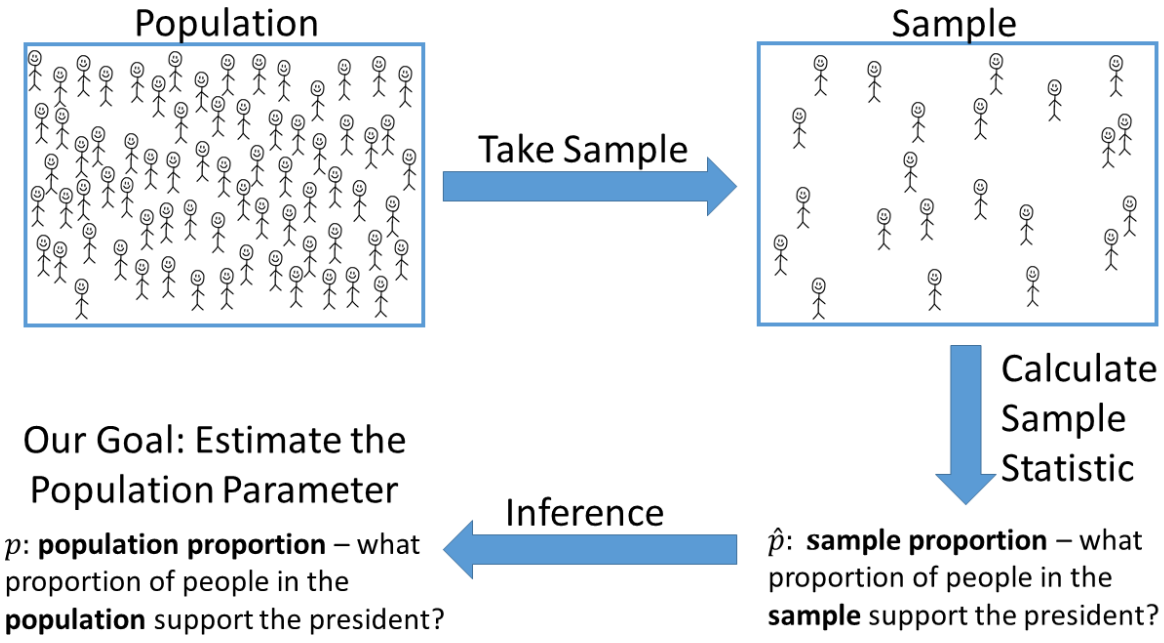
# Hypothesis Tests and Confidence Intervals for Population Proportions

## Overview

In the 2016 presidential elections, President Trump won 46.1% of the popular vote (46.1% of people who cast a vote for president voted for Trump).

The Pew Research Center recently conducted a nationally representative poll where they asked respondents, *"Do you approve or disapprove of the way Donald Trump is handling his job as President?"* 23 people in the sample did not answer this question; among the 4145 participants who answered this question, 1700 approved.

**What can the data from this sample tell us about the population?**




### Types of inference:

1. **Point Estimate:** Our best guess at the value of the population parameter. Often, this is the value of the sample statistic.
2. **Confidence Intervals:** Informally, a range of values for the population parameter that are consistent with the data in the sample.
3. **Hypothesis Tests:** Do the data offer evidence that the the population parameter is different from a specified value?

# Statistical Model

Let $X$ = Number of people in our sample who approve of the way Donald Trump is handling his job as President.

We might use the model

$X \sim \text{Binomial}(4145, p)$

We know the sample size, $n = 4145$.

We want to estimate $p$, the proportion of the population who approve of the way Donald Trump is handling his job as President.

## Checking Conditions (Think Before You Compute)

- Is the sample **representative** of the population? (always required)
  - Were we careful in **randomly selecting** the people to include in the sample?
  - Are there any forms of **bias** in the sampling? In this case, 23 people did not respond (there is some non-response). Does this matter?
- Are the the people in the sample **independent** from each other? (required for all methods we will use in this class)
  - Does knowing one person's response change the information we have about another person's response?
  - Closely related to the question of randomization above.
- Is the variable we have recorded a **count of the number of people in our sample with a certain characteristic**? (required to use the Binomial model)

**"All models are wrong but some are useful." - George Box**

The question is: **Are the assumptions close enough to being satisfied that we think we can say something useful with these data and this model?**

Proceed cautiously and acknowledge the limitations of your analysis.

# Computations in R

```
library(readr)
library(dplyr)
library(mosaic)

pew_poll <- read_csv("http://www.evanlray.com/stat140_s2018/lecture/20180308_p_htest/pew_poll.csv")
head(pew_poll)
```

```
## # A tibble: 6 x 2
##   respondent_id approve_president
##           <int>             <chr>
## 1          3068        Disapprove
## 2          2925        Disapprove
## 3          1834        Disapprove
## 4           508           Approve
## 5          2016           Approve
## 6          1890           Approve
```

**Method 1: If we have a data frame with each individual's response**

```
binom.test(~ approve_president, data = pew_poll, p = 0.461)
```

```
##
##
##
## data:  pew_poll$approve_president  [with success = Approve]
## number of successes = 1700, number of trials = 4145, p-value =
## 4.696e-11
## alternative hypothesis: true probability of success is not equal to 0.461
## 95 percent confidence interval:
##  0.3951063 0.4252859
## sample estimates:
## probability of success
##              0.4101327
```

The argument `p = 0.461` specifies the value of the population in the null hypothesis of a hypothesis test (see below).

**Method 2: If we have a count of how many people responded each way**

```
tally(~ approve_president, data = pew_poll)
```

```
## approve_president
##    Approve Disapprove       <NA>
##       1700       2445         23
```

```
binom.test(x = 1700, n = 4145, p = 0.461)
```

```
##
##
##
## data:  1700 out of 4145
## number of successes = 1700, number of trials = 4145, p-value =
## 4.696e-11
## alternative hypothesis: true probability of success is not equal to 0.461
## 95 percent confidence interval:
##  0.3951063 0.4252859
## sample estimates:
## probability of success
##              0.4101327
```

# Set-Up and Interpretation of Inference

## 1. Point Estimate

Your best guess of the population parameter.

## 2. Confidence Interval

A range of values for the population parameter that are consistent with the data in the sample.

## 3. Hypothesis Test

Do the data offer evidence that the the population parameter is different from a specified value?

46.1% of voters voted for Trump. Do the data offer evidence that the proportion of the American public who now approve of the job he is doing as president different from this value?

**(a) State null and alternative hypotheses**

**(b) Write down the p-value and its interpretation**

**(c) Compare the p-value to a significance level and draw a conclusion**

A small p-value casts doubt on the null model/hypothesis used to perform the calculation.

- A p-value of .10 or less is generally considered to be some evidence against the null model/hypothesis.
- A p-value of .05 or less is generally considered to be fairly strong evidence against the null model/hypothesis.
- A p-value of .01 or less is generally considered to be very strong evidence against the null model/hypothesis.
- A p-value of .001 or less is generally considered to be extremely strong evidence against the null model/hypothesis.