FinalReport

Bijah LaFollette

2023-12-04

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

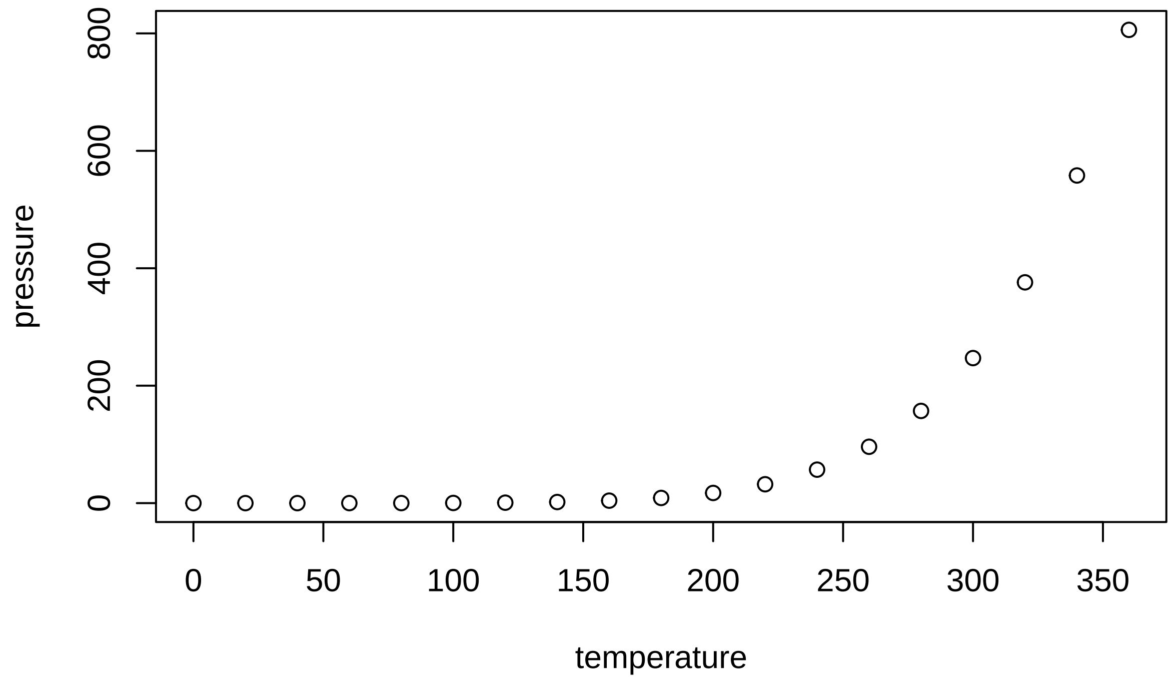
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

#INTRODUCTION: For my project I have decided to examine data related to elections from 1998 to 2020 to see what variables have caused polls to be inaccurate. Going into the project,I assumed that polls were, generally speaking, biased towards democrats given the inaccurate polls during the 2016 election. More broadly, as a result of the shock of the 2016 election, public confidence in polls and election predictions generally has plummeted. This has meant that some voters completely discopnect themselves from viewing political news before elections because they don't trust polls and the media. To understand why polls can be inaccurate, I looked at two main explanatory variables: sample size and the way the polls were conducted.

```
library(infer)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
Rawpolls <- read_csv("raw-polls.csv")
```

```
## Rows: 9559 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (13): race, location, type_simple, type_detail, pollster, polldate, cand...
## dbl (16): poll_id, question_id, race_id, year, pollster_rating_id, samplesiz...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

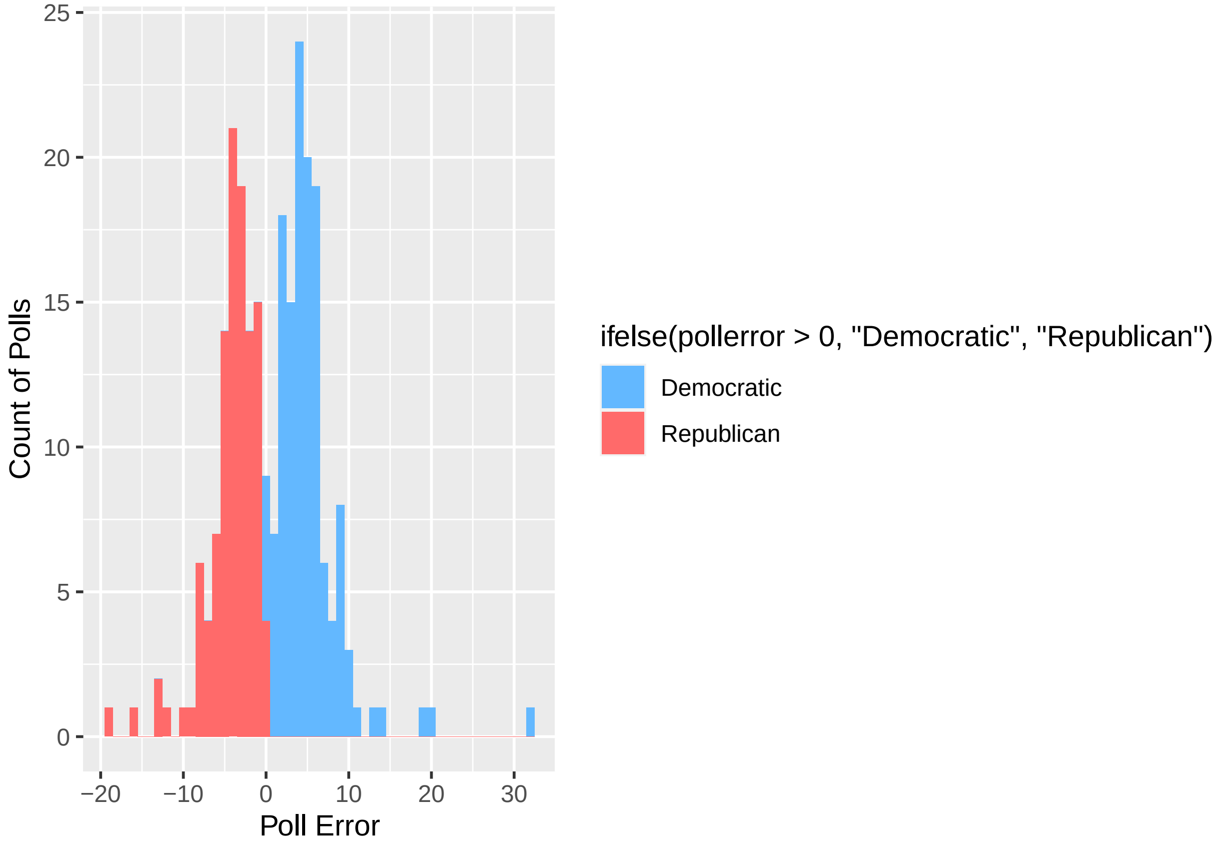
```
mediabias <- read_csv("pollster-ratings.csv")
```

```
## Rows: 453 Columns: 23
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (12): Pollster, NCPP / AAPOR / Roper, Live Caller With Cellphones, Metho...
## dbl (11): Pollster Rating ID, # of Polls, Predictive    Plus-Minus, Simple A...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
Rawpolls <- Rawpolls |>
    mutate(bias_direction = if_else(bias >= 0, "left_bias", "right_bias"),
    partisanerror_size = case_when(
      bias > 0 & bias < 3 ~ "leftlean",
      bias >= 3 & bias <= 7 ~ "moderateleft",
      bias > 7 ~ "hardleft",
      bias > -3 & bias < 0 ~ "leanright",
      bias <= -3 & bias >= -7 ~ "moderateright",
      bias < -7 ~ "hardright",  # Added missing operator (<) to compare bias with -7
      TRUE ~ NA_character_  # Add a catch-all condition if none of the above criteria match
    )
    )

    Rawpolls <- Rawpolls |>
  mutate(
    bias_direction = if_else(bias >= 0, "left_bias", "right_bias"),
    partisanerror_size = case_when(
      bias > 0 & bias < 3 ~ "leftlean",
      bias >= 3 & bias <= 7 ~ "moderateleft_bias",
      bias > 7 ~ "hardbias_left",
      bias > -3 & bias < 0 ~ "leanright",
      bias <= -3 & bias >= -7 ~ "moderateright_bias",
      bias < -7 ~ "hardbias_right",  # Added missing operator (<) to compare bias with -7
      TRUE ~ NA_character_  # Add a catch-all condition if none of the above criteria match
    )
  )

      Rawpolls <- Rawpolls |>
      mutate(swingstate = if_else(margin_poll < 6 & margin_poll > -6, "swingrace", "blowout"),
             sampletype = case_when(samplesize < 500 ~ "smallsample",
                                    samplesize >= 500 & samplesize <= 750 ~ "mediumsample",
                                    samplesize > 750 ~ "Large_sample"))

biastotal <- Rawpolls |>
  filter(year %in% c(2016, 2018, 2020, 2022)) %>%
  group_by(pollster, bias_direction) %>%
  summarize(pollerror = mean(bias)) %>%
  drop_na(pollerror) %>%
  ggplot(mapping = aes(x = pollerror, fill = ifelse(pollerror > 0, "Democratic", "Republican"))) +
  geom_histogram(binwidth = 1) +
  scale_fill_manual(values = c("steelblue1", "indianred1")) +
  labs(x = "Poll Error", y = "Count of Polls")
```

```
## `summarise()` has grouped output by 'pollster'. You can override using the
## `.groups` argument.
```

```
biastotal
```



As we can see, polling errors over the last 20 years are fairly evenly distrubited between democrats and republicans. But why are there errors in the first place? My research question: Is there a relationship between the way polls are conducted and the accuracy of the polls and how does sample size affect accuracy? My hypothesis was that polls that had larger sample sized and those that were not conducted by live phone would prove to be the most accurate. This is because of "cancellation fear" that many Trump supporters, particularly females, had in 2016 that made them afraid to admit support for candidate Trump to a live person. If I observe that polls with a larger sample size had lower mean errors, then this would support my hypothesis. If I observe that polls that were conducted by Interactive Voice Response (IVR) (an automated telephone system where a person speaks to a robot who recitates pre-recorded messages or texts-to-speech) polls were more accurate than live or mail polls, this would also support my hypothesis.

This study will be informative for the purposes of identifying what types of polls are most accurate and if large sample sizes are actually necessary. Polls with large sample sizes can be more expensive to conduct; If we cannot prove that larger sample sizes increase accuracy then we can show polling organizations that they are not necessary.

##Data and Research Method

I analyze data from elections from 1998-2020, which includes data from congressional and federal elections. I gained this data from 538's database on elections. My data comes from two different datasets – one that focuses on different election results / poll error / and another that focuses on the methodologies of the pollster. The datasets include variables such as year, race, location, sample size, the margin predicted by the polls and the actual margin in the election. The ‘error' column is the ” Absolute value of the difference between the actual and polled result. This is calculated as `abs(margin_poll - margin_actual) "biasis calculated only for races in which the top two finishers were a Democrat and a Republican. It is calculated as`margin_poll - margin_actual‘. Positive values indicate a Democratic bias (the Democrat did better in the poll than the election). Negative values indicate a Republican bias.”

#Data and Research Method The first data anslysis topic I covered was comparing errors for polls with a small sample size and polls with a large sample size. Small sample type is defined as those samples where the sample size was less than 500 (25th percentile of sample size) and the large_sample was defined as those samples where the sample size was greater than 850 (75th percentile).

```r
samples.error <- Rawpolls |>
  group_by(sampletype) |>
  summarize(avg.error = mean(error)) |>
  pivot_wider(names_from = sampletype, values_from = avg.error)  |>
  mutate(ATE = Large_sample - smallsample) |>
  select(c(Large_sample, smallsample, ATE)) |>
  knitr::kable(col.names = c("Large Sample", "Small Sample", "ATE"), digits = 3)
samples.error
```

Large Sample	Small Sample	ATE
4.708	7.409	-2.701

In this first plot, we can see that polls with a large sample size averaged a 4.7 error while polls with a small sample size averaged a 7.409 error. This means that large sample size polls were 2.7 points more accurate. As we can see the treatment effect of having a larger sample size reduced polling error by 2.7 points for this dataset. To see if this is statistically significant or just do to random chance I ran a p value test under the null hypothesis that there should be no difference in polling error for small vs large sample sizes.

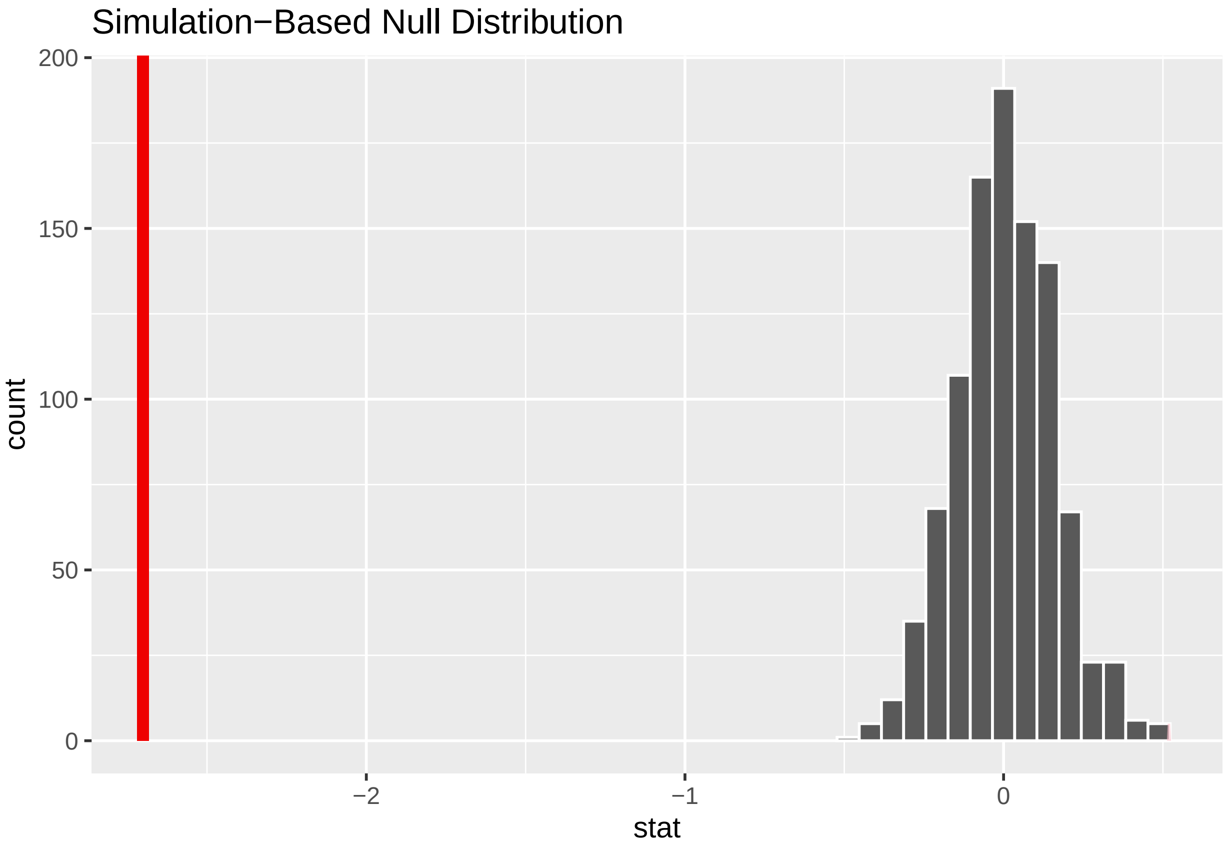
```r
Rawpolls
```

```
## # A tibble: 9,559 x 33
##    poll_id question_id race_id  year race       location type_simple type_detail
##      <dbl>       <dbl>   <dbl> <dbl> <chr>      <chr>    <chr>       <chr>
##  1   54373       87909    1455  1998 1998_Gov-~ NY       Gov-G       Gov-G
##  2   26255       87926    1456  1998 1998_Gov-~ OH       Gov-G       Gov-G
##  3   26026       31266    1736  1998 1998_Sen-~ NV       Sen-G       Sen-G
##  4   26013       31253    1738  1998 1998_Sen-~ NY       Sen-G       Sen-G
##  5   63632      117103    1738  1998 1998_Sen-~ NY       Sen-G       Sen-G
##  6   26255       31495    1741  1998 1998_Sen-~ OH       Sen-G       Sen-G
##  7   64053      117875    1966  1998 1998_Hous~ ID-1     House-G     House-G
##  8   64053      117876    1967  1998 1998_Hous~ ID-2     House-G     House-G
##  9   28268       33546    8661  1998 1998_Hous~ US       House-G     House-G
## 10   28267       33545    8661  1998 1998_Hous~ US       House-G     House-G
## # i 9,549 more rows
## # i 25 more variables: pollster <chr>, pollster_rating_id <dbl>,
## #   polldate <chr>, samplesize <dbl>, cand1_name <chr>, cand1_party <chr>,
## #   cand1_pct <dbl>, cand2_name <chr>, cand2_party <chr>, cand2_pct <dbl>,
## #   cand3_pct <dbl>, margin_poll <dbl>, electiondate <chr>, cand1_actual <dbl>,
## #   cand2_actual <dbl>, margin_actual <dbl>, error <dbl>, bias <dbl>,
## #   rightcall <dbl>, comment <chr>, partisan <chr>, bias_direction <chr>, ...
```

```
ate1 <- Rawpolls |>
  filter(sampletype %in% c("Large_sample", "smallsample")) |>
  specify(error ~ sampletype) |>
  calculate(stat = "diff in means", order = c("Large_sample", "smallsample"))
ate1
```

```
## Response: error (numeric)
## Explanatory: sampletype (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 -2.70
```

```
ate_rawpolls_dust <- Rawpolls |>
  filter(sampletype %in% c("Large_sample", "smallsample")) |>
  specify(error ~ sampletype) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in means", order = c("Large_sample", "smallsample"))

ate_rawpolls_dust |> visualize() +
  shade_p_value(obs_stat = ate1, direction = "both")
```



```
ate1_pvalue <- ate_rawpolls_dust |>
  get_p_value(obs_stat = ate_rawpolls_dust, direction = "both")
```

```
## Warning: The first row and first column value of the given ‘obs_stat‘ will be
## used.

## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of ‘reps‘ chosen in the ‘generate()‘ step.
## See ‘?get_p_value()‘ for more information.
```

```
ate1_pvalue
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

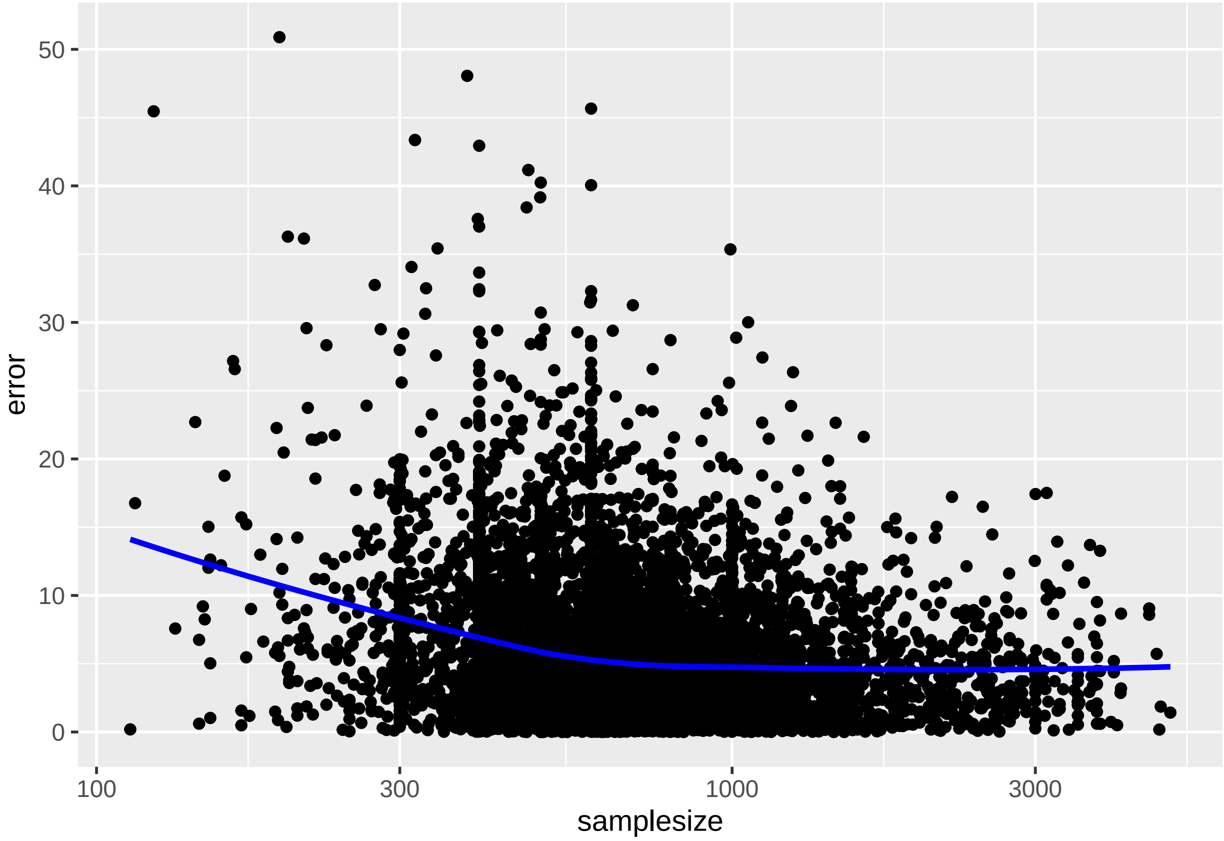
This shows that under the null hypothesis where the sample size has no effect on polling error, the chance that we would observe a result where large sample sizes were 2.7 points more accurate is, highly, unlikely. The P value is 0, and it does not even fall on the distribution of possible outcomes. In fact, based on the distributon above, it is only likely that we could observe a difference of 0.5 on either direction.

As a result, we can reject the null hypothesis. this means that the 2.7 ATE calculated above is statistically signifgant. Further evidence of sample sizes effect on poll accuracy comes from a regression I did later.

#RESULTS

```
Rawpolls |>
  filter(samplesize < 5000) |>
  ggplot(mapping = aes(x = samplesize, y = error)) +
  geom_point() +  # Scatterplot of samplesize vs. error
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  scale_x_log10()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
Rawpolls |>
  filter(samplesize > 1000)
```

```
## # A tibble: 1,568 x 33
##    poll_id question_id race_id  year race       location type_simple type_detail
##      <dbl>       <dbl>   <dbl> <dbl> <chr>      <chr>    <chr>       <chr>
##  1   26208       31448    1723  1998 1998_Sen-~ CT       Sen-G       Sen-G
##  2   54342       87878    1450  1998 1998_Gov-~ MN       Gov-G       Gov-G
##  3   26050       87927    1456  1998 1998_Gov-~ OH       Gov-G       Gov-G
##  4   26050       31290    1741  1998 1998_Sen-~ OH       Sen-G       Sen-G
##  5   54292       87828    1443  1998 1998_Gov-~ IL       Gov-G       Gov-G
##  6   26083       31323    1728  1998 1998_Sen-~ IL       Sen-G       Sen-G
##  7   54215       87751    1436  1998 1998_Gov-~ CA       Gov-G       Gov-G
##  8   54247       87783    1439  1998 1998_Gov-~ FL       Gov-G       Gov-G
##  9   54259       87795    1440  1998 1998_Gov-~ GA       Gov-G       Gov-G
## 10   54293       87829    1443  1998 1998_Gov-~ IL       Gov-G       Gov-G
## # i 1,558 more rows
## # i 25 more variables: pollster <chr>, pollster_rating_id <dbl>,
## #   polldate <chr>, samplesize <dbl>, cand1_name <chr>, cand1_party <chr>,
## #   cand1_pct <dbl>, cand2_name <chr>, cand2_party <chr>, cand2_pct <dbl>,
## #   cand3_pct <dbl>, margin_poll <dbl>, electiondate <chr>, cand1_actual <dbl>,
## #   cand2_actual <dbl>, margin_actual <dbl>, error <dbl>, bias <dbl>,
## #   rightcall <dbl>, comment <chr>, partisan <chr>, bias_direction <chr>, ...
```

This visualization shows that the optimal number of particpants is around 750. After that, there are decreasing returns to increasing poll sample size, as the slope of the line begins to flatten. Past a 1,000 poll sample size (which is 16.4% % of polls in the data), there is very little improvement in reducing poll error. This suggests that extremely large polls can be a waste of time and resources and should be discontinued for future elections. Instead new ideas about how to improve polling could be better focused by thinking about how polls are conducted.

Moving on to my discussion of the way in which the polls were conducted, the following graph summarizes mean errors for 17 different polling methods. The graph also contains error bars which, at the bottom, represent the average error - a standard deviation and the top bar which represents the average error plus a standard deviation. This gives viewers a sense of how spread out the errors were for each polling method.

```
mediabias_summary <- mediabias %>%
  group_by(Methodology) %>%
  summarize(avg.error = mean(`Simple Average Error`),
            sd.error = sd(`Simple Average Error`),
            diff = avg.error - mean(`Simple Average Error`))

  mediabias.plot <- ggplot(mediabias_summary, aes(x = Methodology, y = avg.error)) +
    geom_bar(stat = "identity", fill = "skyblue", alpha = 0.7) +
    geom_errorbar(aes(ymin = avg.error - sd.error, ymax = avg.error + sd.error),
                  width = 0.3, position = position_dodge(0.9)) +
    labs(x = "Methodology", y = "Average Error", title = "Average Error by Methodology") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

##Multi regression

```
mediabias
```

```
## # A tibble: 453 x 23
##    Pollster             `Pollster Rating ID` `# of Polls` `NCPP / AAPOR / Roper`
##    <chr>                               <dbl>        <dbl> <chr>
##  1 Monmouth University                   215          108 yes
##  2 Selzer & Co.                          304           48 yes
##  3 ABC News/The Washin~                    3           73 yes
##  4 Siena College/The N~                  448           59 yes
##  5 Field Research Corp~                   94           25 yes
##  6 Marquette Universit~                  195           12 yes
##  7 Muhlenberg College                    219           29 yes
##  8 Marist College                        183          183 yes
##  9 Data Orbital                           73            9 yes
## 10 National Journal                      224           12 yes
## # i 443 more rows
## # i 19 more variables: `Live Caller With Cellphones` <chr>, Methodology <chr>,
## #   ‘Banned by 538‘ <chr>, ‘Predictive    Plus-Minus‘ <dbl>, ‘538 Grade‘ <chr>,
## #   ‘Mean-Reverted Bias‘ <chr>, ‘Races Called Correctly‘ <chr>,
## #   ‘Misses Outside MOE‘ <chr>, ‘Simple Average Error‘ <dbl>,
## #   ‘Simple Expected Error‘ <dbl>, ‘Simple Plus-Minus‘ <dbl>,
## #   ‘Advanced Plus-Minus‘ <dbl>, ‘Mean-Reverted Advanced Plus Minus‘ <dbl>, ...
```

```
model <- lm(data = mediabias, `Simple Average Error` ~ `Methodology`)
model
```

```
##
## Call:
## lm(formula = ‘Simple Average Error‘ ~ Methodology, data = mediabias)
##
## Coefficients:
##                      (Intercept)              MethodologyIVR/Live
##                           7.4167                          -0.3588
##         MethodologyIVR/Live/Text            MethodologyIVR/Online
##                          -2.5667                          -1.0167
##       MethodologyIVR/Online/Live  MethodologyIVR/Online/Live/Text
##                          -1.3267                          -2.4667
##       MethodologyIVR/Online/Text              MethodologyIVR/Text
##                          -2.0167                          -3.1667
##              MethodologyLandline                  MethodologyLive
##                           5.8167                          -1.6205
##                 MethodologyLive*             MethodologyLive/Text
##                          -0.2734                          -1.1167
##                  MethodologyMail                MethodologyOnline
##                          -2.3167                          -1.5167
##           MethodologyOnline/Live      MethodologyOnline/Live/Text
##                          -0.9957                          -1.1767
##           MethodologyOnline/Text
##                           3.9833
```

#CONCLUSION In conclusion, the data shows that polls, historically, do not necessarily benefit one party systemically more than the other.