ECDFs in ggplot2

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

First let's simulate some data - we'll make some normally distributed height values and plot a histrogram.

```
library(ggplot2)
```

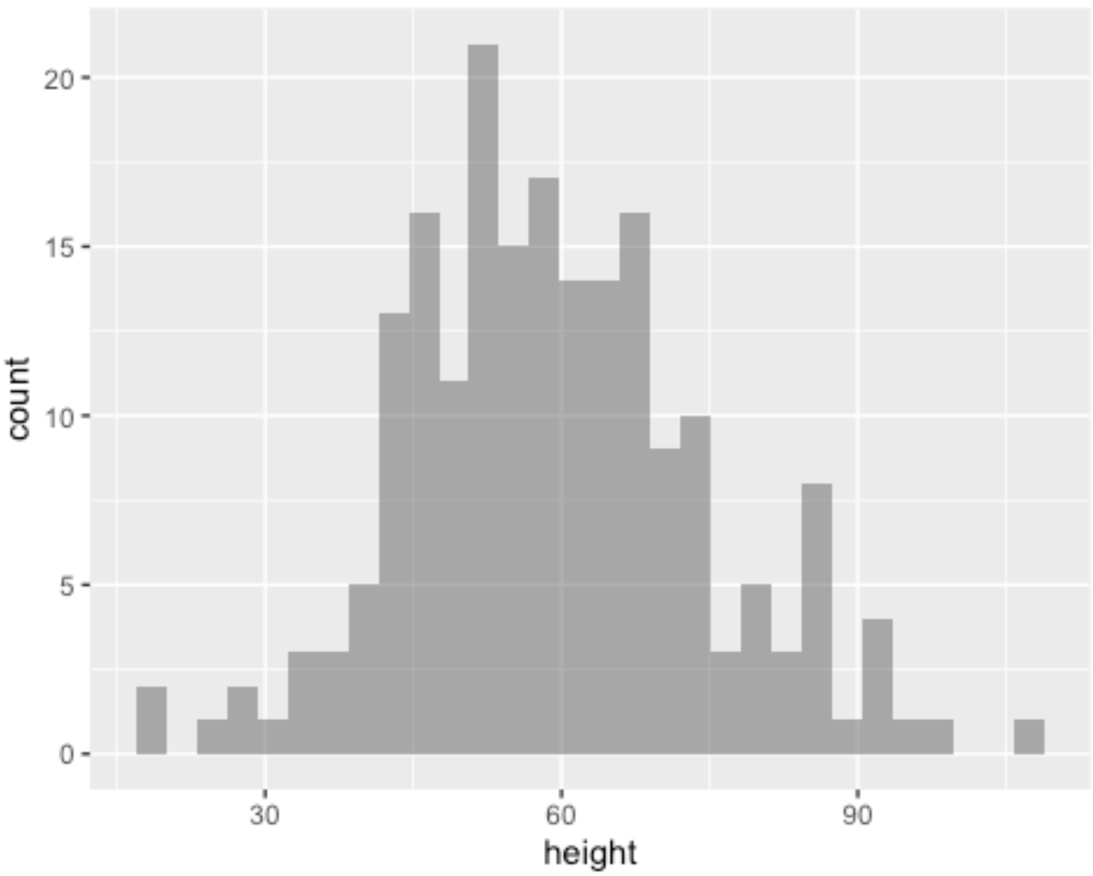
```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2020a
.
## 1.0/zoneinfo/America/New_York'
```

```
set.seed(1234)
df <- data.frame(height = round(rnorm(200, mean=60, sd=15)))
head(df)
```

```
##   height
## 1     42
## 2     64
## 3     76
## 4     25
## 5     66
## 6     68
```

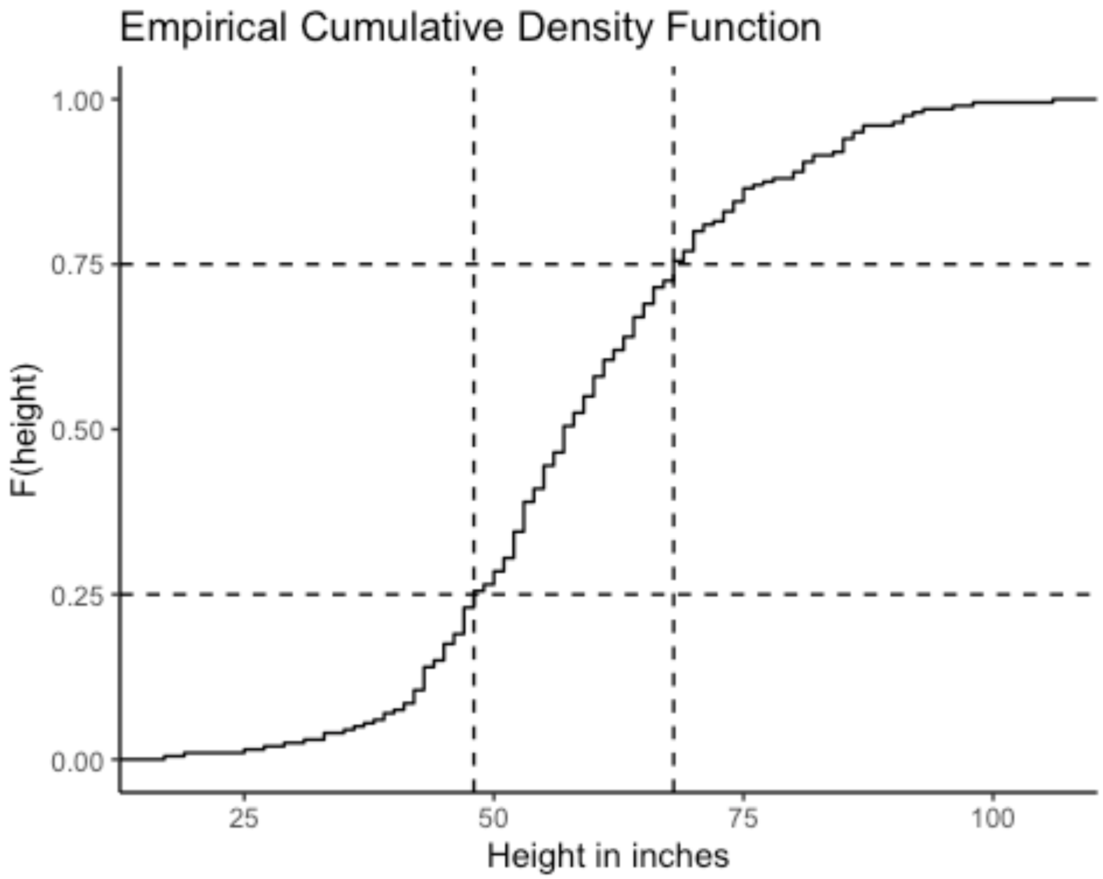
```
ggplot(df, aes(height)) +
  geom_histogram(alpha = 0.5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



No let’s create an empirical cummulative density function plot of the same data. We’ll use the “step” geom in ggplot2. And we’ll add some reference lines at the upper and lower quartiles.

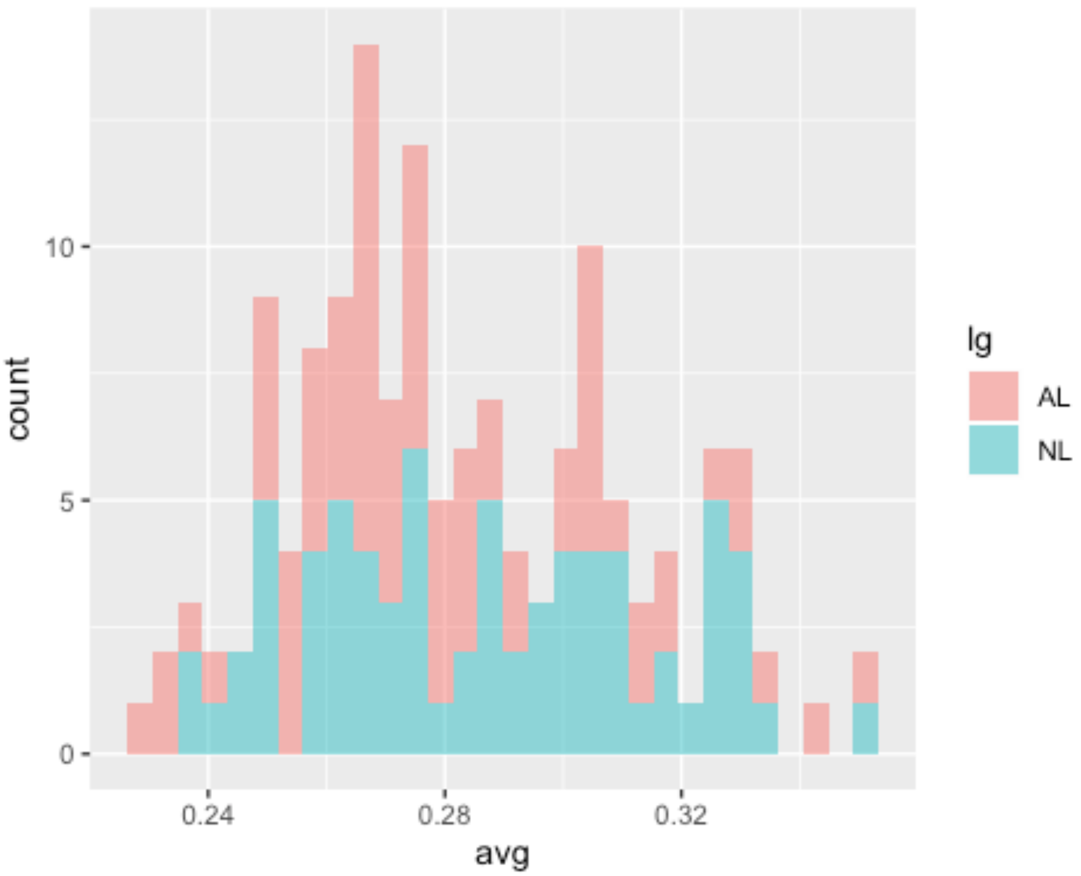
```r
# Basic ECDF plot
ggplot(df, aes(height)) +
  stat_ecdf(geom = "step") +
  labs(title="Empirical Cumulative Density Function",
       y = "F(height)", x="Height in inches")+
  geom_vline(aes(xintercept = quantile(height)[2]), linetype="dashed") + #fir
st quartile
  geom_vline(aes(xintercept = quantile(height)[4]), linetype="dashed")+ #thir
d quartile
  geom_hline(yintercept=0.25, linetype="dashed")+
  geom_hline(yintercept=0.75, linetype="dashed")+
  theme_classic()
```



Now lets this do this with some real data. We'll use the TopHitters data set from 2001 (in the gcookbook library - install this if you don't have it already).

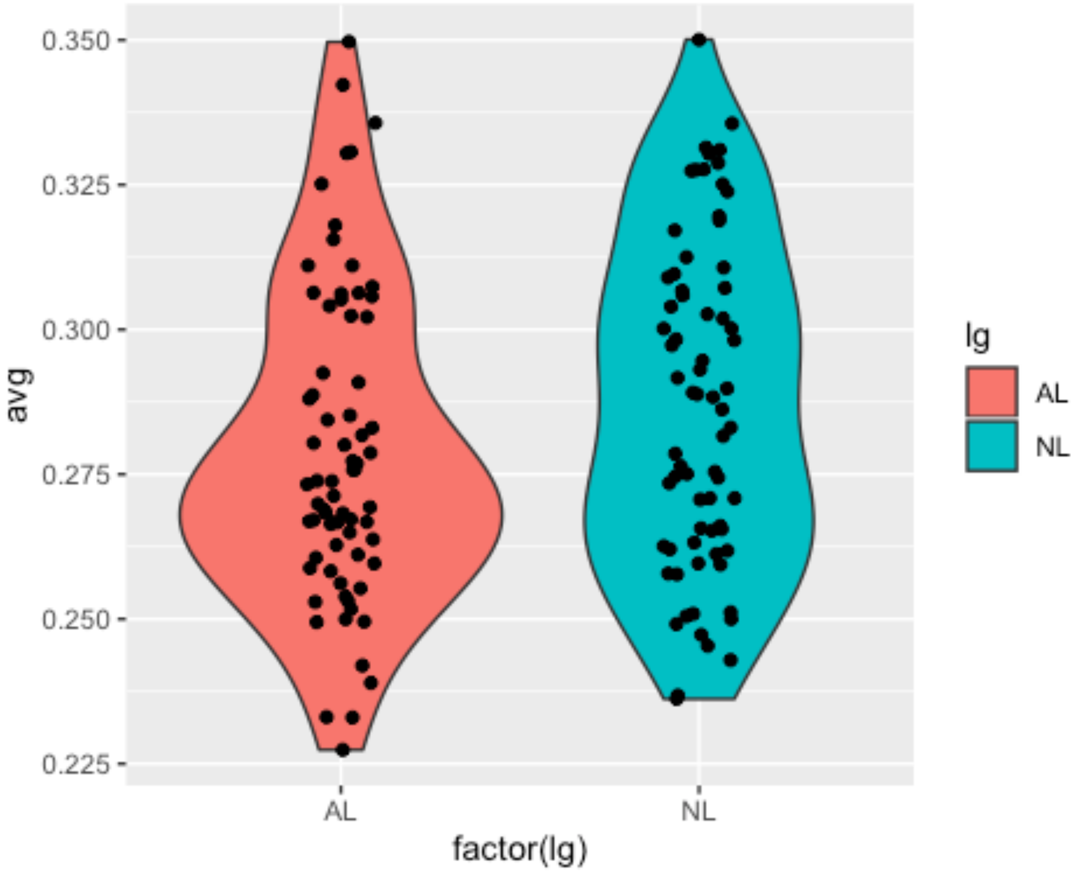
First let's get the data and look at a histogram.

```
library(gcookbook)
#View(tophitters2001)

#histogram
ggplot(tophitters2001, aes(avg, fill=lg)) +
  geom_histogram(alpha = 0.5)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

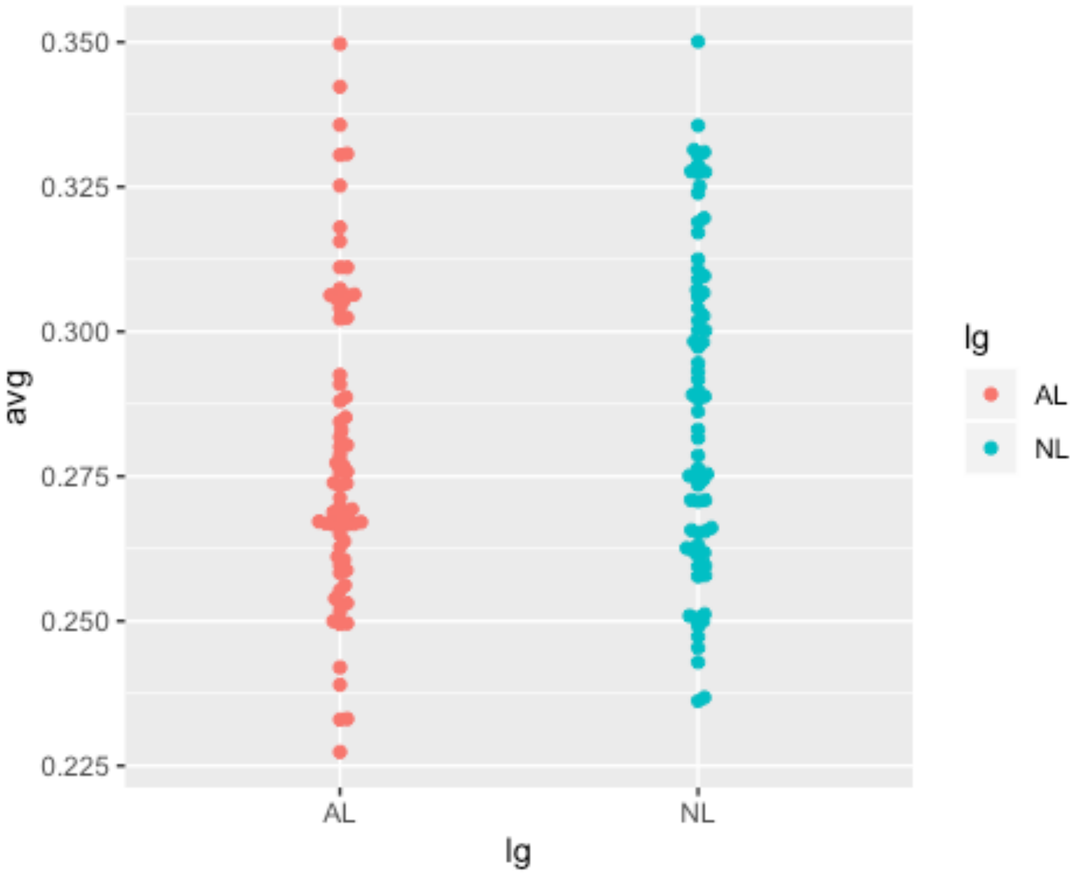


We see the distribution of batting averages for the top MLB hitters in 2001 colored by league. Now let’s look at some other distribution representations. The swarmplot requires ggbeeswarm.

```r
#violin plot
ggplot(tophitters2001, aes(factor(lg), avg)) +
  geom_violin(aes(fill=lg)) +
  geom_jitter(width=0.10)
```

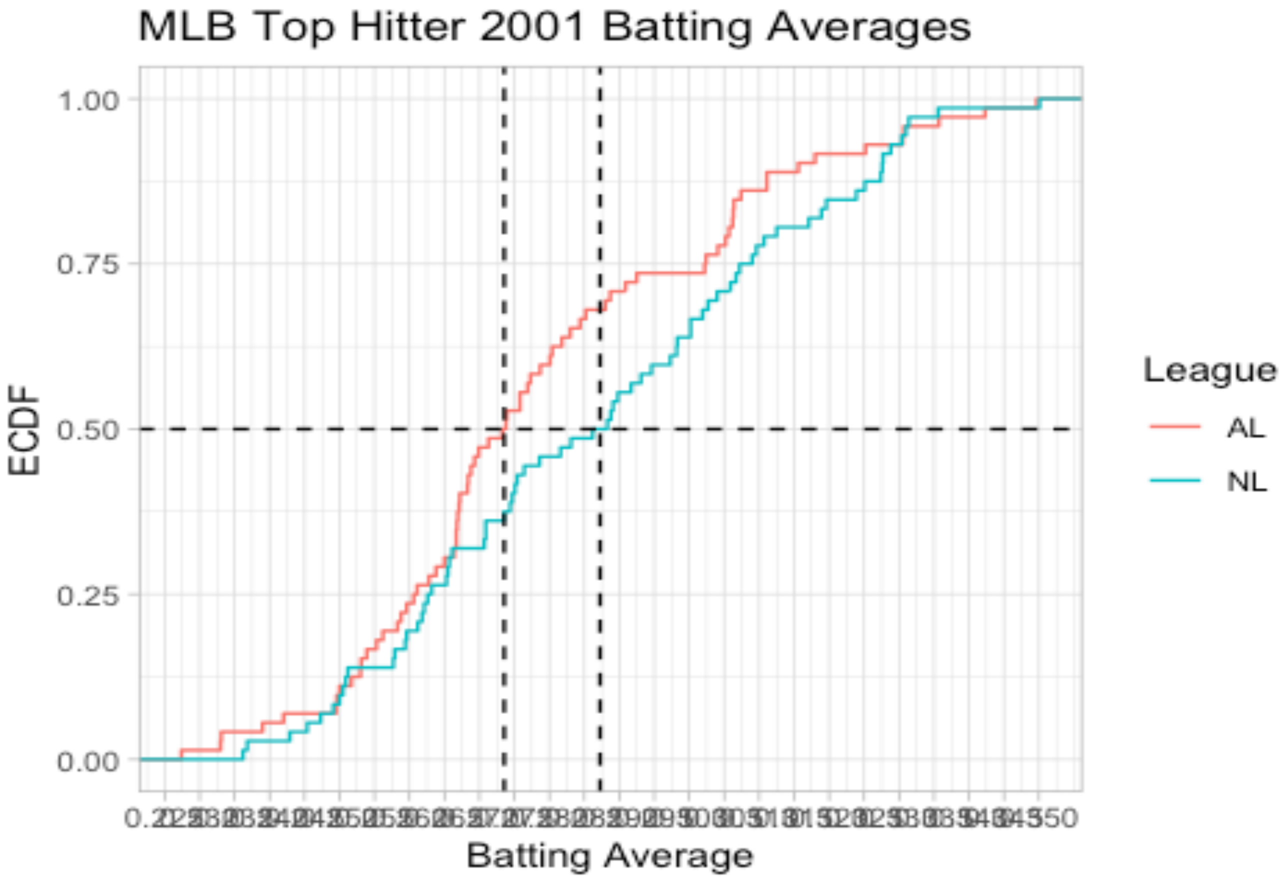


```
#Swarmplot
#install.packages("ggbeeswarm")
library(ggbeeswarm)
ggplot(tophitters2001, aes(lg, avg)) +
  geom_beeswarm(aes(color=lg))
```



Finally, let’s show the difference in ECDFs for the 2 leagues. What can you say about top hitter distributions across the two leagues? Expand the plot size in R-Studio for a better look.

```
##ECDF for TopHitters2001
ggplot(tophitters2001, aes(avg)) +
  stat_ecdf(geom = "step", aes(color=lg)) +
  geom_hline(aes(yintercept=0.5), linetype="dashed") +
  geom_vline(data=subset(tophitters2001, lg=="AL"), aes(xintercept=quantile(a
vg)[3]),
             linetype="dashed") +
  geom_vline(data=subset(tophitters2001, lg=="NL"), aes(xintercept=quantile(a
vg)[3]),
             linetype="dashed") +
  labs(title="MLB Top Hitter 2001 Batting Averages",
       y = "ECDF", x="Batting Average") +
  scale_x_continuous(breaks=seq(0.220, 0.350, 0.005),
                     labels=function(x){sprintf("%.3f", x)}) +
  scale_color_discrete("League") +
  theme_light()
```

MLB Top Hitter 2001 Batting Averages
ECDF
1.00
0.75
0.50
0.25
0.00
Batting Average
League
AL
NL