BIOS: 4120 Lab 15

May 1-2, 2018

Survival Analysis

In lecture this week, we briefly discussed the concepts of survival analysis. Today's lab will focus on Kaplan-Meier survival curves and the log-rank test. We will examine the aplastic anemia dataset using R. The dataset contains five variables:

. Trt: Whether the patient received Methotrexate (MTX) or Methotrexate and cyclosporine (MTX+CSP).
. Time_gvhd: Time until graft-versus-host disease. Measured in days.
. Status_gvhd: What happened at the end of Time_gvhd. The patient was either censored (0) or developed graft-versus-host disease (1).
. Time: Time until death. Measured in days.
. Status: What happened at the end of Time. The patient was either censored (0) or died (1).

Two common endpoints in survival analysis are Overall Survival and Progression-Free Survival. In this dataset, Overall Survival is time since randomization until death and Progression-Free Survival is time since randomization until GVHD.

Kaplan-Meier Estimates

A **survival function** is a function of time, and is defined as the probability of the event in question not occurring by time t (i.e., the patient surviving until time t or later).

Ex: S(10) = .95 means there is a 95% chance of surviving until day 10 (or equivalently, only a 5% chance of dying by day 10).

The most popular way to estimate survival functions is using Kaplan-Meier estimates. To estimate the Overall Survival function, do the following:

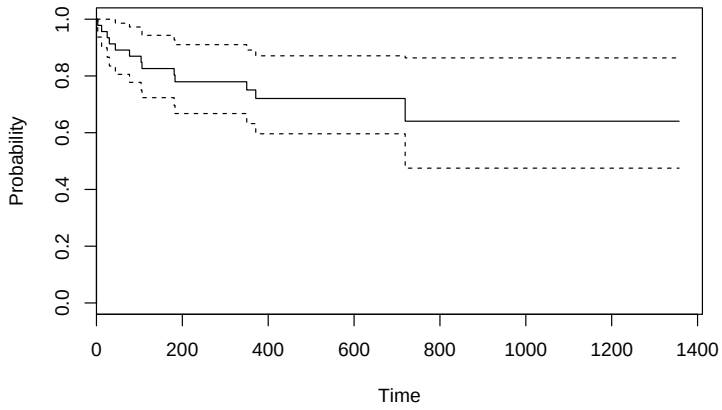
```
library(survival)
anemia <- read.delim("http://myweb.uiowa.edu/pbreheny/data/anemia.txt")
S <- with(anemia, Surv(Time,Status!=0))
fit <- survfit(S~1)
```

The survfit function calculates the survival curve that we learned how to compute in class. Recall if you know the time of death and number of subjects at risk, we can calculate survival probability. For example, here is the survival probability estimated at the first five times when a death occurred and the cumulative product of survival used to estimate the survival curve:

```
##      time n(t) d(t) [n(t)-d(t)]/n(t) cumproduct
## [1,]    3   46    1           0.9783     0.9783
## [2,]   12   45    1           0.9778     0.9565
## [3,]   25   44    1           0.9773     0.9348
## [4,]   30   43    1           0.9767     0.9130
## [5,]   44   42    1           0.9762     0.8913
```

To plot the entire estimated survival curve, use:

```
plot(fit, ylab = "Probability", xlab = "Time")
```



This is the Kaplan-Meier survival function estimate of the survival function, ignoring the different treatment groups.

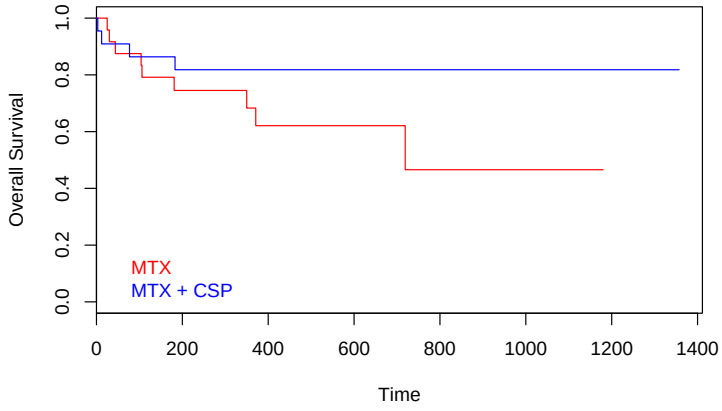
Why do the confidence intervals seem to get wider as time progresses?

What do the steps represent?

What is the median survival time?

We can also stratify by treatment group and examine both survival estimates.

```
fit2 <- with(anemia, survfit(S~Trt))
plot(fit2, ylab = "Overall Survival", xlab = "Time", col =
c("red","blue"))
legend("bottomleft", c("MTX","MTX + CSP"), text.col = c("red","blue"),
bty = "n")
```



Log-rank test

If you are curious, here is the code you would use to conduct a log-rank test to determine if treatment type significantly improves survival

```
survdiff(Surv(anemia$Time, anemia$Status) ~ anemia$Trt)
```

```
## Call:
## survdiff(formula = Surv(anemia$Time, anemia$Status) ~ anemia$Trt)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## anemia$Trt=MTX       24        9     6.45     1.007      2.01
## anemia$Trt=MTX+CSP   22        4     6.55     0.992      2.01
##
##  Chisq= 2  on 1 degrees of freedom, p= 0.156
```

Does the p-value indicate survival is significantly different between the two treatment groups?

Final Review

Note: While we tried to cover to a variety of topics from the course, there is not enough time in lab for the review to cover all of the topics. Don't solely rely on the review material in this lab when studying for the final exam.

Types of Bias:

Selection bias Instead of random sampling, certain subgroups of the population were more likely to be included than others.

Nonresponse bias Nonresponders can differ from responders in many important ways

Perception bias The perception of benefit from a treatment (placebo effect)

Confounding Confounding is a major source of bias. In order to avoid confounding, we conduct randomized controlled experiments so that the control and treatment groups are as similar as possible.

Errors:

```
##                H0 True H0 False
## Reject               A        B
## Fail to Reject       C        D
```

Type I Error

A Type I error is committed when a true null hypothesis is rejected.
In terms of disease detection (where the null hypothesis is no disease), this is a false positive.
In the table above, this is A.

Type I Error Rate (α)

The Type I error rate is the proportion of true hypotheses that were rejected.
In the table above, this is A/(A+C).

Type II Error

A Type II error is committed when a false null hypothesis is not rejected.
In the table above, this is D.

Type II Error Rate (β)

The Type II error rate is the proportion of false null hypotheses that failed to be rejected.
In the table above, this is D/(B+D).

Probability:

$$\text{Addition Rule: } P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\text{Complement Rule: } P(A^C) = 1 - P(A)$$

$$\text{Multiplication Rule: } P(A \cap B) = P(A)P(B|A)$$

$$\text{Law of Total Probability: } P(A) = P(A \cap B) + P(A \cap B^c)$$

$$\text{Bayes' Rule: } P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A^c)P(B|A^c)}$$

Sensitivity

Sensitivity is the probability of a patient testing positive for a disease given that the patient has the disease. This is often denoted by $P(+|D)$, where $+$ indicates a positive test result and D indicates having the disease.

Specificity

Specificity is the probability of a patient testing negative for a disease given that the patient does not have the disease. This is often denoted by $P(-|D^c)$, where $-$ indicates a negative test result and D^c indicates not having the disease.

Hypothesis Testing:

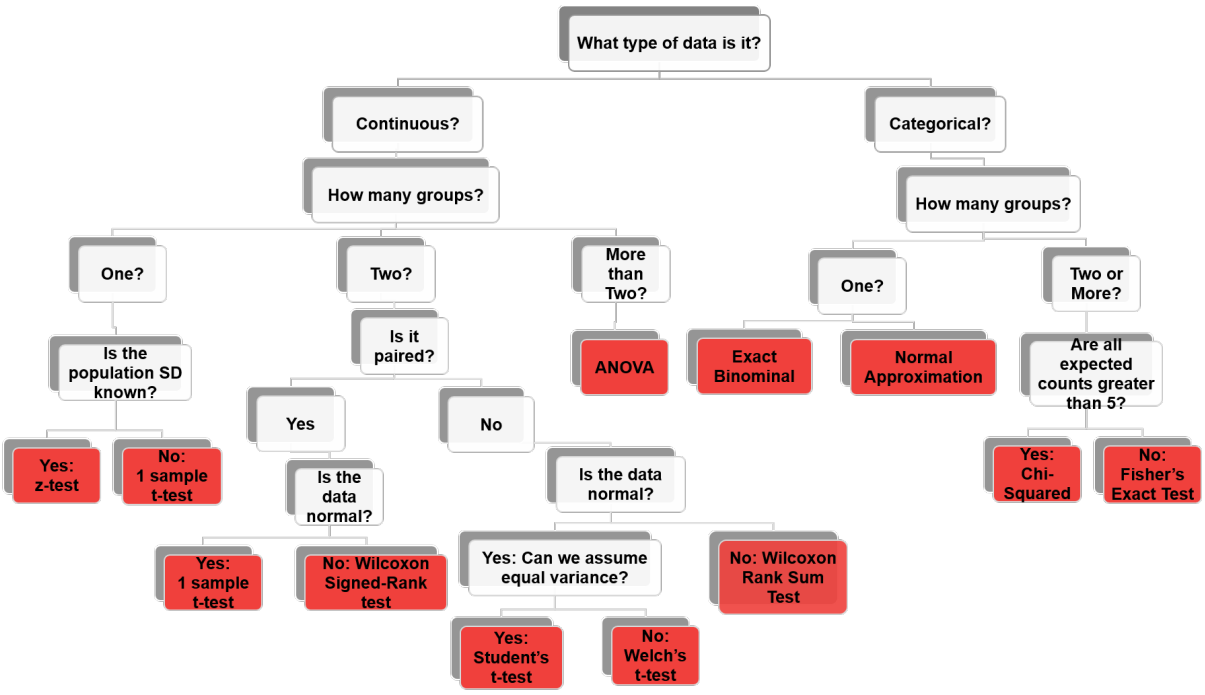


Figure 1: Flowchart Hypothesis Testing

Practice Problems

Problem 1

1256 individuals were tested in a saliva-based screening test for HIV. We know that 368 of the individuals tested have HIV, and 358 of them tested positive in the saliva-based screening test. Overall there were a total of 360 positive test results.

a. Construct a contingency table for the data.
b. What is the sensitivity of the saliva test?
c. What is the specificity?

Problem 2

The distribution of LDL cholesterol levels in a certain population is approximately normal with mean 90 mg/dl and standard deviation 8 mg/dl.

a. What is the probability an individual will have a LDL cholesterol level above 100 mg/dl?

b. Suppose we have a sample of 5 people from this population. What is the probability that at least one of them having levels above 100 mg/dl?

Problem 3

A psychologist was interested in exploring whether or not male and female college students have different driving behaviors. She focused on the fastest speed ever driven by an individual to see if the mean fastest speed driven by male college students differs from than the mean fastest speed driven by female college students. She surveyed 34 male college students and 29 female college students. The mean for males was 105.5 mph while the mean for females was 90.9 mph. The two samples had a pooled standard deviation of 16.9.

a. Conduct a t-test comparing the two groups.

b. Construct a 95% confidence interval for this difference.

Problem 4

A team from Yale School of Medicine took a look at 1,433 people diagnosed with intracranial meningioma, the most commonly diagnosed brain tumor in the United States. Researchers compared these patients to a test group of 1,350 people without tumors. Participants offered self-reported lifetime dental X-ray histories. Researchers then analyzed the X-rays that these two groups had undergone.

a. What type of study is this? What type of test would you perform?

b. If the odds ratio calculated for this study turned out to be (0.64, 1.15), what would you conclude?

Answers

Problem 1

a.

```
##        HIV No HIV
## Test + 358      2
## Test -  10    886
```

b. $358/368 = 0.973$

c. $886/888 = 0.998$

Problem 2

a.

```
(z <- (100-90)/8)
```

```
## [1] 1.25
```

```
(p <- pnorm(z,lower.tail=FALSE))
```

```
## [1] 0.1056498
```

When you use the table, you get the area below that value. In order to get the probability of an LDL level above 100, subtract this value from 1.

b.

```
1-dbinom(0,5,p)
```

```
## [1] 0.4278128
```

Use the binomial distribution to calculate this by hand: $\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$

Problem 3

a. $SE = 16.9 * \sqrt{\frac{1}{34} + \frac{1}{29}} = 4.272$

$t = \frac{105.5-90.9}{4.272} = 3.42$

From the table, the p-value with $df = 61$ is between 0.001 and 0.005.

b. $(105.5 - 90.9) \pm 2.00(4.272)$

14.6 ± 8.544

$(6.06, 23.14)$

Problem 4

a. Retrospective; Chi-square or Fisher's Exact

b. The confidence interval contains 1, so this would suggest that it is not significant.