확률분포함수 probability density function 개요

측정형 변수의 실증적 empirical 분포함수 - 분포의 적합성 검정

한 개 변수의 히스토그램(확률분포함수)

미국 MLB 타자 연봉 데이터 (http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/baseball.csv)

타자 연봉 및 능력 데이터

```
baseball<-read.csv('http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/
baseball.csv')
names(baseball);dim(baseball)
baseball.subset<-baseball[c(1:30),]
```

```
> names(baseball);dim(baseball)
 [1] "Player_Name"      "Team"             "TimesatBat"       "Hits"
 [5] "HomeRuns"         "Runs"             "RBIs"             "Walks"
 [9] "YearsinMLB"       "CareerTimesatBat" "CareerHits"       "CareerHome"
[13] "CareerRuns"       "CareerRBIs"       "CareerWalks"      "League"
[17] "Division"         "Position"         "PutOuts"          "Assists"
[21] "Errors"           "Salary"           "League2"
[1] 322  23
```

선수 안타수 분포함수

(1) hist() 함수 이용

breaks= 막대(bin)의 개수 설정 (10~15), probability=T (확률분포함수), F(y-축 - 빈도)

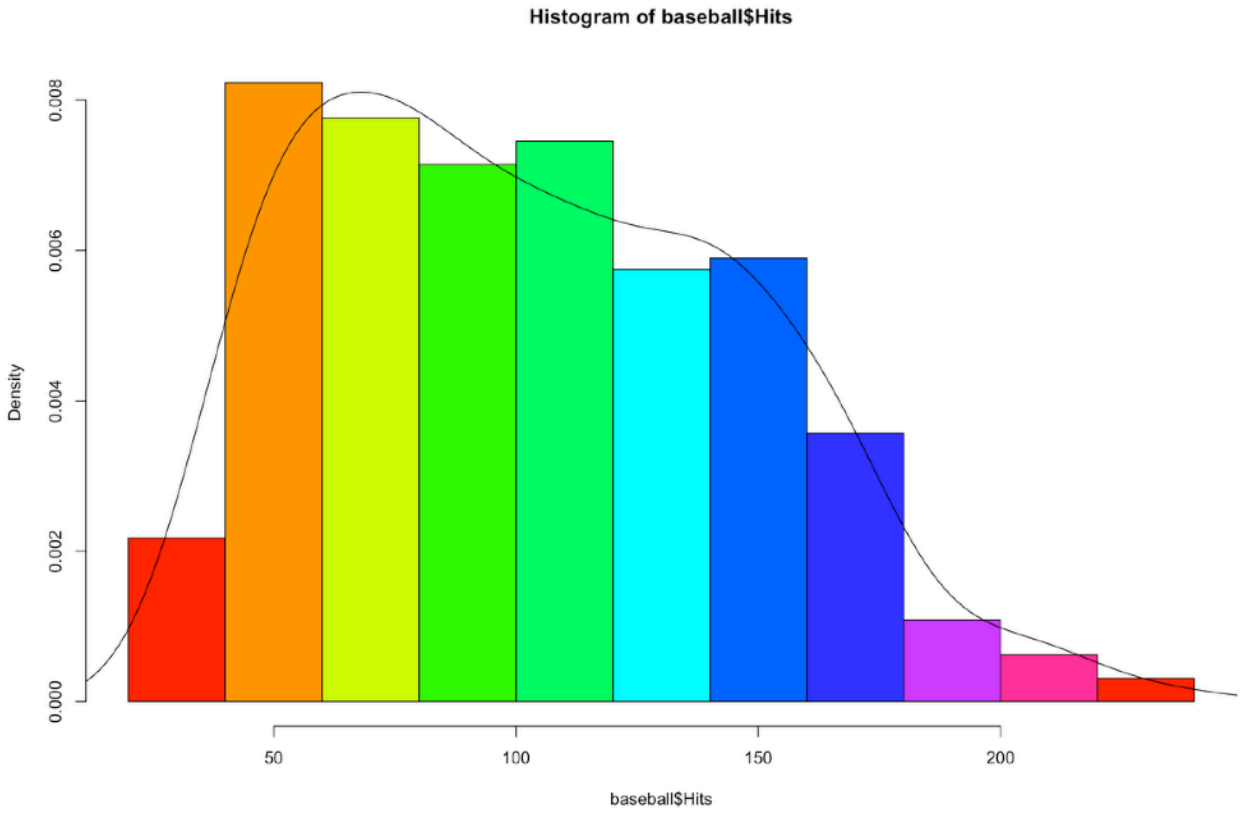
rainbow(빈의 개수) - 빈 색깔을 서로 다르게 표현

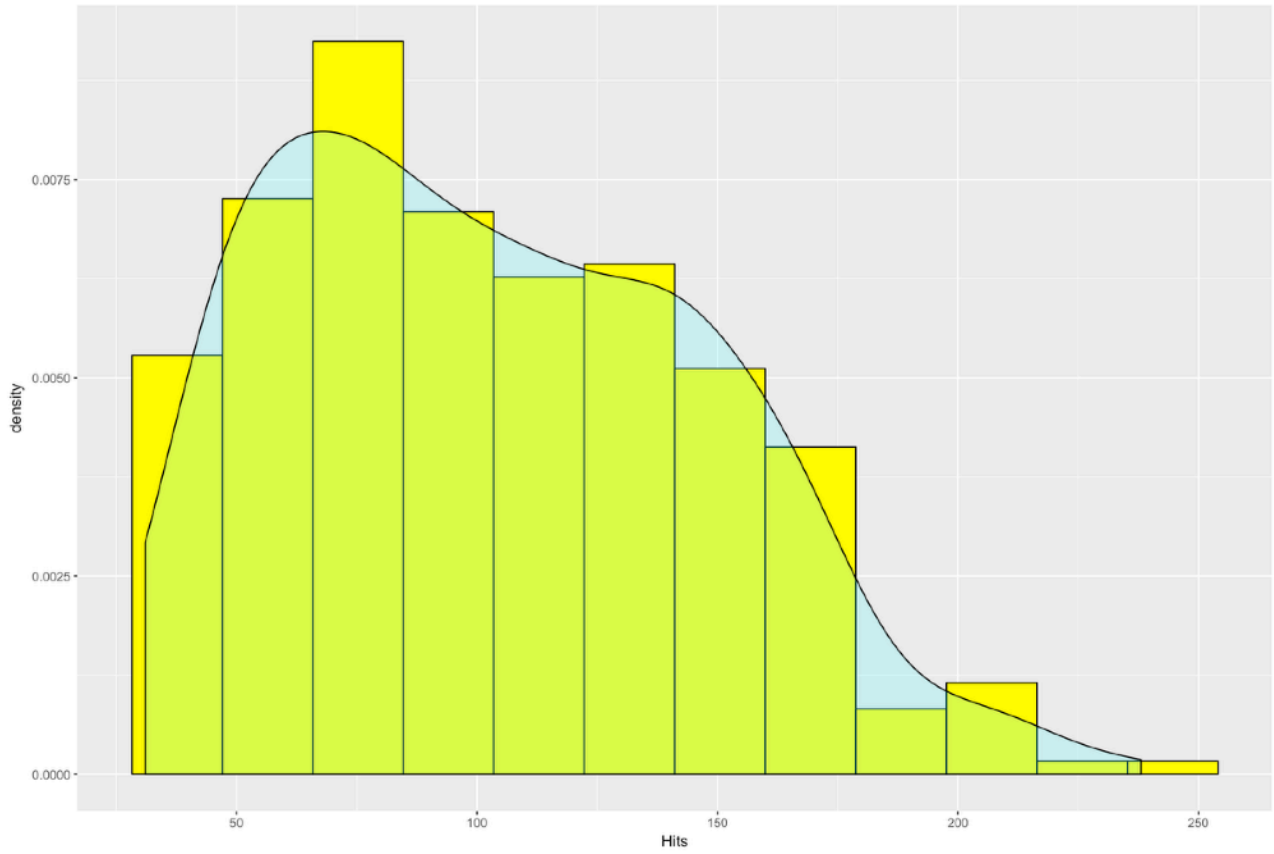
```
den<-density(baseball$Hits)
hist(baseball$Hits,breaks=10,col=rainbow(10),probability=TRUE)
lines(den)
```

(2) ggplot() 함수 이용

y=..density.. <=> y-축을 확률로 설정, bins= 막대 개수 설정, alpha= 불투명 정도 1이면 불투명, 작을수록 투명도 높아짐

```
ggplot(baseball,aes(x=Hits)) +
  geom_histogram(aes(y=..density..),colour="black",fill='yellow',bins=12)+
  geom_density(alpha=.2, fill='cyan')
```


Histogram of baseball$Hits
Density
0.000
0.002
0.004
0.006
0.008
50
100
150
200
baseball$Hits


density
0.0000
0.0025
0.0050
0.0075
50
100
150
200
250
Hits

집단 변수가 존재하는 경우

리그 (National, American) 타자 안타수

집단별 평균을 구한다. 이를 이용하여 수평선을 그린다. bins= 막대 개수,
theme(legend.position='right') : 범례 표시를 그래프 오른쪽에 위치

```
library(plyr)
mu <- ddply(baseball,'League',summarise,grp.mean=mean(Hits))
head(mu)
ggplot(baseball,aes(x=Hits,color=League)) +
  geom_histogram(fill="white",position='dodge',bins=15)+
  geom_vline(data=mu, aes(xintercept=grp.mean, color=League),
             linetype="dashed")+
  theme(legend.position='right') +
  ggtitle('PDF of Hits by League') +
  theme_bw()
```