participation            | 1 | 2 | 3 | 4 | 5 |
| c. Provided useful information  | 1 | 2 | 3 | 4 | 5 |
| d. Utilized appropriate teaching methods, i.e., audiovisual, handouts, lectures | 1 | 2 | 3 | 4 | 5 |
6. Please rate each workshop faculty member:

Name	Knowledge of Subject	Organization/Delivery
K. K. Gordon Lan	1 2 3 4 5	1 2 3 4 5

1. Are you currently working in a clinical trial? (Yes) (No)

2. What is your job title? \_\_\_\_\_

3. Do you have any suggested topics for workshops at future meetings? If so, please list below:

\_\_\_\_\_  
\_\_\_\_\_

4. What aspect of the workshop did you like best?

\_\_\_\_\_  
\_\_\_\_\_

5. What aspect of the workshop would you change if this workshop were offered again?

\_\_\_\_\_  
\_\_\_\_\_

6. Additional Comments: \_\_\_\_\_

\_\_\_\_\_

# Design and Interim Data Analysis of Clinical trials

Gordon Lan, Johnson & Johnson

The Society for Clinical Trials 31<sup>st</sup> Annual Meeting  
Short Course Part 1  
May 16, 2010  
Baltimore, Maryland

## Interim data analysis

In 1960s and 1970s, many NIH-sponsored clinical trials were design as fixed studies. Data were monitored periodically by a group of scientists (Policy Advisory Board, Data safety and monitoring Board, Data Monitoring Committee). Statistical procedyes

Group sequential methods: Pocock (1977), O'Brien-Fleming (1979), Alpha spending functions (1983).

1980s: NIH started to use sequential designs for clinical trials.

1990s: The pharmaceutical industry started to use sequential design.

## Outline

1. Design of a fixed study (Chapter 2)  
Sample size estimation for a fixed design;  
B-value and the trend of the data.
2. Conditional power and predictive power (Chapter 3)
3. Group sequential methods (GSM)  
Classical GSM, Pocock 1977 and O'Brien-Fleming 1979  
Spending function  
Computing boundary and drift for sample size evaluation  
(Chapter 4)
4. Survival data analysis (Appendix) ---Lecture 2

Reference: *Statistical Monitoring of Clinical Trials: A unified approach*  
By Proschan, Lan and Wittes; Springer 2006.

Software available -- <http://www.medsch.wisc.edu/landemets/>  
(window version, 1d98)

## Distribution theory for a one-sample problem

Why start with a one-sample problem?

The mathematics behind a one-sample problem is very straightforward and easy to understand. Extension of the idea (not the mathematics) to the two-sample case needs only slight modifications.



Compare a new treatment T with a control treatment C.

Suppose  $Y_T \sim N(\mu_T, \sigma^2)$  and  $Y_C \sim N(\mu_C, \sigma^2)$ , then  
 $X = (Y_T - Y_C) / \sigma\sqrt{2} \sim N(\Delta, 1)$ , where  $\Delta = (\mu_T - \mu_C) / \sigma\sqrt{2}$ .

In other words, if we pair responses  $Y_T$  and  $Y_C$ , and “standardized” the difference by  $X = (Y_T - Y_C) / \sigma\sqrt{2}$ , then the 2-sample problem becomes an 1-sample problem.

$X$  has mean  $\Delta$  and variance 1. A positive response  $\Delta$  favors the new treatment. To simplify our discussion, we assume the  $X$ 's are normally distributed. The theory applies to responses different from normal if the sample size is “LARGE”.

“Trend of the data” – The partial sum process

Let  $X_1, X_2, \dots, X_n, \dots$  be iid  $N(\Delta, 1)$ . Define  $S_n = X_1 + X_2 + \dots + X_n$ .

Then  $ES_n = n \Delta$  and  $\text{Var}(S_n) = n$ .

The expectation is a linear function of the variance.

**Good news:** This linear relationship gives us an easy tool to “predict” the future outcome conditional on accumulating data.

**Bad news:** The prediction depends on the treatment effect  $\Delta$  which is unknown to us. In addition, we evaluate  $Z$  instead of  $S$ .

$$Z(n) = S_n / \sqrt{n}.$$

Example: To design a clinical trial, we test the hypothesis  
 $H_0: \Delta = 0$  versus  $H_a: \Delta > 0$ .

If we take an one-sided  $\alpha = 1.96$  and 85% power ( $\beta = 0.15$ ),  
how many patients do we need?

How many patients do we need to reach a 85% power?

$$Z(N) = S_N / \sqrt{N}.$$

$$EZ(N) = N \Delta / \sqrt{N} = \sqrt{N} \Delta.$$

Let us assume that the treatment effect  $\Delta = \Delta_1 = 0.2$ .

Solve for N from the equation:

$$EZ(N) = \sqrt{N} \Delta_1 = z_\alpha + z_\beta = 1.96 + 1.04 = 3, N = 225.$$

For a given  $\Delta = \Delta_1$ , the drift parameter is  $\theta = E(Z) = \sqrt{N}\Delta_1$ .

To evaluate sample size N, solve N from

$$\theta = E(Z) = \sqrt{N}\Delta_1 = z_\alpha + z_\beta = 1.96 + 1.04 = 3.0.$$

---

$\Delta_1 =$	0.5	<b>0.2</b>	0.1	0.05	0.01	$\rightarrow 0$
N =	36	<b>225</b>	900	3600	90000	$\rightarrow \infty$

---

A fundamental equation for sample size evaluation:

$$\theta = EZ = z_\alpha + z_\beta.$$

(Change  $z_\beta$  to **0.84** for 80% and **1.28** for 90%.)

Comments:

$\theta = EZ(N) = \sqrt{N}\Delta$  is called the Drift parameter.

This drift parameter, depending on  $N$  and  $\Delta$ , is unknown to us in practice. However, for any given value of  $\Delta$ ,  $\theta = EZ(N)$  is known.

Under the null hypothesis,  $\Delta = 0 = \theta$ .

Under alternative  $\Delta = \Delta_1 > 0$ ,  $\theta > 0$ .

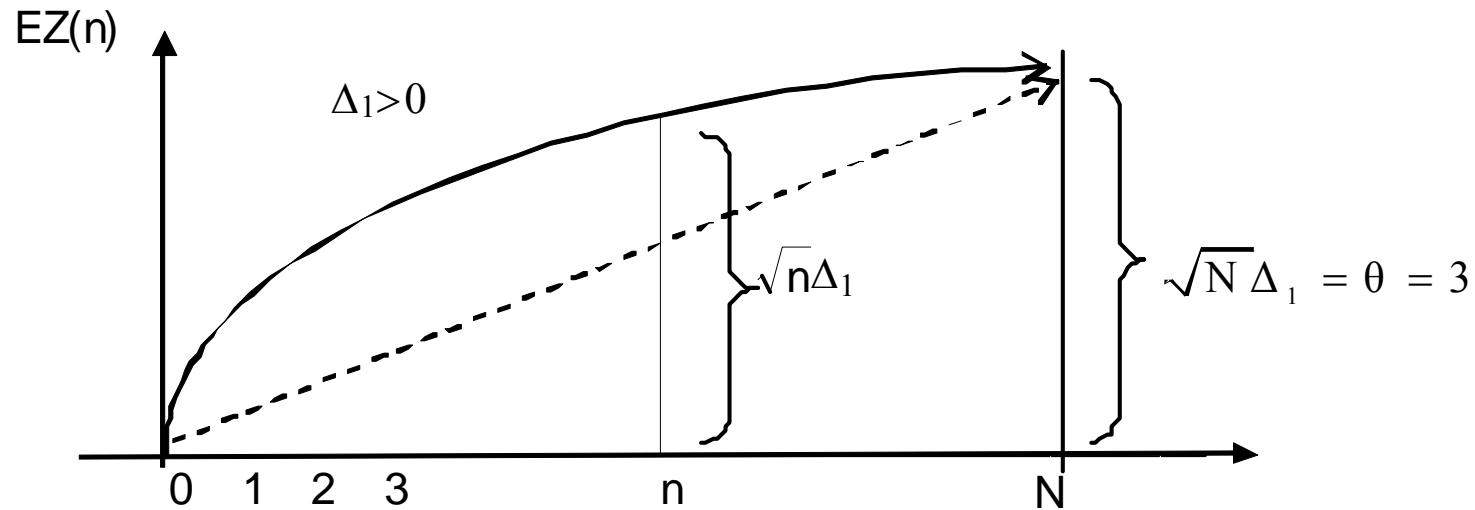
(If the treatment is beneficial, the drift is positive.)

The trend of the data =  $\Delta$   
(Partial sums)

Interim analysis

	$X_1, X_2, \dots, X_n,$	$X_{n+1}, \dots, X_N$
Unconditional	random	random
Conditional	fixed	random
	$S_N$	$= S_n + (S_N - S_n)$
Conditional	$ES_N$	$= n\Delta + (N-n)\Delta$
	$Var(S_N)$	$= n + (N-n)$
Unconditional	$E_C(S_N)$	$= S_n + (N-n)\Delta$ ( $\Delta=?$ )
Variance	$Var_C(S_N)$	$= N-n$

The trend of the data =  $\theta$  (B-values)



$(n, Z(n)) \rightarrow (\tau, Z_\tau) \rightarrow (\tau, B_\tau)$  where  $\tau = n/N$  &  $B_\tau = Z_\tau \sqrt{\tau}$ .

## Example:

---

$$\Delta=0.2, \theta = \sqrt{N}\Delta=3 \Rightarrow N=225.$$

$$\text{when } n=90, Z_{.4}=2.846$$

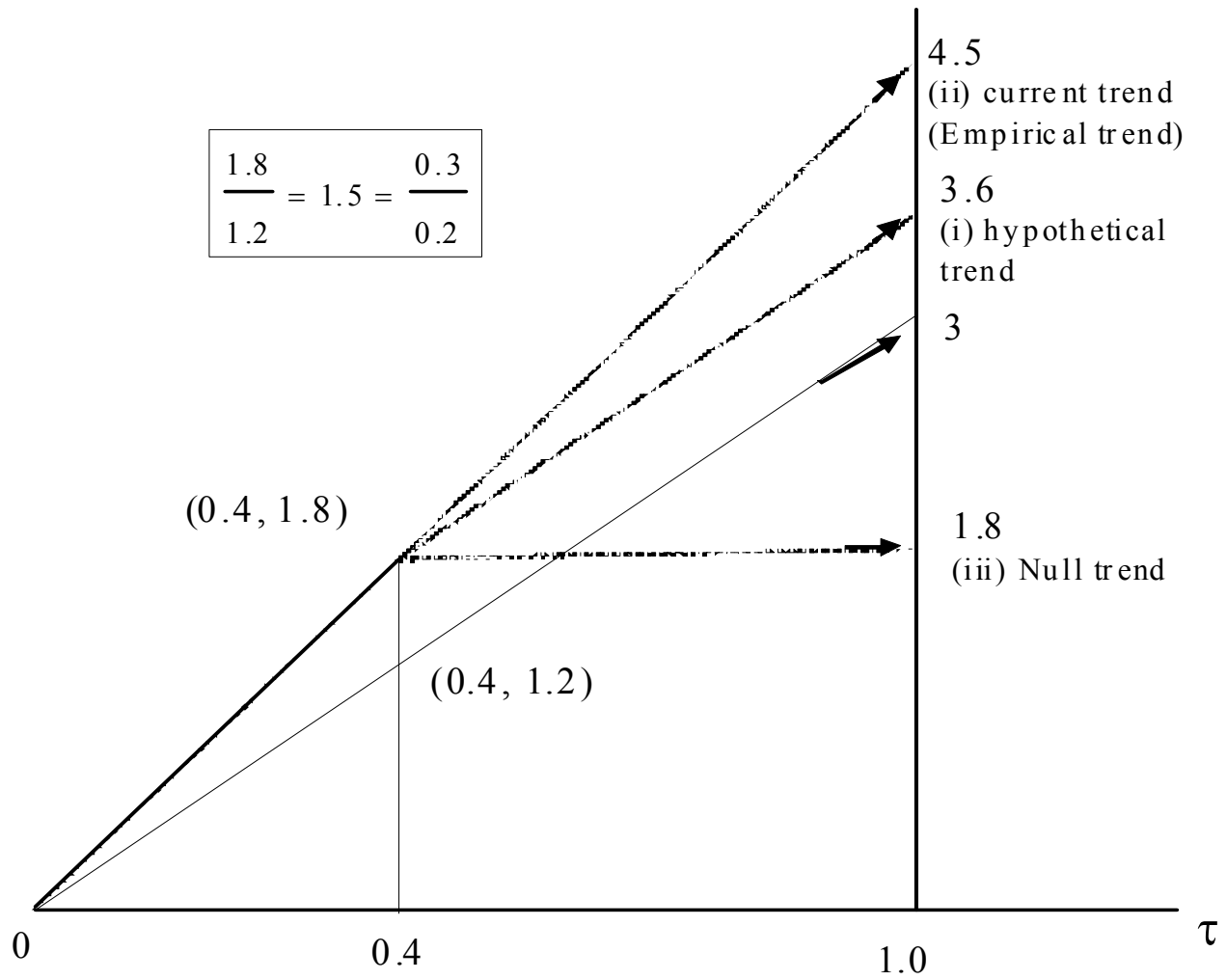
$$CP(\theta) = P(Z_1 \geq 1.96 | Z_{.4}=2.846, \theta) = ?$$

$$\tau = 90/225 = 0.4; B_{\tau} = 2.846 \sqrt{\tau}=1.8.$$

Note that  $Z_1 = B_1$ . To evaluate  $CP(\theta)$ , let us do it in TWO steps.

1. Find  $E_C(Z_1)$ .
2. Find  $P_C(Z_1 \geq 1.96)$ .





$$\begin{aligned}
\text{(i)} \quad & P[Z_1 = B_1 \geq 1.96 | B_4 = 1.8, \theta = 3] \\
& = P\left[\frac{Z_1 - 3.6}{\sqrt{.6}} \geq \frac{1.96 - 3.6}{\sqrt{.6}} \mid B_4 = 1.8, \theta = 3\right] \\
& = P[N(0,1) \geq -2.12] \\
& = 0.9830
\end{aligned}$$

$$\begin{aligned}
\text{(ii)} \quad & P[Z_1 \geq 1.96 | B_4 = 1.8, \theta = 4.5] \\
& = P\left[\frac{B_1 - 4.5}{\sqrt{.6}} \geq \frac{1.96 - 4.5}{\sqrt{.6}} \mid B_4 = 1.8, \theta = 4.5\right] \\
& = P[N(0,1) \geq -3.28] \\
& = .9995
\end{aligned}$$

$$\begin{aligned}
\text{(iii)} \quad & P[B_1 \geq 1.96 | B_4 = 1.8, \theta = 0] \\
& = P\left[N(0,1) \geq \frac{1.96 - 1.8}{\sqrt{.6}}\right] \\
& = P[N(0,1) \geq 0.21] \\
& = 0.4168
\end{aligned}$$

## Two-sample comparisons, Comparison of two means


$$H_o: \mu_x = \mu_y \quad \text{vs} \quad H_a: \mu_x > \mu_y$$


$$X_1, X_2, \dots, X_M \quad \text{iid} \quad N(\mu_x, \sigma^2) \quad N = M + M = 2M$$


$$Y_1, Y_2, \dots, Y_M \quad \text{iid} \quad N(\mu_y, \sigma^2)$$

$$Z_{(N)} = \frac{\bar{X}_M - \bar{Y}_M}{\sigma \sqrt{1/M + 1/M}} = \frac{\sum_1^M X_i - \sum_1^M Y_i}{\sigma \sqrt{M+M}} = \frac{\sum_1^M (X_i - Y_i)}{\sigma \sqrt{N}}$$

$$\theta = EZ_{(N)} = \frac{\mu_x - \mu_y}{\sigma} \sqrt{\frac{N}{4}} = \frac{\mu_x - \mu_y}{\sigma} \sqrt{N} \sqrt{\frac{1}{2} \times \frac{1}{2}}$$

  
**treatment  
difference**

  
**sample  
size**

  
**two-  
sample  
factor**

Set  $EZ = \theta = z_\alpha + z_\beta$  and solve for N.

After  $m_1$  X's and  $m_2$  Y's have been observed,

$$Z_\tau = \frac{\bar{X}_{m_1} - \bar{Y}_{m_2}}{\sigma \sqrt{\frac{1}{m_1} + \frac{1}{m_2}}} \quad \text{where } \tau = \frac{\left(\frac{1}{m_1} + \frac{1}{m_2}\right)^{-1}}{\left(\frac{1}{M} + \frac{1}{M}\right)^{-1}}.$$

Note that  $\tau \approx \tau^* = \frac{m_1 + m_2}{N} = \frac{n}{N}$ .

## Examples:

$$N = 800 = 400 + 400.$$

$$(i) \ n = 500 = 250+250, \ \tau^* = .625, \ \tau = .625.$$

$$(ii) \ n = 500 = 230+270, \ \tau^* = .625, \ \tau = .621.$$

$$(iii) \ n = 500 = 200+300, \ \tau^* = .625, \ \tau = .600.$$

$$N = 900 = 600 + 300.$$

$$(i) \ n = 540 = 360+180, \ \tau^* = .600, \ \tau = .600.$$

$$(ii) \ n = 540 = 345+195, \ \tau^* = .600, \ \tau = .623.$$

$$(iii) \ n = 500 = 300+240, \ \tau^* = .600, \ \tau = .667.$$

## References:

Lan and Zucker 1993 Stat. in Medicine

Lan, Reboussin and DeMets 1994 Comm. Stat. A

We pick a  $\Delta_1$  to design a study. During interim, what if the observed  $\Delta^\wedge$  is quite different from  $\Delta_1$ ?

## Sample size re-estimation

What is the test statistic after sample size re-estimation

A hot topic in adaptive design and will not be covered in this short course.

For a fixed design, the use of conditional power to stop early for benefit WILL INFLATE the alpha level.

Reason: A fixed design spends all alpha at the end.  $P(Z_1 \geq 1.96) = 0.025$ .

Any interim analysis for benefit has to spend additional alpha.

## Early stopping of clinical trials

Use CP (or PP) for futility stopping.

To stop early for benefit, we use a group sequential design.

Stop early for benefit if a one-sided upper boundary is crossed.



## Repeated Significance Tests & Group Sequential Methods (One-sided version of the original article)

Armitage, et al. 1969 JRSS

$$P(Z_1 \geq 1.96) = .025$$

$$P(Z_{.5} \geq 1.96 \text{ or } Z_1 \geq 1.96) = 0.043$$

---

K	1	2	3	4	5	...	$\infty$
$P(\text{Type I error})$	.025	.043	.055	.064	.072	...	1

Pocock boundary (1977 Biometrika)  
For  $\alpha = .025$  (one-sided)

K	1	2	3	4	5	8	12
c.v. or boundary	1.96	2.16	2.28	2.34	2.41	2.50	2.58

## O'Brien-Fleming boundary (1979 Biometrics)

$$K = 5, \quad \alpha = .025 \quad (\text{one sided})$$

$$P \left( \begin{array}{l} B_{.2} \geq 2.04 \quad \text{or} \quad B_{.4} \geq 2.04 \quad \text{or} \quad B_{.6} \geq 2.04 \\ \text{or} \quad B_{.8} \geq 2.04 \quad \text{or} \quad B_1 \geq 2.04 \end{array} \right) = .025$$

$$B_{.2} \geq 2.04 \quad \Leftrightarrow \quad Z_{.2} \geq \frac{2.04}{\sqrt{.2}} = 4.56$$

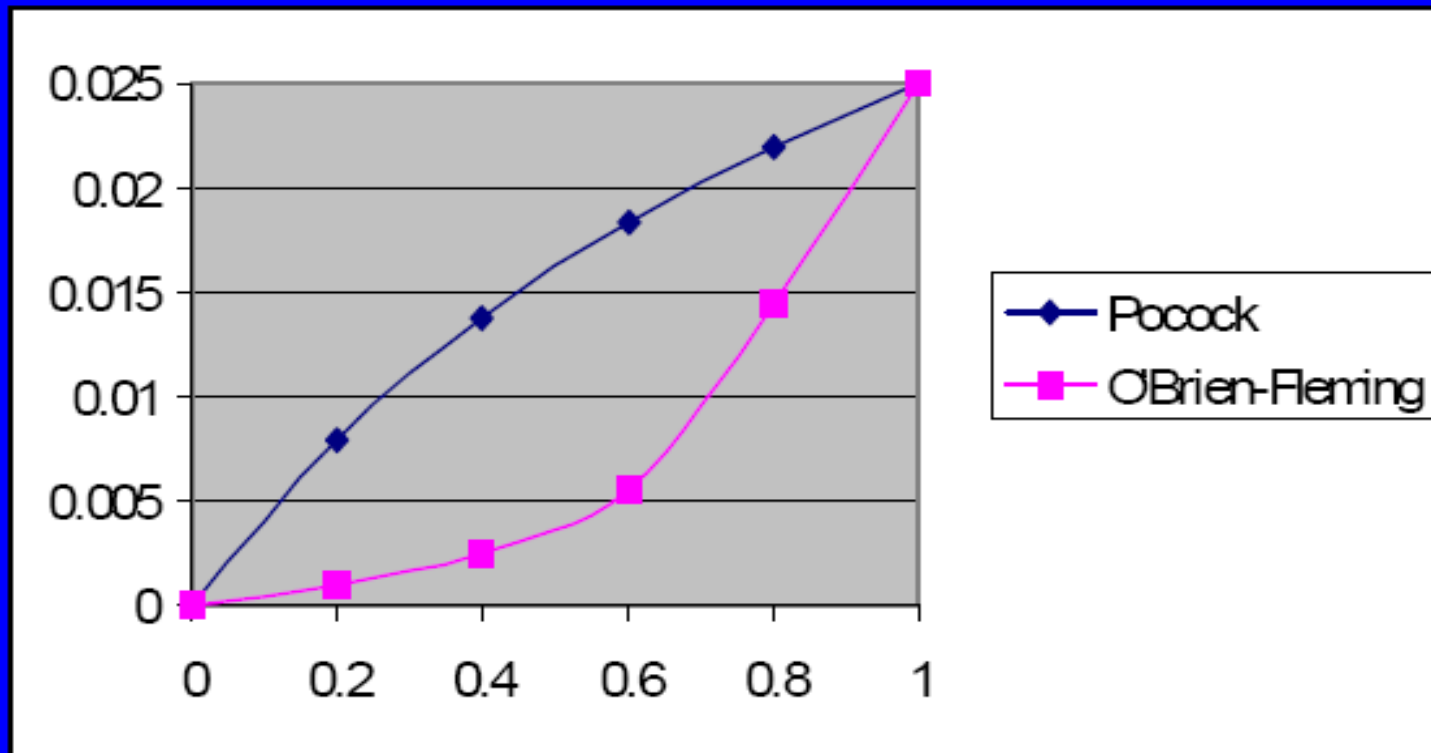
$$B_{.4} \geq 2.04 \quad \Leftrightarrow \quad Z_{.4} \geq \frac{2.04}{\sqrt{.4}} = 3.23$$

$$B_{.6} \geq 2.04 \quad \Leftrightarrow \quad Z_{.6} \geq \frac{2.04}{\sqrt{.6}} = 2.63$$

$$B_{.8} \geq 2.04 \quad \Leftrightarrow \quad Z_{.8} \geq \frac{2.04}{\sqrt{.8}} = 2.28$$

$$B_1 \geq 2.04 \quad \Leftrightarrow \quad Z_1 \geq 2.04$$

$k = 5, \alpha = .025$



$$\alpha_1^*(\tau) = 2 - 2\Phi(Z_{0.5\alpha} / \sqrt{\tau}) \approx \text{O'Brien-Fleming}$$

$$\alpha_2^*(\tau) = \alpha \log(1 + (e-1)\tau) \approx \text{Pocock}$$

## Alpha spending functions

$\alpha^*(\tau)$  is a non- decreasing function defined on  $[0,1]$   
with  $\alpha^*(0) = 0$ ,  $\alpha^*(1) = \alpha$ .

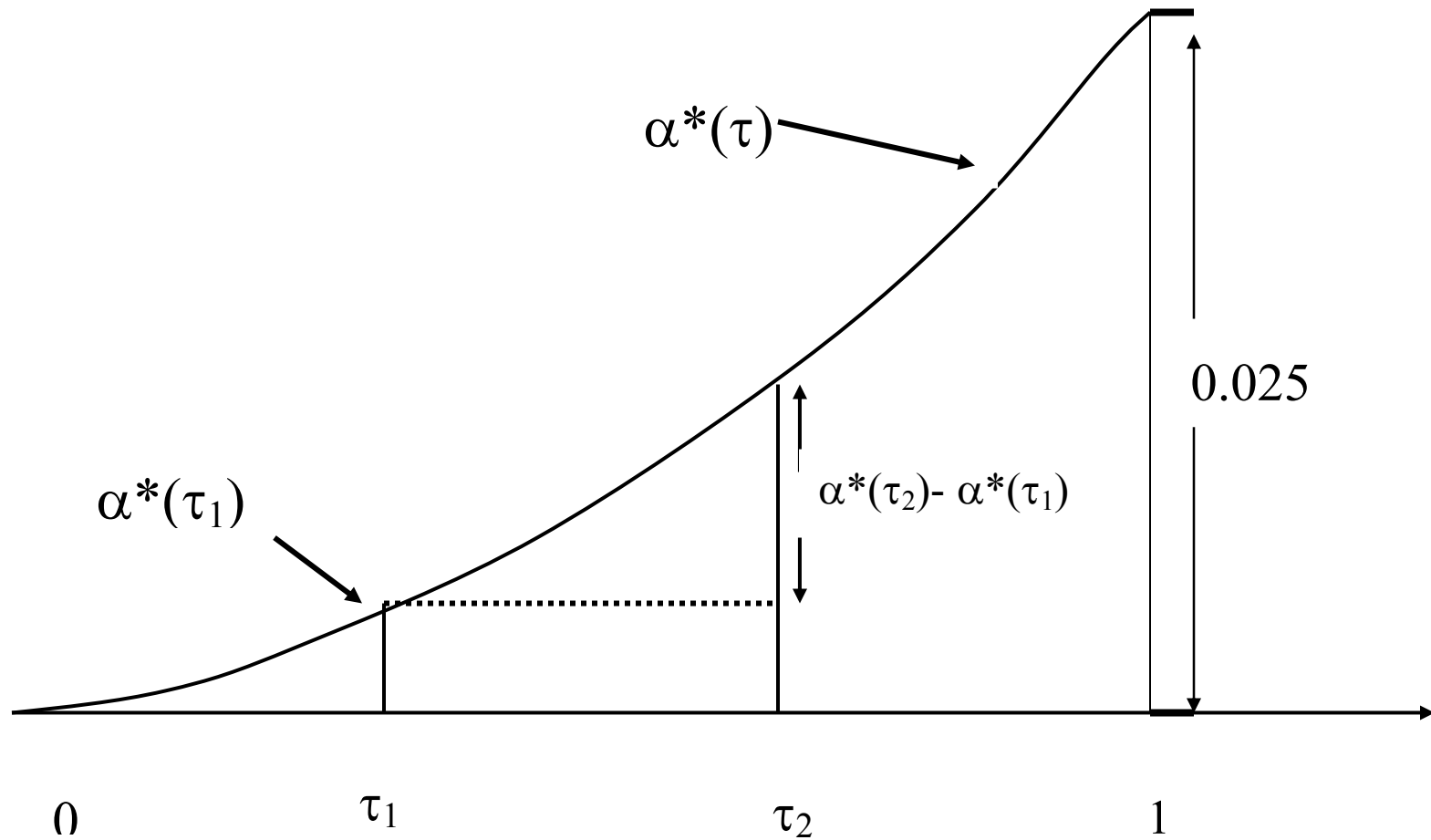
At  $\tau_1$  , find  $b_1$  such that (under  $H_0$ )  $P(Z\tau_1 \geq b_1) = \alpha^*(\tau_1)$ .

At  $\tau_2$  , find  $b_2$  such that  $P(Z\tau_1 < b_1 \text{ and } Z\tau_2 \geq b_2) = \alpha^*(\tau_2) - \alpha^*(\tau_1)$ .

At  $\tau_3$ , .....

(Wisconsin software)

# Alpha-spending function



For a fixed design, sample size  $N$  can be evaluated from  
 $\theta = EZ = z_\alpha + z_\beta$ .

When  $N$  is derived from the above equation for a sequential design, the power of the design will be less than  $1-\beta$ , or, we need a larger  $\theta$  to reach the power  $1-\beta$ .

The value of  $\theta$  required to reach desired power for a sequential design depend on the boundary chosen. ([Wisconsin software](#))

Comparison of two means,  $N=N/2+N/2$ :

Treatment effect  $\Delta = (\mu_X - \mu_Y)/\sigma$ ,

$$\theta = EZ = E \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{1/0.5N + 1/0.5N}} = \Delta \sqrt{\frac{N}{4}}.$$

For  $\Delta = 0.1$ ,  $\theta = 3 \Rightarrow N=3600$  for 85% power.

Suppose we modify this to a sequential design.

Trade off:

We may stop the trial early, but we will lose power.

(An example will be given later using the Wisconsin software.)



The Wisconsin software (window version) can do the following:

- I. Bounds – compute boundary from a spending function.
- II. Drift: Compute drift parameter  $\theta$  from given boundary and desired power.
- III. Probability: Probability of boundary crossing for a given boundary and drift.
- IV. Confidence (intervals).

This window version is a front end of the interactive program ld98. Unfortunately, the window version is NOT compatible with Office 2007. (???)

## I. Bounds (choose “Bounds” under “Compute”)

Interim analyses K (default=5) → may be changed

Information times (def = equally spaced) → user input

Test boundary (def = two sided symmetric)

→ one-sided, two-sided asymmetric

Overall alpha (def = 0.05) → 0.025, e.g.

Spending function (def = O’Brien-Fleming)

→ Pocock, Power family, Hwang-Shih-DeCani

Truncation bounds (def=no) (an example will be given later)

Example: Change to one-sided 0.025, hit “calculate” button.

Output:

	Time	Upper Bound	Nominal upper alpha	Cum alpha
1	0.2	4.8769	.00000	.00000
2	0.4	3.3569	.00039	.00039
3	0.6	2.6803	.00368	.00381
4	0.8	2.2898	.01102	.01221
5	1	2.0310	.02113	.02500

Change “Truncate bounds” from “no” to “yes” with truncation point = 3.5, hit “calculate” button.

New boundary values are:

3.5000, 3.5000, 2.6893, 2.2915, 2.0317  
(4.8769, 3.3569, 2.6803, 2.2898, 2.0310)

### III. Probability (choose “Probability” under “Compute”)

For the boundary derived above, if the drift parameter (input) is  $\theta = EZ = z_\alpha + z_\beta = 3$ , then the last column of the output looks like this: .01545

.06083

.35986

.65963

.84223 = power

We need to increase the  $\theta$  value to attain power=0.85.

## II. Drift (choose “Drift” under “Compute”)

Same input as before, input power = 0.85, hit “calculate” button.  
Output: drift = 3.033.

For given treatment effect, solve for sample size from equation

$$EZ = z_{\alpha} + z_{\beta} = 3.033.$$

## Constant boundaries (Classical GSM) versus spending functions

Example: Pocock boundary

$\tau =$	.4	0.7	1.0
Constant	2.194		2.194
Boundary	2.274	2.274	2.274

What is the boundary value when  $\tau = 0.4$ ?

Spending	2.224	2.305	2.310
Function	2.224		2.165

### Conditional power example:

$\tau=$	0.2	0.4	0.6	0.8	1.0
boundary (Z)	3.50	3.50	2.69	2.29	2.03
Z-value	0.93	1.37	2.36		
Boundary (B)	1.565	2.214	2.084	2.048	2.03
B-value	0.416	0.866	<b>1.828</b>		

$$\theta^{\wedge}=1.828/0.6=3.047$$

$\tau=$	0.2	0.4	1
boundary(B)	2.048-1.828=0.220	2.03-1.828=0.202	
boundary (Z)	0.492	0.319	-8
CP	0.808	0.958	1



Quick review of B-value: Conditional power depends on  $\tau$ ,  $B_\tau$  and the drift parameter  $\theta$ .

$$CP(\tau, B_\tau, \theta) = \Phi\left[\frac{B_\tau + (1-\tau)\theta - 1.96}{\sqrt{1-\tau}}\right] \quad (\text{Eq 1})$$

- (i) It is easy to evaluate.
- (ii) It communicates easily to clinicians.

It seems to be natural to take  $\theta = \theta_E = B_\tau/\tau$ .  
Under this empirical drift,

$$CP(\tau, B_\tau, \theta_E) = \Phi\left[\frac{B_\tau/\tau - 1.96}{\sqrt{1-\tau}}\right] \quad (\text{Eq 2})$$

However,  $\theta_E = B_\tau/\tau$  is only a point estimate of  $\theta$ .  
If we consider  $\theta_E$  as random, CP becomes PP.

From conditional power (CP) to predictive power (PP)

$\tau=0.4, Z_\tau=1.6.$

-2.0 SD	0.0433
-1.5 SD	0.1353
-1.0 SD	0.3124
-0.5 SD	0.5491
<b>Empirical</b>	<b>0.7690</b>
+0.5 SD	0.9112
+1.0 SD	0.9750
+1.5 SD	0.9950
+2.0 SD	0.9993

## Weighted average of CP ( $\tau=0.4$ , $Z_\tau=1.6$ )

-2.0 SD	0.0433	0
-1.5 SD	0.1353	0
-1.0 SD	0.3124	0.1
-0.5 SD	0.5491	0.2
<b>Empirical</b>	<b>0.7690</b>	<b>0.4</b>
+0.5 SD	0.9112	0.2
+1.0 SD	0.9750	0.1
+1.5 SD	0.9950	0
+2.0 SD	0.9993	0

$$0.1 \times 0.3124 + 0.2 \times 0.5491 + 0.4 \times 0.7690 + 0.2 \times 0.9112 + 0.1 \times 0.9750 = \mathbf{0.7284}$$

How can we choose another **fixed** drift, say  $\theta_M$ , to replace  $\theta_E$  to evaluate the chance of a positive study?

- (i)  $\theta_M$  depends on  $\theta_E$ ; and
- (ii)  $\theta_M$  depends on the “accuracy” of  $\theta_E$  as a point estimate of  $\theta$ .

**Do we expect  $\theta_M \geq \theta_E$  or  $\theta_M \leq \theta_E$ ?**

## Predictive power (considers $\theta$ as random)

Note that  $\theta_E = B_\tau/\tau$  is a point estimate of  $\theta$ . If we consider  $\theta$  as random with distribution function  $G$

$$\begin{aligned} \text{PP} &= \text{PP}[\tau, B_\tau, G(\theta)] = \int \text{CP}(\tau, B_\tau, \theta) dG(\theta) \\ &= \int \text{CP}(\tau, B_\tau, \theta) g(\theta) d\theta \end{aligned}$$

(Note that we did not introduce a prior distribution and went directly to the posterior distribution of  $\theta$ .)

A reasonable choice of  $G$  (for a fixed  $n$ )

Since  $\bar{X}_n$  is  $N(\mu, 1/n)$ ,

let us consider  $\mu$  to be  $N(\bar{X}_n, 1/n)$ .

This is equivalent to  $\theta \sim N(\theta_E, 1/\tau)$ .

Conceptually, this is similar to calling

$[\bar{X}_n \mp 1.96\sqrt{1/n}]$  a 95% c.i. for  $\mu$ .

If  $G^*$  is taken to be  $N(\theta_E, 1/\tau)$ , then

$$PP(\tau, B_\tau, G^*) = \Phi\left[\frac{(\theta_E - 1.96)\sqrt{\tau}}{\sqrt{1-\tau}}\right].$$

Compare this expression with

$$CP(\tau, B_\tau, \theta_E) = \Phi\left[\frac{\theta_E - 1.96}{\sqrt{1-\tau}}\right].$$

Reference: Lan KKG, Hu P, Proschan MA (2009) “A conditional power approach to the evaluation of predictive power.” *Statistics in Biopharmaceutical Research*; 1: 131-136.

$$G^* \sim N(\theta_E, 1/\tau)$$

$$PP(\tau, B_\tau, G^*) = \Phi\left[\frac{(\theta_E - 1.96)\sqrt{\tau}}{\sqrt{1-\tau}}\right]$$

If we modify the empirical drift  $\theta_E$  to

$$\theta_M = \frac{(1-\sqrt{\tau})(B_\tau + 1.96\sqrt{\tau})}{(1-\tau)\sqrt{\tau}}, \text{ then}$$

$$CP(\tau, B_\tau, \theta_M) = PP(CP(\tau, B_\tau, \theta_{G^*})).$$



In other words, if we replace the empirical drift  $\theta_E$  by  $\theta_M$ , the conditional power becomes the predictive power.

By doing this, we don't have to introduce PP as an integral to the clinicians.

Quick summary:

Under the empirical trend  $\theta_E = B_\tau / \tau$ :

If  $CP > 50\%$ ,  $CP > PP > 50\%$ .

If  $CP < 50\%$ ,  $CP < PP < 50\%$ .

If we replace the empirical drift  $\theta_E = B_\tau / \tau$  by the modified drift  $\theta_M$ , then  $CP = PP$ .

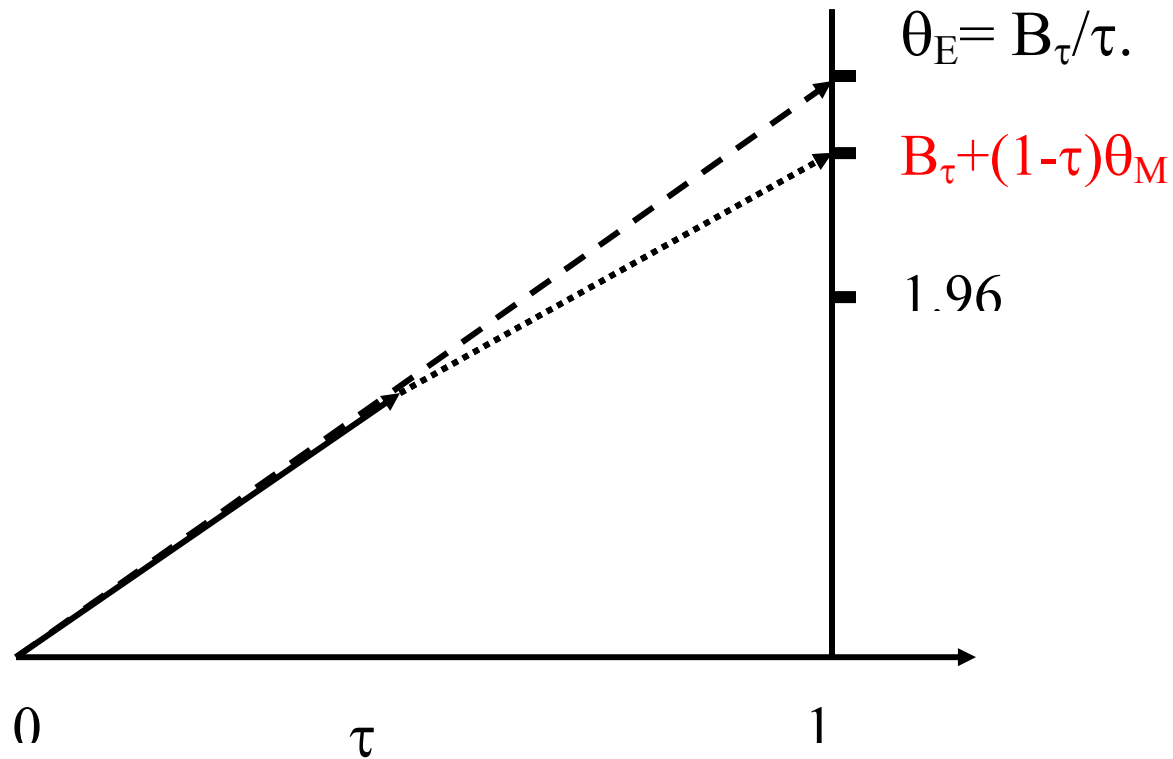
\* $E_C Z_1 = B_\tau + (1 - \tau)\theta_M$  is somewhere between 1.96 and  $\theta_E$ .

\*The critical value 1.96 may be replaced by any other positive number.

$$\theta_E = B_\tau + (1-\tau)\theta_C = B_\tau/\tau.$$

When  $\theta_E > 1.96$ ,  $CP > PP > 50\%$ .

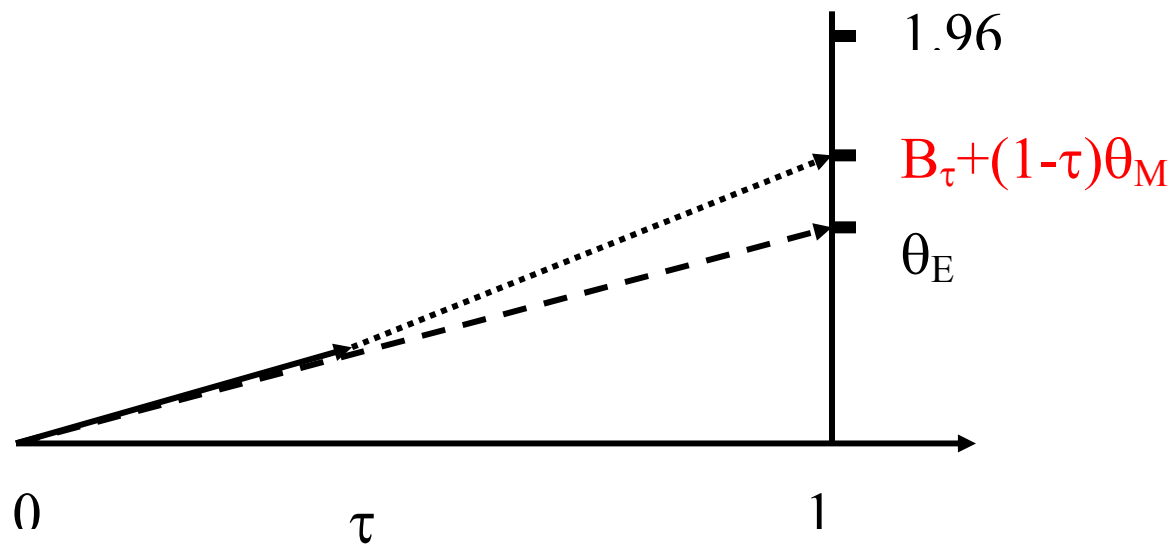
There is a drift  $\theta_M < \theta_E$  so that  $CP(\tau, B_\tau, \theta_M) = PP$ .



$$\theta_E = B_\tau + (1-\tau)\theta_E = B_\tau/\tau.$$

When  $\theta_E < 1.96$ ,  $CP < PP < 50\%$ .

There is a drift  $\theta_M > \theta_E$  so that  $CP(\tau, B_\tau, \theta_M) = PP$ .



# Comparison of Survival Distributions: From Theory to Practice

K. K. Gordon Lan, Johnson & Johnson

The Society for Clinical Trials 31<sup>st</sup> Annual Meeting  
Short Course Part 2  
Baltimore, Maryland

May 16, 2010

# Outline

1. Brief review of some stochastic processes  
Z (one-sample), Z (two-sample) and Z(logrank)
2. Life table, hazard, Kaplan-Meier
3. Linear rank statistics; logrank
4. Mantel-Haenszel procedure
5. Proportional Hazards Assumption,  
practical problems, misinterpretation.
6. Sample size estimation for a survival trial
7. (Time permitting) Wilcoxon statistic, Mann-Whitney statistic, U-statistics

## Why start with a one-sample problem?

The mathematics behind a one-sample problem is very straightforward and easy to understand.

Extension of the idea (not the mathematics) to the two-sample case needs only slight modifications.

## One-sample hypothesis testing:

Let  $X_1, X_2, \dots$  be iid  $N(\Delta, 1)$ .

Test  $H_0: \Delta = 0$  versus  $H_a: \Delta > 0$ .

The test is  $Z(N) = (X_1 + X_2 + \dots + X_N) / \sqrt{N}$  and we reject  $H_0$  if  $Z(N) \geq 1.96$ .

During interim analysis with  $n$  observations, we may compute  $Z(n) = (X_1 + X_2 + \dots + X_n) / \sqrt{n}$ .

$$E[Z(N)] = N\Delta / \sqrt{N} = \Delta\sqrt{N}.$$

$\{ Z(n) \}$  is a stochastic process.



Change of scale: Define  $\tau = n/N$ . Re-write  $Z(n)$  as  $Z_\tau$ .

Suppose we are going to evaluate  $Z$  three times at 0.3, 0.8 and 1.0. Then  $\{Z_{0.3}, Z_{0.8}, Z_{1.0}\}$  is a discrete stochastic process.

Stochastic process versus final  $Z$ -value:

\*Suppose  $Z_{0.3}$  is observed, can we use it to predict  $Z_{1.0}$ ?

\*In a sequentially designed study, the DSMB evaluates the interim  $Z$  and make decision to modify the study design or stop the study early.

One-sample case:  $\{Z_{0.3}, Z_{0.8}, Z_{1.0}\}^1$

Two-sample case:  $\{Z_{0.3}, Z_{0.8}, Z_{1.0}\}^2$

Survival studies (logrank):  $\{Z_{0.3}, Z_{0.8}, Z_{1.0}\}^S$

$$\{Z_{0.3}, Z_{0.8}, Z_{1.0}\}^1 \sim \{Z_{0.3}, Z_{0.8}, Z_{1.0}\}^2 \sim \{Z_{0.3}, Z_{0.8}, Z_{1.0}\}^S$$

Under  $H_0$ .

Under  $H_a$  if Proportional Hazards Assumption is valid.

What is logrank test? What if the PHA is violated?

These topics will be discussed later.

From one-sample to two-sample:

$$\begin{aligned} & (X_1 + \dots + X_n) - (Y_1 + \dots + Y_n) \\ &= (X_1 - Y_1) + \dots + (X_n - Y_n) \\ &= D_1 + D_2 + \dots + D_n \end{aligned}$$

Mathematically, it is easy to understand why a one-sample process  $\{Z_\tau\}$  is similar to two-sample process  $\{Z_\tau\}$  if the sample sizes for  $X$  and  $Y$  are the same.

What if they are different?

In practice, are the observations  $(X, Y, D)$  iid?

1. The “sicker” patients get into the study earlier.
2. Modification of the inclusion/exclusion criteria affects the iid’ness.
3. The clinical centers may need a learning period to administer a new procedure.

Keep this in mind when you design a clinical trial.

In the one-sample case:  $E[Z(N)] = N\Delta/\sqrt{N} = \Delta\sqrt{N}$ .

In the two-sample case,  $EZ(N) = \frac{\mu_X - \mu_Y}{\sigma} \sqrt{\frac{N}{4}} = \Delta \sqrt{\frac{N}{4}}$ .

$EZ = \Delta \sqrt{(\text{information})}$       **(DRIFT PARAMETER)**

In the one-sample case,  $N = \text{sample size} = \text{information}$ .

In the two-sample case,  $\text{information} = N/4$ .

In the survival setting, we may compare two means of survival times, OR, use a linear rank statistic to compare two survival distributions.

## Sample size and power for a one-sample problem

Let  $X_1, X_2, \dots$  be iid  $N(\Delta, 1)$ .

Test  $H_0: \Delta = 0$  versus  $H_a: \Delta > 0$ .

$Z(N) = (X_1 + X_2 + \dots + X_N) / \sqrt{N}$ .

Then  $E[Z(N)] = N\Delta / \sqrt{N} = \Delta\sqrt{N}$ .

Therefore, power  $\uparrow$  with  $N$  if  $\Delta > 0$ .

More patients (information)

→ more power (iid observations, survival???)PHA)

What happens if the  $X$ 's are independent but the means vary?  
 $X_1 \sim N(\Delta_1, 1)$ ,  $X_2 \sim N(\Delta_2, 1)$ ..... are independent.

For example,  $X_1 \sim N(1, 1)$ ,  $X_2 \sim N(1, 1)$ ,  $X_3 \sim N(1, 1)$ ,  
 $X_4 \sim N(0, 1)$ ,  $X_5 \sim N(0, 1)$

$Z(1) \sim N(1, 1)$ ,  $Z(2) \sim N(1.41, 1)$ ,  $Z(3) \sim N(1.73, 1)$ ,  
 $Z(4) \sim N(1.5, 1)$ ,  $Z(5) \sim N(1.34, 1)$ .

Rule of thumb: If  $\mu_{n+1} > 0.5 \left( \sum_1^n \Delta_i \right) / n$ ,

then  $EZ(n+1) > EZ(n)$ .

It can be shown that if the proportional hazards assumption (PHA) is violated,  $\{Z_{\log\text{rank}}\}$  computed sequentially over time behaves like  $\{Z(n)\}$  with different  $\Delta$ 's.



# Life table

t	N	d	q	p	S
0-1	1000	50	0.05	0.95	0.95
1-2	950	38	0.04	0.96	0.912
2-3	912	30	0.0329	0.9631	0.882
3-4	882				
⋮					



Hazard (discrete version)

# Life Table

Proportional hazards assumption  
(discrete version)

Placebo

t	q	p	S
0-1	0.05	0.95	0.95
1-2	0.04	0.96	0.912
2-3	0.0329	0.9631	0.882
3-4			
⋮			

↓ treatment effect =20% risk reduction

Treatment

t	q*	p*	S*
0-1	0.05 x 0.8	0.96	0.96
1-2	0.04 x 0.8	0.968	0.9293
2-3	0.0329 x 0.8	0.97368	0.90482
3-4			
⋮			

## From discrete to continuous

t	N	d	q	p	S
0-1	1000	50	0.05	0.95	0.95
1-2	950	38	0.04	0.96	0.912
2-3	912	30	0.0329	0.9631	0.882
3-4	882				
⋮					

Life Table  
 $t \rightarrow t + \Delta t$

$q(t)$   
 $\lambda(t) \Delta t$

Consider time interval from  $t$  to  $t+\Delta t$  ( $\Delta t$  may not be 1):

$$q(t) = P(t \leq T < t + \Delta t \mid t \leq T),$$

and the hazard function is defined as

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid t \leq T)}{\Delta t}.$$

Survival time  $T$  with distribution function  $F$ , density function  $f$ .

$F(t) = P[T \leq t]$ ;  $f(t) = dF(t)/dt$ ;  $S(t) = 1 - F(t) = P[T > t]$ .

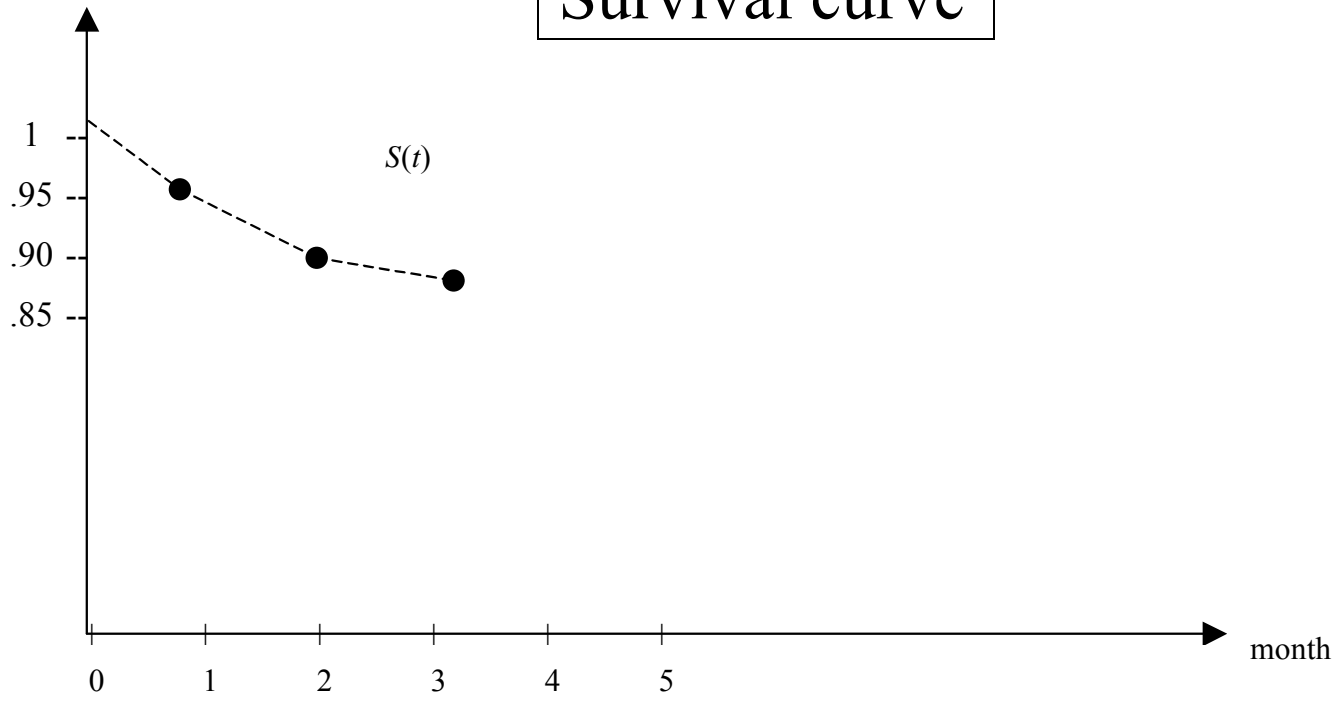
$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid t \leq T)}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{S(t) \Delta t} = \frac{f(t)}{S(t)}\end{aligned}$$

# Life Table (simplified version)

t	N	d	q	p	S
0-1	1000	50	0.05	0.95	0.95
1-2	950	38	0.04	0.96	0.912
2-3	912	30	0.0329	0.9631	0.882
3-4	882				
⋮					

↑  
**Conditional Probability**  
**= The discrete version of**  
**hazard**

# Survival curve



If  $\lambda(t) = \lambda$ , then  $S(t) = e^{-\lambda t}$ ;  
 $F(t) = 1 - e^{-\lambda t}$  and  $f(t) = \lambda e^{-\lambda t} = \lambda(t)S(t)$ .

Rewrite  $S(t) = e^{-\lambda t}$  as  $e^{-\int_0^t \lambda dt} = e^{-\Lambda(t)}$ .

$\Lambda(t)$  is called the cumulative hazard function.

In general, if the hazard function is  $\lambda(t)$ , then

$$S(t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(u) du} ;$$

$$F(t) = 1 - S(t) = 1 - e^{-\int_0^t \lambda(u) du} \text{ and}$$

$$f(t) = \lambda(t)S(t)$$



## Compare mean (average) survival times

Treatments: A and B

Average survival time of treatment A patients = 5.2 years.

Average survival time of treatment B patients = 4.7 years.

### Censor data

Parametric distributions (Exponential; Weibull,...)

Non-parametric: (linear) rank tests

Time-to-event measurements (waiting time)

Event: death, 2<sup>nd</sup> MI, kidney transplant ....

## Ranks in linear rank statistics

Tennis players	Mary	Bill	Becky	Gordon	Pam
Rank	1	2	3	4	5

Team1: Bill and Gordon

Team2: Mary, Becky and Pam

Which team is better?

## Scores in linear rank statistics

Tennis players	Mary	Bill	Becky	Gordon	Pam
Rank	1	2	3	4	5
Price (Score)	\$2.50	\$1.25	\$1.00	\$0.25	\$0.00
Average score = \$1					
Centered score	\$1.50	\$0.25	\$0.00	-\$0.75	-\$1.00
Net winnings					

Linear rank statistic:  $S = \sum (\text{Net winnings of team 2 players})$   
 $= -\sum (\text{Net winnings of team 1 players})$

Net winnings of the girls' team:

$$\$1.50 + \$0 + (-\$1.00) = \$0.50.$$

Net winnings of the boys' team:  $-\$0.50$ .

The girls' team is "better".

Is it significantly better? ( $Z = S/\sqrt{\text{Variance}}$ )

What if we change from one set of scores to another set of scores?

(Winner takes all!)

## Ranks in a linear rank statistic

Player	Mary	Bill	Becky	Gordon	Pam
Rank in skill	1	2	3	4	5

## Survival analysis and video games:

Player	Mary	Bill	Becky	Gordon	Pam
Survival time	695	477	354	321	217
Rank of order statistics	5	4	3	2	1

## From tennis competition to video game

Tennis competition: Rank #1 is the best.

Video game: Rank #1 is the WORST.

In a video game, in addition to the ranking of the players, we also measure the time-to-event or survival times.

In clinical trials, Team 1 = Control, Team 2 = New compound or new treatment procedure.

## Savage scores

	Scores		Centered Scores
Pam	(217)	$1/5$	-1
Gordon	(321)	$1/5 + 1/4$	-1
Becky	(354)	$1/5 + 1/4 + 1/3$	-1
Bill	(477)	$1/5 + 1/4 + 1/3 + 1/2$	-1
Mary	(695)	$1/5 + 1/4 + 1/3 + 1/2 + 1/1$	-1

Boys's team:  $(1/5 + 1/4 - 1) + (1/5 + 1/4 + 1/3 + 1/2 - 1)$   
= - \$ 0.27

Girls' team: \$0.27 (Savage statistic)

The Girls' team is better.

## Censored survival times

Player	Mary	Bill	Becky	Gordon	Pam
Survival time	695	477	354+	321	217
Rank	4,5	3,4	3,4,5	2	1

Also, how can we modify the scores?

	Scores	Centered Scores
Pam	(217) $1/5$	-1
Gordon	(321) $1/5 + 1/4$	-1
Becky	(354) $1/5 + 1/4 + 1/3$	-1
Bill	(477) $1/5 + 1/4 + 1/3 + 1/2$	-1
Mary	(695) $1/5 + 1/4 + 1/3 + 1/2 + 1/1$	-1



In medical studies, we may have 1000 patients and 850 censored survival times.

Also, evaluation of the  $\text{Var}(S)$  will be VERY messy.

## From scores to payments:

	Scores		Centered Scores
Pam	(217)	1/5	-1
Gordon	(321)	1/5 + 1/4	-1
Becky	(354)	1/5 + 1/4 + 1/3	-1
Bill	(477)	1/5 + 1/4 + 1/3 + 1/2	-1
Mary	(695)	1/5 + 1/4 + 1/3 + 1/2 + 1/1	-1

	Scores		Centered Scores
Pam	(217)	1/5	-1
Gordon	(321)	1/5 + 1/4	-1
Becky	(354)	1/5 + 1/4 + 1/3	-1
Bill	(477)	1/5 + 1/4 + 1/3 + 1/2	-1
Mary	(695)	1/5 + 1/4 + 1/3 + 1/2 + 1/1	-1

## Censored data

	Scores	Centered scores
Pam	(217) $1/5$	-1
Gordon	(321) $1/5 + 1/4$	-1
Becky	(354 <sup>+</sup> ) $1/5 + 1/4$	
Bill	(477) $1/5 + 1/4 + 1/2$	-1
Mary	(695) $1/5 + 1/4 + 1/2 + 1/1$	-1

Net winnings

Boys' team:  $(1/5 + 1/4 - 1) + (1/5 + 1/4 + 1/2 - 1) = - \$ 0.60$

Girls' team: \$0.60 (Savage statistic or **logrank** statistic)

Under proportional hazards model, the (locally) optimal score function is  $\phi(t) = -\log(1-t)$ . There are two sets of scores derived from this score function.

$$\text{Savage score} = \frac{1}{N} + \frac{1}{N-1} + \dots + \frac{1}{N-i+1}$$

$$\text{Logrank score} = -\log\left(1 - \frac{i}{N+1}\right).$$

When N is LARGE, use  $\int \frac{1}{1-x} dx = -\log(1-x)$  to show

Savage score  $\approx$  logrank score.

The Savage statistic and the logrank statistic are asymptotically equivalent.

Alternative: Parametric location shift; Lehmann alternative...  
For a specific alternative, there is an optimal score function  $\phi$  defined on the unit interval. There are two ways to define scores from a given score function:

(1) Approximate scores; (2) Exact scores.

The corresponding two statistics are asymptotic equivalent.

References:

1. Hajek and Sidek (1967). Theory of rank tests. Academic Press, New York.
2. Randles and Wolfe (1979). Introduction to the theory of nonparametric statistics. John Wiley & Sons, New York.

## Reference for the payment approach

1. Lan and Wittes (1985), “Rank tests for survival analysis: A comparison by analogy with games”. *Biometrics* 41, 1063-1069.
2. Proschan, Lan and Wittes (2006).  
Statistical Monitoring of Clinical Trials:  
A Unified Approach. Springer. Appendix 1.

## Mantel-Haenszel Procedure (1959)

Population	Exposed(E)	Unexposed( $\bar{E}$ )	
Disease (D)	A	B	$N_1 = A + B$
No Disease ( $\bar{D}$ )	C	D	$N_2 = C + D$
	$M_1 = A + C$	$M_2 = B + D$	T

$$RR = \frac{P(D | E)}{P(D | \bar{E})} = \frac{A / M_1}{B / M_2} = \frac{0.2}{0.1} = 2$$

$$OR = \frac{P(D | E) / P(\bar{D} | E)}{P(D | \bar{E}) / P(\bar{D} | \bar{E})} = \frac{\frac{A}{M_1} / \frac{C}{M_1}}{\frac{B}{M_2} / \frac{D}{M_2}} = \frac{AD}{BC} = \frac{0.2 / 0.8}{0.1 / 0.9} = 2.25$$

$$(OR)' = \frac{P(E | D) / P(E | \bar{D})}{P(E | \bar{D}) / P(\bar{E} | \bar{D})} = \frac{\frac{A}{N_1} / \frac{C}{N_1}}{\frac{B}{N_2} / \frac{D}{N_2}} = \frac{AD}{BC} = OR$$

Sample	Exposed(E)	Unexposed( $\bar{E}$ )	
Disease (D)	a	b	$n_1 = a + b$
No Disease ( $\bar{D}$ )	c	d	$n_2 = c + d$
	$m_1 = a + c$	$m_2 = b + d$	T

Observations:

(1) For rare diseases,  $RR \approx OR = (OR)'$

$P(D|E) = 0.01$ ,  $P(D|\bar{E}) = 0.005$ ,

$$RR = \frac{0.01}{0.005} = 2,$$

$$OR = \frac{0.01 / 0.99}{0.005 / 0.995} = 2.01.$$

(2) RR may not be estimable in retrospective studies.

(3)  $RR = 1 \Leftrightarrow OR = 1$ ,

$RR > < 1 \Leftrightarrow OR < > 1$ .



## 2 X 2 Tables (partial odds ratio versus odds ratio)

Male	Treatment	Placebo	
R	0.3	0.1	Partial OR
NR	0.7	0.9	= $.3 \times .9 / .1 \times .7 = 3.86$

Female	Treatment	Placebo	
R	0.9	0.7	Partial OR
NR	0.1	0.3	= $.9 \times .3 / .7 \times .1 = 3.86$

Population	Treatment	Placebo	
R	0.6	0.4	
NR	0.4	0.6	OR = $.6 \times .6 / .4 \times .4 = 2.25$

(Example given by Prof. Gary Koch, UNC)

## More examples:

1)

S=1	
400	500
600	1250
OR <sub>S=1</sub> =1.667	

S=2	
70	126
75	225
OR <sub>S=2</sub> =1.667	

1+2	
470	626
675	1475
OR=1.641	

2)

200	197
156	233
OR <sub>S</sub> =1.516	

1000	63
1644	157
OR <sub>S</sub> =1.516	

1200	260
1800	390
OR=1	

3)

40	60
60	90
OR <sub>S</sub> =1	

10	50
50	250
OR <sub>S</sub> =1	

50	110
110	340
OR=1.4	

4)

194	21
706	79
OR <sub>S=1</sub> =1.033	

6	29
94	871
OR <sub>S=2</sub> =1.917	

200	50
800	950
OR=4.75	

5)

110	380
390	2620
OR <sub>S=1</sub> =1.945	

90	20
1410	980
OR <sub>S=2</sub> =3.128	

200	400
1800	3600
OR=1	

$H_0: \phi = 1$  (**partial** odds ratio) vs.  $H_a: \phi \neq 1$

$S=1$		$S=2$		$S=k$
$a_1$		$a_2$		$a_k$
$n_{11}$	$n_{12}$	$n_{21}$	$n_{22}$	$n_{k1}$
$m_{11}$	$m_{12}$	$m_{21}$	$m_{22}$	$m_{k1}$
$N_1$				$N_k$

$$Z_{MH} = \frac{\sum (a_i - E a_i)}{\sqrt{\sum \text{Var}(a_i)}} \quad \text{where} \quad \left\{ \begin{array}{l} E a_i = \frac{m_{i1} n_{i1}}{N_i} \\ \text{Var}(a_i) = \frac{m_{i1} m_{i2} n_{i1} n_{i2}}{N_i^2 (N_i - 1)} \end{array} \right.$$

$$\hat{\phi}_s(MH) = \frac{\sum (a_i d_i / N_i)}{\sum (b_i c_i / N_i)}$$

Partition the time interval under study into many, many very small sub-intervals.

Consider the sub-interval  $[t, t + \Delta t)$ .

When 1 event occurred in this sub-interval:

	$D$	$\bar{D}$	
X	$\delta_t$		$m_t$
Y			$n_t$
	1	$N_t - 1$	$m_t + n_t = N_t$

$\delta_t = 1$  if “event” is X,  
otherwise,  $\delta_t = 0$ .

Observed – Expected =  $\delta_t - m_t/N_t$ .

$\Sigma (O-E) =$  cumulative difference

When 0 event occurred in the sub-interval:

	D	$\bar{D}$	
X			$m_t$
Y			$n_t$
	0	$N_t$	$m_t+n_t=N_t$

Observed – Expected = 0 – 0 = 0.

“Mantel-Haenszel” the 2X2 tables over time.

Ref: Mantel (1966). “Evaluation of survival data and two new rank order statistics arising in its consideration.” *Cancer Chemotherapy Reports* 50,163-170.

The Mantel-Haenszel statistic (assume no ties):

At  $T_{(i)}$ : (concept of at risk, Wilcoxon)

	D	$\bar{D}$	
X	$\delta_i$		$m_i$
Y			$n_i$
	1	$N_i - 1$	$m_i + n_i = N_i$

$\delta_i = 1$  if  $T_{(i)}$  is a X  
 $= 0$  otherwise

$$S = \sum \left( \delta_i - \frac{m_i}{N_i} \right)$$

$$S(t) = \sum_{T_{(i)} \leq t} \left( \delta_i - \frac{m_i}{N_i} \right) \quad \text{(log rank)}$$

When there is no censoring and no ties, the Mantel-Haenszel statistic  $S = S(\infty)$  becomes the Savage statistic.

Ties can be handled.

Time interval  $[t, t+\Delta t)$ :

	D	$\bar{D}$
X	$P_1$	$Q_1$
Y	$P_2$	$Q_2$

When  $\Delta t$  is “very small”,

$Q_1 \cong 1 \cong Q_2$  and

Relative Risk =  $P_1/P_2$

$\cong$  Odds ratio =  $P_1Q_2/P_2Q_1$ .

When  $\Delta t \rightarrow 0$ , Odds ratio  $\rightarrow$  Relative Risk.

Time interval  $[t, t+\Delta t)$ :

	D	$\bar{D}$
X	$P_1 = .02 \times 0.0001$ $= .00002$	$Q_1 = .99998$
Y	$P_2 = .01 \times 0.0001$ $= .00001$	$Q_2 = .99999$

$$HR = .02/.01 = 2; RR = .00002/.00001 = 2$$

$$\approx (.00002/.99998)/(.00001/.99999) = OR$$

$$\text{As } \Delta t \rightarrow 0, HR = RR \rightarrow OR$$



## (1) Early stopping for the comparisons of two means

$$H_o : \mu_x = \mu_y \quad \text{vs} \quad H_a : \mu_x > \mu_y$$

$$X_1, X_2, \dots, X_{1000} \quad iid \quad N(\mu_x, \sigma^2)$$

$$Y_1, Y_2, \dots, Y_{1000} \quad iid \quad N(\mu_y, \sigma^2)$$

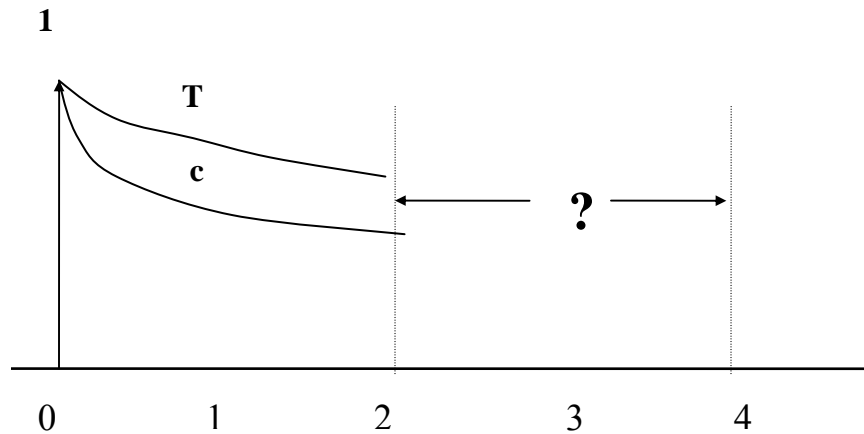
$$X_1, X_2, \dots, X_{500} \quad \text{plus} \quad X_{501}, \dots, X_{1000}$$

$$Y_1, Y_2, \dots, Y_{500} \quad Y_{501}, \dots, Y_{1000}$$

$$Z_{0.5} = \frac{\bar{X}_{500} - \bar{Y}_{500}}{\sigma_{1000} \sqrt{\frac{1}{500} + \frac{1}{500}}}$$

$$Z_1 = \frac{\bar{X}_{1000} - \bar{Y}_{1000}}{\sigma_{2000} \sqrt{\frac{1}{1000} + \frac{1}{1000}}}$$

## (2) Comparisons of two survival curves



Proportional Hazards Model :  $\frac{\lambda_c(t)}{\lambda_T(t)} = r$

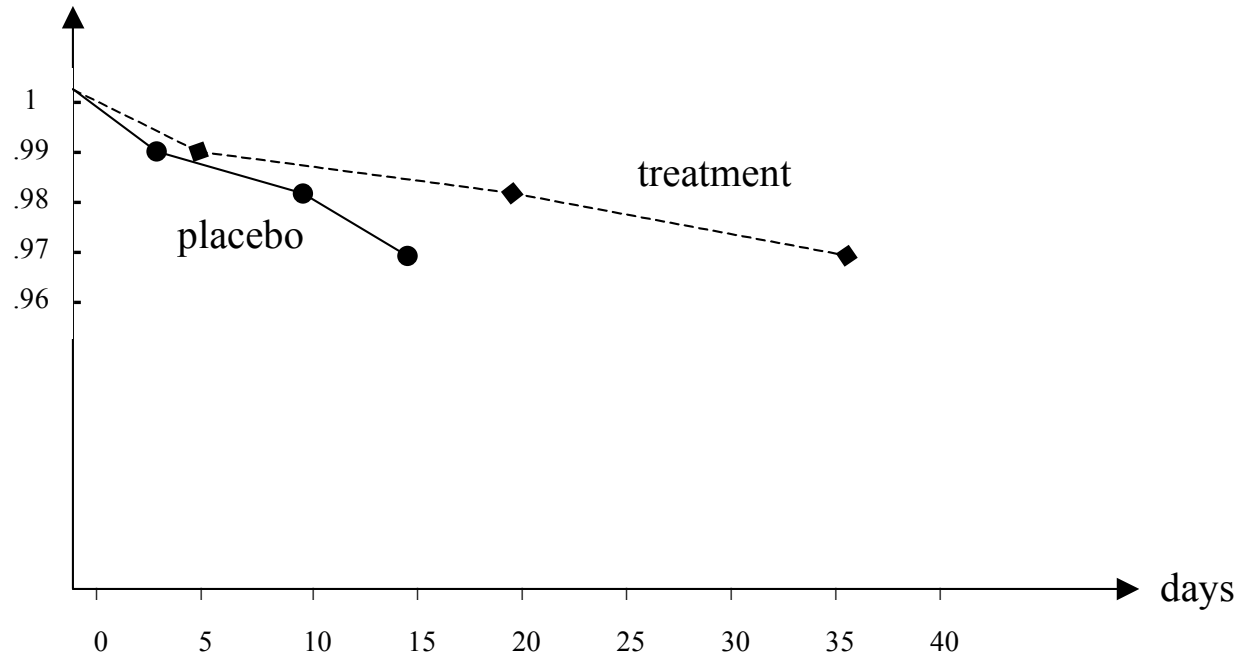
Primary Endpoint = Mortality (**which treatment is better?**)

The BHAT (Beta-Blocker Heart Attack Trial) was a randomized, double-blind multicenter clinical trial of propranolol versus placebo in patient. The primary objective was to determine if long-term administration of propranolol in this population would result in a significant reduction in total mortality over the follow-up period. (BHAT Preliminary Report, JAMA 81)

“The treatment is better than the placebo if it reduces mortality.” (**12, 18 or 36 months?**)

Definition #1: The treatment delays the occurrence of death.

	placebo	treatment
1%	3 days	5 days
1%	10 days	20 days
1%	15 days	35 days



Definition #2: The treatment reduces hazard (all the time).

Use the Kaplan-Meier Curves to define better:  
A better treatment prolongs life.



Use logrank test to define better:  
A better treatment reduces hazard.

BHAT: Kaplan-Meier (T) > Kaplan-Meier (P) for 30 months.  
The two hazard functions crossed each other several times. (later)

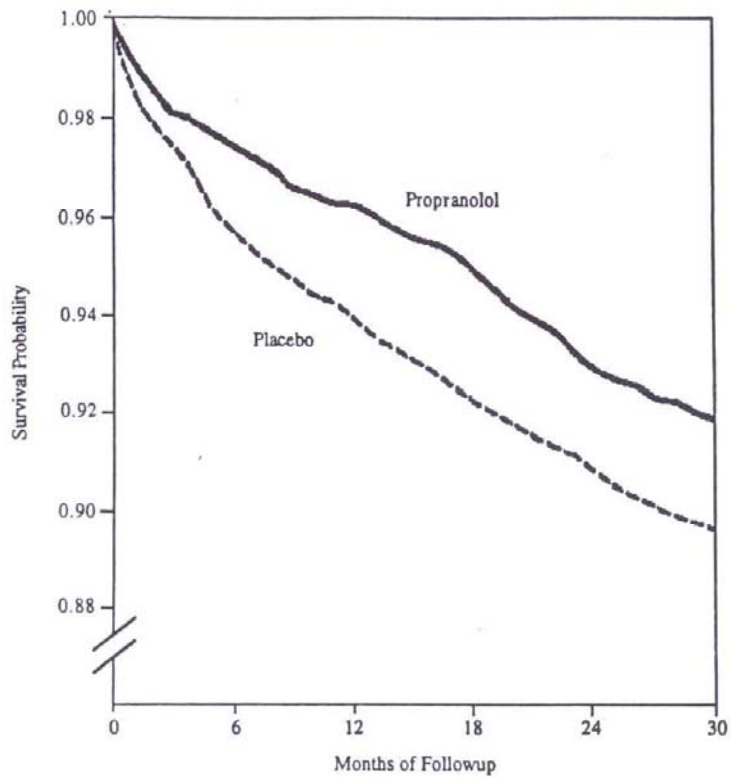


Fig. 2. BHAT survival curves.

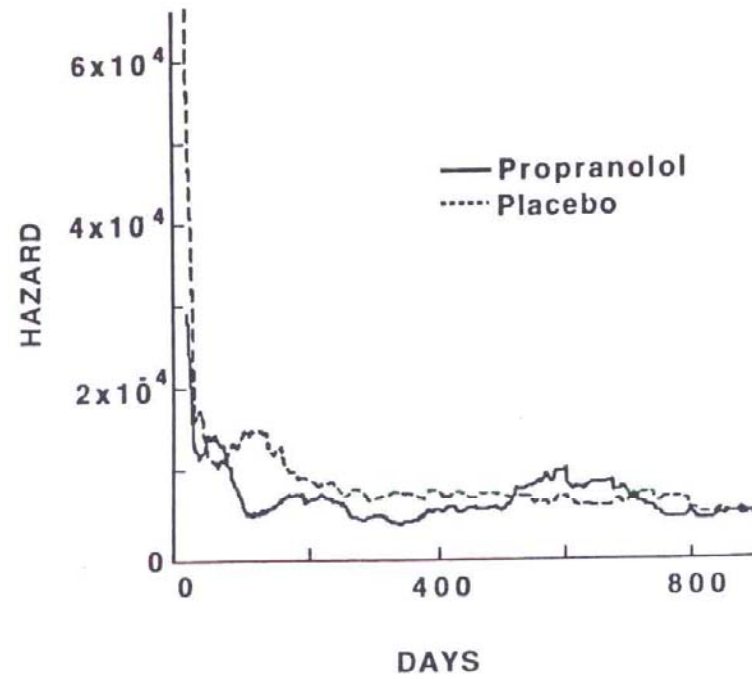


Fig. 3. BHAT hazard curves.



## Sample size evaluation for a survival trial

Under the proportional hazards model,

hazard ratio  $=\lambda_C(t)/\lambda_T(t) = \text{HR}$  and

$$E[ Z_{\text{logrank}} ] = (\log \text{HR}) \sqrt{D/4},$$

$D$  = expected number of events in the trials.

To reach a power of  $100(1-\beta)\%$ , solve  $D$  from

$$E[\text{logrank } Z] = (\log \text{HR}) \sqrt{D/4} = z_\alpha + z_\beta.$$

Note that in practice,  $D$  is unknown and not observable. Therefore, we have to use observed number of events to replace  $D$ . Recruit  $N \geq D$  patients and follow them until  $D$  events are observed.



For the comparison of two means, solve for N from

$$EZ(N) = \Delta \sqrt{N / 4} = z_{\alpha} + z_{\beta}.$$

For the logrank tset,

$$EZ_{\text{logrank}} = \log (\lambda_C / \lambda_T) \sqrt{D / 4} = \Delta \sqrt{D / 4} .$$

D = expected number of events.

Solve for D from

$$EZ_{\text{logrank}} (D) = \Delta \sqrt{D / 4} = z_{\alpha} + z_{\beta}.$$

**Based on contiguous alternative and asymptotic theory.**

Example:  $HR = \log(\lambda_C/\lambda_T) = 1.25$ ,  $\alpha = 0.025$  (one-sided), power = 85% or  $\beta = .15$ .

$$E[Z_{\text{logrank}}] = \ln(1.25) \sqrt{D/4} = 1.96 + 1.04 = 3$$

$$\Rightarrow D \approx 724 \text{ (events)}$$

Recruit  $N \geq 724$  patients and follow them until 724 events are observed.

EXAMPLE: Assume

1. 70% of the control patients survive 30 months;
2. the treatment reduce hazard by 20%;
3. survival times follow exponential distributions.

$$0.7 = e^{-30\lambda} \rightarrow \lambda = \lambda_C = 0.01189 \rightarrow \lambda_T = 0.00951.$$

To evaluate median survival time for the control group:

$$e^{-M\lambda} = 0.5 \rightarrow M = M_C = \text{median survival time for C} = 58.30.$$

$$M_T = 72.88.$$

EaST output on next page

# Survival Design for Untitled1

	Plan1			Plan2		
<b>Plan Id</b>	Plan1			Plan2		
<b>1-Sided or 2-Sided Test</b>	1-Sided			1-Sided ✓		
<b>Significance Level (Alpha)</b>	0.025			0.025 ✓		
<b>Power (1-Beta)</b>	0.85			0.85 ✓		
<b>Subject Accrual Per Unit Time</b>	300			200		
<b>Median Time (Control)</b>	58.30057			58.30057 ✓		
<b>Median Time (Treatment)</b>	72.87571			72.87571 ✓		
<b>Assigned Fraction (Treatment)</b>	0.5			0.5		
<b>Planned Number of Looks</b>	1			1 ✓		
<b>Stop Early to Reject &gt;&gt;</b>						
<b>Boundary Shape to Reject H0</b>						
<b>Boundary Shape to Reject H1</b>						
	<b>Min</b>	<b>Max</b>		<b>Min</b>	<b>Max</b>	
<b>Committed Accrual (Duration)</b>	2.41	18	22.16	3.61	24	27.38
<b>Committed Accrual (Subjects)</b>	722	5400	6648	722	4800	5476
<b>Maximum Study Duration</b>	22.71			27.66		
<b>Maximum Number of Events</b>	728			728		
	<b>If H0</b>	<b>If H1</b>	<b>If H1/2</b>	<b>If H0</b>	<b>If H1</b>	<b>If H1/2</b>
<b>Expected Accrual (Subjects)</b>						
<b>Expected Study Duration</b>	21.36	22.71		26.13	27.66	
<b>Expected Number of Events</b>						
<b>Note Pad</b>						

Another software for survival trial design

STOPP

## The Wilcoxon statistic (1945)

$X_1, X_2, \dots, X_m$  (m X's and n Y's)

$Y_1, Y_2, \dots, Y_n$

$T_1, T_2, \dots, T_m, T_{m+1}, \dots, T_N; N = m+n$

$T_{(1)} < T_{(2)} < \dots < T_{(i)}, \dots < T_{(N)}$

Wilcoxon score = rank  $\approx$  rank/(N+1);  $\phi(t) = t$ .

$a_i = a_{Ni} = i$

Centered Wilcoxon score =  $i - \frac{N+1}{2}$ .

A modified Wilcoxon score is  $a_i = a_{Ni} = \frac{i}{N+1}$ , and the

corresponding centered score is  $\frac{i}{N+1} - 0.5$ .

Score = Rating

### Mann-Whitney (1947) {U-statistic}

$$MW = \{\# \text{ of } (X, Y) \text{ pairs } \ni X < Y\} - \frac{1}{2}mn$$

$\equiv$  Centered Wilcoxon statistic

Rank sum of the Y's (m X's and n Y's)

$$\begin{aligned} &= (\# \text{ of } X\text{'s} < \text{the smallest } Y) + 1 \\ &+ (\# \text{ of } X\text{'s} < \text{the second smallest } Y) + 2 \\ &+ \dots \\ &+ (\# \text{ of the } X\text{'s} < \text{the largest } Y) + n \end{aligned}$$

Centered Wilcoxon statistic (m X's and n Y's)

$$\begin{aligned} &= \text{Rank sum of the Y's} - n(m+n+1)/2 \\ &= \{\# \text{ of } (X, Y) \text{ pairs}\} + n(n+1)/2 - n(m+n+1)/2 \\ &= \{\# \text{ of } (X, Y) \text{ pairs}\} - mn/2 = MW \end{aligned}$$

## The Wilcoxon payments

The “loser” at  $T_{(i)}$ , pays \$1 to every competing player. When there is no censoring,

$$a_1 = 1 - N$$

$$a_2 = 2 - (N - 1)$$

⋮

$$a_i = i - (N - i + 1) = 2i - (N + 1) = 2 \left\{ i - \frac{N + 1}{2} \right\}.$$

⋮

⋮

$$S = \sum a_i \text{'s the } Y\text{'s} = 2(\text{centered Wilcoxon statistic}).$$



## Gehan statistic

Mann-Whitney = MW

$$= \{\# \text{ of } (X, Y) \text{ pairs } \ni X < Y\} - \frac{mn}{2}.$$

Since  $mn = \#\{X < Y\} + \#\{Y < X\}$ ,

$$\#\{X < Y\} - \#\{Y < X\} = \#\{X < Y\} - [mn - \#\{X < Y\}] = 2MW.$$

Gehan =  $\#\{X < Y\} - \#\{Y < X\}$ ; when there are censored observations, ignore pairs when the order of X, Y cannot be determined. {5 & 3+; 5+ & 7+}

### The Wilcoxon payments:

The “loser” at  $T_{(i)}$ , pays \$1 to every competing player.

The Wilcoxon and Gehan statistics (assume no ties): At  $T_{(i)}$ :

	D	$\bar{D}$	
X	$\delta_i$		$m_i$
Y			$n_i$
	1	$N_i-1$	$m_i+n_i=N_i$

$\delta_i = 1$  if  $T_{(i)}$  is a X  
 $= 0$  otherwise

$$S = \sum N_i \left( \delta_i - \frac{m_i}{N_i} \right)$$

$$S^* = \sum \frac{N_i}{N+1} \left( \delta_i - \frac{m_i}{N_i} \right) \hat{=} \sum \widehat{S(T_{(i)})} \left( \delta_i - \frac{m_i}{N_i} \right)$$

When there is censoring,  $N_i/(N+1)$  estimates the survival curve of TAC.

## Peto-Peto-Prentice version of the Wilcoxon statistic

Reference for the topic:

Lan KKG and Wittes JT. Rank tests for survival analysis: A comparison by analogy with games. *Biometrics* 1985; **41**: 1063-1069.