Agenda

1. Inference for a single proportion
2. Inference for a difference of two proportions

Inference for a Single Proportion Recall from last time, we discussed three ways to construct the *null distribution* of the sample proportion $\hat{p}$.

```r
n <- 123
p_0 <- 1/2
p_hat <- 77/123
require(mosaic)
```

1. Simulation: Use the computer to *simulate* the null distribution.

   (a) Assumptions: independence

   (b) Pros: few assumptions, no math, can simulate very complex situations with a little programming skill

   (c) Cons: requires computer (impossible before 1970), does not always return the same answer

```r
library(oilabs)
sim <- data_frame(soda = c("Coke", "Pepsi")) %>%
  rep_sample_n(size = n, replace = TRUE, reps = 10000) %>%
  group_by(replicate) %>%
  summarize(N = n(), coke = sum(soda == "Coke")) %>%
  mutate(coke_pct = coke / N)
qplot(data = sim, x = coke_pct, geom = "density")
```

2. Probability Theory: Use mathematics to *compute* the null distribution.

   (a) Assumptions: independence, probability model

   (b) Pros: gives exact sampling distribution

   (c) Cons: only the simplest situations can be solved in closed form, may be hard to detect mistakes

Last time we proved that if $X \sim Bernoulli(p)$ is a random variable giving the preference of any given person, and $Y = X_1 + X_2 + \cdots + X_n \sim Binomial(n, p)$ is a random variable giving the number of people among n who prefer Coke, and $\hat{p} = Y/n$ is a random variable giving the proportion of people among n who prefer Coke, then:

$$\begin{aligned} \mathbb{E}[X] &= p & Var[X] &= p(1-p) \\ \mathbb{E}[Y] &= np & Var[Y] &= np(1-p) \\ \mathbb{E}[\hat{p}] &= p & Var[\hat{p}] &= \frac{p(1-p)}{n} \end{aligned}$$

The binomial distribution is a well-known discrete probability distribution, but its density function is cumbersome to work with, and so it is hard to compute binomial probabilities by hand. It is, of course, easy to do with R.

```r
plotDist("binom", params = list(size = n, prob = p_0))
```

The binomial distribution depends on two parameters: the sample size n and the proportion p. We won't talk much more about the binomial distribution in this class (to learn more, take MTH 153 or MTH 246).

3. Normal Approximation: Use statistical theory to *approximate* the null distribution.

   (a) Assumptions: independence, normality, $np > 10$ and $n(1-p) > 10$

   (b) Pros: uses familiar normal distribution, approximation is usually pretty good, possible to compute without computers (kind of)

   (c) Cons: requires more assumptions, not exact

   Since the binomial distribution can be cumbersome to work with, and because under very mild conditions it is approximately normal, scientists often use a normal distribution to approximate the null distribution for a single proportion. Consider the random variable X defined above, and note that the sample proportion $\hat{p}$ can be thought of as the mean of a random sample of n observations of X. The Central Limit Theorem implies that:

$$\mathbb{E}[\bar{X}] = \mu_X = p\,, \qquad Var[\bar{X}] = \frac{\sigma_X^2}{n} = \frac{p(1-p)}{n}\,.$$

   In particular, this implies that $SE_{\hat{p}} = \sqrt{Var[\bar{X}]} = \sqrt{\frac{p(1-p)}{n}}$. Thus, we can use this formula for the standard error to *approximate* the null distribution.

```
se_p0 <- sqrt(p_0 * (1-p_0) / n)
plotDist("norm", params = list(mean = p_0, sd = se_p0))
```

   For a variety of reasons both historical and practical, the normal approximation is the method you are mostly 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 null 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.

1. Simulation: The p-value can be obtained using the `pdata` function, since the null distribution comes from simulated **data** in our workspace.

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

## [1] 0.0038
```

2. Probability Theory: The p-value can be obtained using the `pbinom` function, since the null distribution follows a **binom**ial distribution.

```
2 * pbinom(p_hat * n, size = n, prob = p_0, lower.tail = FALSE)

## [1] 0.003731446
```

3. Normal Approximation: The p-value can be obtained using the `pnorm` function, since the null distribution follows a **norm**al distribution.

```
2 * pnorm(p_hat, mean = p_0, sd = se_p0, lower.tail = FALSE)

## [1] 0.005187149
```

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

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 75 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 Empirical 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 125)*)

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?

Difference of two proportions In many cases we will also want to make inferences about the difference between two proportions. Continuing the line of reasoning that we pursued last time, let X be a binomial random variable that gives the number of hits that Williams will accrue in n_1 at-bats if his true batting average is p_1, and let Y be another binomial random variable that gives the number of hits that Gwynn will accrue in n_2 at-bats if his true batting average is p_2. Then we can define a new random variable

$$Z = \frac{X}{n_1} - \frac{Y}{n_2}$$

that gives the difference in their respective batting averages. Using linearity of expectation, we can compute the expected value of the difference:

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{X}{n_1} - \frac{Y}{n_2}\right] = \frac{1}{n_1} \cdot \mathbb{E}[X] - \frac{1}{n_2} \cdot \mathbb{E}[Y] = \frac{1}{n_1} \cdot n_1 p_1 - \frac{1}{n_2} \cdot n_2 p_2 = p_1 - p_2$$

and the variance:

$$\begin{aligned} Var[Z] = Var\left[\frac{X}{n_1} - \frac{Y}{n_2}\right] &= \frac{1}{n_1^2} \cdot Var[X] + \frac{1}{n_2^2} \cdot Var[Y] \\ &= \frac{1}{n_1^2} \cdot n_1 \cdot p_1(1-p_1) + \frac{1}{n_2^2} \cdot n_2 \cdot p_2(1-p_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \end{aligned}$$

Just as before, this proves that the standard error is $SE_Z = SE_{\widehat{p_1 - p_2}} = \sqrt{SE^2_{\hat{p}_1} + SE^2_{\hat{p}_2}}$. Once again, we'll typically use the normal approximation to the null distribution.

1. Using the normal approximation again, test the hypothesis that Williams and Gwynn had the same true batting averages in 1941 and 1994, respectively.

2. Since we are testing the hyopthesis that $p_1 = p_2$, it is more appropriate to use the *pooled* estimate of the standard error (see page 133). Perform this test.

3. Discuss the extent to which you think the performances of Williams and Gwynn were importantly different.