Agenda

1. Mid-semester Assessment Feedback
2. Randomness and Probability
3. Inference for a single proportion

Randomness

- Key Idea: if a process is *random*, then individual outcomes can't be predicted
- BUT: the distribution of outcomes, in the long run, is often quite regular and predictable.
  - Equally likely outcomes (e.g., drawing any card, rolling a die)
  - Unequally likely outcomes (e.g., drawing a face card vs. non-face card, men's heigths)
- The Law of Large Numbers guarantees that as more observations are collected, as $n \to \infty$, the proportion ($\hat{p}_n$) of occurances with a particular outcome converges to the probability (p) of that outcome.

Example: It's Unfair Of 50 students in a class, 15 are SDS majors and 35 are non-SDS majors. Four competitors will represent Smith at DataFest and are (allegedly) chosen by chance. The 4-person team turns out to be 3 SDS majors and 1 non-SDS major. The non-SDS majors cry foul! What is the chance that three or more of the four drawn will be SDS majors (assuming students are picked with replacement)?

- Possible outcomes: 0 SDS (none), 1 SDS, 2 SDS, 3 SDS, 4 SDS (all)
- Event A: 3 or more SDS majors end up on the team (combination of two outcomes: 3 SDS + 4 SDS)
- How would you simulate this random event (Event A) with cards?

```
library(mosaic)

major <- c(rep("SDS", 15), rep("non-SDS", 35))
the_class <- data.frame(major)

sim <- do(100) * filter(sample_n(the_class, 4, replace = TRUE), major == "SDS") %>%
  summarise(n = n()) %>%
  mutate(event_yes = ifelse(n > 2, 1, 0))

sim %>%
  summarise(mean(event_yes))

##   mean(event_yes)
## 1             0.1

true_prob <- dbinom(3, 4, .3) + dbinom(4, 4, .3)
```

Inference for a Single Proportion Consider the following problem: In a survey of a simple random sample of 123 people 77 say they prefer Coke over Pepsi. Then a point estimate for the proportion of people who prefer Coke over Pepsi is $\hat{p} = 77/123 = 0.624$.
In order to make inferences about the unknown value of p, the true proportion of those in population who prefer Coke, we have to construct the sampling distribution of $\hat{p}$. The center, shape, and spread of the sampling distribution of the proportion will enable us to put the observed $\hat{p}$ in context, build confidence intervals, and conduct hypothesis tests.
There are at least three different ways to approximate the sampling distribution of $\hat{p}$:

1. *Simulation:* This is one of the central themes of this course. For example, to test the null hypothesis that $p_0 = 0.5$, we simulate many random draws from this distribution, and see where $\hat{p}$ lies in this simulated distribution.

```
n <- 123
p_0 <- 1/2
p_hat <- 77/123

library(mosaic)
library(oilabs)

outcomes <- data_frame(soda = c("Coke", "Pepsi"))

sim <- outcomes %>%
  rep_sample_n(size = n, replace = TRUE, reps = 10000) %>% #Null dist
  group_by(replicate) %>%
  summarize(N = n(), coke = sum(soda == "Coke")) %>%
  mutate(coke_pct = coke / N)

ggplot(data = sim, aes(x = coke_pct)) +
  geom_density()
```

   It is important to recognize that by drawing more and more samples, we get a more refined understanding of the sampling distribution, but it remains only an approximation.

   The p-value can be obtained using the `pdata` function, since the sampling distribution comes from simulated data in our workspace.

```
2 * pdata(~coke_pct, q = p_hat, data = sim, lower.tail = FALSE)

## [1] 0.0032
```

2. *Normal Approximation:* Since the binomial distribution can be cumbersome to work with, and because under very mild conditions it is approximately normal, statisticians most often use a normal distribution to approximate the sampling distribution for a single proportion. If the number of individuals who prefer Coke follows a binomial distribution with parameters n and p, then it follows from probability theory that the standard deviation of the proportion who prefer Coke is $SE_p = \sqrt{\frac{p(1-p)}{n}}$. Thus, we can use this formula to estimate the standard error of the sampling distribution and conduct our hypothesis test (replacing p with p_0 in hypothesis testing, or with $\hat{p}$ for confidence intervals).

```
se_p <- sqrt(p_0 * (1-p_0) / n)

ggplot(sim, aes(x = coke_pct)) +
  stat_function(fun = dnorm, args = c(mean = p_0, sd = se_p))
```

The p-value can be obtained using the `pnorm` function, since the sampling distribution follows a normal distribution.

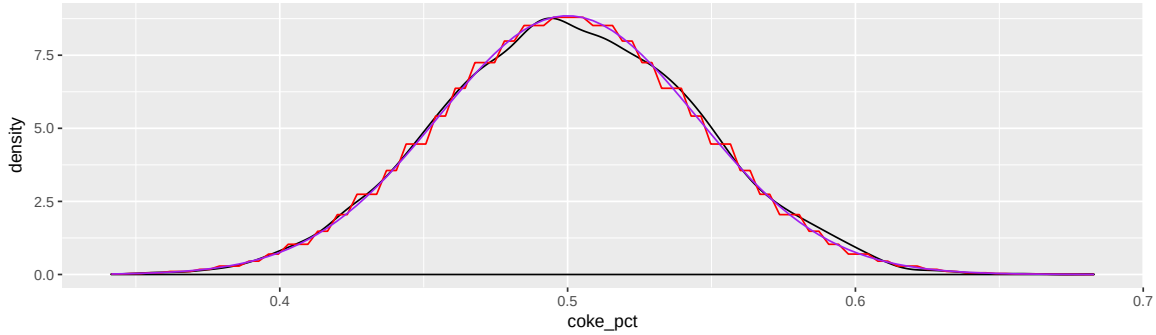
```
2 * pnorm(p_hat, mean = p_0, sd = se_p, lower.tail = FALSE)

## [1] 0.005187149
```

(Note: For historical reasons we often find the z-score for $\hat{p}$ when using the normal approximation). For a variety of reasons both historical and practical, the normal approximation is the method you are most likely to see in your future work, and thus it will be the focus of our attention here.

Note that the p-value is slightly different in each case (since our approximation of the sampling distribution is different in each case), but it is very close, and in each case we will easily reject the null hypothesis that $p = 0.5$ at the 5% level.

What Can Go Wrong? Most of the time, the sampling distribution for a proportion will be quite normal. In the previous example, the fit was excellent.

```
ggplot(data = sim, aes(x = coke_pct)) +
  geom_density() +
  stat_function(fun = dbinom_p, args = c(size = n, prob = p_0), col = "red") +
  stat_function(fun = dnorm, args = c(mean = p_0, sd = se_p), col = "purple")
```



However, if $np < 10$ or $n(1-p) < 10$, then the normal approximation is likely not sufficiently good. Suppose that we had only sampled 12 people instead of 123.

Exercise: Batting Averages, redux Previously, we considered Ted Williams' batting average of .406 in 1941, which is unmatched in 72 years and counting. In 1994, Tony Gwynn of the San Diego Padres hit .394, but a strike by the player's union shortened the season after only 116 games. Thus, Gwynn accumulated 165 hits in 419 at-bats, whereas Williams had 185 hits in 456 at-bats. Let's assume that Gwynn had an unknown, fixed true batting average of p in 1994.

1. The league average batting average in 1994 was .277. Use the normal approximation to test – at the 5 level – the hypothesis that Gwynn was a league-average hitter. Do you reject or fail to reject? (*Hint: If you don't have a computer to compute the p-value, find the z-score and approximate using the 68-95-99.7 Rule*)

2. Use the normal approximation to find a 95 confidence interval for Gwynn's true batting average p. (*Hint: Be sure to use $\hat{p}$ when computing the standard error! (see page 124)*)

3. Does the confidence interval that you found contain the hypothesized proportion of .277? Does it contain .400?

4. A sportswriter claims that Gwynn does not deserve to be mentioned in the same breath as Williams, because Williams hit .400, but Gwynn did not. Does your analysis refute or support this claim?