

Topic 5. Visualization and Univariate Graphics

Josh Clinton

Today's Agenda

- ▶ Introduction to data visualization!
- ▶ Guiding principles
- ▶ The glorious world of `ggplot`
- ▶ Bargraphs and histograms

Motivation: Communicating Data is Essential

- ▶ Data does not exist in a vacuum – it is always interpreted in relationship to something.
- ▶ All data-science should be question driven! What is the question you are asking?
- ▶ What is the answer that your visualization is providing?
- ▶ Does your visualization communicate the relationship cleanly and accurately?
- ▶ Humans infer causality (much too quickly!).

CHALLENGER EXAMPLE

Your visuals must tell an accurate story

- ▶ Tables and graphs are essential for visualization.
- ▶ Visualizations must be stand-alone (if possible).
- ▶ Visualizations must be well-labeled!

NOTE: Rule of thumb: show it to someone without explanation. If they are confused, re-do!

Dimensions of Visualization

You have several “dimensions” to use when presenting information

- ▶ Horizontal (x-axis) location
- ▶ Vertical (y-axis) location
- ▶ Size of data points
- ▶ Shape of data points
- ▶ Color

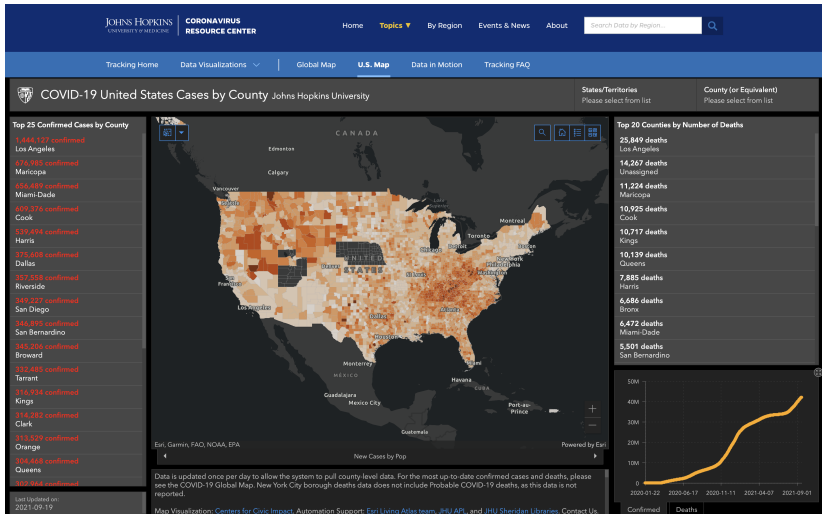
Dimensions of Visualization

You have several “dimensions” to use when presenting information

- ▶ Horizontal (x-axis) location
- ▶ Vertical (y-axis) location
- ▶ Size of data points
- ▶ Shape of data points
- ▶ Color
- ▶ Map each variable to at most one dimension.
- ▶ Be intuitive – don't assign small numbers large dots, etc.
- ▶ Don't be misleading.

MORE EXAMPLES

So is this a good visualization of the pandemic in the US?



<https://coronavirus.jhu.edu/us-map>

Visualization using 'ggplot

- ▶ Everything in an R visualization can be controlled.
- ▶ Graphs themselves are an object that can be saved and altered.
- ▶ Start with a blank “canvass” and then you add the visuals.
- ▶ Actually, you start with the question you want the visualization to answer.

Visualization First Steps

First graph is usually a summary of the data: what does it look like?

- ▶ Central tendency? (Where is most data located?)
- ▶ Variation? (Range? Dispersion? Skew?)

Application: 2020 Election

Data We Are Using



American Association for Public Opinion Research

[My AAPOR](#) [Membership](#) [Publications/Media](#) [Conference/Events](#) [Standards/Ethics](#) [Education](#)

[Home](#) —> [About Us](#) —> [Leadership](#) —> [Committees and Taskforces](#)

Committees and Taskforces

[Return to committee list](#)

Task Force on 2020 Pre-Election Polling

AAPOR Members

[Log in](#) to view email addresses for all committee members.

This committee of survey research and election polling experts reviews and gathers information on the 2020 pre-election polls to evaluate the accuracy of 2020 pre-election polling for both the primaries and the general election on the presidential race and other races.

Members

Joshua D. Clinton - Chair
Term Expires December 2021

Loading Polling Data

```
library(tidyverse)  
load(file="data/Pres2020.PV.Rdata")
```

What do we have here. . .

```
glimpse(Pres2020.PV)
```

```
## Rows: 528
## Columns: 16
## $ poll.id      <dbl> 1942, 1941, 1940, 1939, 1938, 1937, 1936,
## $ Geography    <chr> "NAT", "NAT", "NAT", "NAT", "NAT", "NAT",
## $ Poll         <chr> "Economist/YouGov", "Research Co.", "Ipsos
## $ StartDate    <chr> "10/31/2020", "10/31/2020", "10/29/2020",
## $ EndDate      <chr> "11/2/2020", "11/2/2020", "11/2/2020", "1
## $ DaysinField  <dbl> 3, 3, 5, 1, 1, 3, 3, 3, 4, 4, 2, 5, 5, 14
## $ MoE          <dbl> NA, 3.10, 3.70, 1.70, 3.20, NA, 1.00, NA,
## $ Mode         <chr> "Online", "Online", "Online", "Online", N
## $ SampleSize   <dbl> 1363, 974, 914, 5174, 1008, 1360, 799401,
## $ Biden        <dbl> 53, 53, 52, 52, 48, 53, 52, 53, 52, 52, 4
## $ Trump        <dbl> 43, 44, 45, 46, 42, 43, 46, 41, 41, 42, 4
## $ DemCertVote  <dbl> 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 5
## $ RepCertVote  <dbl> 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 4
## $ Winner       <chr> "Dem", "Dem", "Dem", "Dem", "Dem", "Dem",
## $ Funded       <chr> "Economist", "Research Co.", "Reuters",
## $ Conducted    <chr> "YouGov", "Research Co.", "Ipsos", "Swaya
```

What is our question?

How did public polling on the 2020 presidential election vary over the course of the election?

So the relationship of interest how *polling results* vary over *time*

But what do we mean by *"polling results"*?

- ▶ National Popular Vote vs. State specific support?
- ▶ Support for Biden and Trump? Difference in support between Biden and Trump?

And also, What do we mean by *time*?

- ▶ Day of the year? Proximity to Election Day?
- ▶ Results by day? week? month?

Start by defining some variables we need

```
Pres2020.PV <- Pres2020.PV %>%  
  mutate(margin = Biden - Trump)
```

```
summary(Pres2020.PV$margin)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	-6.000	6.000	8.000	8.021	10.000	17.000

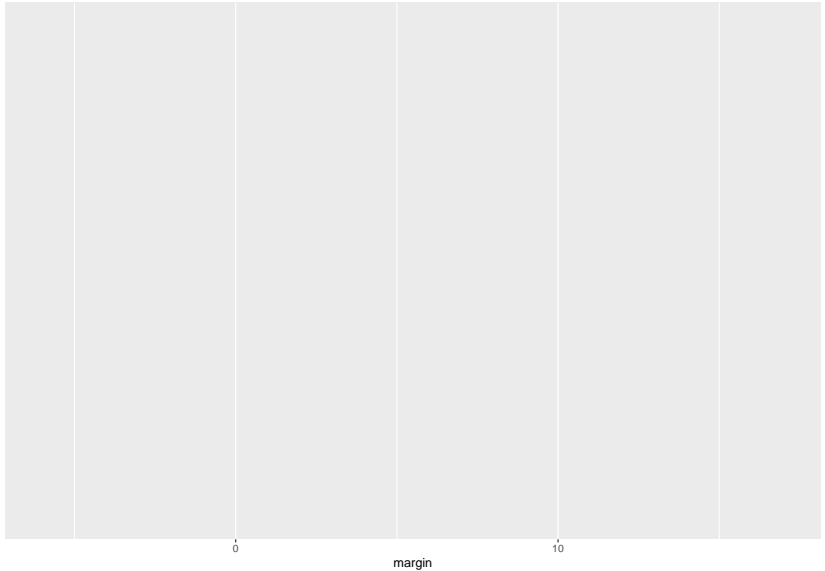
Visualizing margin

```
g <- Pres2020.PV %>%  
  ggplot(aes(x = margin))
```

- ▶ plots can be objects
- ▶ data defines the dataframe being used
- ▶ aesthetics define the variables being used

The Canvass

g



What do we want to convey about margin

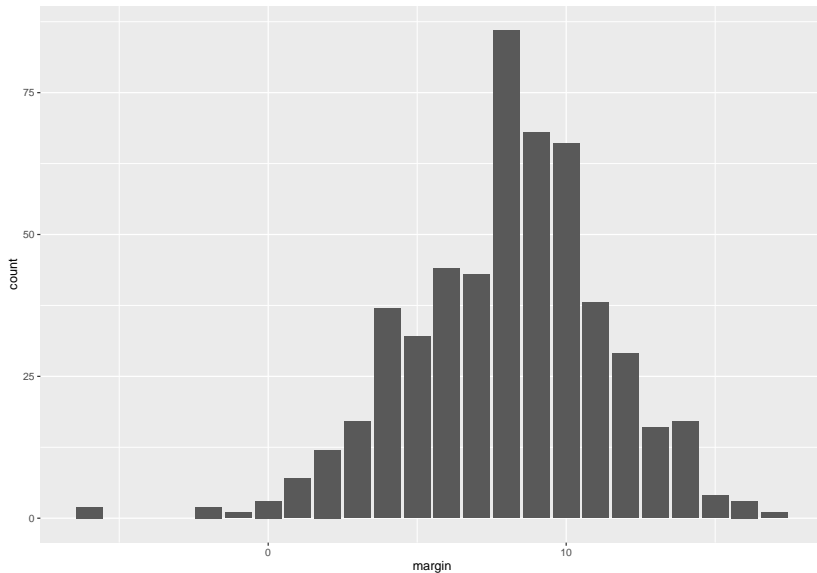
We only have 1 variable (x-axis), but 2 options depending on whether categorical/discrete or continuous!

- ▶ `geom_barplot` - discrete
- ▶ `geom_histogram` - continuous

NOTE: No pie-graphs!

Barplot

```
g + geom_bar()
```



Barplot

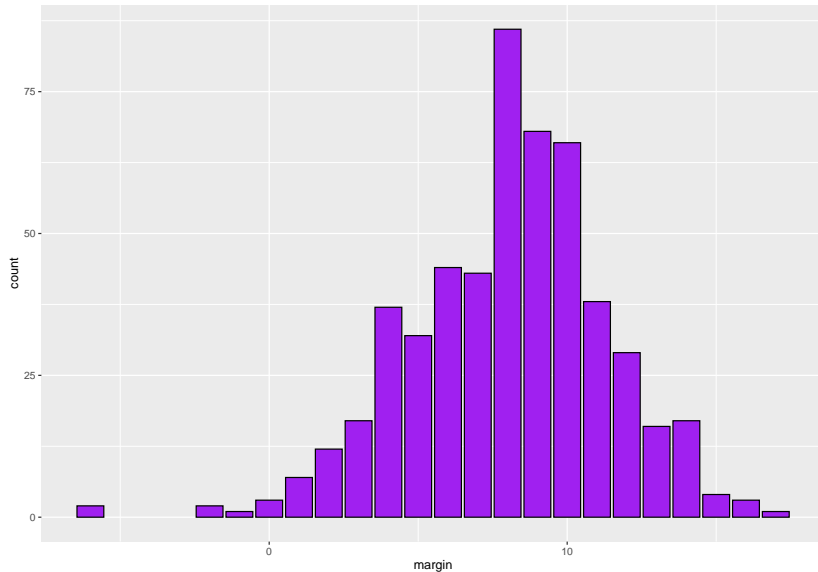
Used for discrete variables - one “bar” for each value

```
g + geom_bar(fill = "purple", color = "black")
```

- ▶ fill is the color of the bars
- ▶ color is the border of the bars

Barplot

```
g + geom_bar(fill = "purple", color = "black")
```



Adding labs

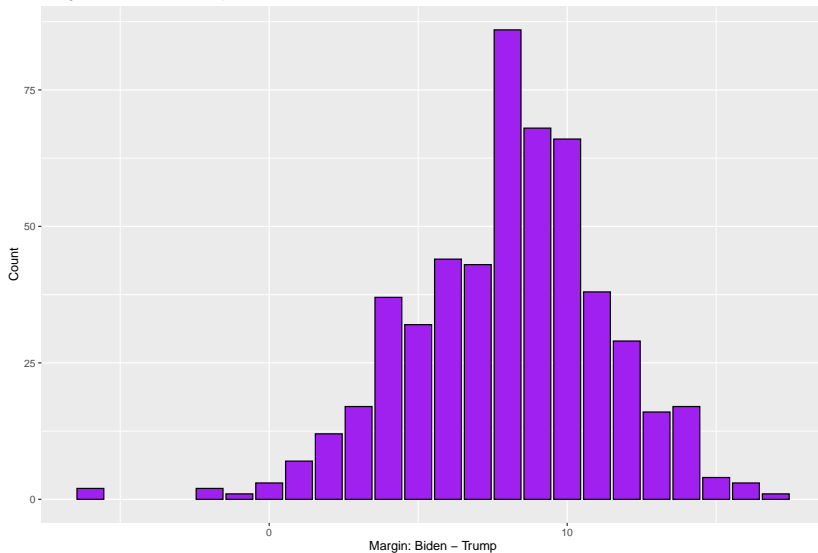
```
g <- g + geom_bar(fill = "purple", color = "black") +  
  labs(title = "Margin in 2020 National Popular Vote Polls") +  
  labs(x = "Margin: Biden - Trump") +  
  labs(y = "Count")
```

- ▶ Note can add multiple layers at the same time
- ▶ Never label using variable names!

Adding labs

gg

Margin in 2020 National Popular Vote Polls



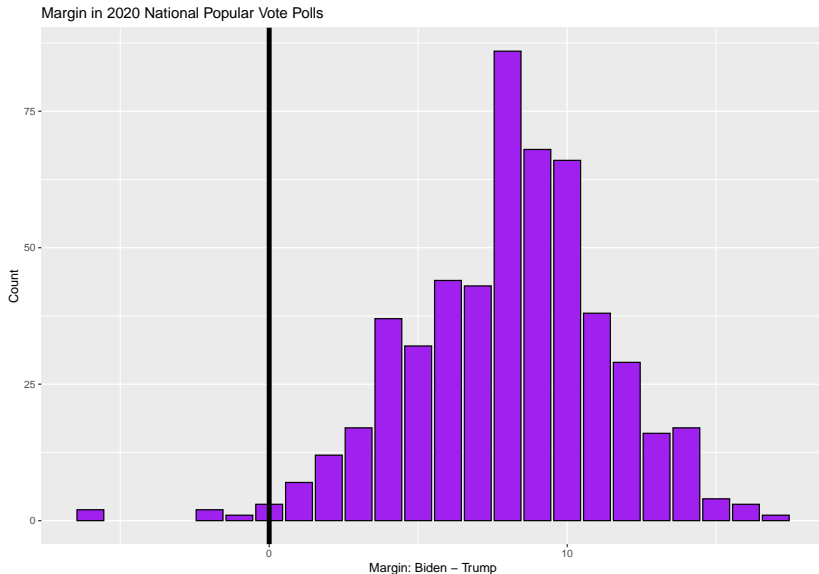
Adding a (linear) line to the plot

```
g + geom_vline(xintercept = 0, lwd=2)
```

- ▶ + `geom_abline(intercept = A , slope = B)`: add a line with y-intercept A and slope B
- ▶ + `geom_vline(xintercept = A)`: add a vertical line with x-intercept A
- ▶ + `geom_hline(yintercept = A)`: add a horizontal line with y-intercept A

Adding a line

```
g + geom_vline(xintercept = 0, lwd=2)
```



Saving graphs

- ▶ Can save manually using R-Studio (bad).
- ▶ Can save using a graphical device.

```
pdf(file="2020MarginBarplot.pdf")  
g  
dev.off()
```

Histogram

Used for continuous variables - divide values into “bins” and plot bins.

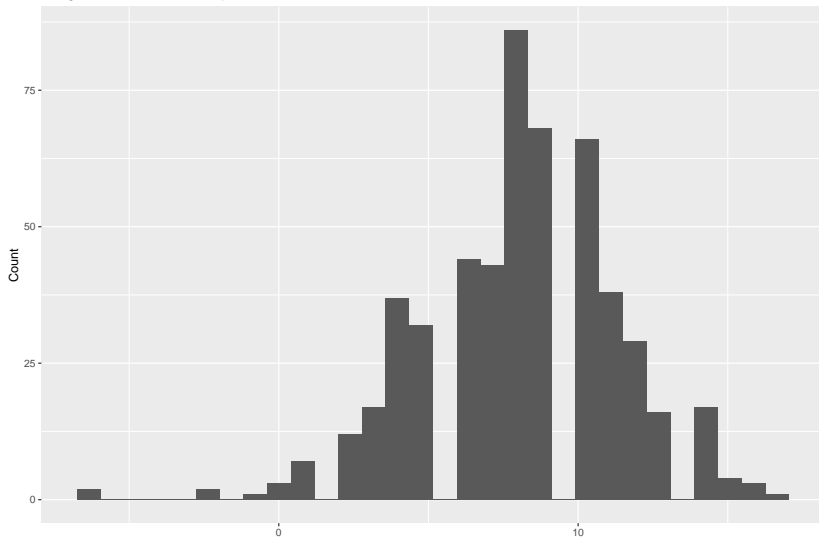
```
h <- Pres2020.PV %>%  
  ggplot(aes(x = margin)) +  
  labs(title = "Margin in 2020 National Popular Vote Polls") +  
  labs(x = "Margin: Biden - Trump") +  
  labs(y = "Count")
```

Take a look at default

```
h + geom_histogram()
```

`stat_bin()` using `bins = 30`. Pick better value with `binwi

Margin in 2020 National Popular Vote Polls



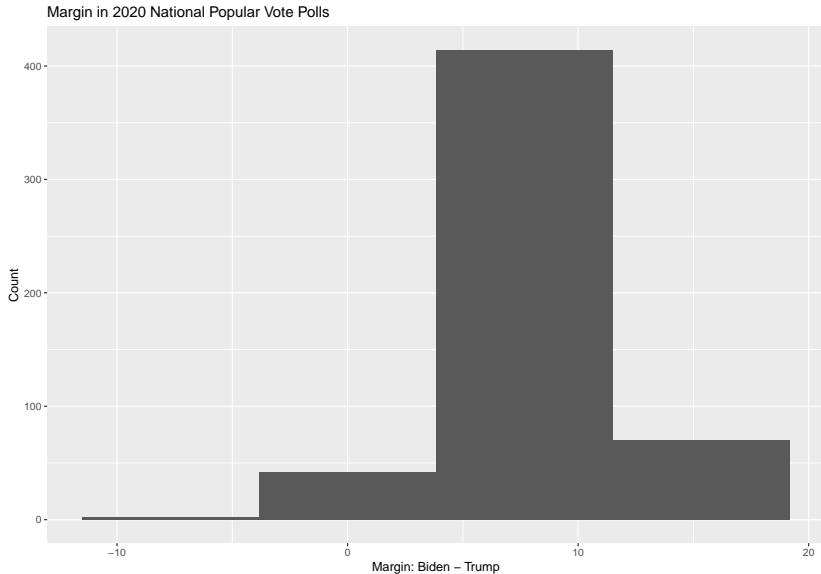
Let's fix..

```
h + geom_histogram(bins = 10)
```

- ▶ bins defines how many bars you want
- ▶ Default is the count of how many observations fall into each bin.
- ▶ Can also plot the density by including `aes(y = ..density..)`

Histogram (Count)

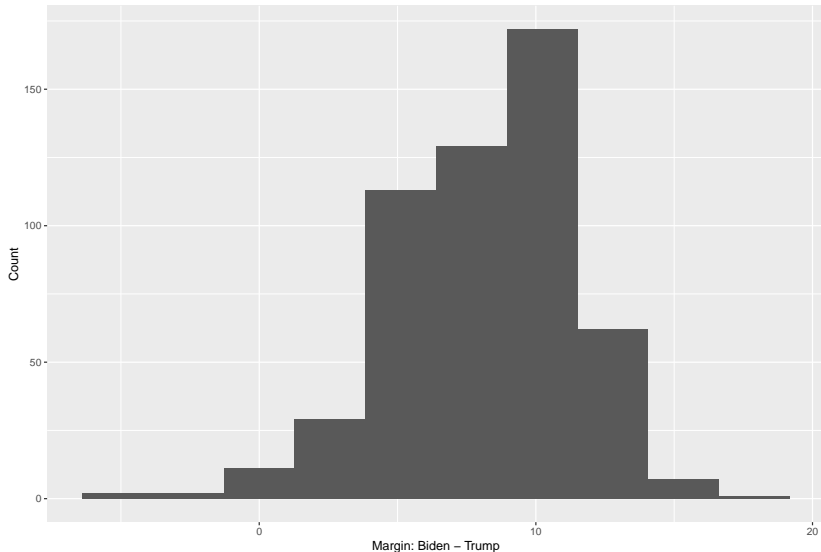
```
h + geom_histogram(bins = 4)
```



Histogram (Count)

```
h + geom_histogram(bins = 10)
```

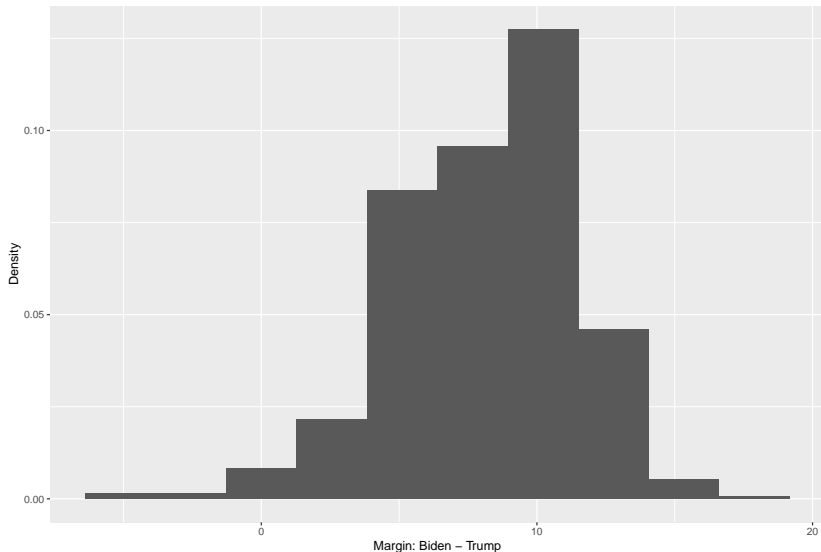
Margin in 2020 National Popular Vote Polls



Histogram (Density)

```
h + geom_histogram(bins = 10, aes(y = ..density..)) +  
  labs(y = "Density")
```

Margin in 2020 National Popular Vote Polls



Density? What is that?

- ▶ Density is not a proportion – it is the “area under the curve”. (Integrals!)
- ▶ So the proportion is: width \times height
- ▶ Primarily used when comparing distributions of different variables: densities always integrate/sum to 1.

Plotting Single Variables

- ▶ If discrete: `geom_bar`
- ▶ If continuous: `geom_histogram`
- ▶ If want to compare variation across variables (on same ggplot) - density.
- ▶ Always label!
- ▶ Next Steps: Try plotting the distribution of Trump and Biden to make sure you can do it.
- ▶ Next time: plotting bivariate relations!