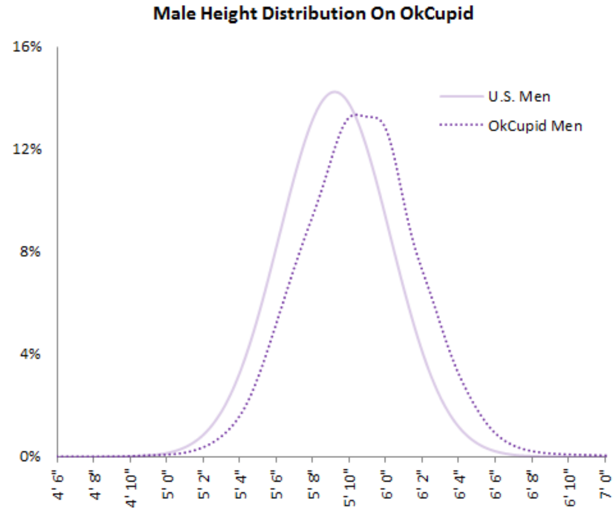
Week 4 Day 4

Human height

Problem 1 Consider the distribution of reported male height for users of the online dating site OkCupid as shown below. What observations can you make from this data graphic?



Problem 2 Assume that the distribution of heights of adult women is approximately normal with mean 162 cm and standard deviation 6 cm.

(a) What percentage of women are taller than 175 cm?

(b) Between what heights do the middle 95% of women fall?

(c) What percentage of women are shorter than 156 cm?

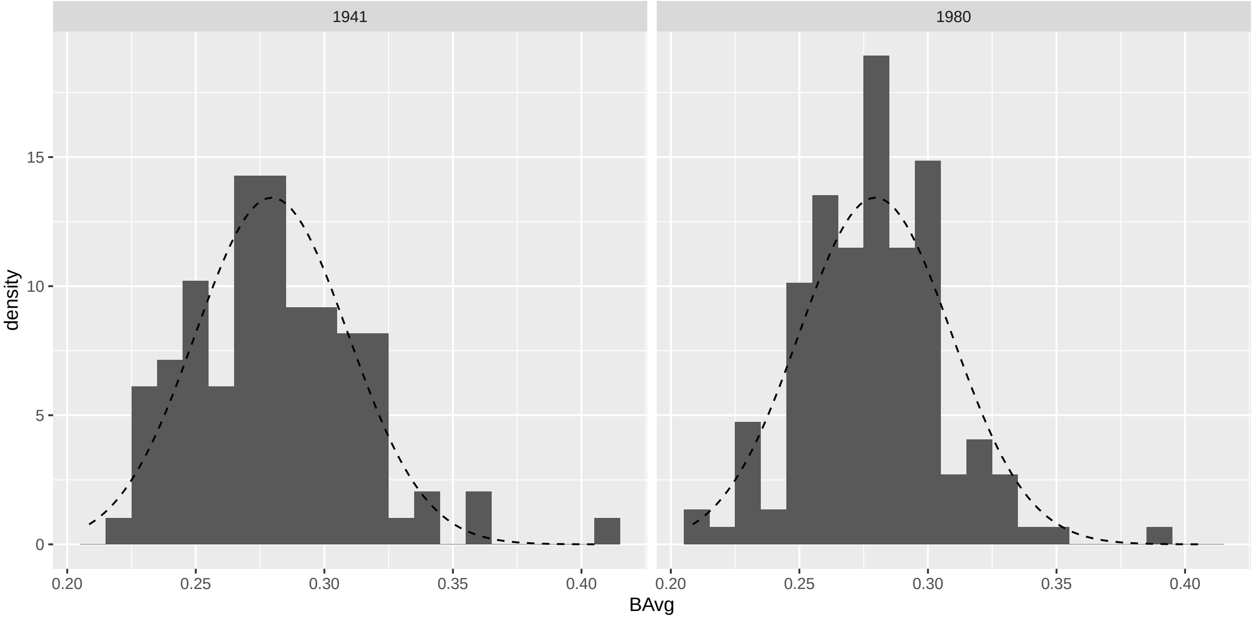
(d) A professor claims that about 51% of women are between 156cm and 165cm inches tall. Is this claim accurate?

MLB Batting Averages

[Adapted from Ben Baumer]

In 1941, Ted Williams of the Boston Red Sox hit .406, famously getting 6 hits in 8 at-bats on the last day of the season. No player in Major League Baseball has hit .400 since. Among the closest attempts was made by George Brett of the Kansas City Royals in 1980, when Brett hit .390. When viewed in relation to his peers, whose performance was more impressive?

playerID	yearID	BAvg
willite01	1941	0.4057018
brettge01	1980	0.3897550



Problem 1 Use the information below to calculate a z-score for both Williams in 1941 and Brett in 1980.

yearID	N	mean_BAvg	sd_BAvg
1941	98	0.2806367	0.0327903
1980	148	0.2788247	0.0275744

Problem 2 I have computed the Z-scores in R. Compare your answer to Problem 1, did you get them correct? Based on the following table, whose performance do you think was more remarkable in the context of his peers? Why? What assumptions are you making?

yearID	best_Z
1941	3.814093
1980	4.022945

Appendix: R code used in this tutorial

```r
# Problem introduction
## compute average
require(Lahman)
library(tidyverse)
library(kableExtra)
mlb <- Batting %>%
  mutate(BAvg = H / AB) %>%
  filter(yearID %in% c(1941, 1980) & AB > 400)

mlb %>%
  filter(BAvg > .36) %>%
  select(playerID, yearID, BAvg) %>%
  kbl(booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"), position = "center")

## display histogram
ggplot(data = mlb, aes(x = BAvg)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.01) +
  facet_wrap(~yearID) +
  stat_function(fun = dnorm, linetype = 2,
                args = list(mean = mean(mlb$BAvg), sd = sd(mlb$BAvg)))

## problem 1
mlb %>%
  group_by(yearID) %>%
  summarize(N = n(), mean_BAvg = mean(BAvg), sd_BAvg = sd(BAvg)) %>%
  kbl(booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"), position = "center")

## problem 2
mlb %>%
  group_by(yearID) %>%
  summarize(best_Z = (max(BAvg) - mean(BAvg)) / sd(BAvg)) %>%
  kbl(booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"), position = "center")
```