Agenda

1. Inference for a single proportion

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) %>%
  group_by(replicate) %>%
  summarize(N = n(), coke = sum(soda == "Coke")) %>%
  mutate(coke_pct = coke / N)
qplot(data = sim, 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] 0.0052
```

   (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

2. Probability Theory: If we assume that each person's preference is independent, and the that the true proportion is fixed, then the number of individuals who will say that they prefer Coke is a random variable that follows a *binomial* distribution.

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

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

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

## [1] 0.003731446
```

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). .

(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

3. 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 elementary probability theory that the standard deviation of the proportion who prefer Coke is $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Thus, we can use this formula for the standard error to estimate the sampling distribution and conduct our hypothesis test.

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

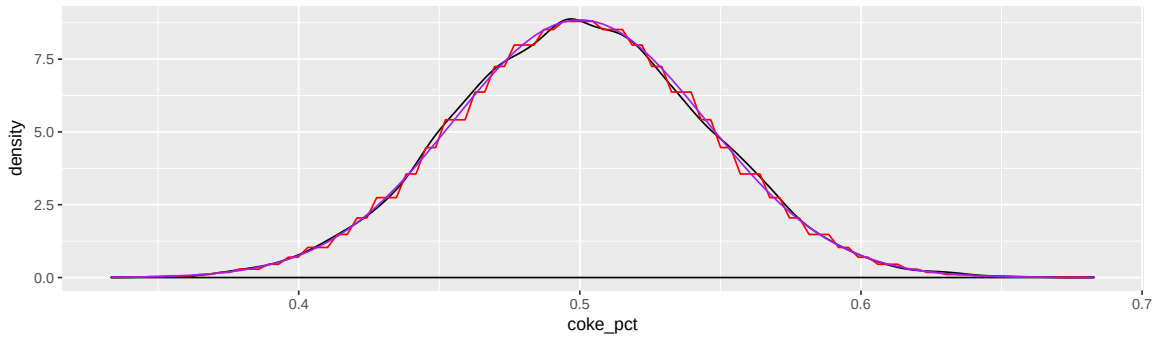
For a variety of reasons both historical and practical, the normal approximation is the method you are mostly likely to see in your future work, and thus it will be the focus of our attention here.

   (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

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.

```
qplot(data = sim, 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.