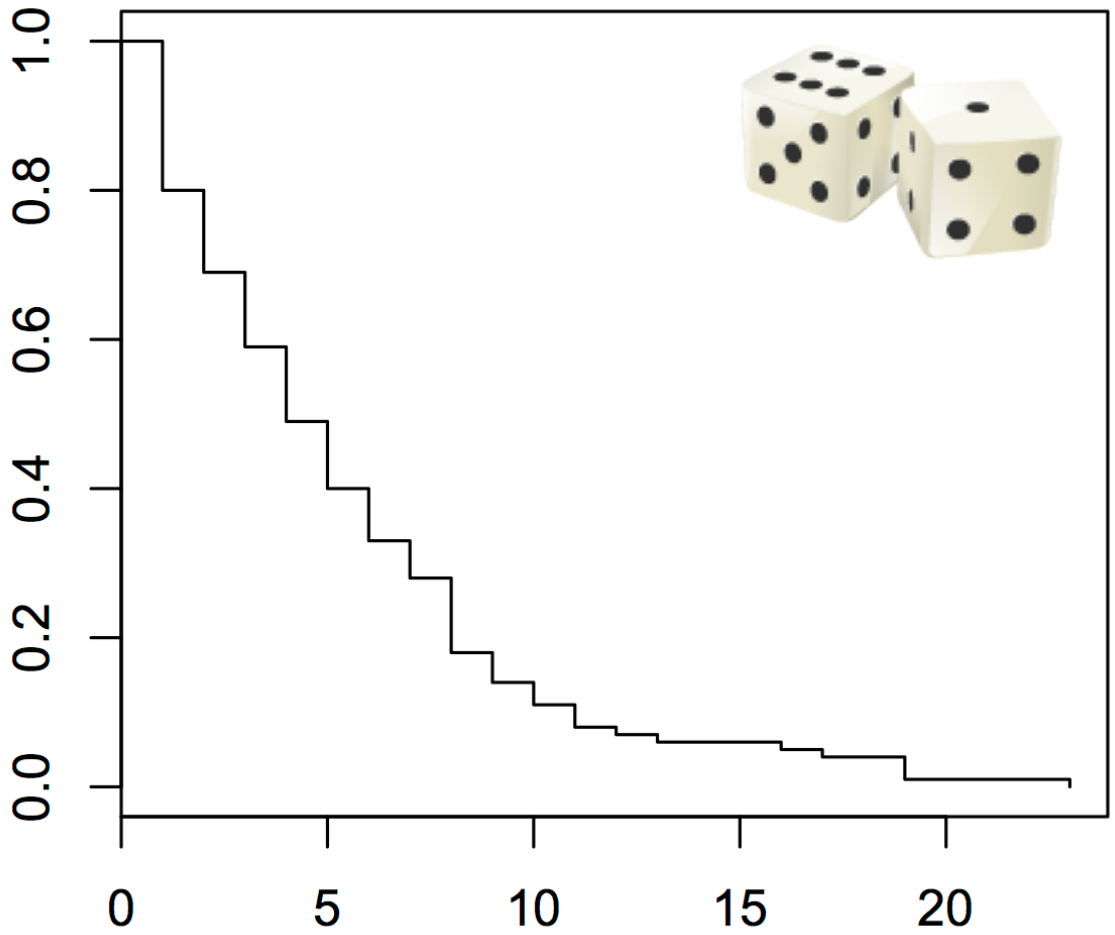
Survival analysis: not just for survival anymore

Analysis of time to an event or rates of events

Events include any binary, non-reversible event or change:

death, recovery, breeding, hatching, fledging, metamorphosis, consumption, breaking, …

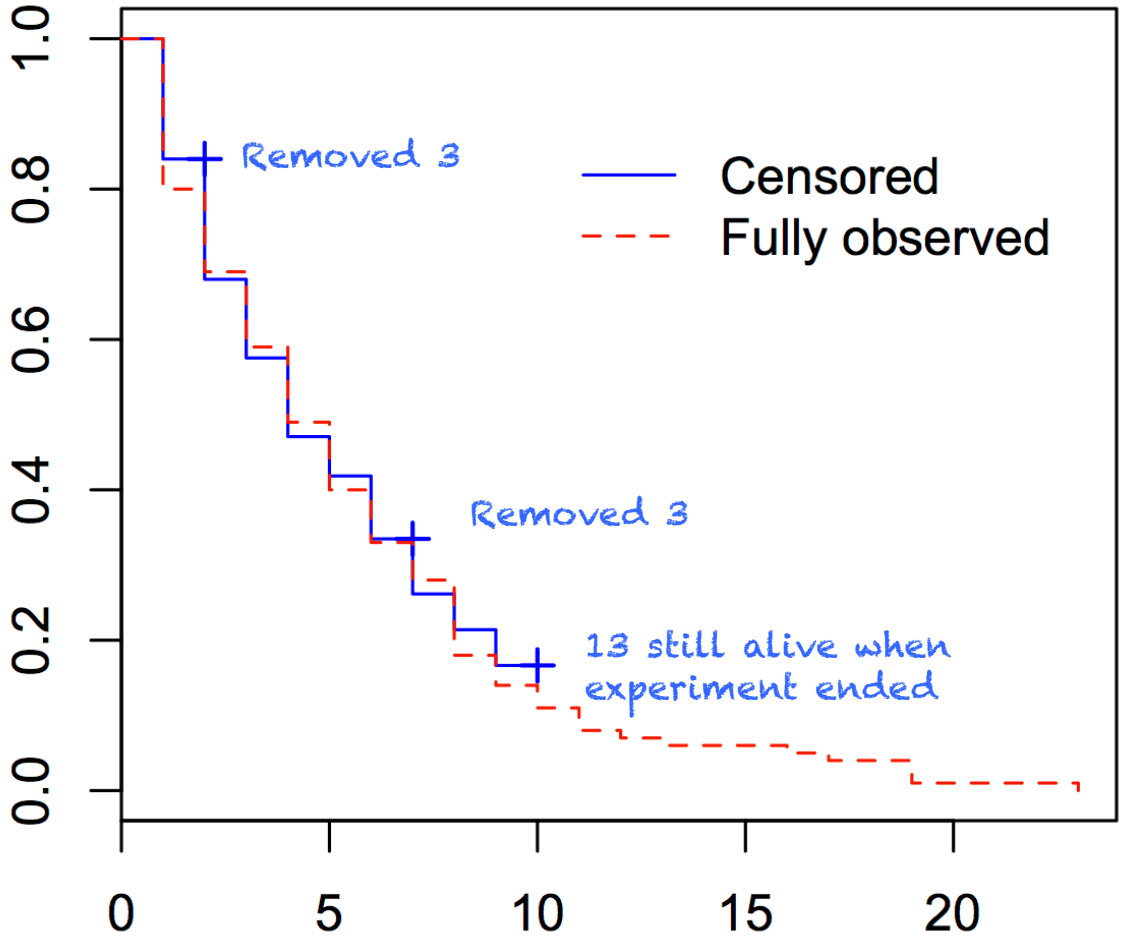
Jesse Brunner, Washington State University



Roll	Died
1	20
2	11
3	10
4	10
5	9
6	7
7	5
8	10
9	4
10	3
11	3
12	1
13	1
16	1
17	1
19	3
23	1

- Rolled 100 dice (=animals)
- If I rolled a "6" the animal "died"
- Rolled until all animals were dead
- Kept track of when each animal died

Average time to death
mean = 5.51 rolls (expect 6 = 1/rate)
median = 4 rolls (expect 4.16 = (1/rate)*ln(2)



Roll	Died	Died Censored	
1	20	16	–
2	11	16	3
3	10	10	–
4	10	10	–
5	9	5	–
6	7	8	–
7	5	7	3
8	10	4	–
9	4	4	–
10	3	1	13
11	3	–	
12	1	–	
13	1	–	
16	1	–	
17	1	–	
19	3	–	
23	1	–	

- Same process as before, <u>but</u>:
- Destructively sampled 3 animals on rolls 2 and 7
- Experiment ended after ten rolls

Average time to death
mean = 4.75 or 3.93 rolls
median = 4 or 3 rolls
depending on whether or not you keep these censored animals in the average!

Kaplan-Meier survival curves

Visualizing survival through time

- Accommodates right censored observations
- Estimates survival probability

Log-rank test (aka Mantel-Haenszel)

Do survival curves differ significantly?

- Built around a contingency table, like a chi-square test
- Can only compare distinct groups

Cox proportional hazard model

Does the *hazard* vary among types of individuals?

- Estimates an underlying hazard (non-parametric)
- Determines whether ratio of hazards among groups (or individuals with certain covariates) differs from 1:1

What is "hazard"?

Parametric survival models (accelerated failure-time)

Does the *hazard* differ between types of individuals?

- *Assumes* some hazard distribution (e.g., exponential, lognormal)
- Like regression; can handle factors and continuous predictors

Kaplan-Meier survival curves

Visualizing survival through time

- Accommodates right censored observations
- Estimates survival probability

$$S(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

$S(t)$ is the probability that an individual's lifetime is $> t$

n_i is the number of individuals at risk (i.e., not dead or censored) just prior to time t_i.
When individuals are censored, they are simply removed from n_i

d_i is the number of deaths occurring at time t_i

The median time to death is the time where $S(t) = 0.5$

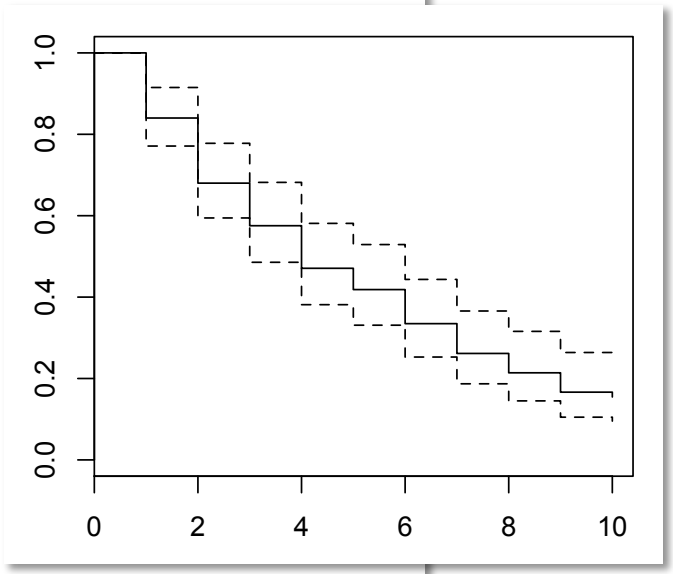
Note that $S(t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution of failures or death, which should make sense. Later, we will relate this to the hazard.

Read in the data

```
setwd(file.choose())
cens <- read.csv("RollDice_censored.csv")
```

Look at the data

```
table(cens)
    Death
Roll  0  1
  1   0 16
  2   3 16
  3   0 10
  4   0 10
  5   0  5
  6   0  8
  7   3  7
  8   0  4
  9   0  4
  10 13  1
```



Create a "survival object"

```
library(survival)
s2 <- Surv(cens$Roll, cens$Death) # note the capital "S"
s2
  [1]  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 [16]  1   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 [31]  2   2   2+  2+  2+  3   3   3   3   3   3   3   3   3   3
 [46]  4   4   4   4   4   4   4   4   4   4   5   5   5   5   5
 [61]  6   6   6   6   6   6   6   6   7   7   7   7   7   7   7
 [76]  7+  7+  7+  8   8   8   8   9   9   9   9  10  10+ 10+ 10+
 [91] 10+ 10+ 10+ 10+ 10+ 10+ 10+ 10+ 10+ 10+
```

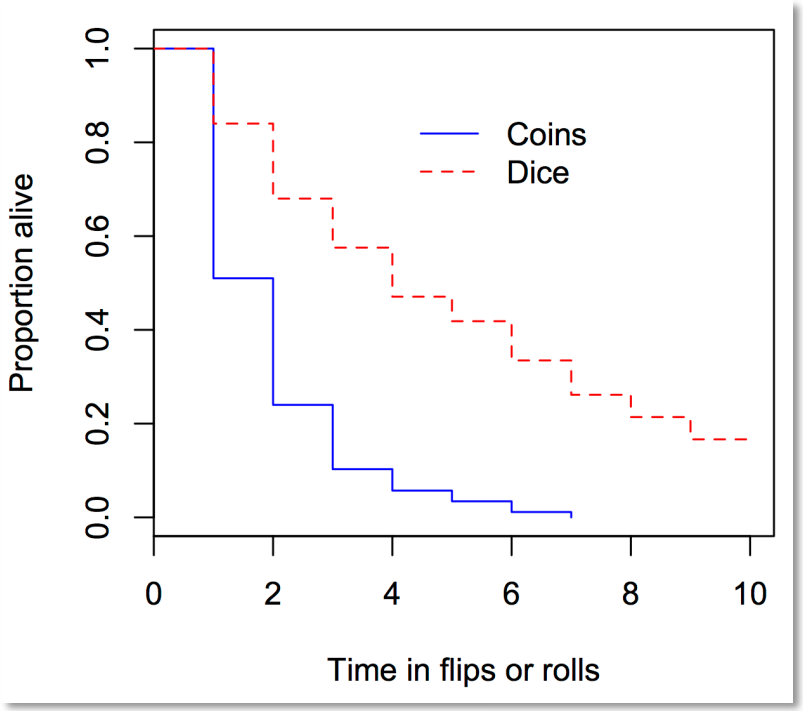
The "+" after a number means the animal was censored

This is the K-M estimate

```
survfit(s2~1) # try summary(survfit(s2~1) )
Call: survfit(formula = s2 ~ 1)
records   n.max n.start  events  median 0.95LCL 0.95UCL
    100     100     100      81       4       3       6
```

Plot the K-M curve

```
plot(survfit(s2~1)) # try including: conf.int = FALSE
```



Comparing survival curves between groups

- Same dice data
- Added in coin flip (heads = dead) for 100 coins

Read in the data

```
dicecoin <- read.csv("DiceCoin.csv")
```

Create the K-M estimates

```
dc <- survfit(Surv(Time, Death) ~ Type, data = dicecoin)
dc
```

```
Call: survfit(formula = Surv(Time, Death) ~ Type, data = dicecoin)
          records n.max n.start events median 0.95LCL 0.95UCL
Type=Coin     100   100     100     97      2       1       2
Type=Die      100   100     100     81      4       3       6
```

Basic plot of K-M curves

```
plot(dc)
```

Nicer plot of K-M curves with labels and different line types & colors...

```
plot(dc, lty = 1:2, col = c("blue", "red"),
         xlab = "Time in flips or rolls",
         ylab = "Proportion alive")
```

and a legend

```
legend(4, 0.9, c("Coins", "Dice"), lty = 1:2,
         col = c("blue", "red"), bty = "n")
```

Are the median's the same? Do the CIs overlap?

Log-rank test (aka Mantel-Haenszel)

Do survival curves differ significantly?

- Built around a contingency table, like a chi-square test
- Can only compare distinct groups

	Group A	Group B	Total
Event	d_{Ai}	d_{Bi}	d_i
No Event	$n_{Ai} - d_{Ai}$	$n_{Bi} - d_{Bi}$	$n_i - d_i$
At Risk	n_{Ai}	n_{Bi}	n_i

The *expected* number of deaths in group A at time i, if both groups are identical, is: $\hat{e}_{Ai} = n_{Ai} \times (d_i / n_i)$

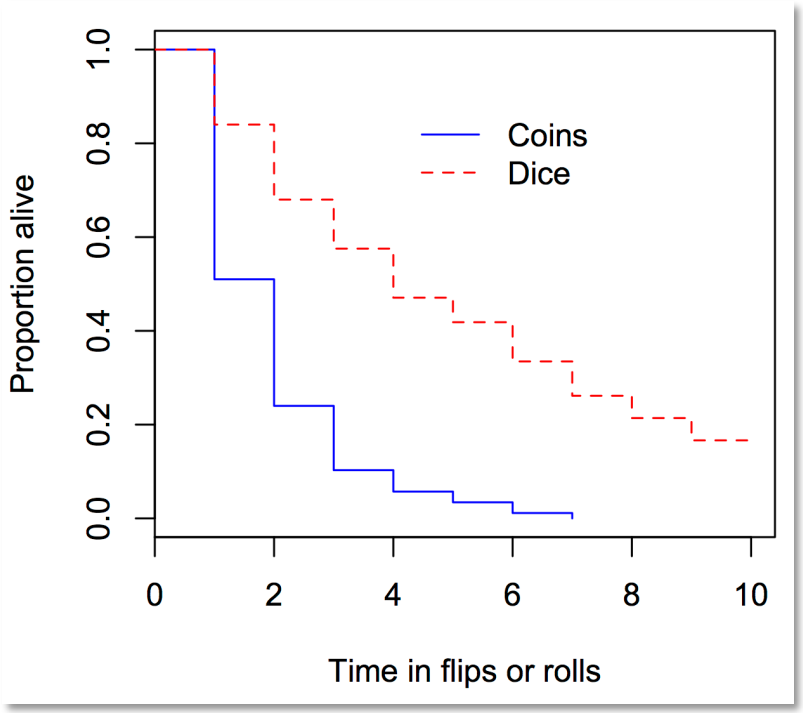
Compare the expected and actual number of deaths in each group at each time i, over each of the m times

$$Q = \frac{\left(\sum_{i=1}^{m} d_{Ai} - \sum_{i=1}^{m} \hat{e}_{Ai}\right)^2}{\sum_{i=1}^{m} \hat{V}(\hat{e}_{Ai})}$$

Log-rank test (aka Mantel-Haenszel)

Do survival curves differ significantly?

- Built around a contingency table, like a chi-square test
- Can only compare distinct groups




```
survdiff(Surv(Time, Death) ~ Type, data = dicecoin)

Call:survdiff(formula = Surv(Time, Death) ~ Type,
data = dicecoin)
             N Observed Expected (O-E)^2/E (O-E)^2/V
Type=Coin  100       97     57.7      26.7      65.3
Type=Die   100       81    120.3      12.8      65.3
 Chisq= 65.3  on 1 degrees of freedom, p= 6.66e-16
```

The code is like regressions

Do the numbers of dead at each time differ from the number expected if coins and dice were the same?

Yes!

Cox proportional hazard model

Does the *hazard* vary among types of individuals?

1) Estimates an underlying "baseline" hazard (non-parametric) based on one group
2) Determines whether being in another group (or having different covariates) changes the hazard from that baseline

So what is hazard, any way?

The hazard function (aka hazard rate, failure rate, or force of mortality) is
the instantaneous rate of occurrence of death or
the probability of dying at time *t* given survival to time *t*
akin to the instantaneous mortality rates population models.

In the Cox proportional hazard model, though, we don't pay much attention to it.
We just estimate a baseline hazard, $h_0(t)$, then see whether it is influenced by other parameters, x_1, x_2, …

$$h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \ldots}$$

Cox proportional hazard model

Does the *hazard* vary among types of individuals?

Again, like regression

```
dc.cox <- coxph(Surv(Time, Death) ~ Type, data = dicecoin)
summary(dc.cox)

     n= 200, number of events= 178
           coef exp(coef) se(coef)      z Pr(>|z|)
TypeDie -1.3479    0.2598   0.1725 -7.814 5.55e-15 ***

        exp(coef) exp(-coef) lower .95 upper .95
TypeDie    0.2598      3.849    0.1853    0.3643

Concordance= 0.677  (se = 0.03 )
Rsquare= 0.27   (max possible= 1 )
Likelihood ratio test= 63.07  on 1 df,   p=1.998e-15
Wald test            = 61.06  on 1 df,   p=5.551e-15
Score (logrank) test = 67.91  on 1 df,   p=2.22e-16
```

You can extract the baseline hazard if you like

```
Basehaz(dc.cox)
      hazard time
1  0.3401289    1
2  0.7302868    2
3  1.0999080    3
4  1.4558872    4
5  1.6903325    5
6  2.1360030    6
7  2.6457717    7
8  3.0297384    8
9  3.5075668    9
10 3.6477072   10
```

We need to work with the exponentiated coefficients, which is the **hazard ratio**

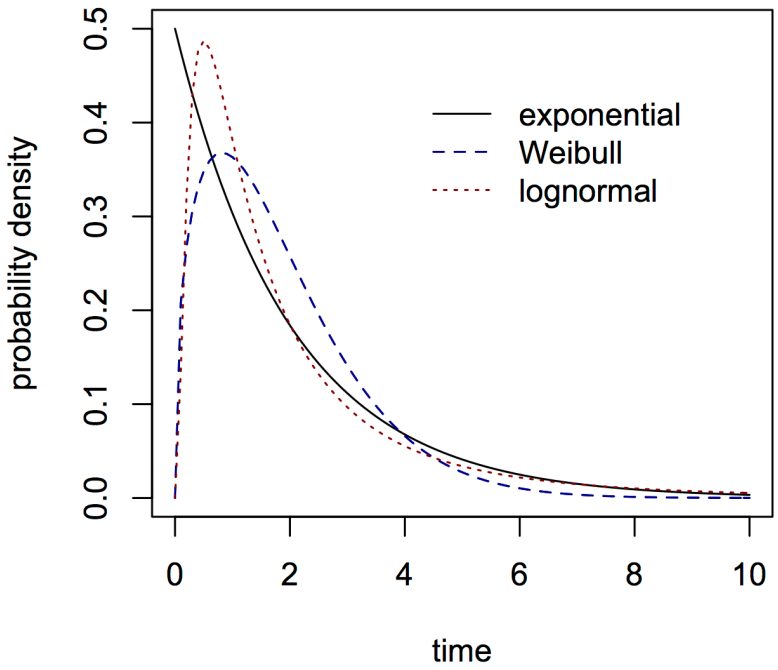
This says that the death rate of dice was about 4x lower than coins... it should have been 3x

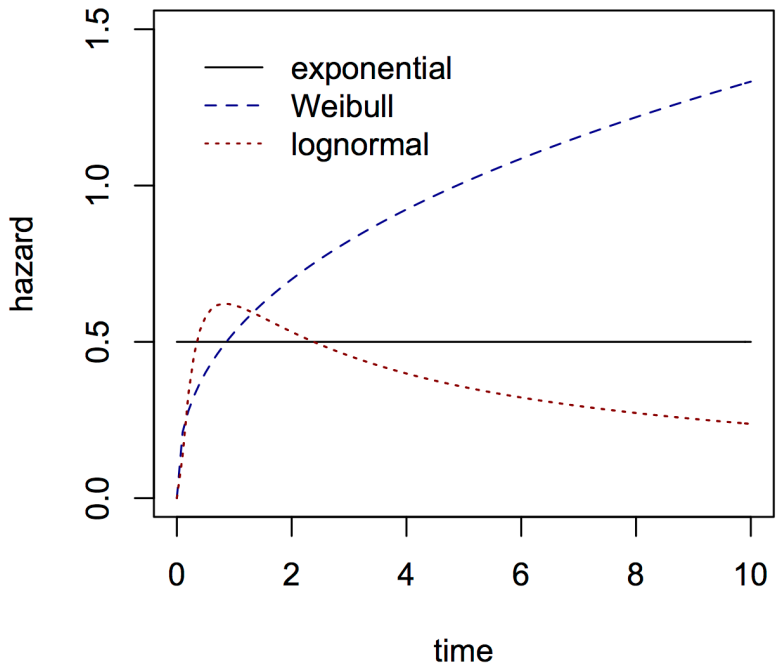
Confidence interval does not overlap 1, meaning equal hazards.

Parametric survival models (accelerated failure-time)

Does the *hazard rate* differ between types of individuals?

- *Assumes* some hazard distribution (e.g., exponential, lognormal)
- The *scale* of the hazard is then a linear function of covariates
    - It's as if we are stretching or accelerating time





Parametric survival models (accelerated failure-time)

Does the *hazard rate* differ between types of individuals?

- *Assumes* some hazard distribution (e.g., exponential, lognormal)
- The scale of the hazard is then a linear function of covariates

Again, like regression, but need to specify distribution

```
dc.aft <- survreg(Surv(Time, Death) ~ Type, data = dicecoin,
                  dist = "exponential")
summary(dc.aft)
```

```
Call:survreg(formula = Surv(Time, Death) ~ Type, data = dicecoin,
    dist = "exponential")
             Value Std. Error    z        p
(Intercept) 0.672      0.102 6.62 3.56e-11
TypeDie     1.097      0.151 7.29 3.21e-13

Scale fixed at 1
Exponential distribution
Loglik(model)= -386.5   Loglik(intercept only)= -412.6
Chisq= 52.22 on 1 degrees of freedom, p= 5e-13
Number of Newton-Raphson Iterations: 4
n= 200
```

These parameters are the **rate** of the exponential

Back-transform to get the rate for coins

```
exp(-0.672)
[1] 0.5106862
```

And for dice

```
exp(-(0.672+1.097))
[1] 0.1705034
```

So we finally get the right hazards!

Kaplan-Meier survival curves

Visualizing survival through time

- Accommodates right censored observations
- Estimates survival probability (can be used to calculate hazard)

```
plot(survfit(Surv(time, event) ~ group))
```

Log-rank test (aka Mantel-Haenszel test)

Do survival curves differ significantly?

- Built around a contingency table of expected number of deaths at time *i* in group *j*
- Like a chi-square test, can only compare distinct groups

```
survdiff(survfit(Surv(time, event) ~ group))
```

Cox proportional hazard model

Does the *hazard* vary among types of individuals?

- Estimates an underlying hazard (non-parametric)
- Determines whether ratio of hazards among groups (or individuals with certain covariates) differs from 1:1
- Assumes hazard is *always* proportional (does not change through time)

```
coxph(Surv(time, event) ~ group + variable)
```

Parametric survival models (accelerated failure-time)

Does the *hazard* differ between types of individuals?

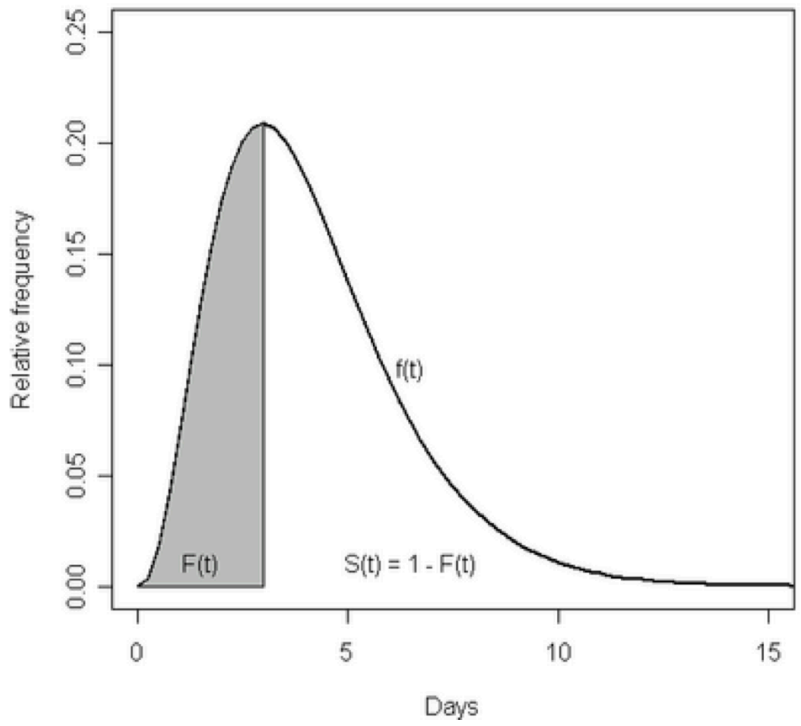
- *Assumes* some hazard distribution (e.g., exponential, lognormal)
- Like regression; can handle factors and continuous predictors
- Covariates can be time-varying
- Saves some degrees of freedom

```
survreg(Surv(time, event) ~ group + variable, dist = "lognormal")
```

Hazard function — $h(t)$
probability of death in the next instant, given survival to time t

Survival function — $S(t)$
probability of surviving beyond time t

Probability density or distribution function (PDF) — $f(t)$
essentially the expected distribution of times to death



These are all related to each other in fairly simple, often very useful ways…

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\partial \ln S(t)}{\partial t}$$

$$f(t) = S(t)h(t)$$

$$S(t) = \exp\left[-\int_0^t h(t)\right] = \exp\left[-H(t)\right]$$

Roll your own!

Make hazard a function of something you think is important!

- Can still accommodate censored observations
- Can infer useful relationships between intrinsic or external variables and the rate or timing of some event
- Fit survival probability to data on number alive (or unmated or still larval or…) using likelihood

How does cold weather influence overwintering survival of nymphal ticks ?

Put ticks in cores in the ground, dug some up every two weeks, counted number still alive.

$$h(t) = a/\left[1+\exp(-4b(T\min(t)-c)\right]$$

$$S(t) = \exp\left[-\int_0^t h(t)\right] = \exp\left[-H(t)\right]$$

A flexible 3-parameter logistic function of Tmin

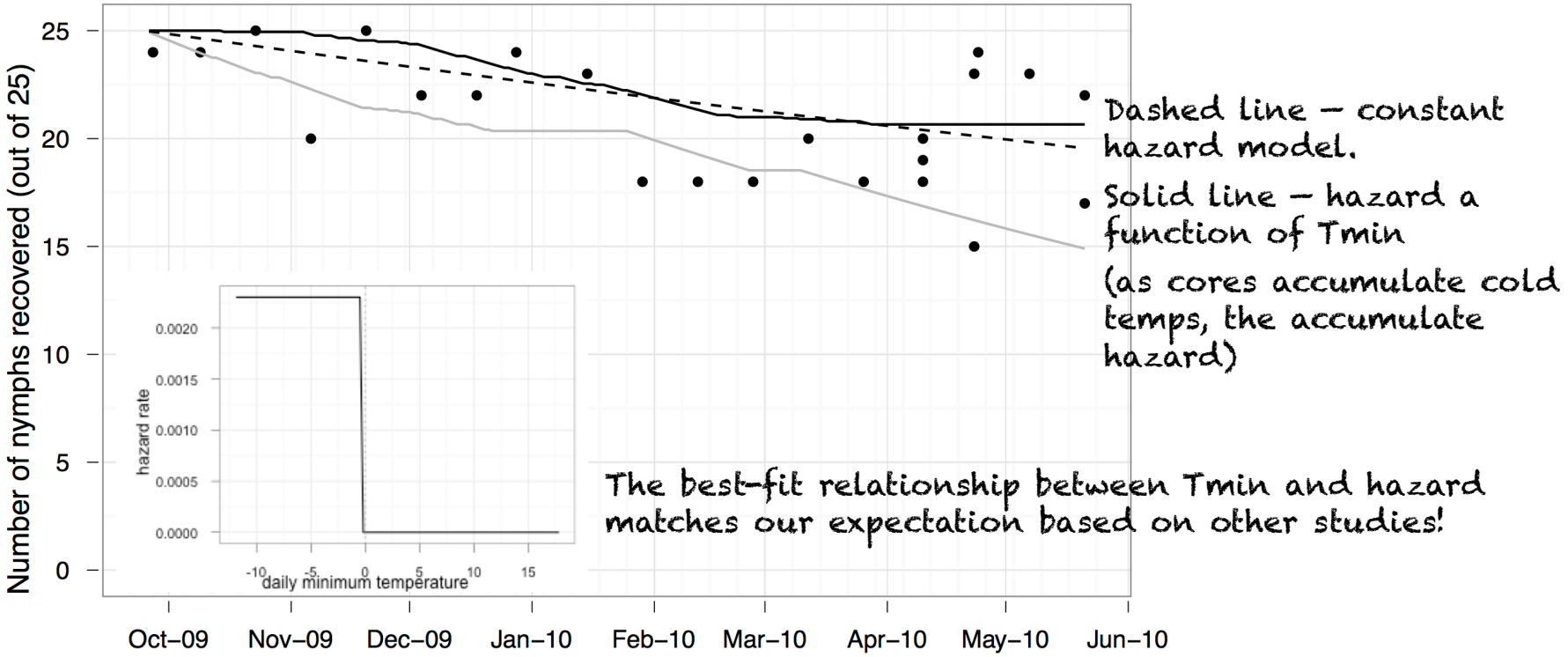
For each core,

1) calculate H(t) as accumulated hazard (due to cold temps)
2) exponentiate negative H(t) to probability of surviving until the core was dug up, S(t)
3) Calculate likilihood of observing x out of n live ticks given S(t)
4) Repeat to find MLE

Roll your own!

Make hazard a function of something you think is important!

- Can still accommodate censored observations
- Can infer useful relationships between intrinsic or external variables and the rate or timing of some event
- Fit survival probability to data on number alive (or unmated or still larval or…) using likelihood



Useful links & guides

- JMP: the help files are very good
- Mark Stevenson's An introduction to survival analysis: http://epicentre.massey.ac.nz/Default.aspx?tabid=77
- Germàn Rodrìguez's Survival analysis class website & pdfs: http://data.princeton.edu/pop509a/
- Charles Franklin's lecture notes (ch 15 & 16) on likelihood, survival functions, hazard rates, and pdfs: http://users.polisci.wisc.edu/franklin/Content/MLE/MLE.htm
- R Vignette for survival package: Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model