phuse·virtual

Single Day Event

Survival Analysis in R vs. SAS

Oct 14th, 2021

Min-Hua Jen, *Eli Lilly*
Mia Qi, *Janssen R&D*

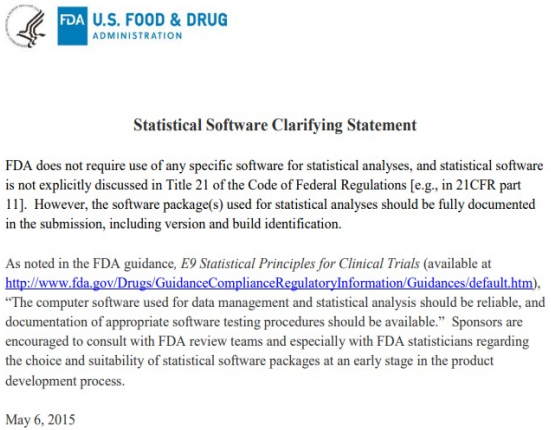
phuse.global

Agenda

- Motivation
  - Clinical statistical reporting in a multilingual world project
- Framework
- Illustration with use cases
- Summary
- Conclusions

Motivation

- The FDA has stated it does not dictate in what language analysis should be done



U.S. FOOD & DRUG ADMINISTRATION

Statistical Software Clarifying Statement

FDA does not require use of any specific software for statistical analyses, and statistical software is not explicitly discussed in Title 21 of the Code of Federal Regulations [e.g., in 21CFR part 11]. However, the software package(s) used for statistical analyses should be fully documented in the submission, including version and build identification.

As noted in the FDA guidance, *E9 Statistical Principles for Clinical Trials* (available at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm), "The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available." Sponsors are encouraged to consult with FDA review teams and especially with FDA statisticians regarding the choice and suitability of statistical software packages at an early stage in the product development process.

May 6, 2015

- SAS vs R
  - Commercial vs open-source
  - Commonly used software
  - Don't always produce the exact same values

- Clinical statistical reporting in a multilingual world project

Clinical statistical reporting in a multilingual world

- Background
- Scope of discrepancies:
  - ✓ Survival analyses
  - ➢ Linear models
  - ➢ CMH
  - ➢ Mixed models
- Public resource: GitHub (https://github.com/phuse-org/CSRMLW)
- White paper in progress

Framework

- Identify the most commonly used survival analysis methods in clinical trials
- Identify the most commonly used code for both R and SAS
- Compare R and SAS results in different scenarios
- Explore the reasons for any discrepancies

Survival analysis in clinical trials

The most commonly used survival methods in clinical trials are:

- Kaplan-Meier (KM) estimators - non-parametric statistics to estimate the survival function
- Cox proportional hazards model - semi-parametric model, commonly used to assess the relationship between the survival time and explanatory variables
- Log-rank test - non-parametric test to compare the survival functions of two or more groups

Other methods

Commonly used but out of scope for this presentation, e.g.:

- Accelerated failure time model
- Competing risk model
- Restricted mean survival time
- Time-dependent Cox model

A mockup for survival analysis in clinical trials

Summary of Survival Data

	[Study Treatment 1]	[Study Treatment 2]
Analysis set: [Analysis set]	###	###
Event – n (%)	### (xx.x%)	### (xx.x%)
Censored – n (%)	### (xx.x%)	### (xx.x%)
Time to event ([months])		
25th percentile (95% CI)	x.xx (xx.x, xx.x)	x.xx (xx.x, xx.x)
Median (95% CI)	xx.xx (xx.xx, xx.xx)	xx.xx (xx.xx, xx.xx)
75th percentile (95% CI)	xx.xx (xx.xx, NE)	xx.xx (xx.xx, NE)
Min, Max	(xx.xx, xx.xx+)	(xx.xx, xx.xx)
Landmark estimates		
[12]-[month] event-free rate (95% CI)	x.xxx (x.xxx, x.xxx)	x.xxx (x.xxx, x.xxx)
[24]-[month] event-free rate (95% CI)	x.xxx (x.xxx, x.xxx)	x.xxx (x.xxx, x.xxx)
[36]-[month] event-free rate (95% CI)	x.xxx (x.xxx, x.xxx)	x.xxx (x.xxx, x.xxx)
p-value [a]		< 0.xxxx
Hazard ratio (95% CI) [b]		x.xxx (x.xxx, x.xxx)

KM estimators

Log-rank test

Cox PH model

Key: + = censored observation, NE = not estimable, CI = confidence interval

[a] p-value is from a [stratified/non-stratified] log-rank test.

[b] Hazard ratio and 95% CI from a [stratified/non-stratified] Cox proportional hazards model.

Example data

- Worcester Heart Attack Study

ID <dbl>	LENFOLY <dbl>	FSTAT <dbl>	AFB <dbl>	GENDER <dbl>
1	5.96	0	1	0
2	5.95	0	0	0
3	6.00	0	0	1
4	0.81	1	0	0
5	5.83	0	0	0
6	0.00	1	0	0
7	5.81	0	0	0
8	4.10	1	0	0
9	2.52	1	0	1
10	5.95	0	0	0

Survival time in years

Explanatory variable
0 → did not have afib
1 → had afib

Censoring variable
0 → censored
1 → death

Stratification factor
0 → male
1 → female

Data source: https://stats.idre.ucla.edu/sas/seminars/sas-survival/

Case 1: Non-stratified analysis

R codes

- survival & survminer packages

```r
# Data manipulation
## change AFB order to use "Yes" as the reference group
## to be consistent with SAS
dat$AFB <- factor(dat$AFB, levels = c(1, 0))

# KM estimators
fit.km <- survfit(Surv(LENFOLY, FSTAT) ~ AFB, data = dat)
## quantile estimates
quantile(fit.km, probs = c(0.25, 0.5, 0.75))
## landmark estimates at 1, 3, 5-year
summary(fit.km, times = c(1, 3, 5))

# Log-rank test from `survminer` package
survminer::surv_pvalue(fit.km, data = dat)

# Cox PH model
fit.cox <- coxph(Surv(LENFOLY, FSTAT) ~ AFB, data = dat)
summary(fit.cox)
```

SAS codes

- PROC LIFETEST & PROC PHREG

```sas
# KM estimators and log-rank test
proc lifetest data=whas500;
time lenfoly*fstat(0);
strata /group = afb;
run;

# Cox PH model
proc phreg data = whas500;
class afb;
model lenfol*fstat(0) = afb/rl;
run;
```

Case 1: R results vs. SAS results

	R results			SAS results	
	AFB=No	AFB=Yes		AFB=No	AFB=Yes
Event - n	168	47		168	47
Censored - n	254	31		254	31
Time to event (years)					
25th percentile (95% CI)	0.94 (0.55, 1.47)	0.26 (0.05, 1.11)		0.94 (0.51, 1.45)	0.26 (0.05, 0.90)
Median (95% CI)	5.91 (4.32, NE)	2.37 (1.27, 4.24)		5.91 (4.31, 6.46)	2.37 (1.15, 3.77)
75th percentile (95% CI)	6.44 (6.44, NE)	6.43 (4.24, NE)		6.44 (6.44, 6.46)	6.43 (4.24, 6.43)
Min, Max	(0.0, 6.5)	(0.0, 6.4)		(0.0, 6.5)	(0.0, 6.4)
Landmark estimates					
1-year event-free rate (95% CI)	0.739 (0.699, 0.782)	0.641 (0.543, 0.757)		0.739 (0.695, 0.779)	0.641 (0.524, 0.736)
3-year event-free rate (95% CI)	0.642 (0.595, 0.691)	0.455 (0.351, 0.589)		0.642 (0.591, 0.687)	0.455 (0.335, 0.567)
5-year event-free rate (95% CI)	0.530 (0.472, 0.595)	0.315 (0.211, 0.470)		0.530 (0.467, 0.589)	0.315 (0.195, 0.442)
p-value		0.0010			0.0010
Hazard ratio (95% CI)		0.583 (0.421, 0.806)			0.584 (0.422, 0.808)

Case 1: Reason #1

- Reason for discrepancies in HR & CI:
  - Default methods for handling ties in Cox model are different:
    - R uses “efron”
    - SAS uses “breslow”
  - Both options are available in R and SAS, so by simply changing the default method, we would expect an identical HR and CI.

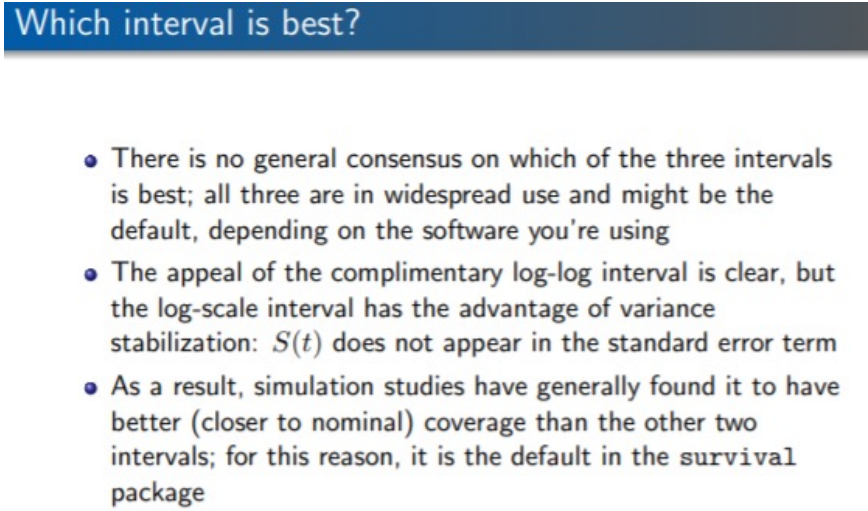
Ties

There are three possible choices for handling tied event times. The Breslow approximation is the easiest to program and hence became the first option coded for almost all computer routines. It then ended up as the default option when other options were added in order to "maintain backwards compatability". The Efron option is more accurate if there are a large number of ties, and it is the default option here. In practice the number of ties is usually small, in which case all the methods are statistically indistinguishable.

Case 1: Reason #2

- Reason for discrepancies in CIs for KM estimators:

- Default methods to calculate the CIs are different:
  - R uses “log”
  - SAS uses “log-log”
- Again, both options are available in R and SAS.



https://myweb.uiowa.edu/pbreheny/7210/f15/notes/9-10.pdf

Change default methods in R and SAS

R codes

- Change to default methods in SAS

```r
# Data manipulation
## change AFB order to use "Yes" as the reference group
## to be consistent with SAS
dat$AFB <- factor(dat$AFB, levels = c(1, 0))

# KM estimators
fit.km <- survfit(Surv(LENFOLY, FSTAT) ~ AFB,
                  conf.type = "log-log", data = dat)
## quantile estimates
quantile(fit.km, probs = c(0.25, 0.5, 0.75))
## landmark estimates at 1, 3, 5-year
summary(fit.km, times = c(1, 3, 5))

# Log-rank test from `survminer` package
survminer::surv_pvalue(fit.km, data = dat)

# Cox PH model
fit.cox <- coxph(Surv(LENFOLY, FSTAT) ~ AFB,
                 ties = "breslow", data = dat)
summary(fit.cox)
```

SAS codes

- Change to default methods in R

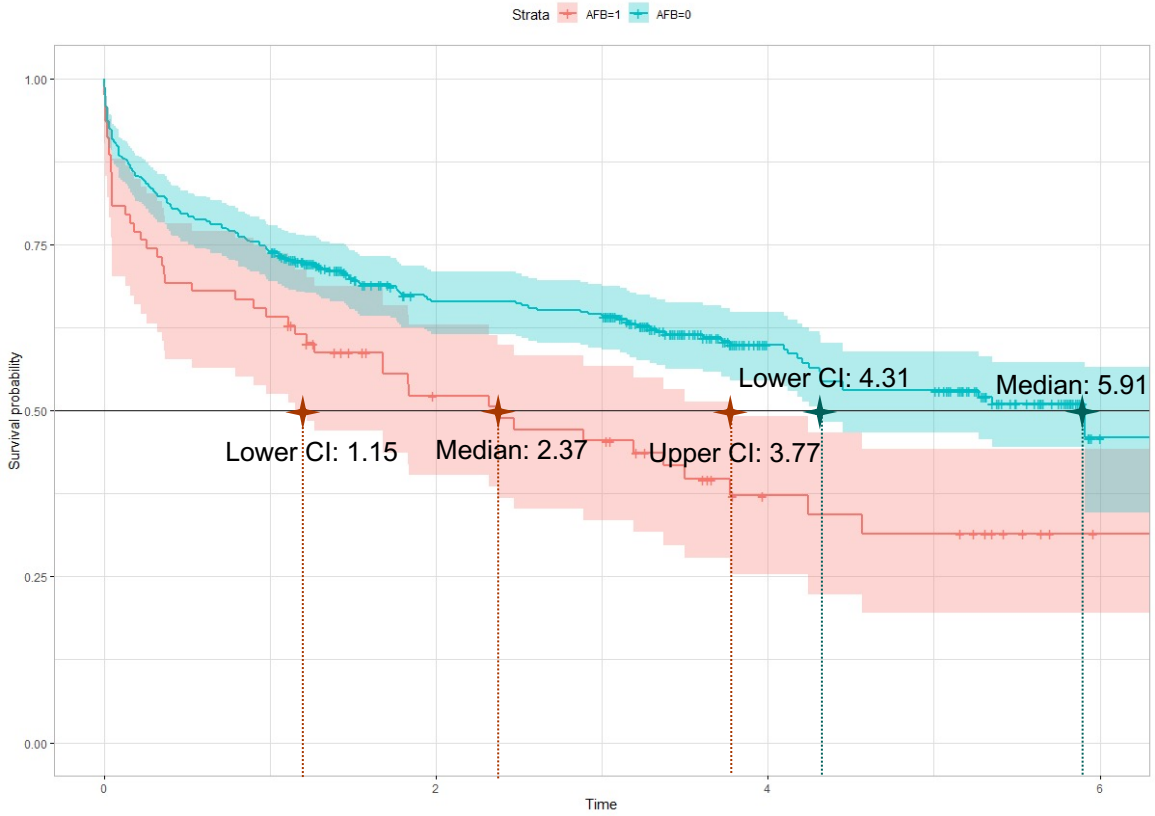
```sas
# KM estimators and log-rank test
proc lifetest data=whas500 conftype = log;
time lenfoly*fstat(0);
strata /group = afb;
run;

# Cox PH model
proc phreg data = whas500;
class afb;
model lenfol*fstat(0) = afb/rl ties = efron;
run;
```

Case 1: Non-stratified analysis updated

- Change the methods in R to SAS default choices:

	R results			SAS results	
	AFB=No	AFB=Yes		AFB=No	AFB=Yes
Event - n	168	47		168	47
Censored - n	254	31		254	31
Time to event (years)					
25th percentile (95% CI)	0.94 (0.51, 1.45)	0.26 (0.05, 0.90)		0.94 (0.51, 1.45)	0.26 (0.05, 0.90)
Median (95% CI)	5.91 (4.31, NE)	2.37 (1.15, 3.77)		5.91 (4.31, 6.46)	2.37 (1.15, 3.77)
75th percentile (95% CI)	6.44 (6.44, NE)	6.43 (4.24, NE)		6.44 (6.44, 6.46)	6.43 (4.24, 6.43)
Min, Max	(0.0, 6.5)	(0.0, 6.4)		(0.0, 6.5)	(0.0, 6.4)
Landmark estimates					
1-year event-free rate (95% CI)	0.739 (0.695, 0.779)	0.641 (0.524, 0.736)		0.739 (0.695, 0.779)	0.641 (0.524, 0.736)
3-year event-free rate (95% CI)	0.642 (0.591, 0.687)	0.455 (0.335, 0.567)		0.642 (0.591, 0.687)	0.455 (0.335, 0.567)
5-year event-free rate (95% CI)	0.530 (0.467, 0.589)	0.315 (0.195, 0.442)		0.530 (0.467, 0.589)	0.315 (0.195, 0.442)
p-value		0.0010			0.0010
Hazard ratio (95% CI)		0.584 (0.422, 0.808)			0.584 (0.422, 0.808)

Case 1: Reason #3



- The kth quantile for a survival curve is the location at which a horizontal line at height p = 1-k intersects the plot of survival probability.
- If the survival curve does not fall to 1-k, then that quantile is undefined.

Case 1: Reason #3 cont'd

Problem Note *64617:* The LIFETEST procedure produces incorrect upper confidence limits for the quartiles for certain data

Details | Hot Fix | About | Rate It

The upper confidence limit for a quartile q_p in PROC LIFETEST is incorrect if these conditions are met:

- an event is observed at the largest time
- the largest event time is also the first time the survival estimate drops below (1-*p*) for the 100pth percentile point, given the pointwise upper bounds for the survival estimate have not dropped below (1-*p*) at previous times before the largest time event

- This discrepancy may exist or may not, it depends on
  - SAS version
  - Data

Case 2 & Case 3

- Case 2: change the subject with largest time in Case 1 data to be censored
- Case 3: stratified analysis by gender
- R and SAS results are identical

Case 4: A special data

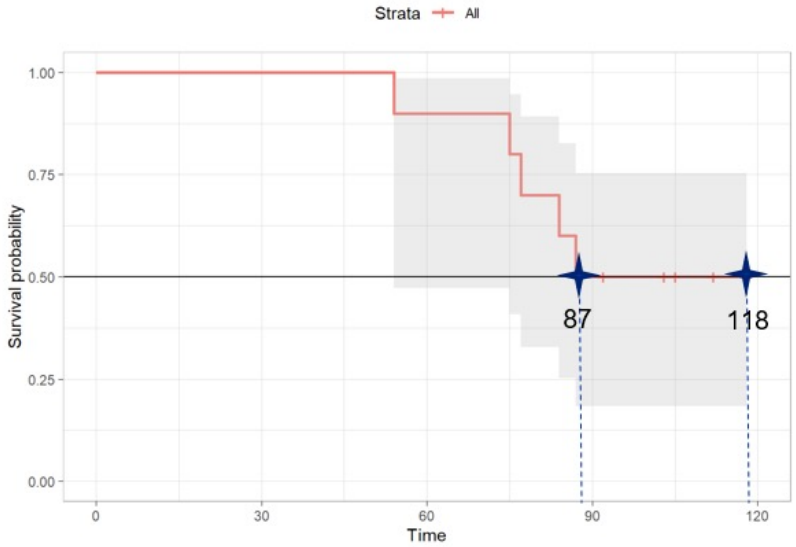
- Suppose a dataset with 10 observations: the first 5 are all events, and the last 5 are all censored.
- Identified in a subgroup analysis from a real clinical trial data.

ID	TIME	CNSR
1	54	1
2	75	1
3	77	1
4	84	1
5	87	1
6	92	0
7	103	0
8	105	0
9	112	0
10	118	0

	R results	SAS results
Event - n	5	5
Censored - n	5	5
Time to event (days)		
25th percentile (95% CI)	77 (54, NE)	77 (54, NE)
Median (95% CI)	102.5 (54, NE)	NE (54, NE)
75th percentile (95% CI)	NE (87, NE)	NE (87, NE)
Min, Max	(54, 118+)	(54, 118+)
Landmark estimates		
80-day event-free rate (95% CI)	0.7 (0.329, 0.892)	0.7 (0.329, 0.892)
100-day event-free rate (95% CI)	0.5 (0.184, 0.753)	0.5 (0.184, 0.753)
120-day event-free rate (95% CI)	0.5 (0.184, 0.753)	NE (NE, NE)

Case 4: Reason #1

- Reason for discrepancies in the median estimates:



- Survival probability is a step function, it is possible for the curve to have a horizontal segment at exactly 1-k, in which case the midpoint of the horizontal segment is returned in R:

(87+118) / 2 = 102.5

Case 4: Reason #1 cont'd

- Reason for discrepancies in the median estimates:

Product-Limit Survival Estimates						
aval		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000		1.0000	0	0	0	10
54.000		0.9000	0.1000	0.0949	1	9
75.000		0.8000	0.2000	0.1265	2	8
77.000		0.7000	0.3000	0.1449	3	7
84.000		0.6000	0.4000	0.1549	4	6
87.000		0.5000	0.5000	0.1581	5	5
92.000	*	.	.	.	5	4
103.000	*	.	.	.	5	3
105.000	*	.	.	.	5	2
112.000	*	.	.	.	5	1
118.000	*	.	.	.	5	0

Let $t_1 < t_2 < \cdots < t_D$ represent the distinct event times.

The general formula for estimating the $100p$th percentile point is

$$q_p = \min\{t_j | \hat{S}(t_j) < 1 - p\}$$

$$q_{.25} = \min\{t_j | \hat{S}(t_j) < 0.75\}$$

If $\hat{S}(t)$ is exactly equal to 0.75 from t_j to t_{j+1}, the first quartile is taken to be $(t_j + t_{j+1})/2$.

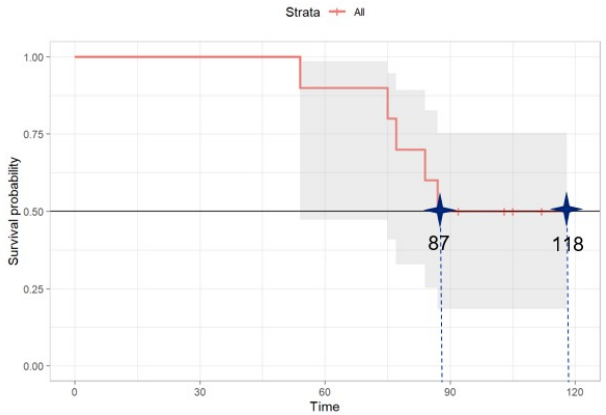
- SAS searches for the smallest **event** time for which the survival estimate is **<** 0.5.
- This does not exist in the dataset, so SAS gives "NE".

Case 4: Reason #2

- Reason for discrepancies in the 120-day event-free rates:

Landmark estimates		
80-day event-free rate (95% CI)	0.7 (0.329, 0.892)	0.7 (0.329, 0.892)
100-day event-free rate (95% CI)	0.5 (0.184, 0.753)	0.5 (0.184, 0.753)
120-day event-free rate (95% CI)	0.5 (0.184, 0.753)	NE (NE, NE)

ID	TIME	CNSR
1	54	1
2	75	1
3	77	1
4	84	1
5	87	1
6	92	0
7	103	0
8	105	0
9	112	0
10	118	0

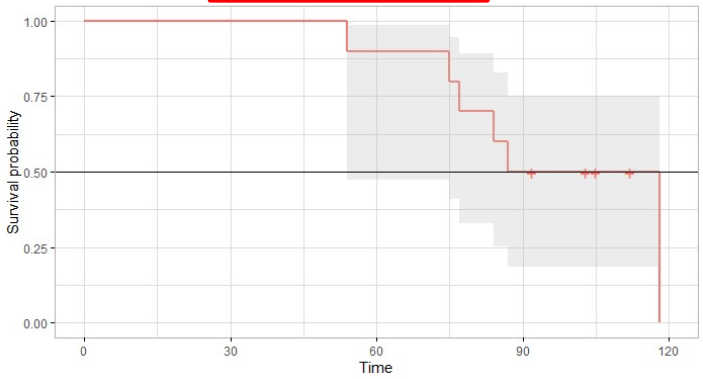


- 120-day is beyond the largest time in data, since it is unknown, SAS gives “NE”, but R uses the rate at the 118-day to estimate the 120-day event free rate.

Case 5

- If change the last observation to be an event

ID	TIME	CNSR
1	54	1
2	75	1
3	77	1
4	84	1
5	87	1
6	92	0
7	103	0
8	105	0
9	112	0
10	118	1



	R results	SAS results
Event - n	6	6
Censored - n	4	4
Time to event (days)		
25th percentile (95% CI)	77 (54, NE)	77 (54, NE)
Median (95% CI)	102.5 (54, NE)	102.5 (54, NE)
75th percentile (95% CI)	118 (87, NE)	118 (87, NE)
Min, Max	(54, 118)	(54, 118)
Landmark estimates		
80-day event-free rate (95% CI)	0.7 (0.329, 0.892)	0.7 (0.329, 0.892)
100-day event-free rate (95% CI)	0.5 (0.184, 0.753)	0.5 (0.184, 0.753)
120-day event-free rate (95% CI)	0 (NE, NE)	0 (NE, NE)

Summary

- Given independent implementations in languages with different structures some discrepancies should be expected.
- Can be due to several reasons including:
  - Different default choices
  - Different algorithms or implementations of common algorithms
  - The choice of statistical estimators
- Most of the time, both R and SAS yield correct results that match
- There could be more discrepancies than the ones explored in this presentation.

Conclusions

- It is to be expected that statistical models and analyses independently implemented in different computer languages may occasionally produce slightly different results.
- Any complex analysis presents software developers and statisticians with multiple choices of algorithms, statistical methodologies and parameterizations.
- While we have concentrated on identifying causes for discrepancies in survival results, other sub-teams in this project are investigating other statistical methodologies.
- A key finding from the project is that transparent and pro-active documentation internally as well as with external agencies is absolutely necessary to make informed decisions about choosing between language implementations.
- A white paper is in progress to cover the high value insights from R vs SAS use cases as examples in the in-depth exposition.

References

- Breheny P. “Inference for the Kaplan-Meier Estimator.” https://myweb.uiowa.edu/pbreheny/7210/f15/notes/9-10.pdf
- Breslow, N. E. (1974) “Covariance Analysis of Censored Survival Data.” Biometrics 30:89–99.
- Efron, B. (1977) “The Efficiency of Cox’s Likelihood Function for Censored Data.” Journal of the American Statistical Association 72:557–565.
- Emmerson J. and Brown J. M. “Understanding Survival Analysis in Clinical Trials.” Clinical Oncology 33:12-14.
- Franklin D. “Our Survival Confidence Intervals are not the Same!” PharmaSUG 2014 - Paper SP10. https://www.pharmasug.org/proceedings/2014/SP/PharmaSUG-2014-SP10.pdf
- Hertz-Picciotto I. and Rockhill B. (1997) “Validity and efficiency of approximation methods for tied survival times in Cox regression.” Biometrics 53:1151-1156.
- Hosmer, D.W. and Lemeshow, S. and May, S. (2008) “Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition.” John Wiley and Sons Inc., New York, NY
- Problem Note 64617: The LIFETEST procedure produces incorrect upper confidence limits for the quartiles for certain data. https://support.sas.com/kb/64/617.html
- R survival package manual
- SAS PROC LIFETEST Documentation
- SAS PROC PHREG Documentation

Survival Analysis in R vs. SAS

Thank You!!!
Questions?