Survival Analysis

PSY 512: Advanced Statistics for Psychological and Behavioral Research 2

Goals

- What is survival analysis?
- When do we use survival analysis?
- One-sample survival analysis
- Two-sample survival analysis
- Multisample survival analysis

What is Survival Analysis?

- Survival analysis is a family of statistical methods designed for the analysis of duration data (i.e., how long until an event occurs?)
  - How long will patients survive after being given a specific terminal diagnosis?
  - How long will former smokers abstain from cigarettes?
  - How long will it take graduate students to finish their degrees?
  - What factors influence how long married couples wait until the birth of their first child?
  - What factors influence when children reach developmental milestones?
- These techniques are also commonly known as "failure analysis"
  - This set of techniques was developed in medicine and engineering

What is Survival Time?

- Survival time refers to a variable which measures the time from a particular starting time (e.g., time initiated the treatment) to a particular endpoint of interest (e.g., attaining certain functional abilities)
  - Start of treatment → Time of death
  - Start of treatment → Development of functional ability
  - Time of marriage → Birth of first child
  - Onset of sexual activity → Orgasm
- It is important to note that for some participants in our studies, a complete survival time may not be available due to censoring

Censoring

- Some participants may not have experienced the "event" at the end of the study
  - Terminal patients may still be alive at the end of the study period
  - Married couples may not have had their first child
  - Graduate students may still be in graduate school
- The exact survival times of these participants are unknown
- These are called *censored observations* or *censored times* and can also occur when individuals are lost to follow-up after a period of study
  - Major causes of censoring
    - Participant is lost to follow-up
    - The study is closed after a fixed period

Regression vs. Survival Analysis

Technique	Predictor Variables	Outcome Variables	Censoring Permitted?
Linear Regression	Categorical or continuous	Continuous	No
Logistic Regression	Categorical or continuous	Categorical (usually binary)	No
Survival Analysis	Time as well as categorical or continuous	Binary	Yes

Regression vs. Survival Analysis

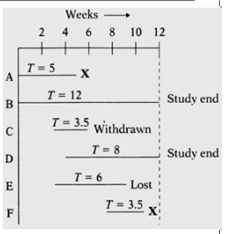
Technique	Mathematical Model	Yields
Linear Regression	Y' = A + BX	Linear Association
Logistic Regression	Ln(P/1-P) = A + BX	Odds Ratio
Survival Analysis	H(t) = ho(t) exp(A + BX)	Hazard Rate

Regression vs. Survival Analysis

- Survival analysis models the time to an event
  - Unlike linear regression, survival analysis has a dichotomous (binary) outcome
  - Unlike logistic regression, survival analysis analyzes the time to an event
- Survival analysis is able to account for censoring
- Can examine survival rate of a single group
- Can compare survival rates of two or more groups
- Assesses relationship between predictors and survival time

Types of Censored Data

- Right Censored Data
  - The end of the interval is unknown because the participant does not experience the event by the end of the study OR the investigator loses contact with the participant (loss to follow-up)
- Left Censored Data
  - The beginning of the interval is unknown for some reason (e.g., inadequate psychiatric records)
  - This is VERY difficult data to analyze and should be avoided whenever possible
  - The approaches that we will discuss are not designed for left censored data



Problems with Censored Data

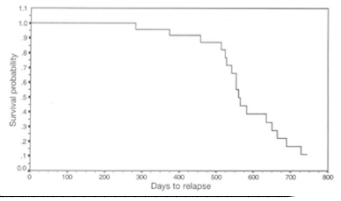
- We are going to follow the convention established by those who commonly use survival analysis and focus on right censored data (i.e., we will use "censored" to refer to "right censored")
- Problems with other approaches to dealing with censored data
  - Ignore the censored durations by only using data for those who experienced the event within the timeframe of the study
    - PROBLEM: Underestimates the time to the event because it systematically excludes those participants with longer times
  - Treat the censored cases as if the event occurred at the end of the timeframe
    - PROBLEM: Underestimates the time to the event by truncating the duration times for those participants with longer times
  - Ignore time and just focus on whether the event occurred within the timeframe of the study
    - PROBLEMS: Causes problems if participants enter the study at different time points (i.e., those entering later have less opportunity to experience the event by the end of the study) and it makes it harder to compare the results of your study with those of other studies of different durations

When to use survival analysis?

- Examples
  - Time to death or clinical endpoint
  - Time in remission after treatment of disease
  - Recidivism rate after addiction treatment
- When one believes that one or more predictor variables may be associated with the time until the occurrence of an event
- It is an especially valuable tool when follow-up is incomplete or variable

What does survival analysis tell us?

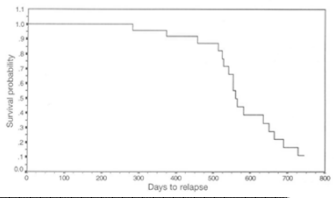
- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

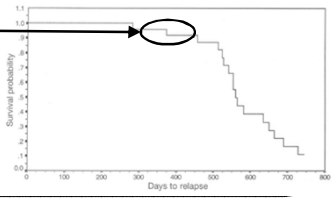
The survival function starts at one and tends to drop toward zero as time passes. The value is one at Time 0 because everyone is cigarette free when they complete the intervention.



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

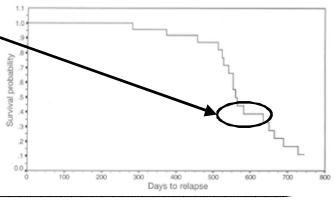
Proportion who were abstinent through the first 400 days was about .92



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

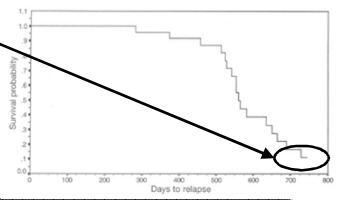
Proportion who were abstinent through the first 600 days was about .39



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

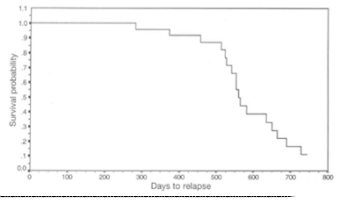
Proportion who were abstinent through the first 745 days was about .11



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

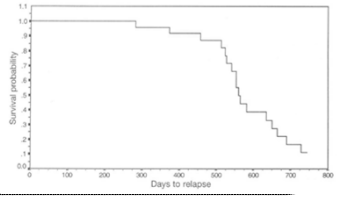
Participants who completed the intervention had a 92% probability of remaining abstinent for at least 400 days, a 39% chance of not smoking for 600 days, and a 11% chance of remaining abstinent for 745 days



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

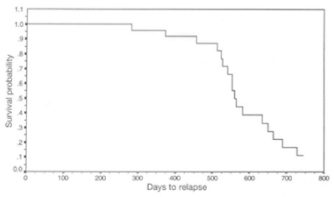
The probability of relapse and the probability of abstinence must add up to 100% so the plot can also be used to estimate relapse at each time point. It is 8% through 400 days, 61% through 600 days, and 89% through 745 days



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

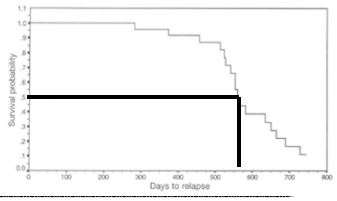
The steepness of the survival function reveals when the probability of relapse is unusually low or high. The plot is relatively flat through 500 days (i.e., low probability of relapse) but it is very steep between 500 and 600 days indicating a high probability of relapse during this period



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

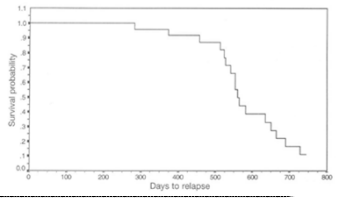
The plot can also be used to estimate the average survival time for participants. Generally the median survival time is used which can be determined by linear interpolation. It is about 560 days in this example.



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

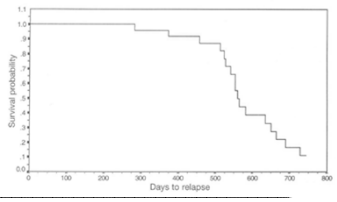
The survival function cannot rise over time because once someone relapses then it cannot be undone



What does survival analysis tell us?

- Survival analysis estimates duration by computing a survival function which estimates the probability that a participant will survive (i.e., experience an event) past a specified time
- Example: Time to relapse for those who quit smoking

The survival function ends at 745 days for this example which means that we do not know the probability of abstaining for 800 or 1,000 days except that it must be less than or equal to the probability of abstaining for 745 days



Assumptions of Survival Analysis

- Participants must be independent
- The event must represent a change from one state to another
  - Events must be mutually exclusive and collectively exhaustive of all outcomes
- Participants are event free when they enter the study
- The survival analysis techniques assume that the time data is continuous
  - Data should be collected across relatively small intervals rather than larger intervals
  - The "best" interval depends on the event time such that days may be appropriate for death rates following terminal diagnosis but minutes may be best if studying time until orgasm
- Censoring should be unrelated to the probability of event occurrence (independent-censoring assumption)
  - This is violated if the participants who drop out of a study are at an unusually low or high probability of experiencing an event

Computing Survival Estimates

- Two essential pieces of information are recorded for each participant
  - The known abstinence time for each participants (recorded in days)
  - The status of the participant on the last day he or she was observed ("event" is used for those who relapsed and "censored" is used for those who did not)

Survival Times for Hypothetical Smoking-Cessation Clients

Client	Day	Status
1	457	Event
2	565	Event
3	461	Censored
4	581	Event
5	541	Event
6	283	Event
7	729	Event
8	553	Event
9	559	Event
10	417	Censored
11	527	Censored
12	481	Censored
13	527	Event
14	399	Censored
15	745	Censored
16	523	Event
17	373	Event
18	689	Event
19	665	Event
20	651	Event
21	635	Event
22	553	Event
23	745	Censored
24	513	Event

Computing Survival Estimates

- The Kaplan-Meier survival function is generally used for this sort of data

Kaplan–Meier Survival Table for Hypothetical Smoking-Cessation Clients

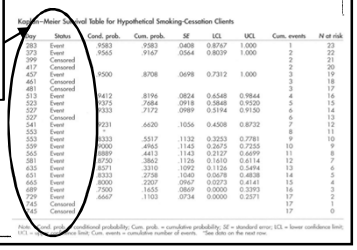
Day	Status	Cond. prob.	Cum. prob.	SE	LCL	UCL	Cum. events	N at risk
283	Event	.9583	.9583	.0408	0.8767	1.000	1	23
373	Event	.9565	.9167	.0564	0.8039	1.000	2	22
399	Censored						2	21
417	Censored						2	20
457	Event	.9500	.8708	.0698	0.7312	1.000	3	19
461	Censored						3	18
481	Censored						3	17
513	Event	.9412	.8196	.0824	0.6548	0.9844	4	16
523	Event	.9375	.7684	.0918	0.5848	0.9520	5	15
527	Event	.9333	.7172	.0989	0.5194	0.9150	6	14
527	Censored						6	13
541	Event	.9231	.6620	.1056	0.4508	0.8732	7	12
553	Event						8	11
553	Event	.8333	.5517	.1132	0.3253	0.7781	9	10
559	Event	.9000	.4965	.1145	0.2675	0.7255	10	9
565	Event	.8889	.4413	.1143	0.2127	0.6699	11	8
581	Event	.8750	.3862	.1126	0.1610	0.6114	12	7
635	Event	.8571	.3310	.1092	0.1126	0.5494	13	6
651	Event	.8333	.2758	.1040	0.0678	0.4838	14	5
665	Event	.8000	.2207	.0967	0.0273	0.4141	15	4
689	Event	.7500	.1655	.0869	0.0000	0.3393	16	3
729	Event	.6667	.1103	.0734	0.0000	0.2571	17	2
745	Censored						17	1
745	Censored						17	0

Note. Cond. prob. = conditional probability; Cum. prob. = cumulative probability; SE = standard error; LCL = lower confidence limit; UCL = upper confidence limit; Cum. events = cumulative number of events. *See data on the next row.

Computing Survival Estimates

- The Kaplan-Meier survival function is generally used for this sort of data

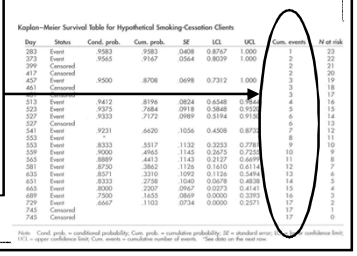
The first step is to sort the cases from shortest duration to longest duration. Censored cases are listed after relapses for any particular length (e.g., 527 days)



Computing Survival Estimates

- The Kaplan-Meier survival function is generally used for this sort of data

For each duration, the cumulative number of events (i.e., relapses) and the number of participants who remain at risk (i.e., under observation) are both recorded. The column "cumulative events" tells us how many events have occurred. One event occurred by Day 283 and three events occurred by Day 457.



Computing Survival Estimates

- The Kaplan-Meier survival function is generally used for this sort of data

This column shows the number of participants who remained under observation after each status change. The number at risk decreases because of an event (i.e., relapse) or censor (i.e., participant drops out of study)

Kaplan–Meier Survival Table for Hypothetical Smoking-Cessation Clients

Day	Status	Cond. prob.	Cum. prob.	SE	LCL	UCL	Cum. events	N at risk
283	Event	.9583	.9583	.0408	0.8767	1.000	1	23
373	Event	.9565	.9167	.0564	0.8039	1.000	2	22
399	Censored						2	21
417	Censored						2	20
457	Event	.9500	.8708	.0698	0.7312	1.000	3	19
461	Censored						3	18
481	Censored						3	17
513	Event	.9412	.8196	.0824	0.6548	0.9844	4	16
523	Event	.9375	.7684	.0918	0.5848	0.9520	5	15
527	Event	.9333	.7172	.0989	0.5194	0.9150	6	14
527	Censored						6	13
541	Event	.9231	.6620	.1056	0.4508	0.8732	7	12
553	Event						8	11
553	Event	.8333	.5517	.1132	0.3253	0.7781	9	10
559	Event	.9000	.4965	.1145	0.2675	0.7255	10	9
565	Event	.8889	.4413	.1143	0.2127	0.6699	11	8
581	Event	.8750	.3862	.1126	0.1610	0.6114	12	7
635	Event	.8571	.3310	.1092	0.1126	0.5494	13	6
651	Event	.8333	.2758	.1040	0.0678	0.4838	14	5
665	Event	.8000	.2207	.0967	0.0273	0.4141	15	4
689	Event	.7500	.1655	.0869	0.0000	0.3393	16	3
729	Event	.6667	.1103	.0734	0.0000	0.2571	17	2
745	Censored						17	1
745	Censored						17	0

Note. Cond. prob. = conditional probability; Cum. prob. = cumulative probability; SE = standard error; LCL = lower confidence limit; UCL = upper confidence limit; Cum. events = cumulative number of events. *See dots on the next row.

Computing Survival Estimates

- The Kaplan-Meier survival function is generally used for this sort of data

The conditional survival probability is the probability of surviving past a particular time for individuals who were at risk for relapse. For example, 23 participants remained in the study after the first event (Day 283). Of these 23 participants, 22 of them survived past the next event (Day 373) which gave a conditional probability of 22/23 = .9565

Kaplan–Meier Survival Table for Hypothetical Smoking-Cessation Clients

Day	Status	Cond. prob.	Cum. prob.	SE	LCL	UCL	Cum. events	N at risk
283	Event	.9583	.9583	.0408	0.8767	1.000	1	23
373	Event	.9565	.9167	.0564	0.8039	1.000	2	22
399	Censored						2	21
417	Censored						2	20
457	Event	.9500	.8708	.0698	0.7312	1.000	3	19
461	Censored						3	18
481	Censored						3	17
513	Event	.9412	.8196	.0824	0.6548	0.9844	4	16
523	Event	.9375	.7684	.0918	0.5848	0.9520	5	15
527	Event	.9333	.7172	.0989	0.5194	0.9150	6	14
527	Censored						6	13
541	Event	.9231	.6620	.1056	0.4508	0.8732	7	12
553	Event						8	11
553	Event	.8333	.5517	.1132	0.3253	0.7781	9	10
559	Event	.9000	.4965	.1145	0.2675	0.7255	10	9
565	Event	.8889	.4413	.1143	0.2127	0.6699	11	8
581	Event	.8750	.3862	.1126	0.1610	0.6114	12	7
635	Event	.8571	.3310	.1092	0.1126	0.5494	13	6
651	Event	.8333	.2758	.1040	0.0678	0.4838	14	5
665	Event	.8000	.2207	.0967	0.0273	0.4141	15	4
689	Event	.7500	.1655	.0869	0.0000	0.3393	16	3
729	Event	.6667	.1103	.0734	0.0000	0.2571	17	2
745	Censored						17	1
745	Censored						17	0

Computing Survival Estimates

- The Kaplan-Meier survival function is generally used for this sort of data

The cumulative survival probability (or just "survival probability') is what was shown in the earlier figure. This is found by multiplying together the conditional probabilities for all events up to and including the event time of interest

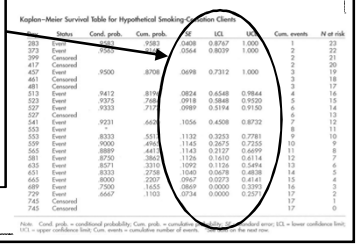
Kaplan–Meier Survival Table for Hypothetical Smoking-Cessation Clients

Day	Status	Cond. prob.	Cum. prob.	SE	LCL	UCL	Cum. events	N at risk
283	Event	.9583	.9583	.0408	0.8767	1.000	1	23
373	Event	.9565	.9167	.0564	0.8039	1.000	2	22
399	Censored						2	21
417	Censored						2	20
457	Event	.9500	.8708	.0698	0.7312	1.000	3	19
461	Censored						3	18
481	Censored						3	17
513	Event	.9412	.8196	.0824	0.6548	0.9844	4	16
523	Event	.9375	.7684	.0918	0.5848	0.9520	5	15
527	Event	.9333	.7172	.0989	0.5194	0.9150	6	14
527	Censored						6	13
541	Event	.9231	.6620	.1056	0.4508	0.8732	7	12
553	Event						8	11
553	Event	.8333	.5517	.1132	0.3253	0.7781	9	10
559	Event	.9000	.4965	.1145	0.2675	0.7255	10	9
565	Event	.8889	.4413	.1143	0.2127	0.6699	11	8
581	Event	.8750	.3862	.1126	0.1610	0.6114	12	7
635	Event	.8571	.3310	.1092	0.1126	0.5494	13	6
651	Event	.8333	.2758	.1040	0.0678	0.4838	14	5
665	Event	.8000	.2207	.0967	0.0273	0.4141	15	4
689	Event	.7500	.1655	.0869	0.0000	0.3393	16	3
729	Event	.6667	.1103	.0734	0.0000	0.2571	17	2
745	Censored						17	1
745	Censored						17	0

Computing Survival Estimates

- The Kaplan-Meier survival function is generally used for this sort of data

Survival probabilities are easy to compute but they are only point estimates for the population value for a particular event time. Therefore, it is often important to compute 95% confidence intervals for each sample probability that estimate the likely range of values for the population survival probability



Kaplan–Meier Survival Table for Hypothetical Smoking-Cessation Clients

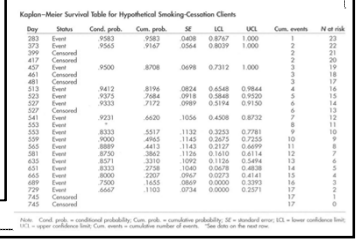
Day	Status	Cond. prob.	Cum. prob.	SE	LCL	UCL	Cum. events	N at risk
283	Event	.9583	.9583	.0408	0.8767	1.000	1	23
373	Event	.9565	.9167	.0564	0.8039	1.000	2	22
399	Censored						2	21
417	Censored						2	20
457	Event	.9500	.8708	.0698	0.7312	1.000	3	19
461	Censored						3	18
481	Censored						3	17
513	Event	.9412	.8196	.0824	0.6548	0.9844	4	16
523	Event	.9375	.7684	.0918	0.5848	0.9520	5	15
527	Event	.9333	.7172	.0989	0.5194	0.9150	6	14
527	Censored						6	13
541	Event	.9231	.6620	.1056	0.4508	0.8732	7	12
553	Event						8	11
553	Event	.8333	.5517	.1132	0.3253	0.7781	9	10
559	Event	.9000	.4965	.1145	0.2675	0.7255	10	9
565	Event	.8889	.4413	.1143	0.2127	0.6699	11	8
581	Event	.8750	.3862	.1126	0.1610	0.6114	12	7
635	Event	.8571	.3310	.1092	0.1126	0.5494	13	6
651	Event	.8333	.2758	.1040	0.0678	0.4838	14	5
665	Event	.8000	.2207	.0967	0.0273	0.4141	15	4
689	Event	.7500	.1655	.0869	0.0000	0.3393	16	3
729	Event	.6667	.1103	.0734	0.0000	0.2571	17	2
745	Censored						17	1
745	Censored						17	0

Note. Cond. prob. = conditional probability; Cum. prob. = cumulative probability; SE = standard error; LCL = lower confidence limit; UCL = upper confidence limit; Cum. events = cumulative number of events. *See data on the next row.

Computing Survival Estimates

- The Kaplan-Meier survival function is generally used for this sort of data

A couple of final notes about the Kaplan-Meier Table. First, survival probabilities are only computed when an event occurs. Second, censored cases are only used in computations that precede the time of censoring (i.e., the Kaplan-Meier method uses what is known about censored data without guessing when an event took place)



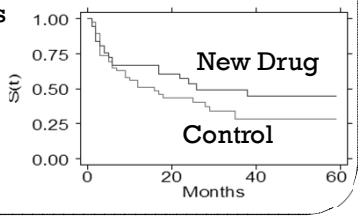
Kaplan–Meier Survival Table for Hypothetical Smoking-Cessation Clients

Day	Status	Cond. prob.	Cum. prob.	SE	LCL	UCL	Cum. events	N at risk
283	Event	.9583	.9583	.0408	0.8767	1.000	1	23
373	Event	.9565	.9167	.0564	0.8039	1.000	2	22
399	Censored						2	21
417	Censored						2	20
457	Event	.9500	.8708	.0698	0.7312	1.000	3	19
461	Censored						3	18
481	Censored						3	17
513	Event	.9412	.8196	.0824	0.6548	0.9844	4	16
523	Event	.9375	.7684	.0918	0.5848	0.9520	5	15
527	Event	.9333	.7172	.0989	0.5194	0.9150	6	14
527	Censored						6	13
541	Event	.9231	.6620	.1056	0.4508	0.8732	7	12
553	Event						8	11
553	Event	.8333	.5517	.1132	0.3253	0.7781	9	10
559	Event	.9000	.4965	.1145	0.2675	0.7255	10	9
565	Event	.8889	.4413	.1143	0.2127	0.6699	11	8
581	Event	.8750	.3862	.1126	0.1610	0.6114	12	7
635	Event	.8571	.3310	.1092	0.1126	0.5494	13	6
651	Event	.8333	.2758	.1040	0.0678	0.4838	14	5
665	Event	.8000	.2207	.0967	0.0273	0.4141	15	4
689	Event	.7500	.1655	.0869	0.0000	0.3393	16	3
729	Event	.6667	.1103	.0734	0.0000	0.2571	17	2
745	Censored						17	1
745	Censored						17	0

Note. Cond. prob. = conditional probability; Cum. prob. = cumulative probability; SE = standard error; LCL = lower confidence limit; UCL = upper confidence limit; Cum. events = cumulative number of events. *See data on the next row.

Two-Sample Survival Analysis

- We can also use survival analysis to determine whether groups differ in the time until an event
- The plot shows the probability of surviving up to 60 months after receiving a terminal diagnosis for two groups (New Drug and Control)

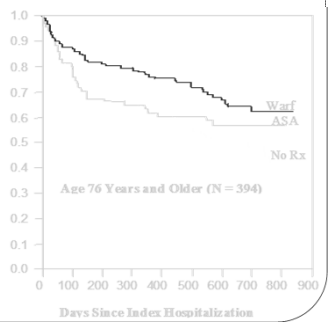


Two-Sample Survival Analysis

- It is possible to determine whether two survival functions were likely to have come from the same population or not
  - This is similar to the use of t-tests or ANOVAs to compare group means but the goal is to compare survival functions rather than average scores
- One approach is to use the log-rank test (also known as the Mantel-Cox statistic)
  - It is a "whole-pattern" test that compares the entire survival function rather than a difference at a particular time point
- Other tests that may be helpful
  - Breslow test: more sensitive for group differences at early time points
  - Tarone-Ware test: more sensitive if the survival functions do not differ by a constant factor (e.g., the functions diverge over time or they intersect)

Multisample Survival Analysis

- It is also possible to determine whether more than two survival functions were likely to have come from the same population or not
- This can be accomplished using an adjusted version of the log-rank test (the degrees of freedom have to be adjusted for more than two groups)
  - Then use pairwise testing similar to ANOVA



Survival Regression

- The previous analyses used categorical predictor variables but you can also examine continuous predictors using Cox regression analysis
  - This allows you to look at associations between one or more continuous or categorical variables and survival
  - Similar to linear regression (yields regression coefficients and predicted values, allows for interactions)
  - Linear regression predicts scores on a continuous outcome variable whereas Cox regression is used to predict the rate of occurrence
    - Also, Cox regression can handle censored data appropriately

Cox Regression

- Survival function – which is referred to as S(t) – defines the probability of surviving longer than time *t*
  - This is what the Kaplan-Meier curves show
  - Cox regression uses a hazard function which is the derivative of the survivor function over time
    - In other words, it is the instantaneous risk of event at time *t* (conditional failure rate)
    - "Hazard" is a neutral term which simply means risk for an event
- Survivor and hazard functions can be converted into each other

Summary

- Survival analyses quantify the time to a single, dichotomous event
- Handles censored data well
- Kaplan-Meier survival curves can be compared statistically and graphically
- Cox proportional hazards models help distinguish individual contributions of predictors