# Homework assignment #3

2021320322 / Minseo Yoon

November 17, 2022

## 1

### 15.5.1

```
take_sample <- function(p, N) {
  x <- sample(c(0, 1), size = N, replace = TRUE, prob = c(1-p, p))
  mean(x)
}
```

### 15.5.2

```
p <- 0.45
errors <- replicate(10000, take_sample(p, 100) - p)
head(errors)
```

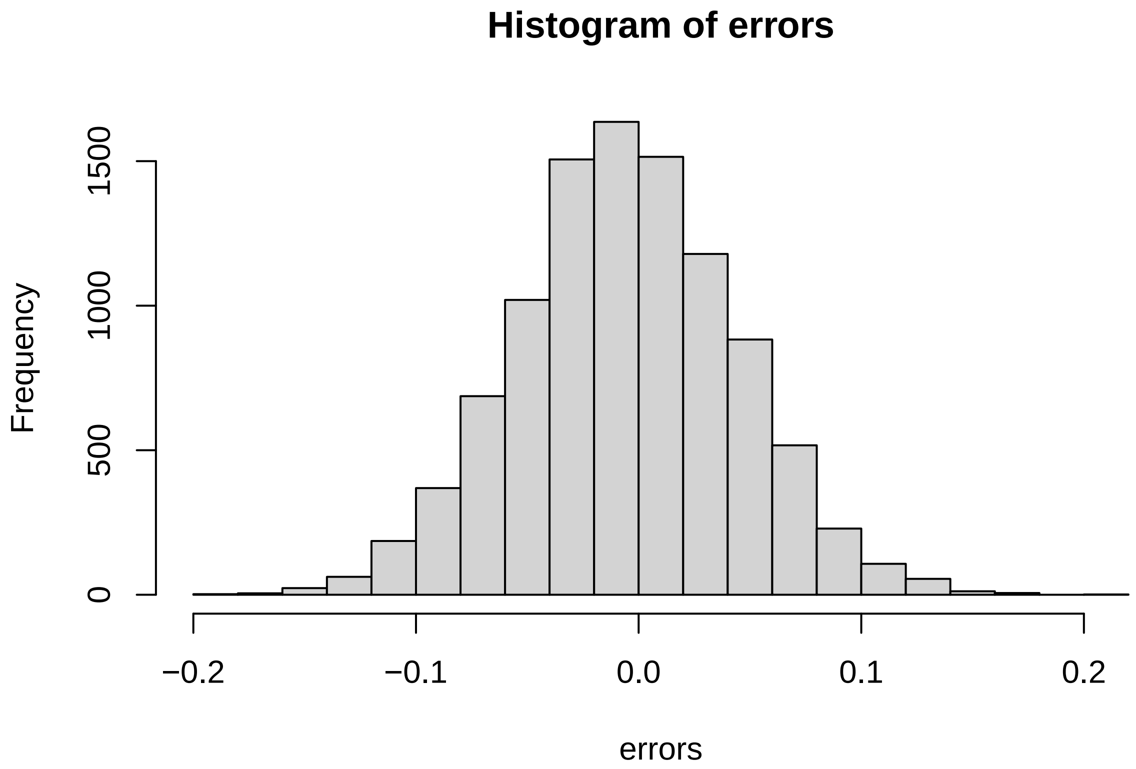
```
## [1]  0.13 -0.03 -0.08  0.07 -0.02 -0.06
```

### 15.5.3

```
mean(errors)
```

```
## [1] -0.000382
```

```
hist(errors)
```



Answer: c. The errors are symmetrically distributed around 0.

**15.5.9**

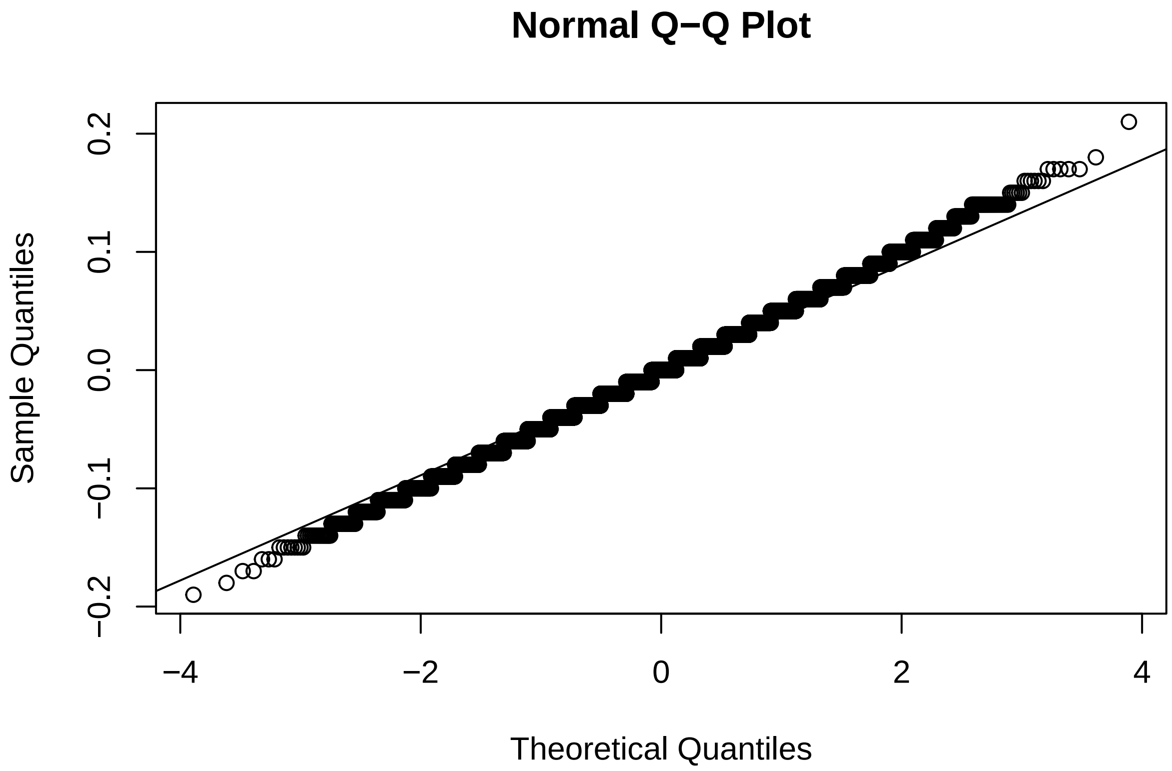
Answer: b. approximately normal with expected value $p$ and standard error $\sqrt{p(1-p)/N}$

**15.5.10**

Answer: b. approximately normal with expected value 0 and standard error $\sqrt{p(1-p)/N}$

**15.5.11**

```
qqnorm(errors); qqline(errors)
```



### 15.5.12

```
p <- 0.45
N <- 100
1 - pnorm(0.5, p, sqrt(p * (1-p) / N))
```

```
## [1] 0.1574393
```

## 2

```
library(dslabs)
data("polls_us_election_2016")
library(tidyverse)
polls <- polls_us_election_2016 %>%
  filter(enddate >= "2016-10-31" & state == "U.S.")
```

### 15.7.1

```r
N <- polls$samplesize[1]
x_hat <- polls$rawpoll_clinton[1] / 100
c(x_hat - qnorm(0.975) * sqrt(x_hat * (1-x_hat) / N),
  x_hat + qnorm(0.975) * sqrt(x_hat * (1-x_hat) / N))
```

```
## [1] 0.4492385 0.4907615
```

### 15.7.2

```r
library(dplyr)
polls <- polls %>%
  mutate(x_hat = polls$rawpoll_clinton / 100, se_hat = sqrt(x_hat * (1-x_hat) / samplesize),
         lower = x_hat - qnorm(0.975) * se_hat,
         upper = x_hat + qnorm(0.975) * se_hat) %>%
  select(pollster, enddate, x_hat, lower, upper)
head(polls)
```

```
##                                                      pollster    enddate  x_hat
## 1                                    ABC News/Washington Post 2016-11-06 0.4700
## 2                                     Google Consumer Surveys 2016-11-07 0.3803
## 3                                                       Ipsos 2016-11-06 0.4200
## 4                                                      YouGov 2016-11-07 0.4500
## 5                                            Gravis Marketing 2016-11-06 0.4700
## 6 Fox News/Anderson Robbins Research/Shaw & Company Research 2016-11-06 0.4800
##       lower     upper
## 1 0.4492385 0.4907615
## 2 0.3744632 0.3861368
## 3 0.3993524 0.4406476
## 4 0.4339199 0.4660801
## 5 0.4624165 0.4775835
## 6 0.4527896 0.5072104
```

### 15.7.3

```r
polls <- polls %>%
  mutate(hit = lower <= 0.482 & 0.482 <= upper)
head(polls)
```

```
##                                                      pollster    enddate  x_hat
## 1                                    ABC News/Washington Post 2016-11-06 0.4700
## 2                                  Google Consumer Surveys 2016-11-07 0.3803
## 3                                                    Ipsos 2016-11-06 0.4200
## 4                                                   YouGov 2016-11-07 0.4500
## 5                                         Gravis Marketing 2016-11-06 0.4700
## 6 Fox News/Anderson Robbins Research/Shaw & Company Research 2016-11-06 0.4800
##       lower     upper   hit
## 1 0.4492385 0.4907615  TRUE
## 2 0.3744632 0.3861368 FALSE
## 3 0.3993524 0.4406476 FALSE
## 4 0.4339199 0.4660801 FALSE
## 5 0.4624165 0.4775835 FALSE
## 6 0.4527896 0.5072104  TRUE
```

### 15.7.4

```
polls %>%
  summarize(mean(hit))
```

```
##   mean(hit)
## 1 0.3142857
```

Answer: 0.3142857

### 15.7.5

Answer: 0.95

### 15.7.6

```
polls <- polls_us_election_2016 %>%
  filter(enddate >= "2016-10-31" & state == "U.S.")  %>%
  mutate(d_hat = rawpoll_clinton / 100 - rawpoll_trump / 100)

N <- polls$samplesize[1]
d_hat <- polls$d_hat[1]
c(d_hat - qnorm(0.975) * sqrt(d_hat * (1-d_hat) / N),
  d_hat + qnorm(0.975) * sqrt(d_hat * (1-d_hat) / N))
```

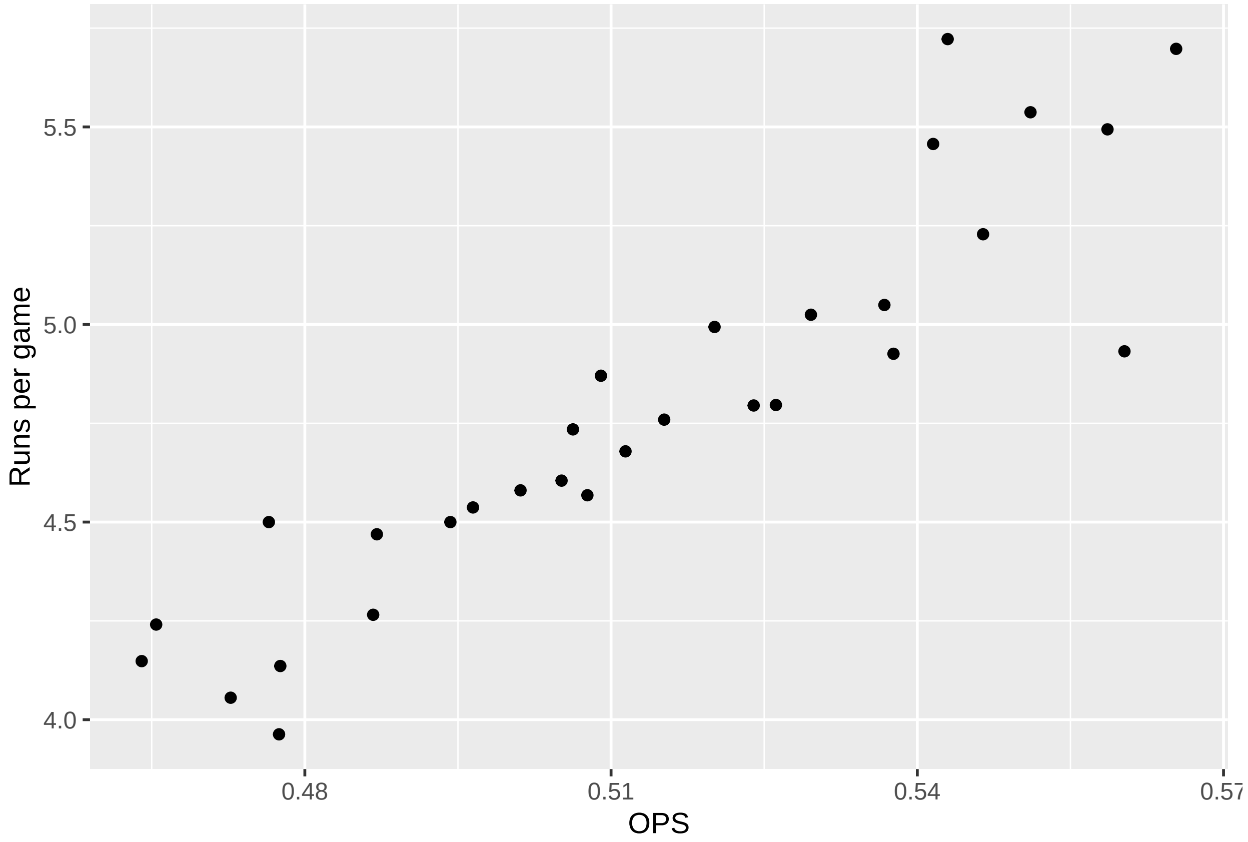
```
## [1] 0.03184851 0.04815149
```

## 3

```
library(Lahman)
library(ggplot2)
```

### 18.10.1

```
Teams %>%
  filter(yearID == 2001) %>%
  group_by(teamID) %>%
  mutate(PA = BB + AB, OPS = BB/PA + (H + X2B + 2*X3B + 3*HR)/AB) %>%
  ggplot(aes(x = OPS, y = R / G)) +
  xlab('OPS') +
  ylab('Runs per game') +
  geom_point()
```

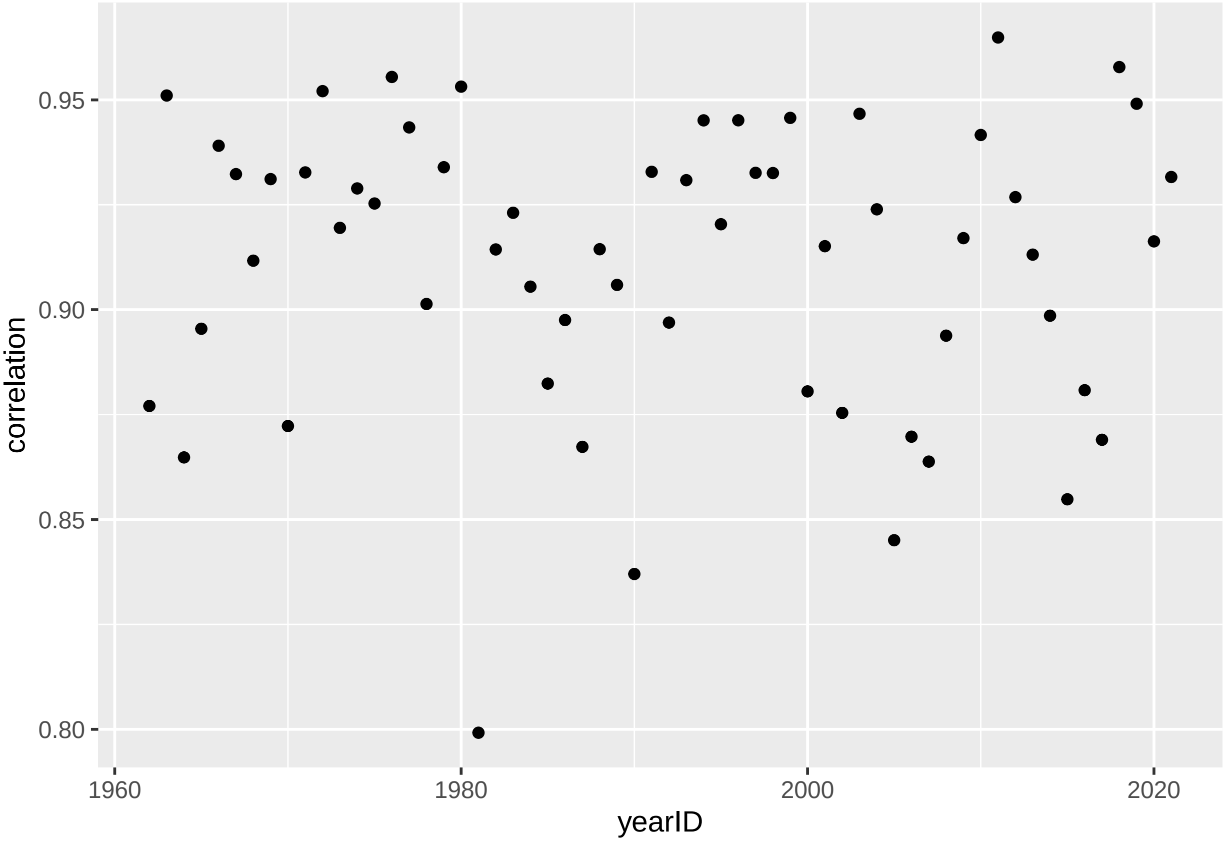


### 18.10.2

```
corr <- Teams %>%
  filter(yearID >= 1962) %>%
  group_by(yearID) %>%
  mutate(PA = BB + AB, OPS = BB/PA + (H + X2B + 2*X3B + 3*HR)/AB) %>%
  summarize(correlation = cor(R/G, OPS))
head(corr)
```

```
## # A tibble: 6 x 2
##   yearID correlation
##    <int>       <dbl>
## 1   1962       0.877
## 2   1963       0.951
## 3   1964       0.865
## 4   1965       0.895
## 5   1966       0.939
## 6   1967       0.932
```

```
corr %>%
  ggplot(aes(x = yearID, y = correlation)) +
  geom_point()
```

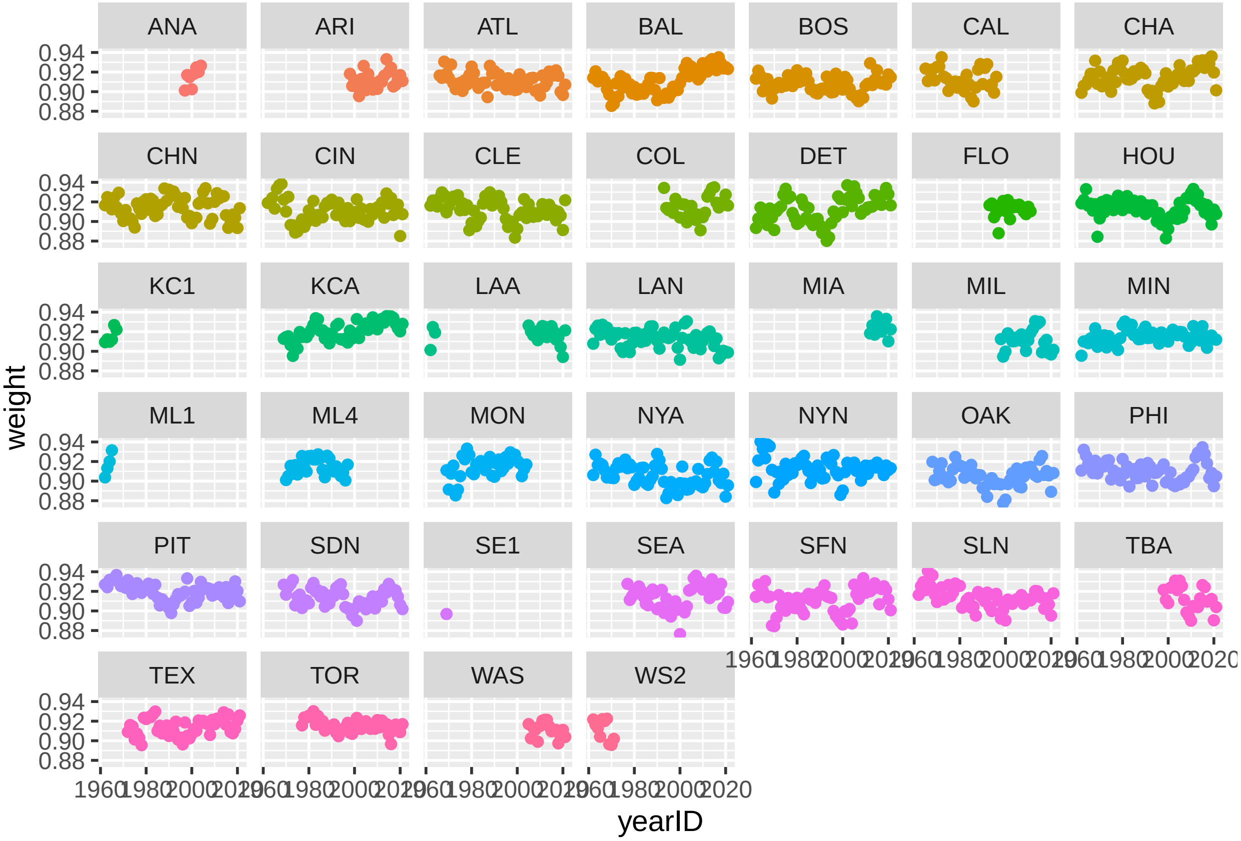


### 18.10.3

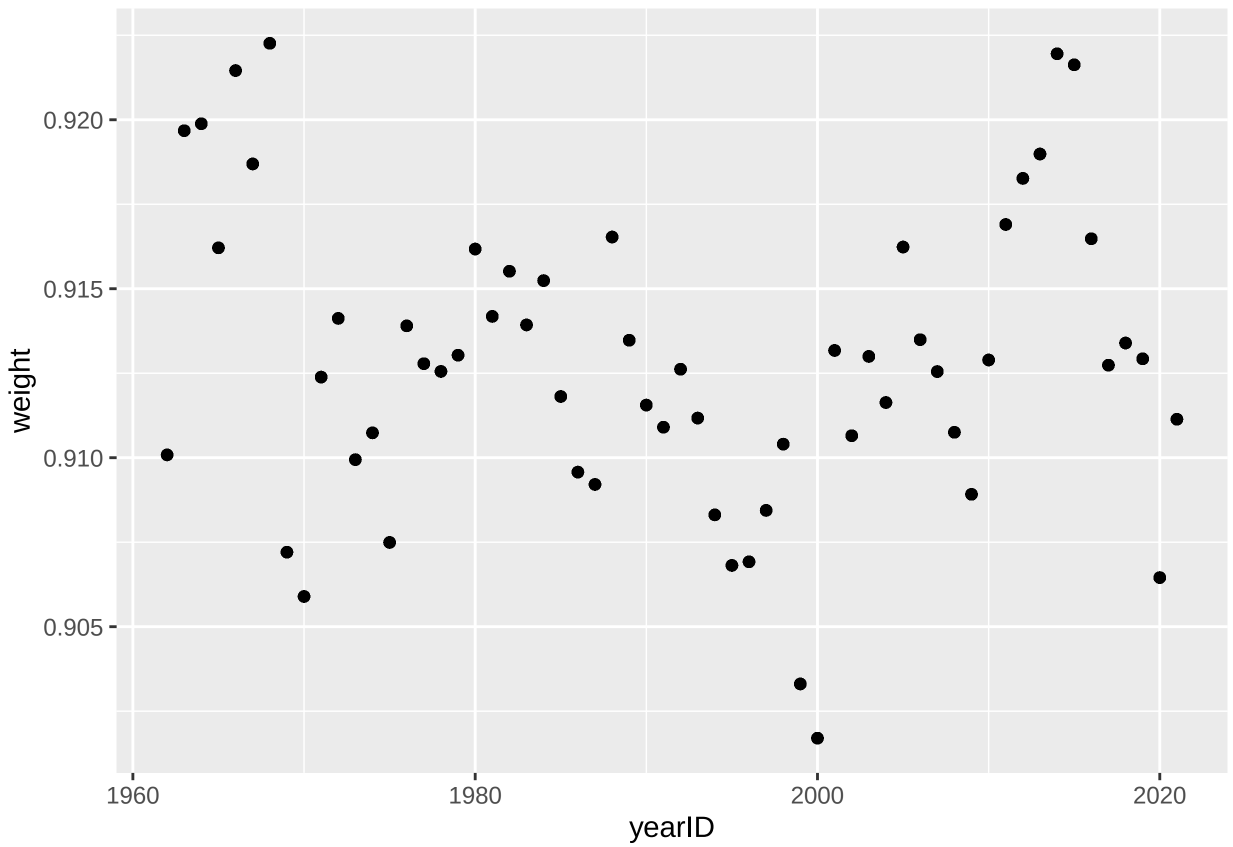
Answer: $\frac{AB}{PA}$

### 18.10.4

```
Teams %>%
  filter(yearID >= 1962) %>%
  group_by(teamID) %>%
  mutate(PA = AB + BB, weight = AB / PA) %>%
  ggplot(aes(x = yearID, y = weight)) +
  geom_point(aes(color = teamID)) +
  theme(legend.position = 'none') +
  facet_wrap(~teamID)
```



```
Teams %>%
  filter(yearID >= 1962) %>%
  group_by(yearID) %>%
  mutate(PA = sum(AB) + sum(BB), weight = sum(AB) / PA) %>%
  ggplot(aes(x = yearID, y = weight)) +
  geom_point()
```



```
Teams %>%
  filter(yearID >= 1962) %>%
  mutate(PA = sum(AB) + sum(BB), weight = sum(AB) / PA) %>%
  summarize(overall_average = mean(weight))
```

```
##   overall_average
## 1       0.9127671
```

Overall average: 0.9127671

### 18.10.5

```
model <- Teams %>%
  filter(yearID >= 1962) %>%
  mutate(BB = BB/G, singles = (H-X2B-X3B-HR)/G, doubles = X2B/G,
         triples = X3B/G, HR=HR/G, R=R/G) %>%
  lm(R ~ BB + singles + doubles + triples + HR, data = .)
model
```

```
##
## Call:
## lm(formula = R ~ BB + singles + doubles + triples + HR, data = .)
##
## Coefficients:
## (Intercept)           BB      singles      doubles      triples           HR
##     -2.5058       0.3658       0.4887       0.7056       1.2398       1.4876
```

```
model$coefficients / model$coefficients[3]
```

```
## (Intercept)          BB     singles     doubles     triples          HR
##  -5.1270368   0.7485091   1.0000000   1.4438035   2.5367386   3.0436979
```

Answer: $0.75 \times BB + singles + 1.44 \times doubles + 2.54 \times triples + 3.04 \times HR$