**Key points**

* Instead of using an urn model where each poll is a random draw from the same distribution of voters, we instead define a model using an urn that contains poll results from all possible pollsters.
* We assume the expected value of this model is the actual spread d=2p−1.
* Our new standard error σ now factors in pollster-to-pollster variability. It can no longer be calculated from p or d and is an unknown parameter.
* The central limit theorem still works to estimate the sample average of many polls X1,...,XN because the average of the sum of many random variables is a normally distributed random variable with expected value d and standard error σ/N−−√.
* We can estimate the unobserved σ as the sample standard deviation, which is calculated with the sd function.

**Code**

Note that to compute the exact 95% confidence interval, we would use qnorm(.975) instead of 1.96.

# collect last result before the election for each pollster

one\_poll\_per\_pollster <- polls %>% group\_by(pollster) %>%

filter(enddate == max(enddate)) %>% # keep latest poll

ungroup()

# histogram of spread estimates

one\_poll\_per\_pollster %>%

ggplot(aes(spread)) + geom\_histogram(binwidth = 0.01)

# construct 95% confidence interval

results <- one\_poll\_per\_pollster %>%

summarize(avg = mean(spread), se = sd(spread)/sqrt(length(spread))) %>%

mutate(start = avg - 1.96\*se, end = avg + 1.96\*se)

round(results\*100, 1)