

Cricket innings and survival analysis

Duncan Golicher

2017-8-14

Introduction

Yesterday I packaged up some cricket data that I had downloaded from ESPN cricinfo and reformatted. I thought it might save some time if anyone wanted the data in a convenient format.

I made a little app to quickly look at batsmen's innings using shiny.

<http://r.bournemouth.ac.uk:3838/cricket/Batsmen/> **Test matches only**

http://r.bournemouth.ac.uk:3838/cricket/all_modes/ **Test, ODI and I2020**

Note, use backspace to clear the search box and then type any letters of the name to find a batsman.

This was just for fun. However in order to not consider the exercise as just for something to fill time over the summer holidays I wondered if the data could have be used to illustrate any statistical concepts when teaching.

As everyone who knows the game is aware, cricket is not a matter of life and death. It is much more important than that. However a cricket innings is in effect a matter of survival. So the data could be used to explain survival analysis.

```
library(devtools)
install_github("dgolicher/crickdata")
#install.packages("survminer")
```

Loading data and subsetting

I downloaded the data in the first place as I was interested in evaluating the evidence for Geoffrey Boycott's gripes about English batsmen not knowing how to occupy the crease any more in test matches. So let's look at four batsmen's career statistics. I'll choose Boycott (opener), Cook (opener) Root (#3-4) and Botham as a lower order slugger for reference.

```
library(crickdata)
data(batting)
batting$completed<-1 ## Will use later
batting$completed[batting$Notout==TRUE]<-0 ## Ones and zeros
batting<-batting[!is.na(batting$Runs),] ## Remove did not bats
library(dplyr)
players<-unique(batting$Player)
#players[grepl("Cook",players)]
root<-batting[grepl("JE Root",batting$Player),]
cook<-batting[grepl("AN Cook",batting$Player),]
boycott<-batting[grepl("Boycott",batting$Player),]
botham<-batting[grepl("Botham",batting$Player),]
```

```
d<-rbind(root,cook,boycott,botham)
d<-d[d$type=="Test",]
d<-d[!is.na(d$Runs),]
d$Player<-sub("\\(ENG\\)","",d$Player)
d$Player<-factor(d$Player)
d$Player<-factor(d$Player,levels(d$Player)[c(2,1,3:4)])
```

Skew in batting figures

One statistical issue to resolve initially is the degree to which the distribution of total runs scored is right skewed. If data are right skewed then the median is less than the mean.

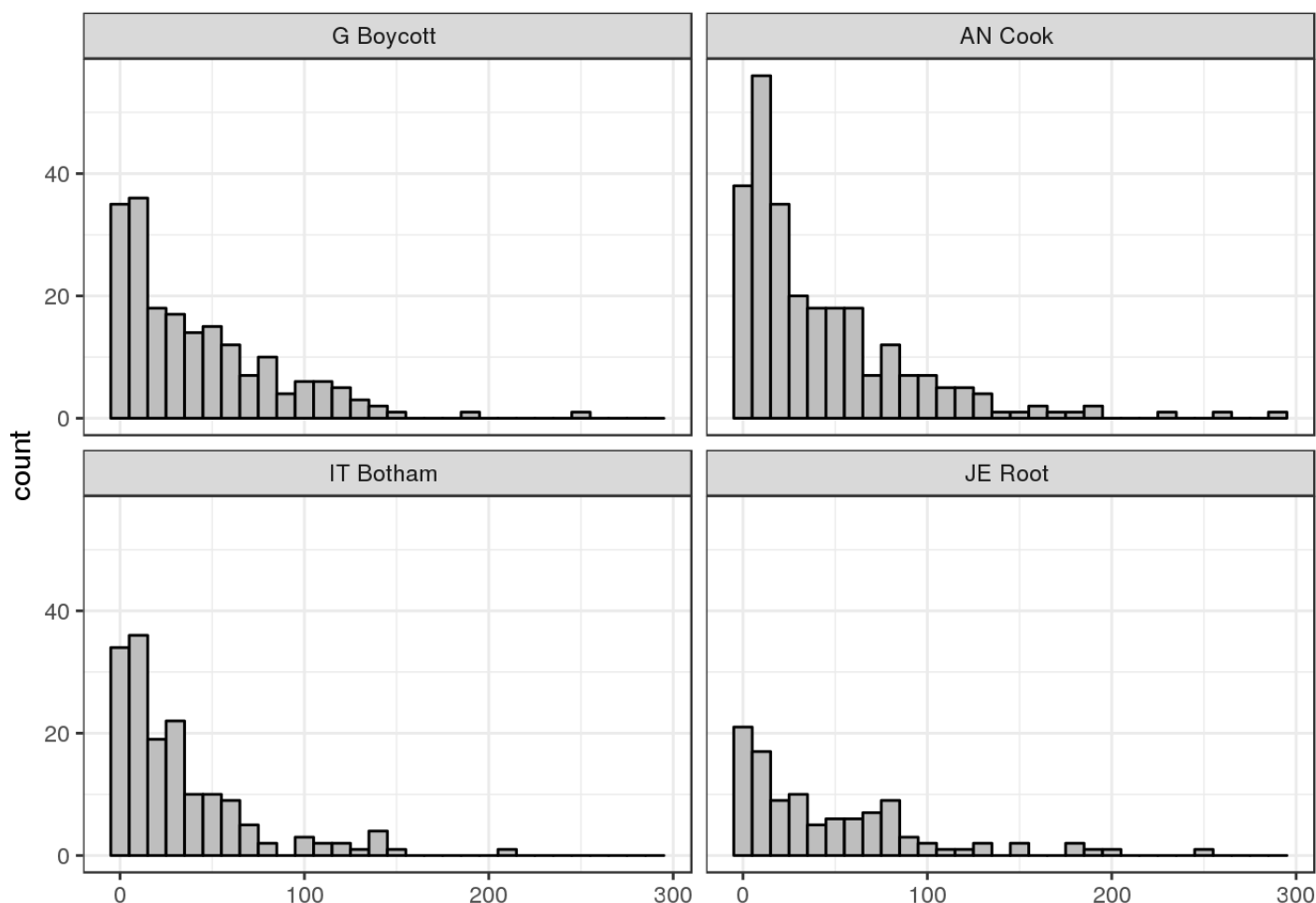
```
library(ggplot2)
theme_set(theme_bw())
```

Histograms

A histogram of total runs scored shows a J shape for all four batsmen, rather than the unimodal hump shape of a normal distribution. This makes sense when you consider the process involved. The normal distribution arises when many additive effects act together to determine the value of a variable. In this case

Cricket innings and survival analysis by Duncan Golicher Last updated 5 months ago

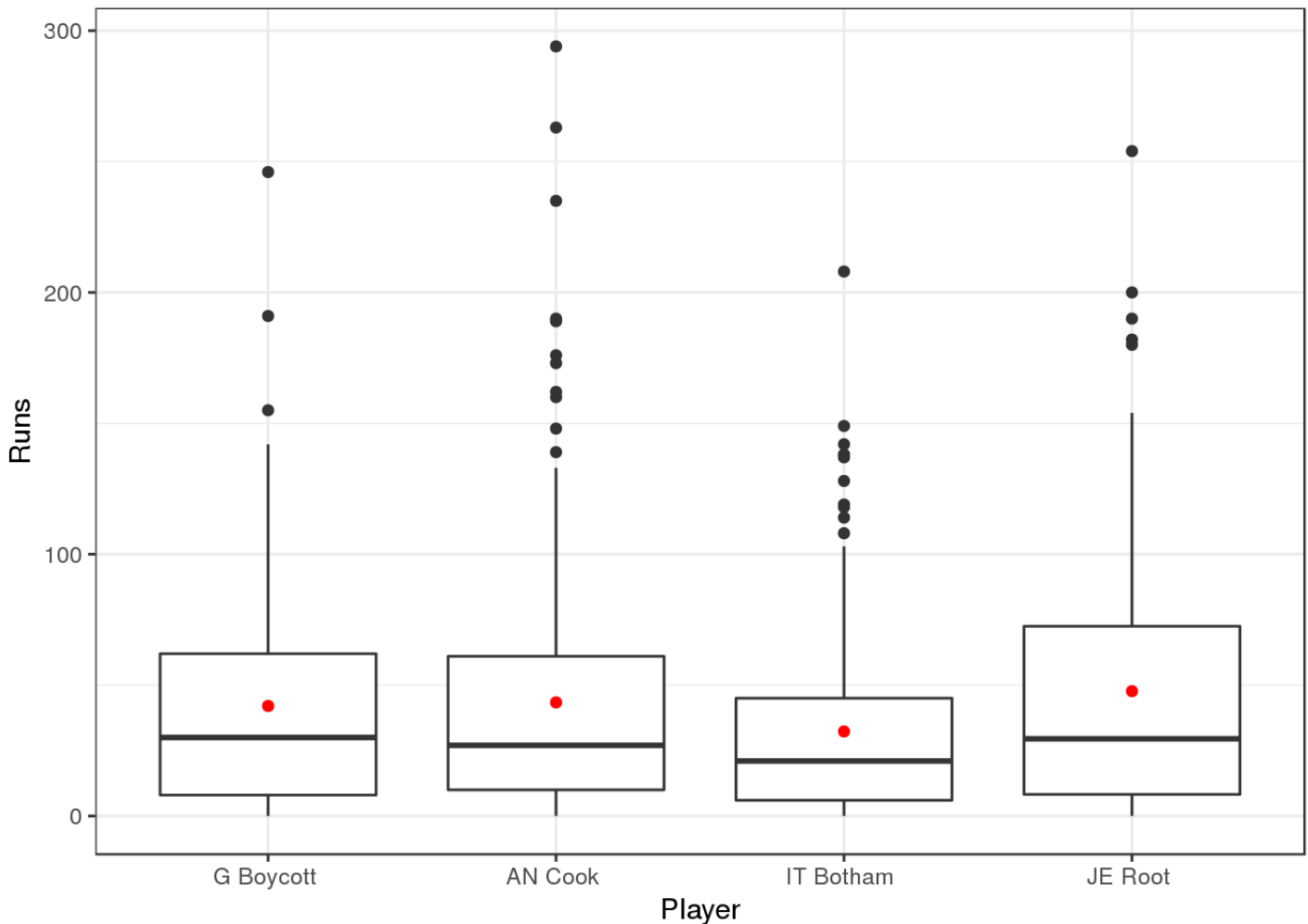
```
g0<-ggplot(d,aes(x=Runs))
g1<-g0+geom_histogram(fill="grey",colour="black",binwidth=10)
g1+facet_wrap("Player")
```



Boxplots with mean marked as red point

A simple way of looking at the differences between the mean and the median, together with the right skewed distribution of the values is to plot boxplots with the mean marked. The mean always falls above the median.

```
g0<-ggplot(d,aes(x=Player,y=Runs))
g1<-g0+geom_boxplot()
g1 +stat_summary(fun.y=mean,geom="point",col="red")
```



You may have realised that there is a potential issue here. When the batting average is provided as a statistic it actually refers to the total number of runs divided by the number of **completed** innings. Innings that leave the batsman not out are not added to the total used in the division. We can look at this as a table

```
d %>% group_by(Player)%>%
  summarise(innings=n(),
    completed=sum(completed),
    mean=round(mean(Runs),1),
    ba=round(sum(Runs)/completed,1),
    median=median(Runs),
    SR=round(mean(SR,na.rm=TRUE),1))
```

```
## # A tibble: 4 x 7
##   Player innings completed mean ba median SR
##   <chr>   <int>   <int> <dbl> <dbl> <dbl> <dbl>
```

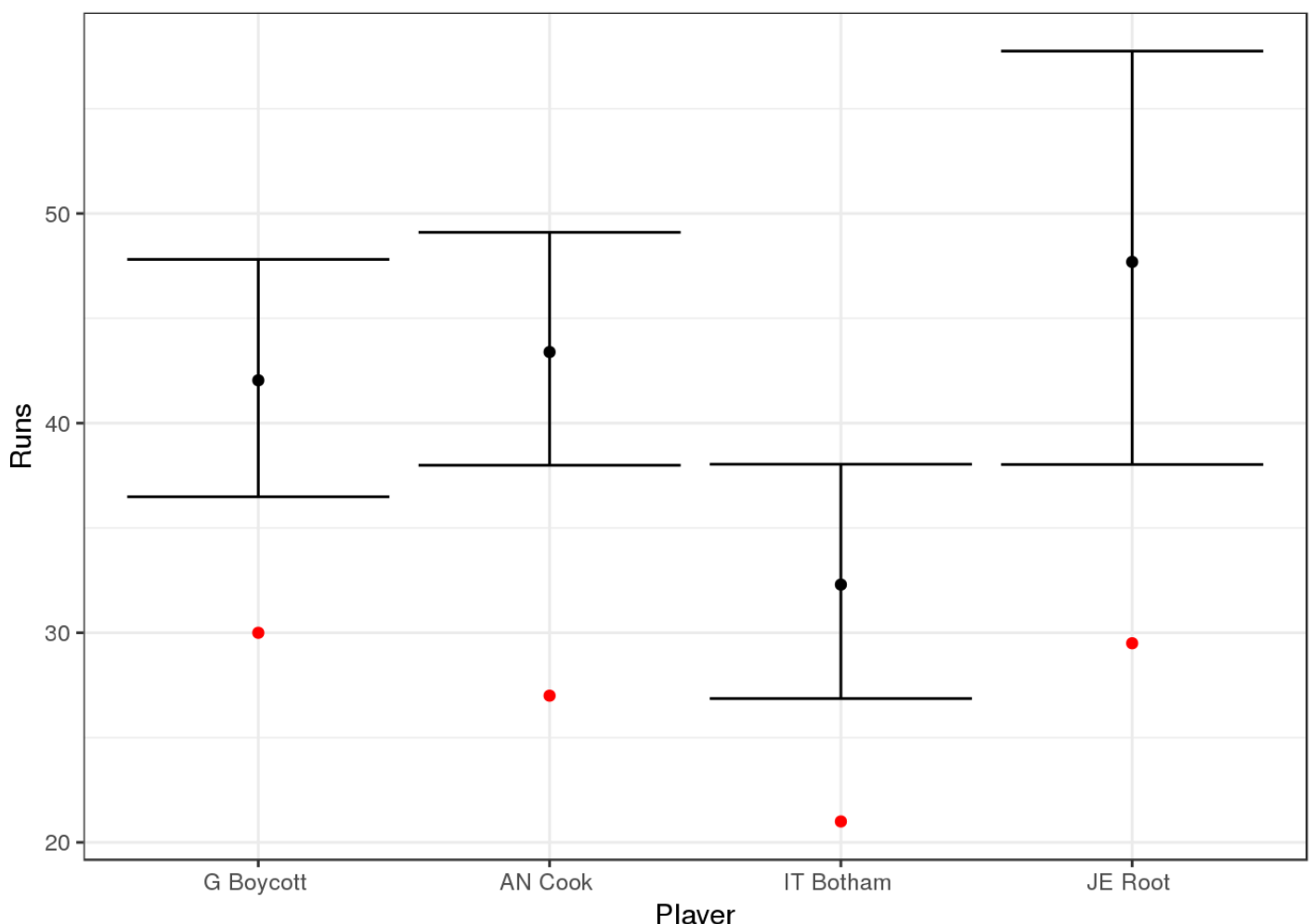
##	<fctr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
## 1	G Boycott	193	170	42.0	47.7	30.0	35.0
## 2	AN Cook	261	246	43.4	46.0	27.0	46.5
## 3	IT Botham	161	155	32.3	33.5	21.0	58.9
## 4	JE Root	106	95	47.7	53.2	29.5	51.8

There is more of a difference between the Boycott's batting average and his mean score than in the case of the other players. Boycott notoriously did not like getting out. There is another issue here, that we will return to. The median doesn't take into account not out scores either. However this is slightly more complex. The median represents the score that we would expect the batsman to make or exceed in 50% of the innings. So it doesn't matter whether a batsman is not out if the score is above the median. It only needs to be taken into account if the score is below the median.

Confidence interval for the means

As the data are skewed the most robust confidence intervals for means are calculated using a bootstrap. This takes repeated random samples from the column of runs and calculates the mean. Some of the samples will not include the highest innings, and others will leave out more of the low scores. This provides a reasonable way of looking at statistically significant differences between the means in this case- It would be possible to do this for the genuine batting average fairly easily in R and compare those, but I'll just take the default for the means using ggplots for speed. The median is shown in red.

```
g0<-ggplot(d,aes(x=Player,y=Runs))
g0<-g0+stat_summary(fun.y=mean,geom="point") +stat_summary(fun.y=median,geom="point",col="red")
g1<-g0+stat_summary(fun.data=mean_cl_boot,geom="errorbar")
g1
```

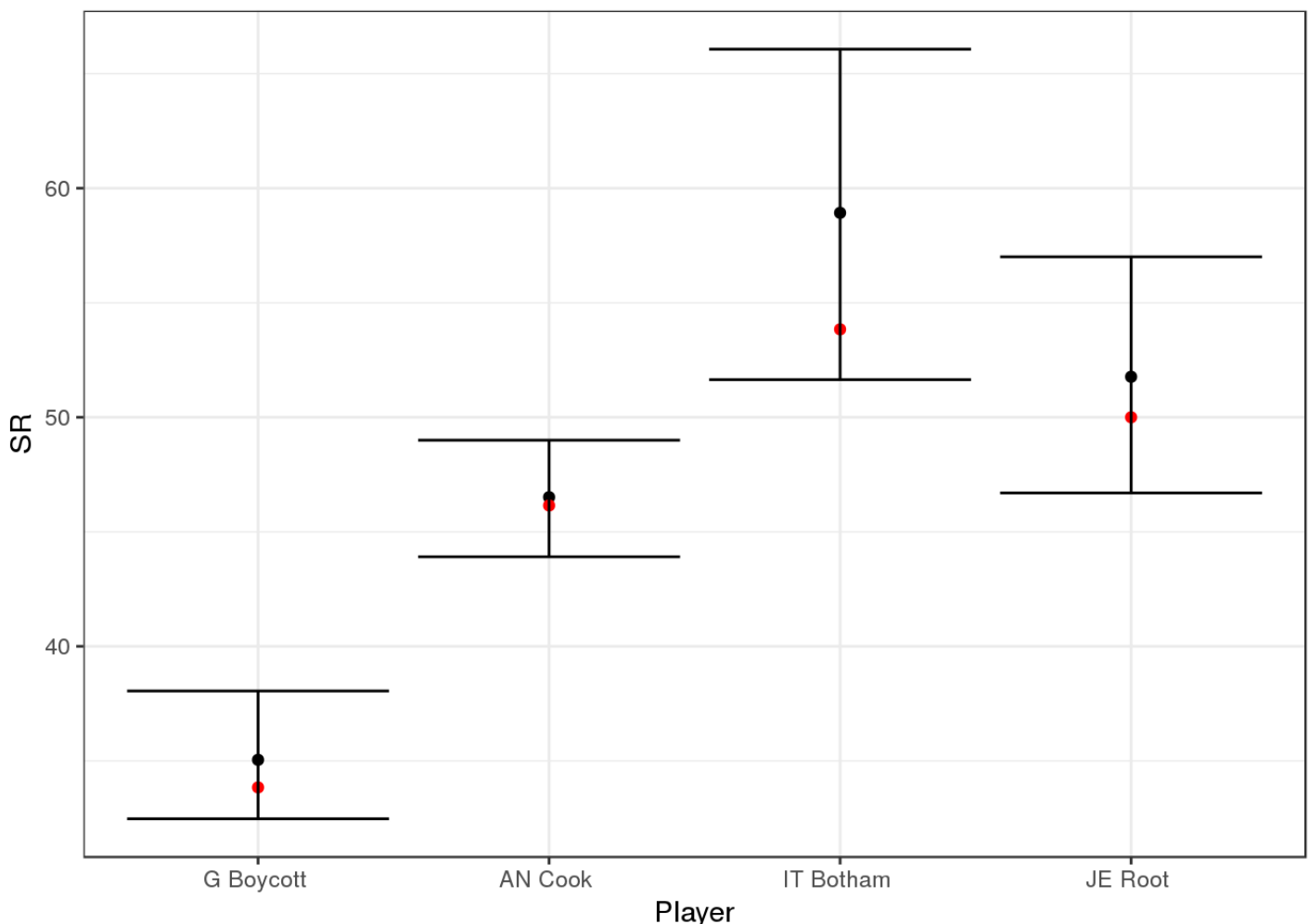


The confidence intervals overlap for the three top order batsmen. In other words if a selection of scores were drawn at random from any of the three sets of figures, the mean score would not depend on the name of the batsman. There is no statistically significant difference between them in terms of mean run totals. However Botham scores significantly fewer runs.

Scoring rate

Although Botham's innings tended to result in a lower mean number of runs, he was famous for scoring them much more quickly than other batsmen at the time. We can use the same bootstrap approach to look at scoring rate measured as runs per hundred balls.

```
g0<-ggplot(d,aes(x=Player,y=SR))
g0<-g0+stat_summary(fun.y=mean,geom="point")+stat_summary(fun.y=median,geom="point",col="red")
g1<-g0+stat_summary(fun.data=mean_cl_boot,geom="errorbar")
g1
```



Interestingly there is in fact no significant difference between Joe Root's scoring rate and that of Ian Botham. However Boycott clearly scored at a significantly slower rate than any of the other players.

Survival analysis

OK. So simple, well known analytical techniques have led to the conclusion that Boycott accumulated as many runs on average as Root and Cook, but did so more slowly. However there were some issues with the approach taken that were not fully resolved. A cricket innings is a matter of survival, and ends when a player is out. This is similar to the progress of a patient suffering from a disease or the failure rate of a

player is out. This is similar to the progress of a patient suffering from a disease or the failure rate of a piece of machinery. Survival analysis is a major field of statistics, but not very commonly encountered in environmental contexts. I have used it to look at survival of tree seedlings in published work.

One of the issues that survival analysis aims to handle cleanly is that of “censoring”. This term refers to incomplete knowledge, rather than missing data. Right censoring occurs when a failure (or death) is not observed by the end of a trial period. In this case, not out innings are a form of right censoring. We do not know what score the batsman might have gone on to make before being out. In the case of Geoffrey Boycott’s infamous 246 not out against India at Headingley in 1967 some people believe that he would still be batting today if tests were endless.

The survival package in R provides many tools for analysing these situations.

```
library(survival)
```

The first step is to set up a Surv object for modelling. This contains the information on censoring needed for the next steps. We can make two objects. One represents the lifetimes in terms of runs scored before failure (being out) and the other represents the numbers of balls faced. The completed column already contains the ones and zeros needed to form these objects.

```
runs<-Surv(d$Runs,d$completed)
balls<-Surv(d$BF,d$completed)
head(runs)
```

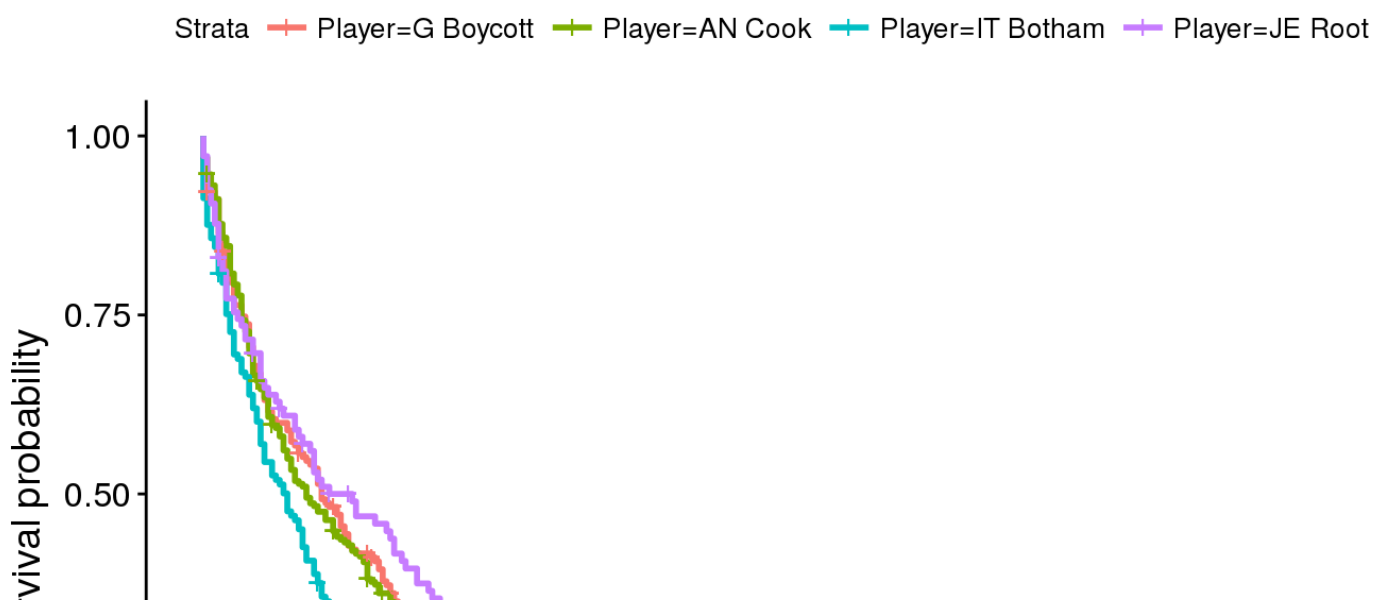
```
## [1] 254 200+ 190 182+ 180 154+
```

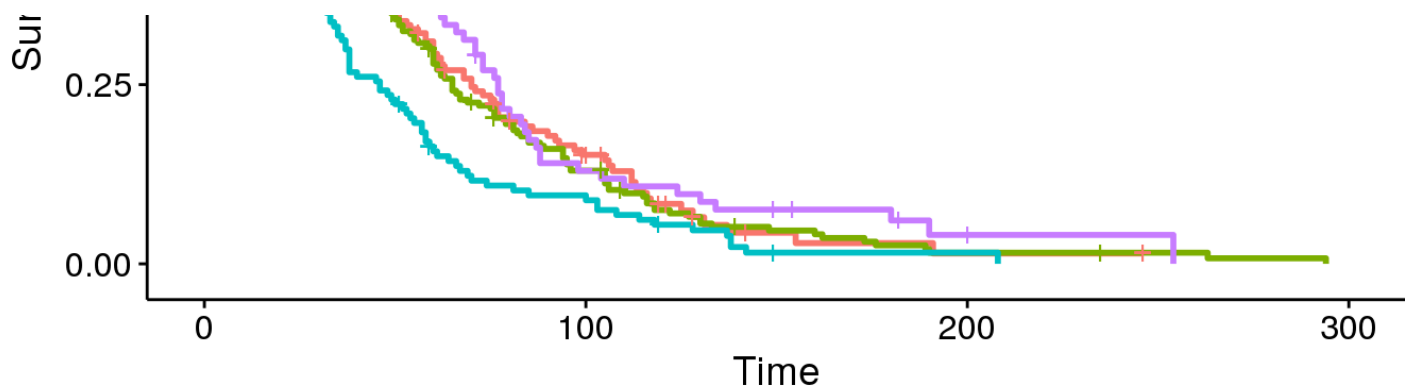
Notice that the object has a plus sign added for not out (right censored) innings.

Kaplan Meier plot

The first step in most survival analyses is to convert the data into a life table and form a Kaplan Meier plot. The survminer package in R provides much nicer plots than the default.

```
library(survminer)
fit <- survfit(runs ~ Player,data=d)
ggsurvplot(fit,conf.int = FALSE)
```





The stepwise pattern of the plot shows that innings do tend to follow a pattern that resembles exponential decay over time. The pattern with respect to runs scored looks to the eye to be remarkably similar for the three top order batsmen, with Joe Root just above the others. Botham looks to the eye to be more likely to be out before reaching a high score, and this is compatible with the previous analysis. Asking R to print the fit now produces the adjusted medians that I mentioned were needed. These are also given confidence intervals.

```
fit
```

```
## Call: survfit(formula = runs ~ Player, data = d)
##
##
##           n events median 0.95LCL 0.95UCL
## Player=G Boycott  193    170     31     25     39
## Player=AN Cook    261    246     27     22     37
## Player=IT Botham  161    155     22     15     27
## Player=JE Root    106     95     39     25     56
```

Cox proportional hazards model

If the underlying baseline risk of being out is assumed to follow a common pattern per run in all cases, which looks very reasonable here, we can formalise the analysis using a Cox Proportional Hazard model.

```
mod<-coxph(runs~d$Player)
summary(mod)
```

```
## Call:
## coxph(formula = runs ~ d$Player)
##
##      n= 721, number of events= 666
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## d$PlayerAN Cook    0.04003   1.04084  0.10005   0.400   0.6891
## d$PlayerIT Botham   0.31625   1.37198  0.11133   2.841   0.0045 **
## d$PlayerJE Root   -0.08836   0.91543  0.12848  -0.688   0.4916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## d$PlayerAN Cook      1.0408    0.9608    0.8555    1.266
## d$PlayerIT Botham     1.3720    0.7289    1.1030    1.707
## d$PlayerJE Root       0.9154    1.0924    0.7116    1.178
##
```

```

## Concordance= 0.536 (se = 0.013 )
## Rsquare= 0.017 (max possible= 1 )
## Likelihood ratio test= 12.14 on 3 df, p=0.006912
## Wald test = 12.68 on 3 df, p=0.005372
## Score (logrank) test = 12.79 on 3 df, p=0.005108

```

The output from a CPH model with many covariates can be challenging to interpret, but in this case it is quite straight forward. The hazard ratio is the exponentiated coefficient in the table. This represents the increase or decrease in risk of an event given a unit change in time (in this case time is measures as progress in terms of runs scored) relative to the baseline. The baseline is Geoffrey Boycott. So neither Cook nor Root are significantly more likely to get out before scoring the next run than Boycott. However the hazard ratio for Botham **is** significantly higher at 1.37. In other words there is a 37% higher chance of Botham getting out before progressing to a higher score than Boycott at every ball faced. Further analysis could fit an exponential model and look more closely at a parametric form of the hazard function, but we'll leave that for now.

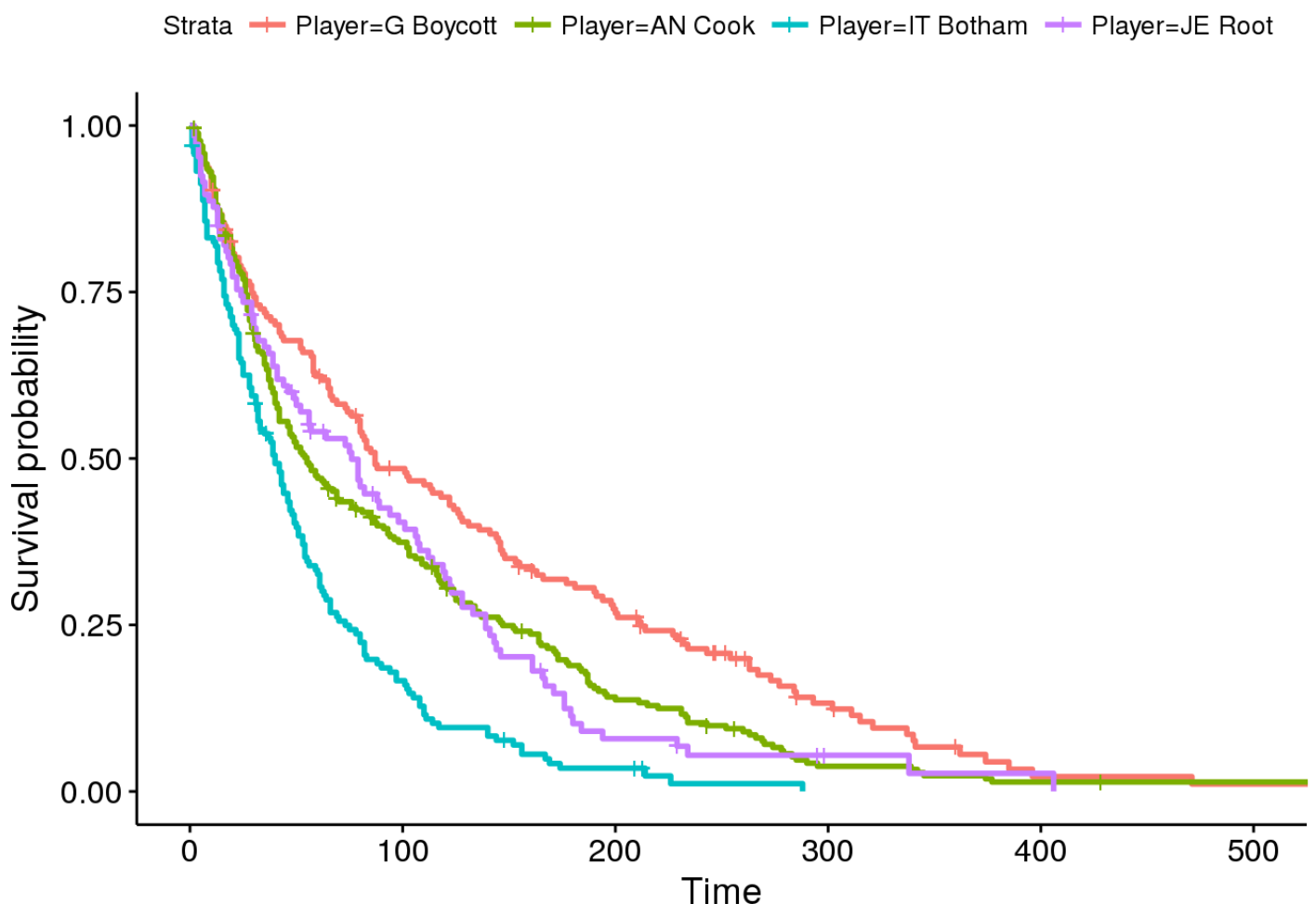
Surviving the next ball

The analysis above is based on survival with relation to progressing the scoring. A more classical form of life and death analysis involves simply surviving at the crease to face another ball. This can be analysed in terms of balls faced in exactly the same way.

```

fit <- survfit(balls ~ Player,data=d)
ggsurvplot(fit,conf.int = FALSE)

```



fit.

```
## Call: survfit(formula = balls ~ Player, data = d)
##
##      20 observations deleted due to missingness
##              n events median 0.95LCL 0.95UCL
## Player=G Boycott    173      152      87      74      126
## Player=AN Cook      261      246      55      42      70
## Player=IT Botham    161      155      40      32      48
## Player=JE Root      106       95      76      50     106
```

There is clear space between Boycott and the other players. The curve is much flatter for Boycott and very steep for Botham.

```
mod<-coxph(balls~d$Player)
summary(mod)
```

```
## Call:
## coxph(formula = balls ~ d$Player)
##
##      n= 701, number of events= 648
##      (20 observations deleted due to missingness)
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## d$PlayerAN Cook    0.3397    1.4045  0.1037  3.275  0.00106 **
## d$PlayerIT Botham  0.8901    2.4354  0.1185  7.513  5.76e-14 ***
## d$PlayerJE Root    0.3742    1.4538  0.1320  2.834  0.00459 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## d$PlayerAN Cook        1.404      0.7120      1.146      1.721
## d$PlayerIT Botham      2.435      0.4106      1.931      3.072
## d$PlayerJE Root        1.454      0.6878      1.122      1.883
##
## Concordance= 0.572 (se = 0.013 )
## Rsquare= 0.076 (max possible= 1 )
## Likelihood ratio test= 55.4 on 3 df,  p=5.648e-12
## Wald test              = 57.83 on 3 df,  p=1.712e-12
## Score (logrank) test = 59.96 on 3 df,  p=5.995e-13
```

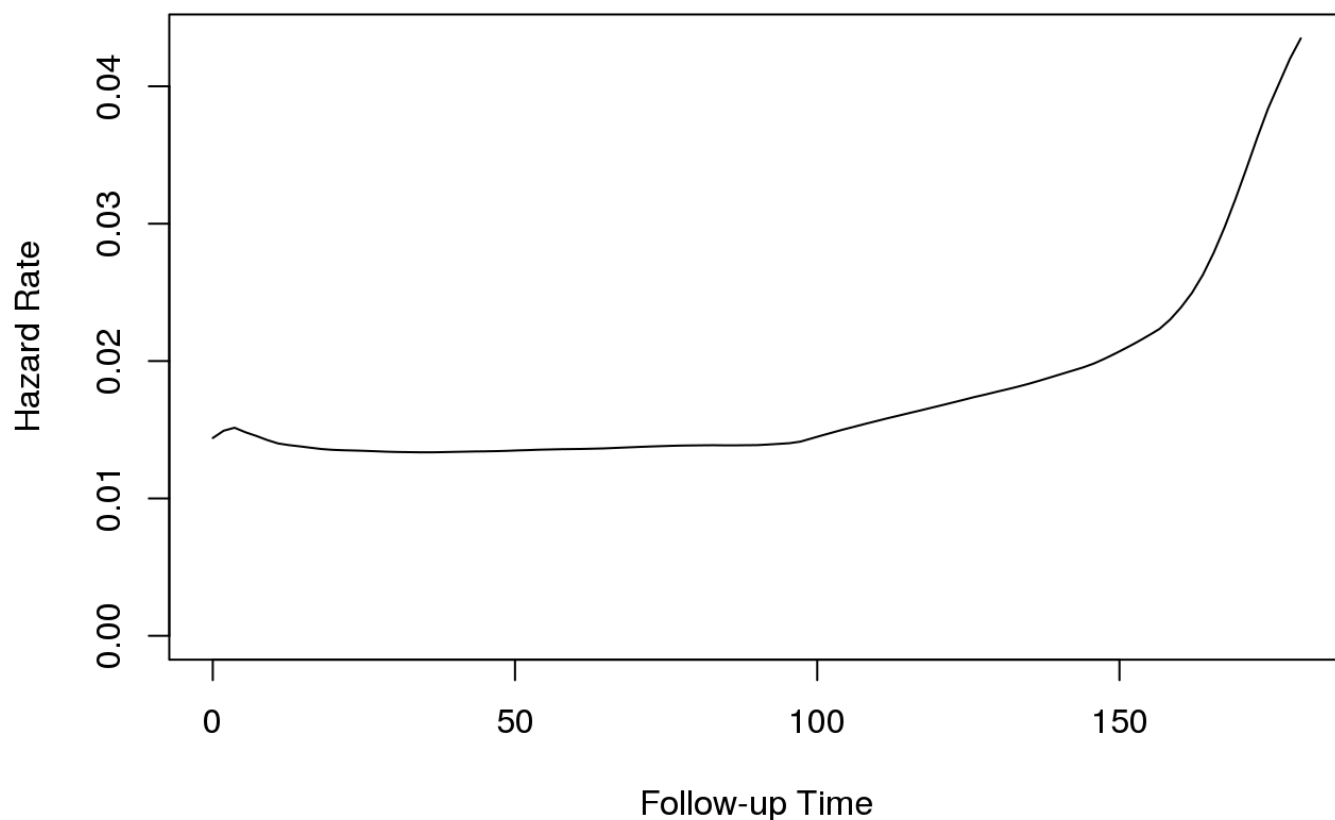
So there we have it. Boycott was indeed statistically significantly more difficult to get out per ball bowled at him than any of the other players. Each ball bowled at Botham had about two and half times the probability of resulting in a wicket than a ball bowled at Boycott. That is the direct interpretation of the exponentiated hazard coefficient of 2.43 in simple words. A ball bowled at Root has a 45% higher chance of taking his wicket than a ball bowled at Boycott. However as the previous analysis showed it wasn't more difficult to get Boycott out per run that he scored than either Root or Cook. Boycott was a survivor, but a very slow scorer.

Did the probability of taking a wicket change over the innings.

There are various ways of looking at this issue and several options in R. A quick way is to graph the hazard using the `muhaz` package. We will do that for two of the players.

Hazard rate for Root

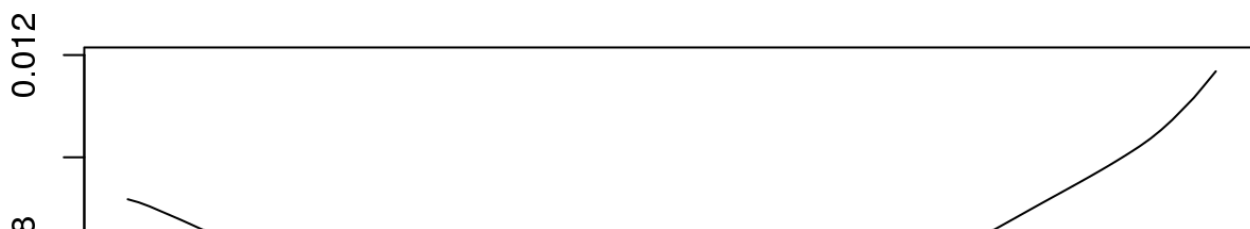
```
library(muhaz)
mod<-muhaz(root$BF,root$completed)
plot(mod)
```

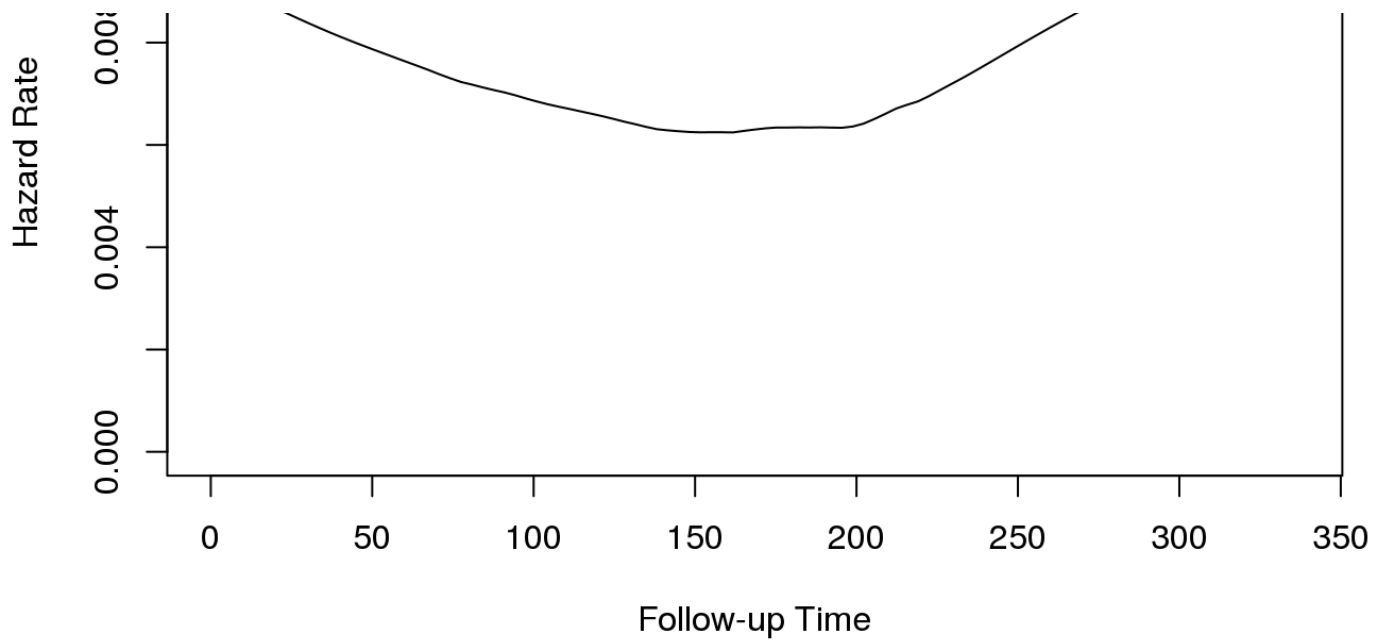


This stays fairly constant, but increases at the end of the innings. Of course there is less data at this point, so if the confidence intervals were added this increase may not be significant.

Hazard rate for Boycott

```
boycott<-na.omit(boycott)
mod<-muhaz(boycott$BF,boycott$completed)
plot(mod)
```





So, the hazard rate for opening batsmen tends to stay reasonably flat throughout the innings. In Boycott's case it falls slightly (presumably after he's seen off the new ball). To put this in context a hazard rate of 0.008 can be used to calculate the probability of getting Boycott out before lunch. If he faces 100 balls in the first session the chances of not getting out each ball is $1 - 0.008 = 0.992$. Raising this to the power 100 gives the probability of still being there i.e. $\text{round}(0.992^{100} * 100, 2) = 44.79\%$. We established that Boycott's strike rate was below 35 so this makes for some rather tedious cricket.