Introduction to survival analysis

Yue Jiang

Duke University

A disclaimer

Today's (and next time's) lectures are introductory surface level treatments of survival analysis. We focus on applications and use cases -- there are no theoretical results presented (even for important subjects like variance estimation).

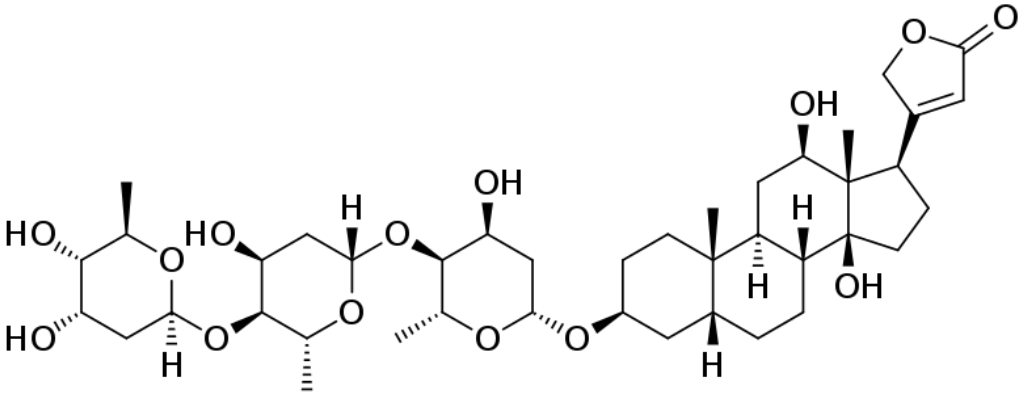
There is much to discuss regarding survival analysis both theoretically and in application. In STA 440, we will focus on using and implementing commonly used methods to tackle real-world datasets instead of focusing on theoretical considerations.

Survival data

In many studies, the outcome of interest is the amount of time from an initial observation until the occurrence of some event of interest.

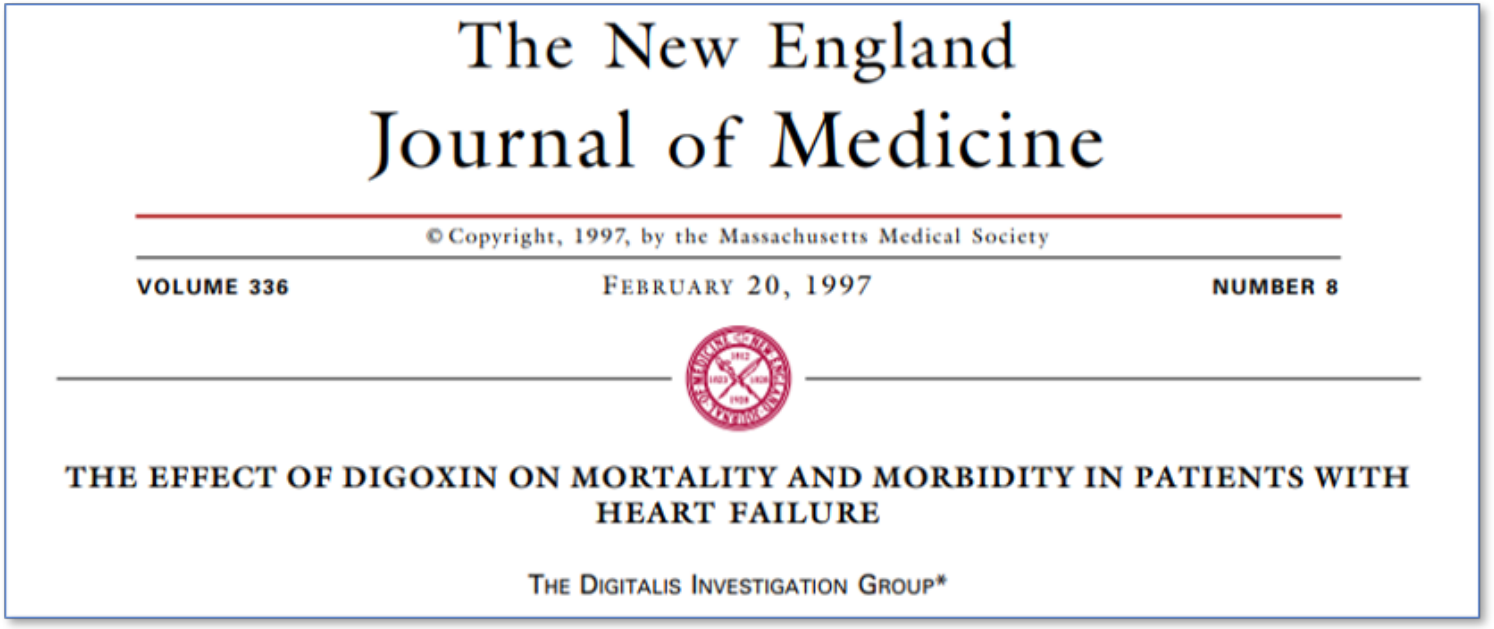
Typically, the event of interest is called a **failure** (even if it's a good thing), and the associated time interval between a starting point and failure the **failure time**, **survival time**, or **event time**.

Digoxin



- Foxgloves have been used in medicine for centuries
- Digoxin (the active ingredient) first isolated in 1930
- Traditionally used for heart arrhythmia and heart failure
- One of the most prescribed drugs globally

The DIG Trial



The New England
Journal of Medicine

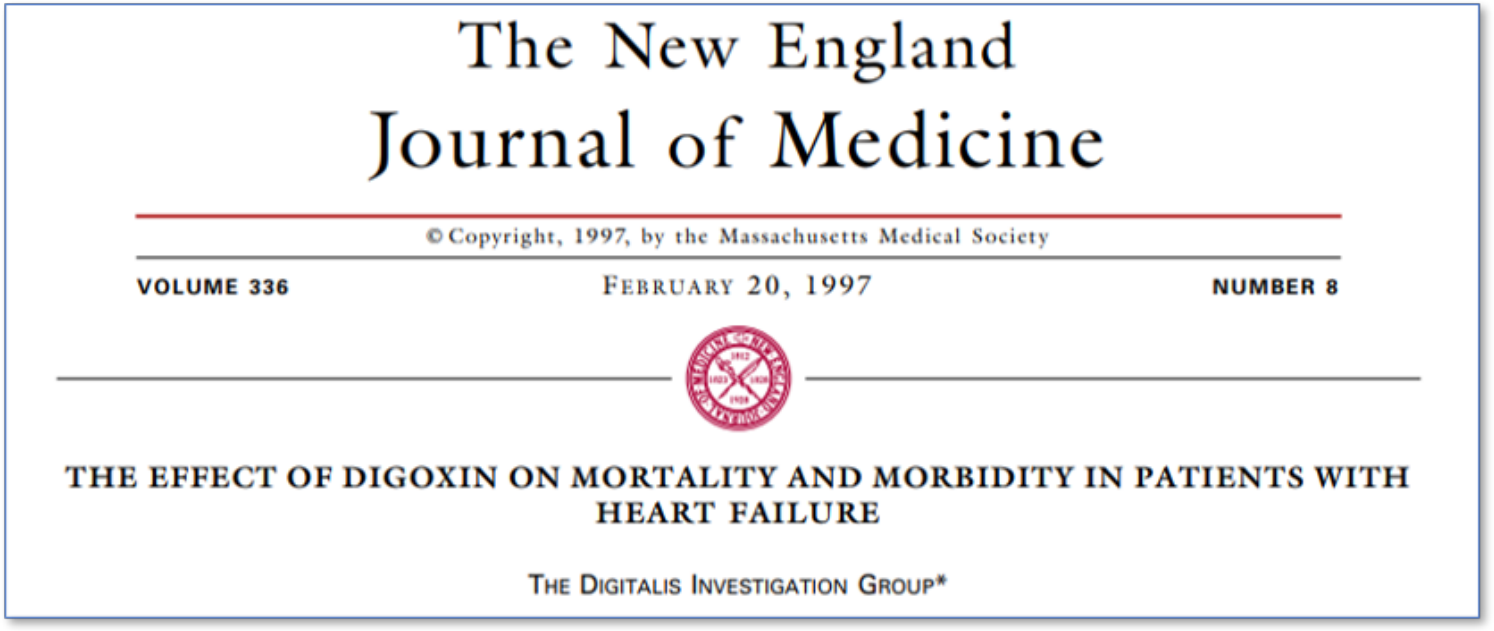
VOLUME 336 FEBRUARY 20, 1997 NUMBER 8

THE EFFECT OF DIGOXIN ON MORTALITY AND MORBIDITY IN PATIENTS WITH HEART FAILURE

THE DIGITALIS INVESTIGATION GROUP*

Investigators compared the **primary outcome** of the number of days from the start of the study to either death or hospitalization from worsening heart failure.

The DIG Trial



The New England
Journal of Medicine

VOLUME 336 FEBRUARY 20, 1997 NUMBER 8

THE EFFECT OF DIGOXIN ON MORTALITY AND MORBIDITY IN PATIENTS WITH HEART FAILURE

THE DIGITALIS INVESTIGATION GROUP*

How would ***you*** investigate this question, comparing the two treatment groups of digoxin vs. placebo?

A naive analysis

Death or hospitalization due to worsening heart failure:

```
dig %>%
  select(ID, TRTMT, DWHF, DWHFDAYS) %>%
  slice(1:10)
```

```
##    ID TRTMT DWHF DWHFDAYS
## 1   1     0    1     1379
## 2   2     0    1     1329
## 3   3     0    1      631
## 4   4     1    0     1157
## 5   5     0    1      191
## 6   6     0    0     1620
## 7   7     1    0      903
## 8   8     1    0     1369
## 9   9     0    0     1747
## 10 10     1    0     1074
```

A naive analysis

```
dig %>%
  filter(DWHF == 1) %>%
  t.test(DWHFDAYS ~ TRTMT, data = .)
```

```
##
##     Welch Two Sample t-test
##
## data:  DWHFDAYS by TRTMT
## t = -6.153, df = 2195.4, p-value = 9.01e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -133.68940  -69.06796
## sample estimates:
## mean in group 0 mean in group 1
##        418.8768        520.2555
```

Are you convinced? What if we made some sort of regression model to account for covariates? Would that be enough?

A naive analysis

```
dig %>%
  count(TRTMT)
```

```
##   TRTMT    n
## 1     0 3403
## 2     1 3397
```

```
dig %>%
  filter(DWHF == 1) %>%
  count(TRTMT)
```

```
##   TRTMT    n
## 1     0 1291
## 2     1 1041
```

Challenges

The unique nature of survival data is that typically not all units are observed until their event times:

- Maybe a patient moved to Fiji and was lost to follow-up
- Maybe a patient never experienced the primary outcome at all because they got hit by a bus
- Maybe the study was only funded to follow patients for two years after enrollment

In these cases, observations are said to be **censored** - we know that they survived until at least their censoring time, but do not know any further information.

Not accounting for censoring in an appropriate way leads to **biased** and/or **inefficient** analyses.

Representing survival data

See live visualization regarding study time vs. patient time.

Representing survival data

Underlying data:

- T: Failure time, a non-negative random variable
- C: Censoring time, a non-negative random variable Observed data for individual i:
- Y_i: $(T_i \wedge C_i)$, the minimum of T_i and C_i
- δ_i: $1_{(T_i \leq C_i)}$, whether we observe a failure

If $\delta_i = 0$, then we have **right-censoring**: the survival time is longer than the censoring time.

Commonly, we assume C_i are *i.i.d.* random variables with some distribution and that the censoring mechanism is *independent* of the failure mechanism.

Our goal is to make inferential statements about T.

Characterizing continuous T

- Density function: $f(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$
- Distribution function: $F(t) = P(T \leq t) = \int_0^t f(s) ds$
- Survival function: $S(t) = P(T > t) = 1 - F(t)$
- Hazard function: $\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$
- Cumulative hazard function: $\Lambda(t) = \int_0^t \lambda(s) ds$

Knowing one is equivalent to knowing the others.

How might you express the hazard function in terms of the density funciton and the survival function?

Survival vs. hazard functions:

Survival (or survivor) function:

$$S(t) = P(T > t)$$

- Non-increasing with $S(0) = 1$ and $\lim_{t \to \infty} S(t) = 0$
- For any given time t, a probability

Hazard function:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

- Instantaneous failure rate, *given* already having survived to time t
- **Not** a probability (for continuous T)
- Non-negative and unbounded for all t
- Often more useful interpretations than survival functions
- Nice analytical properties under right-censoring

Estimating the survival curve

The Kaplan-Meier estimate provides an intuitive *non-parametric* estimate of the survival curve:

- D_i: # who fail at time t_i
- S_i: # who have survived beyond t_i (includes those who were censored exactly at t_i)
- N_i: # at risk of failure at time t_i (i.e., those who did not fail before t_i and were not censored before t_i)

The Kaplan-Meier estimate is

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{D_i}{N_i}\right) = \prod_{i:t_i \leq t} \frac{S_i}{N_i}$$

Estimating the survival curve

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{D_i}{N_i}\right) = \prod_{i:t_i \leq t} \frac{S_i}{N_i}$$

How might we calculate P(survived past $t_1 \cap t_2$)?

P(survive past t_1)P(survive past t_2 | survived past t_1)

...and so on. If an observation is censored, it is no longer at risk of failing at the next failure time and is taken out of the calculation.

Estimating the survival curve

Suppose we had a small study with the following data:

Patient	Event Time	Event Type
1	4.5	Failure
2	7.5	Failure
3	8.5	Censoring
4	11.5	Failure
5	13.5	Censoring
6	15.5	Failure
7	16.5	Failure
8	17.5	Censoring
9	19.5	Failure
10	21.5	Censoring

Estimating the survival curve

t	Risk Set	# Failed	# Censored	$\hat{S}(t)$
0	10	0	0	1
4.5	10	1	0	$1 - \frac{1}{10} = 0.9$
7.5	9	1	0	$0.9 \times (1 - \frac{1}{9}) = 0.8$
8.5	8	0	1	$0.8 \times (1 - \frac{0}{8}) = 0.8$
11.5	7	1	0	$0.8 \times (1 - \frac{1}{7}) = 0.69$
13.5	6	0	1	$0.69 \times (1 - \frac{0}{6}) = 0.69$
15.5	5	1	0	$0.69 \times (1 - \frac{1}{5}) = 0.552$
16.5	4	1	0	$0.552 \times (1 - \frac{1}{4}) = 0.414$
17.5	3	0	1	$0.414 \times (1 - \frac{0}{3}) = 0.414$
19.5	2	1	0	$0.414 \times (1 - \frac{1}{2}) = 0.207$
21.5	1	0	1	$0.207 \times (1 - \frac{0}{1}) = 0.207$

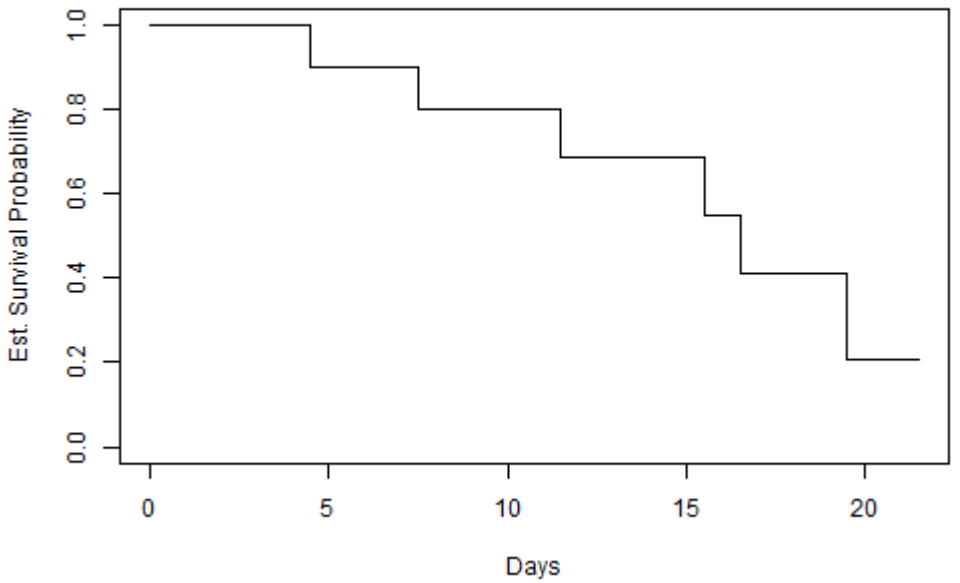
Estimating the survival curve

```
dat
```

```
## # A tibble: 10 x 2
##    times event
##    <dbl> <dbl>
##  1   4.5     1
##  2   7.5     1
##  3   8.5     0
##  4  11.5     1
##  5  13.5     0
##  6  15.5     1
##  7  16.5     1
##  8  17.5     0
##  9  19.5     1
## 10  21.5     0
```

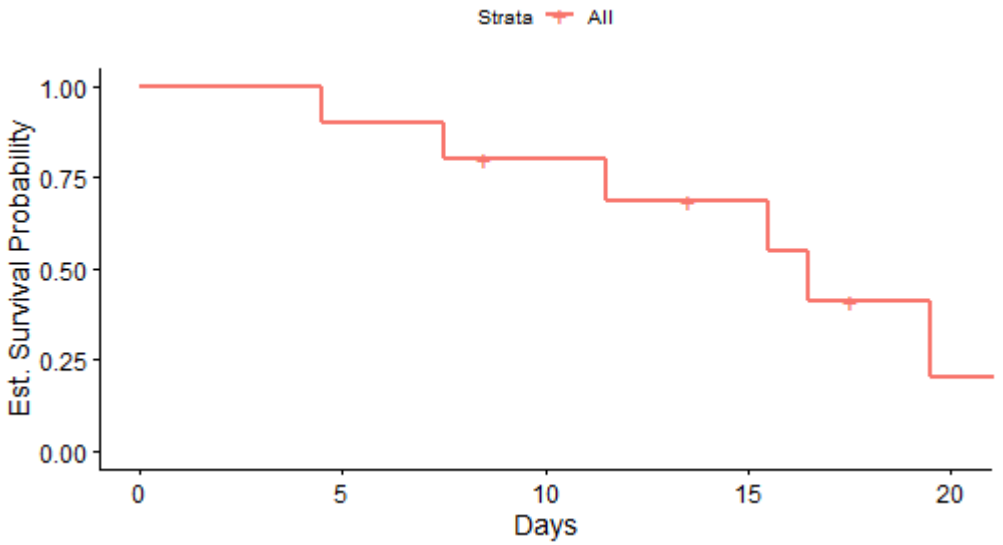
Estimating the survival curve

```
plot(survfit(Surv(times, event) ~ 1, data = dat),
     xlab = "Days", ylab = "Est. Survival Probability",
     conf.int = F)
```



Estimating the survival curve

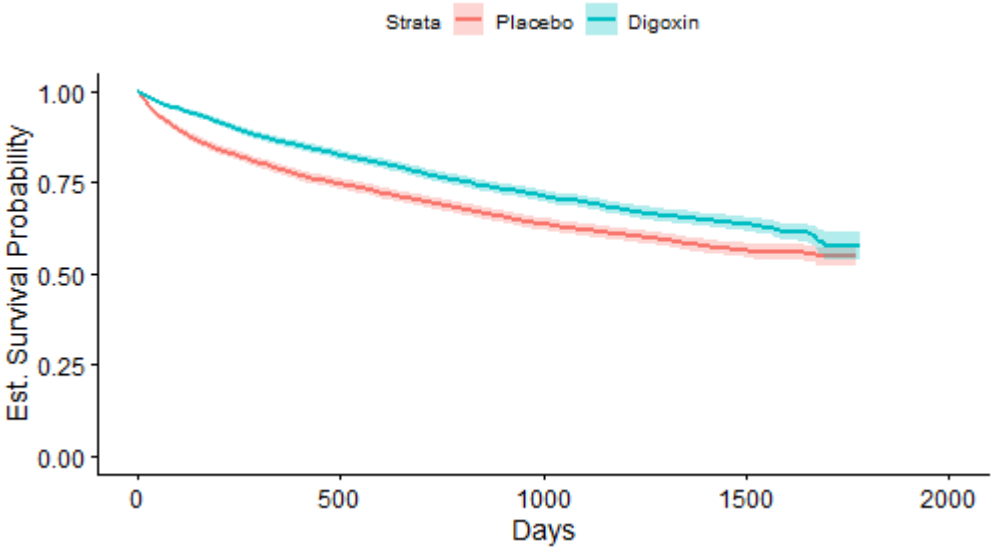
```
library(survminer)
ggsurvplot(survfit(Surv(times, event) ~ 1, data = dat),
     xlab = "Days", ylab = "Est. Survival Probability",
     conf.int = F)
```



Check out the `ggsurvplot` function here.

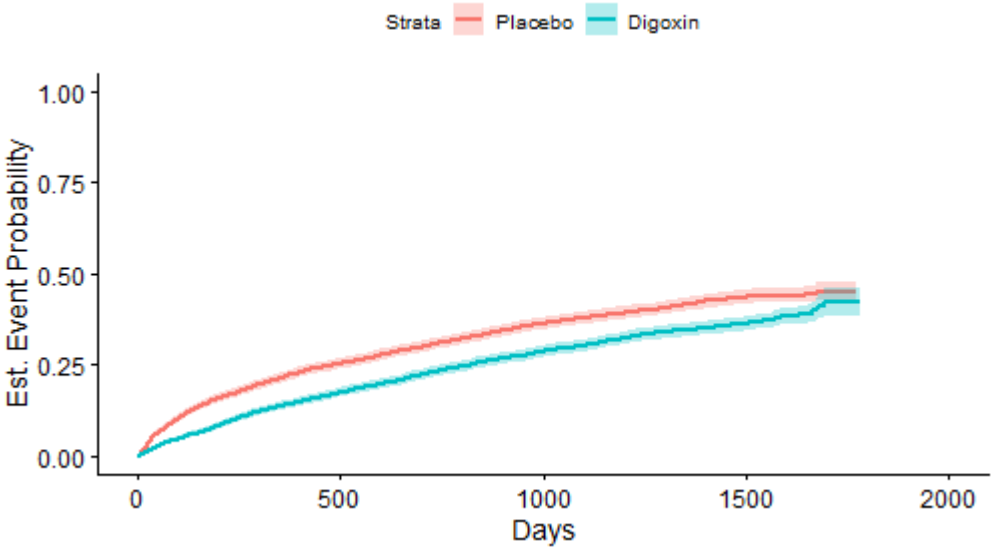
Comparing multiple groups

```
library(survminer)
ggsurvplot(survfit(Surv(DWHFDAYS, DWHF) ~ TRTMT, data = dig),
     xlab = "Days", ylab = "Est. Survival Probability",
     ylim = c(0, 1),
     conf.int = T, censor = F,
     legend.labs = c("Placebo", "Digoxin"))
```



Comparing multiple groups

```
library(survminer)
ggsurvplot(survfit(Surv(DWHFDAYS, DWHF) ~ TRTMT, data = dig),
     xlab = "Days", ylab = "Est. Event Probability",
     ylim = c(0, 1),
     conf.int = T, censor = F, fun = "event",
     legend.labs = c("Placebo", "Digoxin"))
```



Comparing multiple groups

How might we formally test whether there is a difference in the two survival curves?

$$H_0 : S_1(t) = S_2(t)$$
$$H_1 : S_1(t) \neq S_2(t)$$

Comparing multiple groups

The log-rank test constructs 2-by-2 contingency tables (assuming two groups) at each time at which a failure occurs. Then, these tables are combined using Mantel-Haenszel to evaluate whether there is a differnce in the two curves:

	Group 1	Group 2	Total
Deaths at t_i	D_{1i}	D_{2i}	D_i
Survivors past t_i	S_{1i}	S_{2i}	S_i
Total at risk	N_{1i}	N_{2i}	N_i

comparing observed failures D_{1i} against the expected count under H_0.

The test statistic has an asymptotic χ^2_1 distribution under H_0 (for two groups).

Comparing multiple groups

It is most powerful under proportional hazards (check empirically), and not very powerful at all if survival curves cross. Alternatives are available in this situation.

The log-rank test can be extended to adjust for a categorical confounder by considering a stratified version, and can also be extended to test for differences in survival functions across more than 2 groups.

Comparing multiple groups

```
survdiff(Surv(DWHFDAYS, DWHF) ~ TRTMT, data = dig)
```

```
## Call:
## survdiff(formula = Surv(DWHFDAYS, DWHF) ~ TRTMT, data = dig)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## TRTMT=0 3403     1291     1126      24.1      46.6
## TRTMT=1 3397     1041     1206      22.5      46.6
##
##  Chisq= 46.6  on 1 degrees of freedom, p= 9e-12
```