Agenda

1. Homework #4 problem 2.5
2. Hypothesis Testing Continued
3. Central Limit Theorem
4. Normal Distribution

Distribution of Height on OkCupid Consider the distribution of reported male height for users of the online dating site OkCupid.

1. What observations can you make from this data graphic?

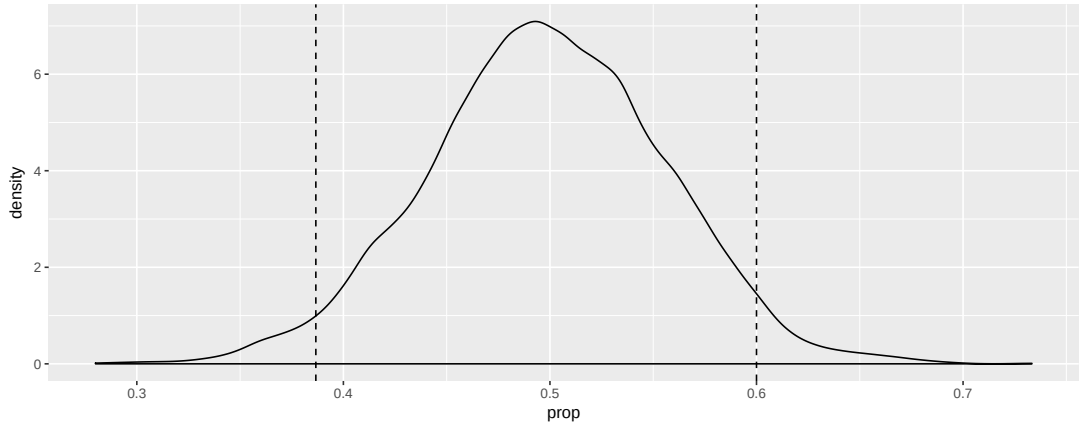
Millenials and Marriage In the national debate on same-sex marriage, it is commonly stated that half of all Americans favor same-sex marriage. In 2014, Pew Research conducted a poll of millenials (Americans born after 1980) and found that 66% answered "yes" when asked: "Do you favor same-sex marriage?" The poll was a random sample of 75 millenials. Does this poll provide convincing evidence that the opinion of millenials is different from those of Americans at large?

1. Write out the *null hypothesis* and the *alternative hypothesis* that are being evaluated, using proper notation.

2. Explain how you could use cards, a coin, or a die to simulate the *null distribution*.

3. What is the value of the observed *test statistic*?

4. In the null distribution below, the dotted vertical lines cut of 2.5% of the distribution in each tail (5% total). Indicate with a solid vertical line the location of the observed test statistic, and shade the area under the curve corresponding to the p-value.



5. The p-value for this test is p = .0064. Using $\alpha = 0.05$, what is your decision regarding the viability of the null hypothesis?

6. Write *one* sentence to your grandpa summarizing what you've learned about the millenials and their opinions on same-sex marriage.

The Normal Distribution

- Central Limit Theorem:
  - The distribution of the sample mean (i.e. the sampling distribution of the mean) will be approximately normal for reasonably large n (at least 30)
  - Provides a mathematical approximation to the simulated null distributions with which we have been working. Consider the practical difficulties of simulating null distributions with a computer!
- Normal distribution has two parameters!

$$N(\mu, \sigma), \qquad \text{has density function } f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- $N(0, 1)$: Standard Normal distribution
- Consider $X \sim N(\mu, \sigma)$. What is the distribution of $Y = X - \mu$? What is the distribution of Y/σ?
- If $x \sim N(\mu, \sigma)$, then $z = \frac{x-\mu}{\sigma} \sim N(0, 1)$.

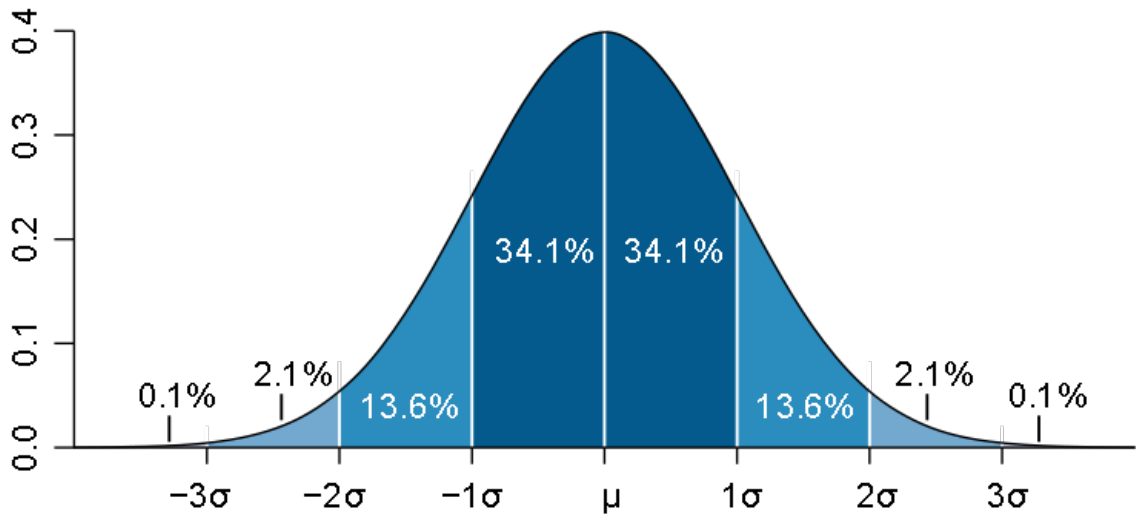
Example: MLB Batting Averages In 1941, Ted Williams of the Boston Red Sox hit .406, famously getting 6 hits in 8 at-bats on the last day of the season. No player in Major League Baseball has hit .400 since. Among the closest attempts was made by George Brett of the Kansas City Royals in 1980, when Brett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

```
library(Lahman)
library(mosaic)

mlb <- Batting %>%
  mutate(BAvg = H / AB) %>%
  filter((yearID == 1941 | yearID == 1980) & AB > 400)

mlb %>%
  filter(BAvg > .36) %>%
  select(playerID, yearID, BAvg)

##    playerID yearID      BAvg
## 1 willite01   1941 0.4057018
## 2 brettge01   1980 0.3897550
```

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

```
mlb %>%
  group_by(yearID) %>%
  summarize(N = n(), mean_BAvg = mean(BAvg), sd_BAvg = sd(BAvg))

## # A tibble: 2 x 4
##   yearID     N mean_BAvg    sd_BAvg
##    <int> <int>     <dbl>      <dbl>
## 1   1941    98 0.2806367 0.03279026
## 2   1980   148 0.2788247 0.02757441
```

2. Whose performance do you think was more remarkable in the context of his peers? Why? What assumptions are you making?



The Empirical Rule for Normal Distributions For any normally distributed variable:

- About 68% of the distribution is contained within 1 standard deviation of the mean.
- About 95% of the distribution is contained within 2 standard deviations of the mean.
- About 99.7% of the distribution is contained within 3 standard deviations of the mean.

Sample Calculations

1. What percentage of the distribution is less than 2 standard deviations, but above the mean?
   - By the rule, about 95% of the population is within two standard deviations of the mean. By symmetry, half of those are above the mean, and half below it. Thus, we estimate that about $95/2 = 47.5\%$ is less than 2 standard deviations above the mean.
   - From the picture, we can calculate the area as about $34.1\% + 13.6\% = 47.7\%$
2. Assume that the distribution of heights of adult women is approximately normal with mean 64 inches and standard deviation 2.5 inches.
   (a) What percentage of women are taller than 5'9"?

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

   (c) What percentage of women are shorter than 61.5 inches?

   (d) A professor claims that about 51% of women are between 61.5 and 65.25 inches tall. Is this claim accurate?