Lecture 3
COMPARISON OF SURVIVAL CURVES

We talked about some nonparametric approaches for estimating the survival function, $\hat{S}(t)$, over time for a group of individuals.

Now we want to compare the survival estimates between two or more groups.

Nonparametric comparisons of groups

Why nonparametric?

- fairly robust
- efficient relative to parametric tests
- often simple and intuitive

Recall(?) **relative efficiency (ARE) of nonparametric versus parametric tests** from Lehmann and Romano 'Testing statistical hypothesis' book. Eg. ARE of Wilcoxon vs. t-test is $3/\pi = 95.5\%$ under normality.

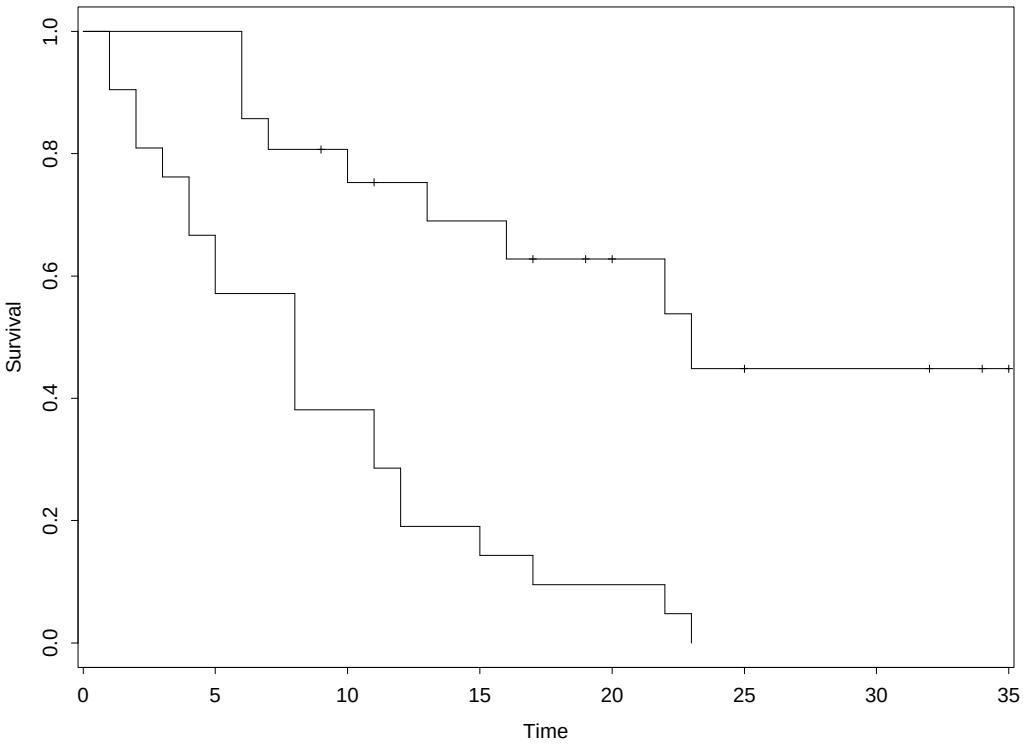

Figure 1: Time to remission of leukemia patients

How do you compare two curves (distributions)?

- the largest distance between the two curves?
- the median survival for each group?
- average hazard? (for exponential distributions, this would be like comparing the mean event times)
- adding up the difference between the two survival estimates over time?

$$\sum_j \left[\hat{S}(t_{jA}) - \hat{S}(t_{jB})\right]$$

- a weighted sum of differences, where the weights reflect the number at risk at each time?
- a rank-based test? e.g., we could rank all the event times, and see whether the sum of ranks for one group was less than the other.

Ex: how do you test two-sample exponential distributions? two-sample binomial distributions?

General Framework for Survival Analysis

(for right-censored data)

We observe $(X_i, \delta_i, \mathbf{Z}_i)$ for individual i, where

- X_i is a possibly censored failure time random variable
- δ_i is the failure/censoring indicator
- $\mathbf{Z}_i$ represents a vector of covariates

Note that $\mathbf{Z}_i$ might be a scalar (a single covariate, say treatment or age) or may be a $p \times 1$ vector (representing several different covariates).

These covariates might be:

- continuous
- discrete
- time-varying (more later)

If $\mathbf{Z}_i$ is a scalar and is binary, then we are comparing the survival of two groups, like in the leukemia example.

More generally though, it is useful to build a model that characterizes the relationship between survival and all of the covariates of interest.

We'll proceed as follows:

- Two group comparisons
- Multigroup and stratified comparisons - stratified logrank
- Failure time regression models
  - Cox proportional hazards model
  - Accelerated failure time model
  - ....
- Prediction, medical AI
- Causal inference...

Two sample tests

- Mantel-Haenszel logrank test
- Peto & Peto's version of the logrank test
- Gehan's Generalized Wilcoxon
- Peto & Peto's and Prentice's generalized Wilcoxon
- Tarone-Ware and Fleming-Harrington classes

Mantel-Haenszel Logrank test

The logrank test is the most well known and widely used.

It also has an intuitive appeal, building on standard methods for binary data. (Later we will see that it can also be obtained as a score test from the Cox Proportional Hazards regression model.)

2×2 **Tables**

First consider the following 2×2 table classifying those with and without the event of interest in a two group setting:

	Event		
Group	Yes	No	Total
0	d_0	$n_0 - d_0$	n_0
1	d_1	$n_1 - d_1$	n_1
Total	d	$n - d$	n

If the margins of this table are considered fixed, then d_0 follows a *Hypergeometric* distribution, under the null hypothesis of no association between the event and group. It follows that

$$E(d_0) = \frac{n_0 d}{n}$$

$$Var(d_0) = \frac{n_0\, n_1\, d(n-d)}{n^2(n-1)}$$

Therefore, under H_0:

$$\chi^2_{MH} = \frac{[d_0 - n_0\, d/n]^2}{\frac{n_0\, n_1\, d(n-d)}{n^2(n-1)}} \underset{\sim}{\text{approx.}} \chi^2_1$$

This is the Mantel-Haenszel statistic and is approximately equivalent to the Pearson's χ^2 test for equal proportions of the two groups given by:

$$\chi^2_P \;=\; \sum \frac{(o-e)^2}{e}$$

Note: Pearson's χ^2 test was derived when only the row margins were considered fixed, and thus the variance in the denominator was replaced by:

$$Var(d_0) \;=\; \frac{n_0\, n_1\, d(n-d)}{n^3}$$

Example: Leukemia data, just counting the number of relapses in each treatment group.

	Fail		
Group	Yes	No	Total
0	21	0	21
1	9	12	21
Total	30	12	42

$$\begin{aligned} \chi^2_p &= 16.8 \quad (p = 0.001) \\ \chi^2_{MH} &= 16.4 \quad (p = 0.001) \end{aligned}$$

But, this doesn't account for the length of survival time itself.

Recall that testing of 2×2 tables can also be phrased in terms of odds ratio:

$$\text{OR} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}.$$

$K \times 2 \times 2$ **Tables**

Now suppose we have K (2×2) tables, all independent, and we want to test for a common group effect, i.e. a common odds ratio not equal to 1.

For the j-th table:

	Event		
Group	Yes	No	Total
0	d_{0j}	$n_{0j} - d_{0j}$	n_{0j}
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
Total	d_j	$n_j - d_j$	n_j

The Cochran-Mantel-Haenszel test [**read about it please**] for a common odds ratio not equal to 1 can be written as:

$$\chi^2_{CMH} = \frac{\{\sum_{j=1}^{K}(d_{0j} - n_{0j} * d_j/n_j)\}^2}{\sum_{j=1}^{K} n_{1j}n_{0j}d_j(n_j - d_j)/[n_j^2(n_j - 1)]}$$

This statistic is distributed approximately as χ^2_1 under the null hypothesis (why).

How does this apply in survival analysis?

Cox & Oakes Table 1.1 Leukemia example

Ordered	Group 0			Group 1		
Event Times	d_j	c_j	r_j	d_j	c_j	r_j
1	2	0	21	0	0	21
2	2	0	19	0	0	21
3	1	0	17	0	0	21
4	2	0	16	0	0	21
5	2	0	14	0	0	21
6	0	0	12	3	1	21
7	0	0	12	1	0	17
8	4	0	12	0	0	16
9	0	0	8	0	1	16
10	0	0	8	1	1	15
11	2	0	8	0	1	13
12	2	0	6	0	0	12
13	0	0	4	1	0	12
15	1	0	4	0	0	11
16	0	0	3	1	0	11
17	1	0	3	0	1	10
19	0	0	2	0	1	9
20	0	0	2	0	1	8
22	1	0	2	1	0	7
23	1	0	1	1	0	6
25	0	0	0	0	1	5

Note that we listed Group 1 for times 1-5 even though there were no events or censorings at those times.

Logrank Test: Formal Definition

The logrank test is obtained by constructing a (2×2) table **at each distinct failure time**, and comparing the failure rates between the two groups, conditional on the number at risk in the groups. The tables are then combined using the Cochran-Mantel-Haenszel test.

Note: The logrank is sometimes called the Cox-Mantel test.

Let $t_1 < ... < t_K$ represent the K ordered, distinct failure times.
At the j-th failure time t_j, we have the following table:

	Die/Fail		
Group	Yes	No	Total
0	d_{0j}	$r_{0j} - d_{0j}$	r_{0j}
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
Total	d_j	$r_j - d_j$	r_j

where d_{0j} and d_{1j} are the number of failures in group 0 and 1, respectively at the j-th failure time, and r_{0j} and r_{1j} are the number at risk at that time, in groups 0 and 1.

First several tables of leukemia data

CMH analysis of leukemia data

TABLE 1 OF TRTMT BY REMISS
CONTROLLING FOR FAILTIME=1

TRTMT REMISS

Frequency| Expected	0	1	Total
0	19	2	21
	20	1	
1	21	0	21
	20	1	
Total	40	2	42

TABLE 2 OF TRTMT BY REMISS
CONTROLLING FOR FAILTIME=2

TRTMT REMISS

Frequency| Expected	0	1	Total
0	17	2	19
	18.05	0.95	
1	21	0	21
	19.95	1.05	
Total	38	2	40

TABLE 3 OF TRTMT BY REMISS
CONTROLLING FOR FAILTIME=3

TRTMT REMISS

Frequency| Expected	0	1	Total
0	16	1	17
	16.553	0.4474	
1	21	0	21
	20.447	0.5526	
Total	37	1	38

TABLE 4 OF TRTMT BY REMISS
CONTROLLING FOR FAILTIME=4

TRTMT REMISS

Frequency| Expected	0	1	Total
0	14	2	16
	15.135	0.8649	
1	21	0	21
	19.865	1.1351	
Total	35	2	37

The logrank test is:

$$\chi^2_{\text{logrank}} = \frac{[\sum_{j=1}^{K}(d_{0j} - r_{0j} * d_j/r_j)]^2}{\sum_{j=1}^{K} \frac{r_{1j} r_{0j} d_j (r_j - d_j)}{[r_j^2 (r_j - 1)]}}$$

It turns out that the tables can be seen as independent, therefore the statistic has an approximate χ^2 distribution with 1 df.

Based on the motivation for the logrank test, which of the survival-related quantities are we comparing at each time point?

- $\sum_{j=1}^{K} w_j \left[\hat{S}_1(t_j) - \hat{S}_2(t_j)\right]$ **?**
- $\sum_{j=1}^{K} w_j \left[\hat{\lambda}_1(t_j) - \hat{\lambda}_2(t_j)\right]$ **?**
- $\sum_{j=1}^{K} w_j \left[\hat{\Lambda}_1(t_j) - \hat{\Lambda}_2(t_j)\right]$ **?**

Calculating logrank statistic step-by-step
Leukemia Example:

Ordered	Group 0		Combined				
Death Times	d_{0j}	r_{0j}	d_j	r_j	e_j	$o_j - e_j$	v_j
1	2	21	2	42	1.00	1.00	0.488
2	2	19	2	40	0.95	1.05	
3	1	17	1	38	0.45	0.55	
4	2	16	2	37	0.86	1.14	
5	2	14	2	35			
6	0	12	3	33			
7	0	12	1	29			
8	4	12	4	28			
10	0	8	1	23			
11	2	8	2	21			
12	2	6	2	18			
13	0	4	1	16			
15	1	4	1	15			
16	0	3	1	14			
17	1	3	1	13			
22	1	2	2	9			
23	1	1	2	7			
Sum						10.251	6.257

$o_j = d_{0j}$

$e_j = d_j r_{0j}/r_j$

$v_j = r_{1j} r_{0j} d_j (r_j - d_j)/[r_j^2 (r_j - 1)]$

$$\chi^2_{\text{logrank}} = \frac{(10.251)^2}{6.257} = 16.793$$

Notes about logrank test:

- The logrank statistic depends on ranks of event times only

- If there are no tied deaths, $d_j = 1$, then the logrank has the form:

$$\frac{[\sum_{j=1}^{K}(d_{0j} - \frac{r_{0j}}{r_j})]^2}{\sum_{j=1}^{K} r_{1j}r_{0j}/r_j^2}$$

  where $d_{0j} = 0$ or 1.

- Numerator can be interpreted as $\sum(o - e)$ where "o" is the observed number of deaths in group 0, and "e" is the expected number, given the risk set and that the two groups have the same rate of failure. So $e = \#\text{deaths} \times$ proportion from group 0 at risk.

- The $(o - e)$ terms in the numerator can be written as

$$\frac{r_{0j}r_{1j}}{r_j}(\hat{\lambda}_{0j} - \hat{\lambda}_{1j})$$

- It does not matter which group you choose to sum over.

  Note that if we summed up (o-e) over the death times for the 6MP group we would get -10.251, and the sum of the variances is the same. So when we square the numerator, the test statistic is the same.

Analogous to the CMH test for a series of tables, the logrank test is most powerful when the "odds ratios" are constant over time points. That is, it is most powerful for **proportional hazards**:

$$\lambda_1(t) = \alpha\lambda_0(t).$$

This is also equivalent to $S_0(t) = S_1(t)^\alpha$. (Why?)

Checking the assumption of proportional hazards:

- check to see if the estimated survival curves cross - if they do, then this is evidence that the hazards are not proportional
- **any other ideas?**

What should be done if the hazards are not proportional?

- If the difference between hazards has a consistent sign, the logrank test tend to do ok.
- Other tests are available that are more powerful against different alternatives.

Left Truncated Logrank Test

Similar to the left truncated KM estimator, when there is left truncation, the logrank test can be obtained by modifying the **risk set** definition.

That is, the risk set now consists of subjects who have entered the study, and have not failed or been censored by a certain time.

Still let r_{0j} be the number at risk from group 0, r_{1j} be the number at risk from group 1, and everything else remains the same.

Gehan's Generalized Wilcoxon Test

First, let's review the Wilcoxon test for uncensored data:

Denote observations from two samples by:

$$(X_1, X_2, \ldots, X_n) \text{ and } (Y_1, Y_2, \ldots, Y_m)$$

Order the combined sample and define:

$$Z_{(1)} < Z_{(2)} < \cdots < Z_{(m+n)}$$

$$R_{i1} = \text{rank of } X_i$$

$$R_1 = \sum_{i=1}^{n} R_{i1}$$

Reject H_0 if R_1 is too big or too small, according to

$$\frac{R_1 - E(R_1)}{\sqrt{\text{Var}(R_1)}} \sim N(0,1)$$

where

$$E(R_1) = \frac{n(m+n+1)}{2}$$

$$\text{Var}(R_1) = \frac{mn(m+n+1)}{12}$$

The **Mann-Whitney** form of the Wilcoxon is defined as:

$$U(X_i, Y_j) = U_{ij} = \begin{cases} +1 & \text{if} \quad X_i > Y_j \\ 0 & \text{if} \quad X_i = Y_j \\ -1 & \text{if} \quad X_i < Y_j \end{cases}$$

and

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}.$$

There is a simple correspondence between U and R_1:

$$R_1 = n(m+n+1)/2 + U/2$$

$$\text{so} \qquad U = 2R_1 - n(m+n+1)$$

Therefore,

$$E(U) = 0$$

$$\text{Var}(U) = mn(m+n+1)/3$$

Extending Wilcoxon to censored data

The Mann-Whitney form leads to a generalization for censored data. Define

$$U(X_i, Y_j) = U_{ij} = \begin{cases} +1 & \text{if} \quad x_i > y_j \text{ or } x_i^+ \geq y_j \\ -1 & \text{if} \quad x_i < y_j \text{ or } x_i \leq y_j^+ \\ 0 & \text{otherwise} \end{cases}$$

Then define

$$W = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}$$

Thus, there is a contribution to W for every comparison where both observations are failures (except for ties), or where a censored observation is greater than or equal to a failure.

The above can also be computed as follows.

[Reading:] First, pool the sample of $(n+m)$ observations into a single group, then compare each individual with the remaining $n+m-1$:

For comparing the i-th individual with the j-th, define

$$U_{ij} = \begin{cases} +1 & \text{if} \quad t_i > t_j \ \text{ or } \ t_i^+ \geq t_j \\ -1 & \text{if} \quad t_i < t_j \ \text{ or } \ t_i \leq t_j^+ \\ 0 & \qquad \textit{otherwise} \end{cases}$$

Then

$$U_i = \sum_{j=1}^{m+n} U_{ij}$$

Thus, for the i-th individual, U_i is the number of observations which are definitely less than t_i minus the number of observations that are definitely greater than t_i. We assume censorings occur after deaths, so that if $t_i = 18^+$ and $t_j = 18$, then we add 1 to U_i.

The Gehan statistic is defined as

$$\begin{aligned} U &= \sum_{i=1}^{m+n} U_i \, \mathbf{1}_{\{i \text{ in group } 0\}} \\ &= W \end{aligned}$$

This puts it in the form of a **linear rank test**, from which the log-rank test got its name.

Under H_0, $U = W$ has mean 0 and variance estimated by

$$\widehat{\mathrm{Var}}(U) = \frac{mn}{(m+n)(m+n-1)} \sum_{i=1}^{m+n} U_i^2$$

Example: (leukemia data)

```
Wilcoxon test for equality of survivor functions
------------------------------------------------

      |  Events                         Sum of
trt   |  observed       expected         ranks
------+-----------------------------------------
0     |        21          10.75           271
1     |         9          19.25          -271
------+-----------------------------------------
Total |        30          30.00             0

           chi2(1) =      13.46
           Pr>chi2 =     0.0002
```

As it turns out, the Wilcoxon test for censored data can be written as a special case of a class of weighted log-rank tests; i.e.

$$U = \sum_{i=1}^{m+n} U_i \mathbf{1}_{\{i \text{ in group } 0\}} = \sum_{j=1}^{K} w_j (d_{1j} - r_{1j} \cdot d_j / r_j)$$

where $w_j = r_j$.

Weighted Log-rank Tests

This general class of tests is like the logrank test, but with weights w_j. The logrank test, Gehan's Wilcoxon test, Peto-Prentice's Wilcoxon and more are included as special cases:

$$\chi^2_w = \frac{\{\sum_{j=1}^{K} w_j(d_{1j} - r_{1j} \cdot d_j/r_j)\}^2}{\sum_{l=1}^{K} \frac{w_j^2 r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}}$$

Test	**Weight** w_j
Logrank	$w_j = 1$
Gehan's Wilcoxon	$w_j = r_j$
Peto/Prentice	$w_j = \hat{S}(t_j)$
Fleming-Harrington	$w_j = [\hat{S}(t_j)]^\rho \ \ (\rho \geq 0)$
Tarone-Ware	$w_j = \sqrt{r_j}$

The F-H family of weighted log-rank tests is sometimes called the G^ρ family; $\rho = 0, 1$ are the most commonly used. This is available in R 'survdiff()'.

Which test should we use?

- All the tests have the correct Type I error for testing the null hypothesis of equal survival, $H_o : S_1(t) = S_2(t)$
- The choice of which test may therefore depend on the alternative hypothesis, which will drive the power of the test.
- The Wilcoxon is sensitive to **early differences** between survival, while the logrank is most powerful under proportional hazards. Compare the relative weights they assign to the test statistic:

$$\begin{aligned} \text{LOGRANK} \quad \textit{numerator} &= \sum_j (o_j - e_j) \\ \text{WILCOXON} \quad \textit{numerator} &= \sum_j r_j (o_j - e_j) \\ \text{or} &= \sum_j \hat{S}(t_j)(o_j - e_j) \end{aligned}$$

- The Wilcoxon also has high power when the failure times are log-normally distributed, with equal variance in both groups but a different mean. It will turn out that this is one special case of the accelerated failure time (AFT) model.
- The Wilcoxon with weights $\hat{S}(t_j)$ is most powerful under the alternative hypothesis of log-logistic model.
- All the above weighted tests will lack power if the **hazards "cross"**.

Counting Process Notation

Let $N(t) = I(X \leq t, \delta = 1)$ be the counting process for one subject;

so either $N(t) = 0$ for all $t > 0$ (if $\delta = 0$),
or $N(t) = 0$ for $0 < t < T$ and 1 for $t \geq T$ (if $\delta = 1$).

Then $N(t) - N(s)$ is the number (0 or 1) of jumps of N in $(s, t]$, and

$$\begin{aligned} \lambda(t)\Delta t &\approx P(t \leq T < t + \Delta t | T \geq t, C \geq t) \\ &= P(N((t + \Delta t)-) - N(t-) = 1 | T \geq t, C \geq t), \end{aligned}$$

so $\lambda(t)$ is the conditional rate at which N jumps in small intervals.

The Stieltjes integral $\int_s^t f(t) dN(t)$ = sum of values of f at the jump time(s) of N in $(s, t]$.

Also define the 'at risk' process $Y(t) = I(X \geq t)$;

so that $Y(t) = 1$ for $0 < t \leq X$ and 0 for $t > X$.

For the two-sample problem: $l = 1, 2$,

let $N_l(t) = \sum_{i=1}^{n_l} N_{li}(t)$, $Y_l(t) = \sum_{i=1}^{n_l} Y_{li}(t)$.

Then (Harrington and Fleming, 1982)

$$\begin{aligned} G^\rho &= \sum_{j=1}^{K} \hat{S}(t_j)^\rho \left(d_{1j} - \frac{r_{1j} \cdot d_j}{r_j} \right) \\ &= \int_0^\infty \hat{S}(u)^\rho \cdot \frac{Y_1(u) Y_2(u)}{Y_1(u) + Y_2(u)} \left\{ \frac{dN_1(u)}{Y_1(u)} - \frac{dN_2(u)}{Y_2(u)} \right\} \end{aligned}$$

with $0/0 = 0$.

Verify: assuming no ties

$$\begin{aligned} & \left\{ \frac{Y_2(u)}{Y_1(u) + Y_2(u)} dN_1(u) - \frac{Y_1(u)}{Y_1(u) + Y_2(u)} dN_2(u) \right\} \\ = \; & \frac{r_{2j}}{r_j} \text{ or } 0 - \frac{r_{1j}}{r_j}, \text{ at time } t_j \\ = \; & 1 - \frac{r_{1j}}{r_j} \text{ or } 0 - \frac{r_{1j}}{r_j} \\ = \; & d_{1j} - \frac{r_{1j}}{r_j} \end{aligned}$$

Asymptotic Relative Efficiency (ARE)

Lehmann and Romano *Testing Statistical Hypothesis* (TSH) book:

Suppose $X_1, ...X_n$ i.i.d.$\sim P_\theta$ a family of distributions satisfying smoothness conditions, $\theta \in \Omega \subset R^1$. $\mathrm{H}_0 : \theta = \theta_0$, $\mathrm{H}_1 : \theta > \theta_0$.

A contiguous sequence is a sequence θ_n, such that $\sqrt{n}(\theta_n - \theta_0) = O(1)$.

For the contiguous sequence θ_n above, let the test statistic T_n be such that:
$\sqrt{n}\{T_n - \mu(\theta_n)\} \stackrel{D}{\to} N(0, \sigma^2)$ under $\mathrm{H}_1 : \theta = \theta_n$,
where $\mu(\cdot)$ has right-hand derivative $\mu'(\theta_0) > 0$.

Theorem 13.2.1 (TSH):
1) .. the test has asymptotic significance level α;
2) the limiting power under the contiguous sequence of alternatives is ..
3) The sample size n such that that power of T_n is at least $1 - \beta$ is

$$n \sim \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{\{(\theta_n - \theta_0)\mu'(\theta_0)\}^2},$$

where $a_n \sim b_n$ means $a_n/b_n \to 1$ as $n \to \infty$.

$\mu'(\theta_0)/\sigma$ is called the efficacy of T.

For two tests T and $\tilde{T}$, $\left\{\frac{\tilde{\mu}'(\theta_0)/\tilde{\sigma}}{\mu'(\theta_0)/\sigma}\right\}^2$ is the **Pitman ARE**.

Example: recall Wald test based on MLE (TSH section 12.4)

Consider $\mathrm{H}_0 : \theta = \theta_0$, $\mathrm{H}_1 : \theta > \theta_0$.

Under H_0, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I(\theta_0)^{-1})$, where $I(\theta_0)$ is the Fisher information, and is estimated by $\hat{\sigma}_n^{-2}$.

Ex. 1) The Wald test that rejects when

$$\sqrt{n}(\hat{\theta}_n - \theta_0) > z_{1-\alpha}\hat{\sigma}_n$$

has asymptotic significance level α;

Ex. 2) under $\theta_n = \theta_0 + h/\sqrt{n}$, where $h > 0$ is a constant, show that the power of the Wald test

$$P(\sqrt{n}(\hat{\theta}_n - \theta_0) > z_{1-\alpha}\hat{\sigma}_n) \to 1 - \Phi(z_{1-\alpha} - h\sqrt{I(\theta_0)})$$

as $n \to \infty$;

Ex. 3) verify the sample size formula on the previous page.

Now assume two sample survival data with sizes n_1 and n_2. Let $n = n_1+n_2$. Assume that $\lim_{n\to\infty} n_l/n = a_l$, $0 < a_l < 1$ for $l = 1, 2$.

Let $H_0 : S_1 = S_2 = S$.

Theorem 2.1 (H+F): Under H_0,

$$\frac{G^\rho}{\sqrt{V}} \xrightarrow{D.} N(0,1)$$

as $n \to \infty$, where V is given in H+F.

A contiguous sequence of alternative hypotheses is H_1^n such that the difference between H_1^n and H_0 is of order $1/\sqrt{n}$.

In particular, let $S_l^n(t)$ $(l = 1, 2)$ be such that

$$\lim_{n\to\infty} S_l^n(t) = S(t)$$

uniformly in $t \in [0, \infty)$, and

$$\gamma(t) = \lim_{n\to\infty} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \frac{\lambda_1^n(t) - \lambda_2^n(t)}{\lambda(t)}$$

exists (also uniform convergence).

Theorem 2.2 (H+F): Under the above sequence of alternatives,

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} G^\rho \xrightarrow{D.} N(\mu_\rho, \sigma_\rho^2),$$

where μ_ρ and σ_ρ^2 are given in H+F.

μ_ρ/σ_ρ is called the Pitman efficiency.

Now consider a specific sequence of alternatives -

Let r.v. T_0 have survival function on $-\infty < t < \infty$:

$$H_\rho(t) = \begin{cases} \exp(-e^t), & \text{if } \rho = 0 \text{ (extreme-value);} \\ (1+\rho e^t)^{-1/\rho}, & \text{if } \rho > 0 \ (\rho = 1 \text{ logistic}). \end{cases}$$

Ex. plot these distributions.

For some θ_0 (eg. $\theta_0 = 0$) let

$$\begin{aligned} \theta_1^n &= \theta_0 + \sqrt{\frac{n_2}{n_1(n_1+n_2)}}, \\ \theta_2^n &= \theta_0 - \sqrt{\frac{n_1}{n_2(n_1+n_2)}}. \end{aligned}$$

Then $\theta_l^n \to \theta_0$ at $1/\sqrt{n}$ rate.

Finally let r.v.

$$T_l^n = g^{-1}(T_0 - \theta_l^n),$$

where $g(\cdot)$ is strictly increasing (eg. $g^{-1}(\cdot) = \exp(\cdot)$).

This way $g^{-1}(T_0 - \theta_0)$ gives the distribution $S(t)$ under the null, and T_l^n gives $S_l(t)$ $(l = 1, 2)$ under the alternatives.

Ex. a) Show that T_l^n has survival function

$$S_l^n(t) = H_\rho(g(t) + \theta_l^n).$$

This is called time-transformed location alternative.

Theorem 2.3 (H+F): For θ_l^n $(l = 1, 2)$ defined above, the test based on $G^\rho/\sqrt{V}$ has maximum (Pitman) efficiency against the contiguous alternatives $S_l^n(t)$ (also defined above, and depends on ρ) among all tests of the form

$$\int_0^\infty K(u) \left\{ \frac{dN_1(u)}{Y_1(u)} - \frac{dN_2(u)}{Y_2(u)} \right\},$$

where $K(\cdot)$ is any stochastic process satisfying regularity conditions.

To understand how the above theorem is used, write $\Delta = \theta_1 - \theta_2$ (omit n in the following).

Ex. b) Show that

$$S_2(t) = S_1[S_1(t)^\rho + \{1 - S_1(t)^\rho\}e^\Delta]^{-1/\rho},$$

or

$$\lambda_2(t) \stackrel{(*)}{=} \lambda_1(t)e^\Delta[S_1(t)^\rho + \{1 - S_1(t)^\rho\}e^\Delta]^{-1}.$$

Corollary: From eq.(*)
$\rho = 0 \Longrightarrow$ Log-rank test is the most efficient when $\lambda_2(t)=\lambda_1(t)e^\Delta$;
$\rho = 1 \Longrightarrow$ Wilcoxon (Peto's) log-rank test is the most efficient under logistic shift alternative.

(**Make sure that you understand the use of this.**)

Since the rank tests are invariant under monotone transformation of time, we can take

$$\begin{aligned} \lambda_1(t) &= 1, \\ \lambda_2(t) &= e^\Delta\{e^{-\rho t} + (1 - e^{-\rho t})e^\Delta\}^{-1}. \end{aligned}$$

Ex. c) Find a monotone transformation $g_1(\cdot)$ such that $g_1(T_1)$ and $g_1(T_2)$ has hazards $\lambda_1(t) = 1$ and $\lambda_2(t)$ above.

See H+F Figure 1.

We can further investigate the ARE within the G^ρ family of weighted log-rank tests:

Under further assumptions that the censoring distributions are the same in the two groups, the asymptotic efficacy for the G^ρ weighted log-rank test under the alternative which is given by the $G^{\rho'}$ distribution simplifies to

$$(\mu_\rho/\sigma_\rho)^2 = \frac{\{\int_0^\infty \pi(u)S(u)^{\rho+\rho'}d\Lambda(u)\}^2}{\int_0^\infty \pi(u)S(u)^{2\rho}d\Lambda(u)},$$

where $\pi(u) = P(T \geq u, C \geq u)$ is the probability of being 'at risk' at time u for any member of the two groups under the null hypothesis, and $S(\cdot)$ and $\Lambda(\cdot)$ are the survival function and the cumulative hazard function, respectively, again all under the null hypothesis.

Ex:

1. for given alternative distribution $G^{\rho'}$, the maximum efficacy of the G^ρ test is achieved when $\rho = \rho'$;
2. if we further assume that the survival function for the censoring distribution $S_c(u) = S(u)^\alpha$, then the asymptotic efficacy becomes $(2\rho + \alpha + 1)(2\rho' + \alpha + 1)/(\rho + \rho' + \alpha + 1)^2$;
3. use the above from part 2. to compute the ARE of Peto's Wilcoxon log-rank test (i.e. G^1) versus the unweighted log-rank test under the proportional hazards alternative assuming $\alpha = .5$ and $\alpha = 0$.

See also H+F Table 1 and Figure 2.

[Reading:] P-sample and stratified logrank tests

We have discussed two-sample problems. In practice, the following can arise:

- There are more than two treatments or groups, and the question of interest is whether the groups differ from each other.

- We are interested in a comparison between two groups, but we wish to adjust for other factor(s) that may affect the outcome

P-sample logrank

This will become rather simple once we are in a general regression setting, but for now –

Suppose we observe data from P different groups, and the data from group p ($p = 1, ..., P$) are:

$$(X_{p1}, \delta_{p1}) \ldots (X_{pn_p}, \delta_{pn_p})$$

We now construct a $(P \times 2)$ table at each of the K distinct death times, and compare the death rates among the P groups, conditional on the number at risk.

Let $t_1, ...t_K$ represent the K ordered, distinct death times. At the j-th death time, we have the following table:

	Die/Fail		
Group	Yes	No	Total
1	d_{1j}	$r_{1l} - d_{1j}$	r_{1j}
.	.	.	.
P	d_{Pj}	$r_{Pj} - d_{Pj}$	r_{Pj}
Total	d_j	$r_j - d_j$	r_j

where d_{pj} is the number of deaths in group p at the j-th death time, and r_{pj} is the number at risk at that time.

The tables are then combined using the CMH approach as before.

Some formal calculations:

Let $\mathbf{O}_j = (d_{1j}, ...d_{(P-1)j})^T$ be a vector of the observed number of failures in groups 1 to $(P-1)$, respectively, at the j-th death time. Given the risk sets r_{1j}, ... r_{Pj}, and the fact that there are d_j deaths, then $\mathbf{O}_j$ has mean:

$$\mathbf{E}_j = (\frac{d_j\, r_{1j}}{r_j},\ \ldots\ , \frac{d_j\, r_{(P-1)j}}{r_j})^T$$

and variance covariance matrix:

$$\mathbf{V}_j = \begin{pmatrix} v_{11j} & v_{12j} & \ldots & v_{1(P-1)j} \\ & v_{22j} & \ldots & v_{2(P-1)j} \\ \ldots & & \ldots & \ldots \\ & & & v_{(P-1)(P-1)j} \end{pmatrix}$$

where the ℓ-th diagonal element is:

$$v_{\ell\ell j} = r_{\ell j}(r_j - r_{\ell j})d_j(r_j - d_j)/[r_j^2(r_j - 1)]$$

and the ℓm-th off-diagonal element is:

$$v_{\ell m j} = r_{\ell j} r_{mj} d_j(r_j - d_j)/[r_j^2(r_j - 1)]$$

The resulting χ^2 test for a single $(P \times 2)$ table would have $(P-1)$ degrees of freedom and is constructed as follows:

$$(\mathbf{O}_j - \mathbf{E}_j)^T \, \mathbf{V}_j^{-1} \, (\mathbf{O}_j - \mathbf{E}_j)$$

Generalizing to K tables

Analogous to what we did for the two sample logrank, we replace the $\mathbf{O}_j$, $\mathbf{E}_j$ and $\mathbf{V}_j$ with the sums over the K distinct death times. That is, let $\mathbf{O} = \sum_{j=1}^{k} \mathbf{O}_j$, $\mathbf{E} = \sum_{j=1}^{k} \mathbf{E}_j$, and $\mathbf{V} = \sum_{j=1}^{k} \mathbf{V}_j$. Then, the test statistic is:

$$(\mathbf{O} - \mathbf{E})^T \, \mathbf{V}^{-1} \, (\mathbf{O} - \mathbf{E})$$

Example:

Time taken to finish a quiz with 3 different noise distractions. All quizes were stopped after 12 minutes.

Noise Level		
Group 1	Group 2	Group 3
9.0	10.0	12.0
9.5	12.0	12^+
9.0	12^+	12^+
8.5	11.0	12^+
10.0	12.0	12^+
10.5	10.5	12^+

```
Log-rank test for equality of survivor functions
------------------------------------------------

      |  Events
group |  observed       expected
------+--------------------------
1     |         6           1.57
2     |         5           4.53
3     |         1           5.90
------+--------------------------
Total |        12          12.00

           chi2(2) =      20.38
           Pr>chi2 =     0.0000


Wilcoxon test for equality of survivor functions
------------------------------------------------------------

      |  Events                        Sum of
group |  observed       expected        ranks
------+---------------------------------------
1     |         6           1.57           68
2     |         5           4.53           -5
3     |         1           5.90          -63
------+---------------------------------------
Total |        12          12.00            0

           chi2(2) =      18.33
           Pr>chi2 =     0.0001
```

Stratified Logrank

Sometimes, even though we are interested in comparing two (or P) groups, we know there are other factors that also affect the outcome. It would be useful to adjust for these other factors in some way.

Example: In randomized clinical trials to compare 2 (or more) treatments, we sometimes need to stratify among different subpopulations because the underlying survival distributions are different, e.g. different stages of colon cancer, different sites of brain tumor.

A **stratified logrank** allows one to compare groups, with the shapes of the hazards being different across strata. But it does make the assumption that the group 1 vs group 2 hazard ratio is the same across strata (i.e. same treatment effect).

In other words: $\frac{\lambda_{1s}(t)}{\lambda_{2s}(t)} = \theta$ where θ is constant over the strata $s = 1, ..., S$.

The approach behind stratified logrank:

Suppose we want to assess the association between survival and a group variable (call this X) that has two different levels. Suppose however, that we want to stratify by a second factor, that has S different levels.

First, divide the data into S separate strata. Within stratum s ($s = 1, ..., S$), proceed as though you were constructing the logrank to assess the association between survival and the variable X. That is, let $t_{1s}, ..., t_{K_s s}$ represent the K_s ordered, distinct death times in the s-th stratum.

At the j-th death time in group s, we have the following table:

	Die/Fail		
X	Yes	No	Total
1	d_{s1j}	$r_{s1j} - d_{s1j}$	r_{s1j}
2	d_{s2j}	$r_{s2j} - d_{s2j}$	r_{s2j}
Total	d_{sj}	$r_{sj} - d_{sj}$	r_{sj}

Let O_s be the sum of the "o"s obtained by applying the logrank calculations in the usual way to the data from group s. Similarly, let E_s be the sum of the "e"s, and V_s be the sum of the "v"s.

The **stratified logrank** is

$$Z = \frac{\sum_{s=1}^{S}(O_s - E_s)}{\sqrt{\sum_{s=1}^{S}(V_s)}}$$

This is a special case of stratified Cox model, and we will talk more about it.

Nursing home example (using Stata):

```
. use nurshome

. gen age1=0

. replace age1=1 if age>85

. sts test age1, strata(gender)

         failure _d:  cens
   analysis time _t:  los


Stratified log-rank test for equality of survivor functions
------------------------------------------------------------

      |  Events
age1  |  observed    expected(*)
------+--------------------------
0     |       795         764.36
1     |       474         504.64
------+--------------------------
Total |      1269        1269.00

(*) sum over calculations within gender

            chi2(1) =       3.22
            Pr>chi2 =     0.0728
```