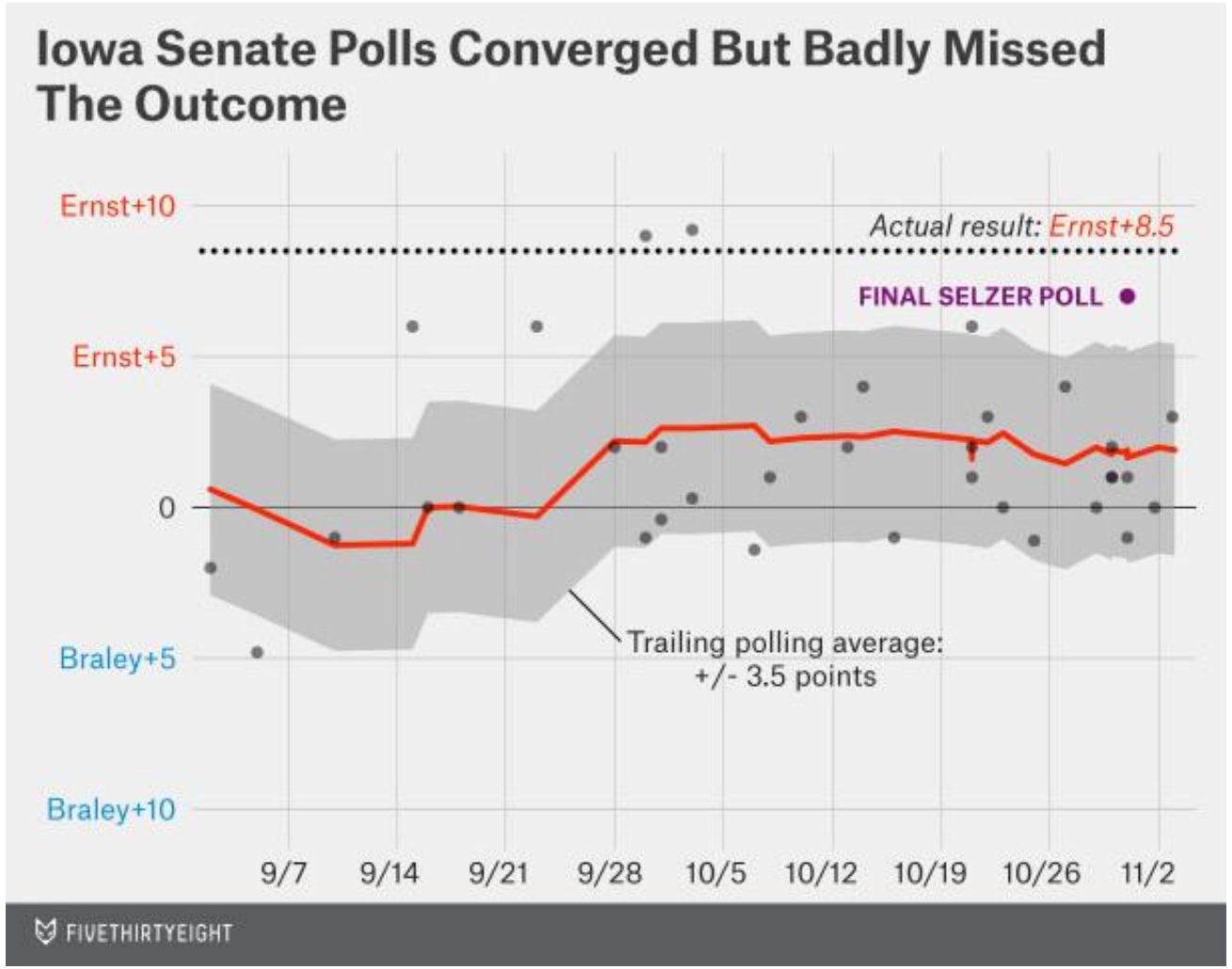
Thinking about Graphs

The Grammar of Graphics and SAS

Reconstructing a graph

From

http://fivethirtyeight.com/features/heres-proof-some-pollsters-are-putting-a-thumb-on-the-scale/



Questions toward reconstruction

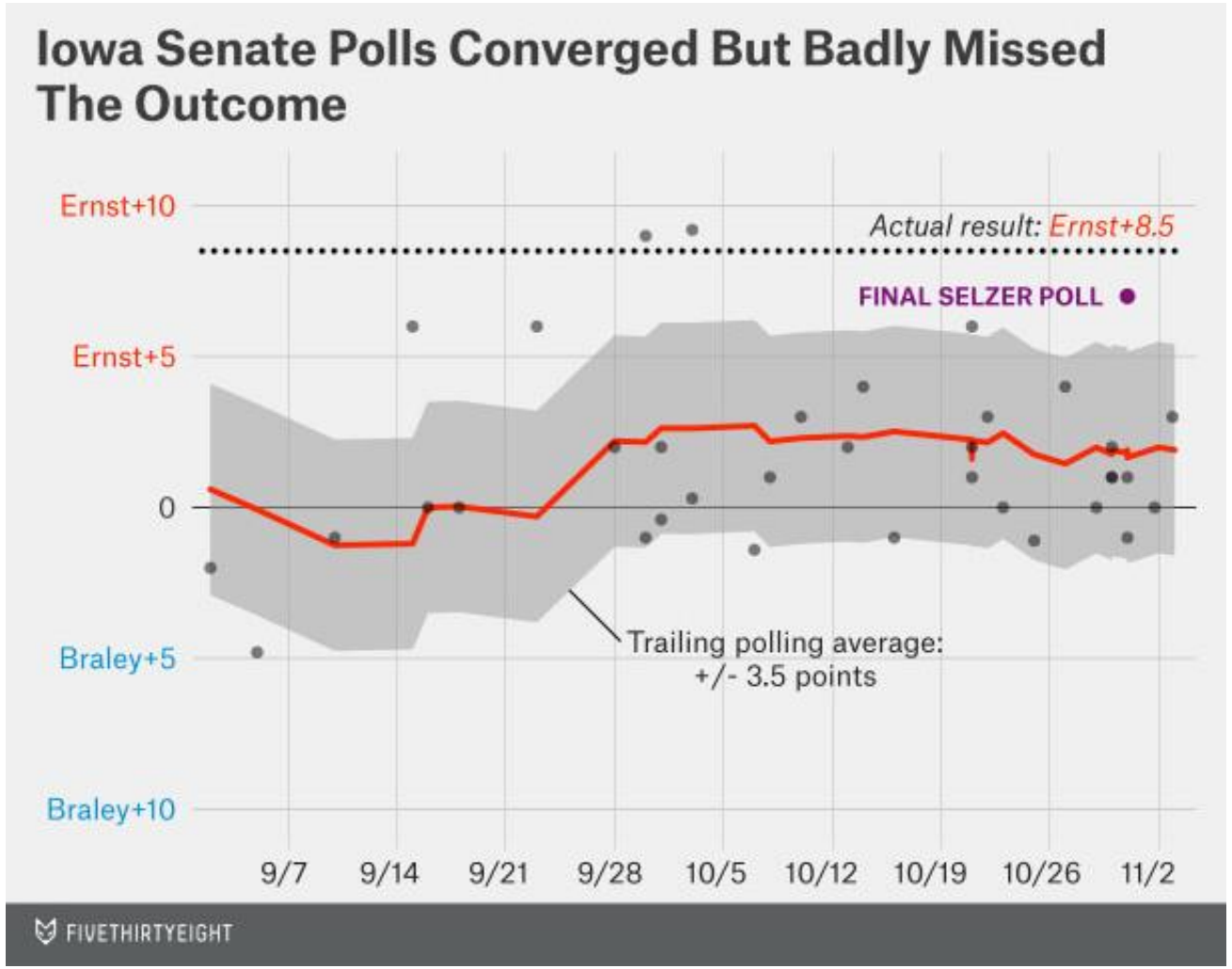
- What are the graphical elements?
- How are they related to data?
- How are they arranged on the screen/paper?
- How are they decorated?

Graphical elements

Points

Line(s)

Area



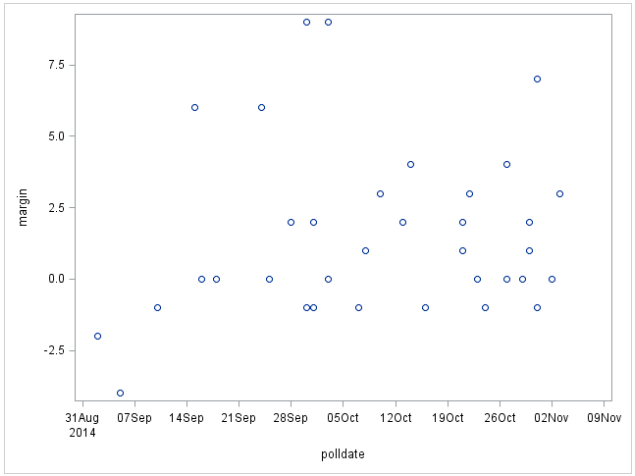
Relation to data

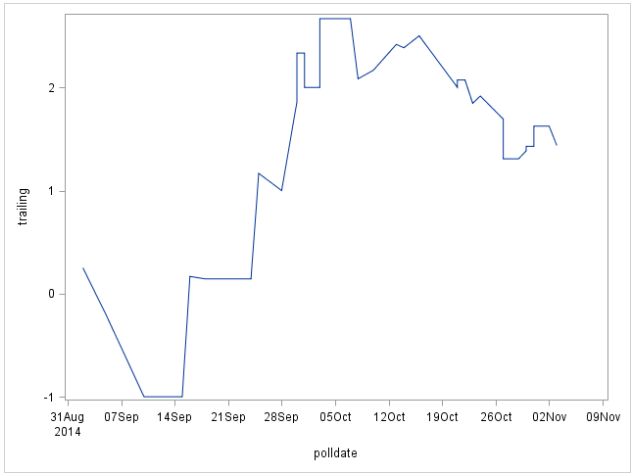
- Points: polling margins versus dates, essentially a scatter plot
- Lines:
  - Grid lines, some emphasized
  - Trailing margin is polling averages versus dates, connected (a.k.a. a line plot)
- Area: a fixed range around the trailing margin

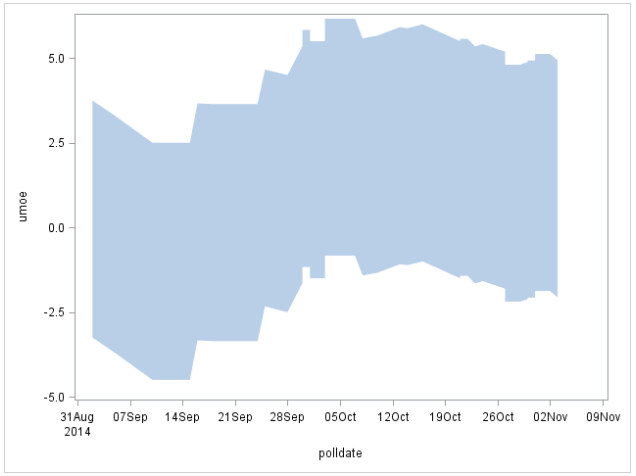
- Given the points, the lines and area can be calculated

Arrangement

- Think in layers, points on top of lines on top of area



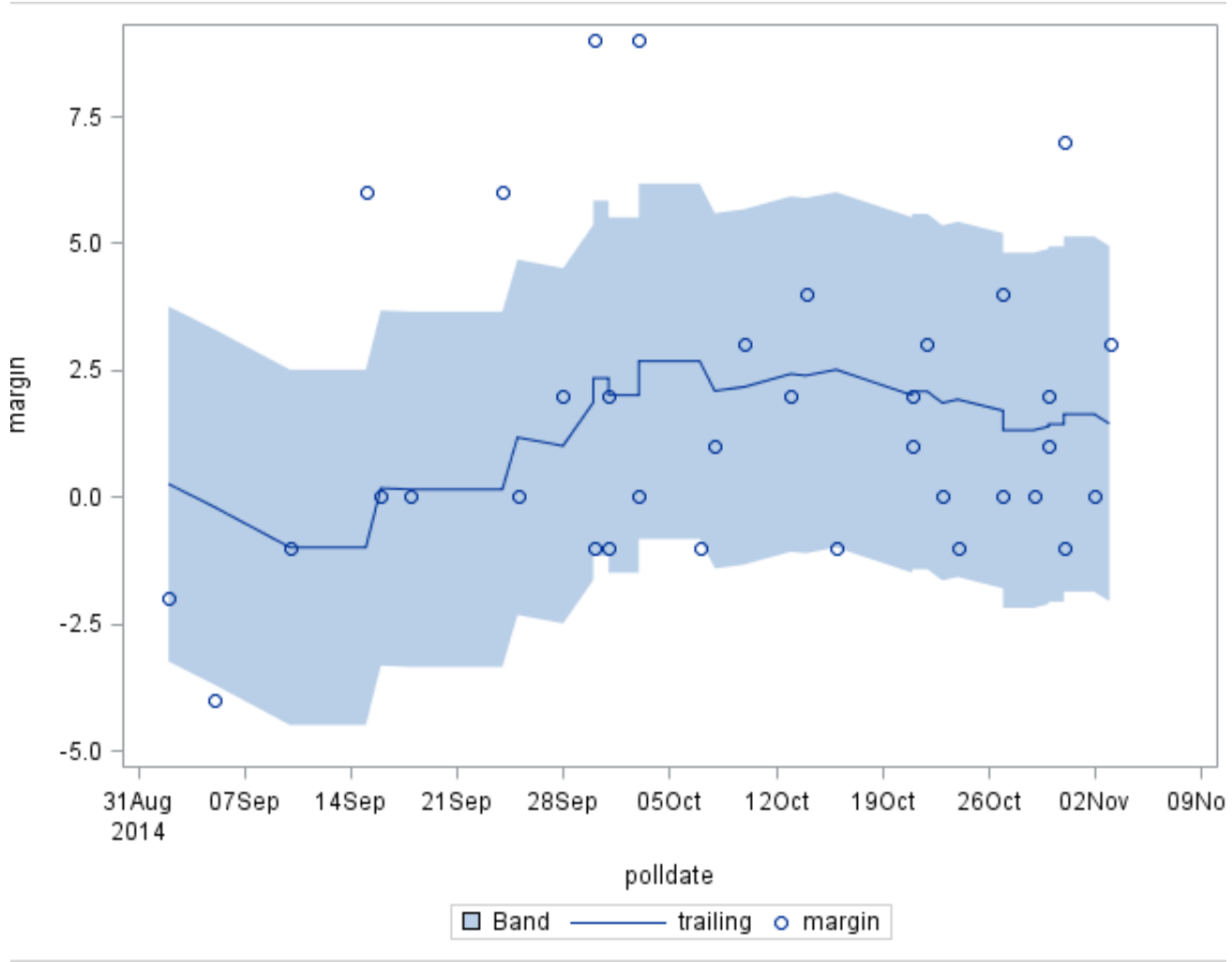




Layered together

Notice the scales now match.

The scales/coordinates are critical to how the elements are aligned on the page, and with each other.

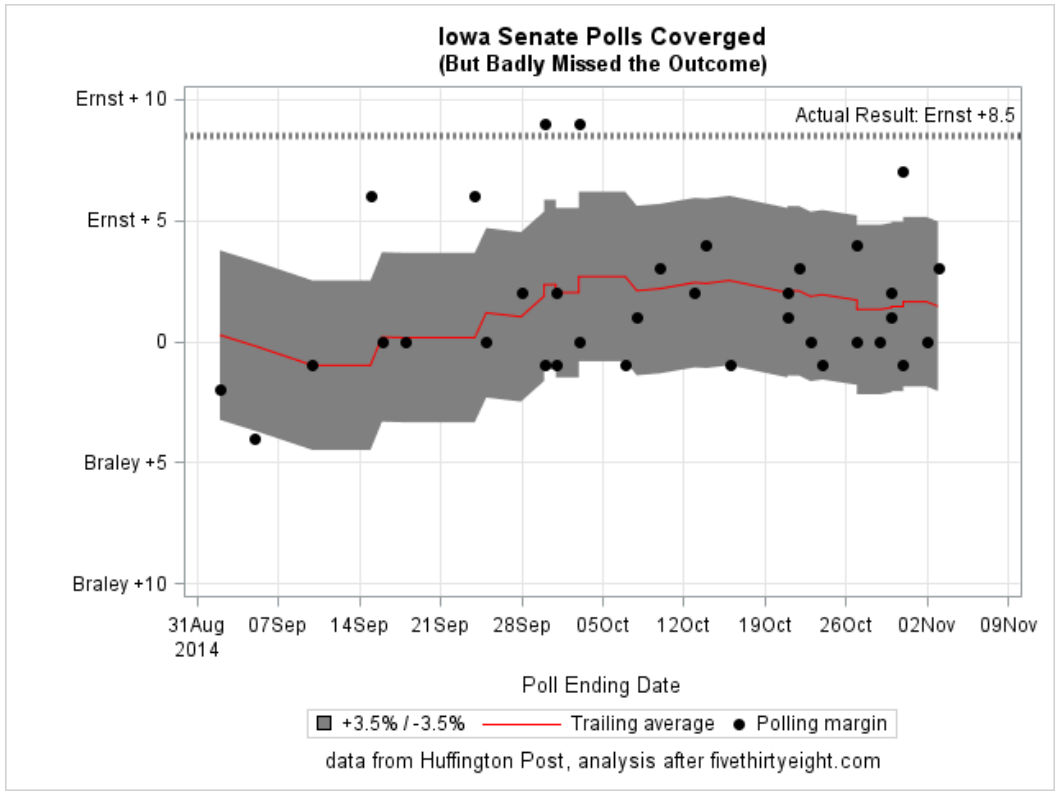


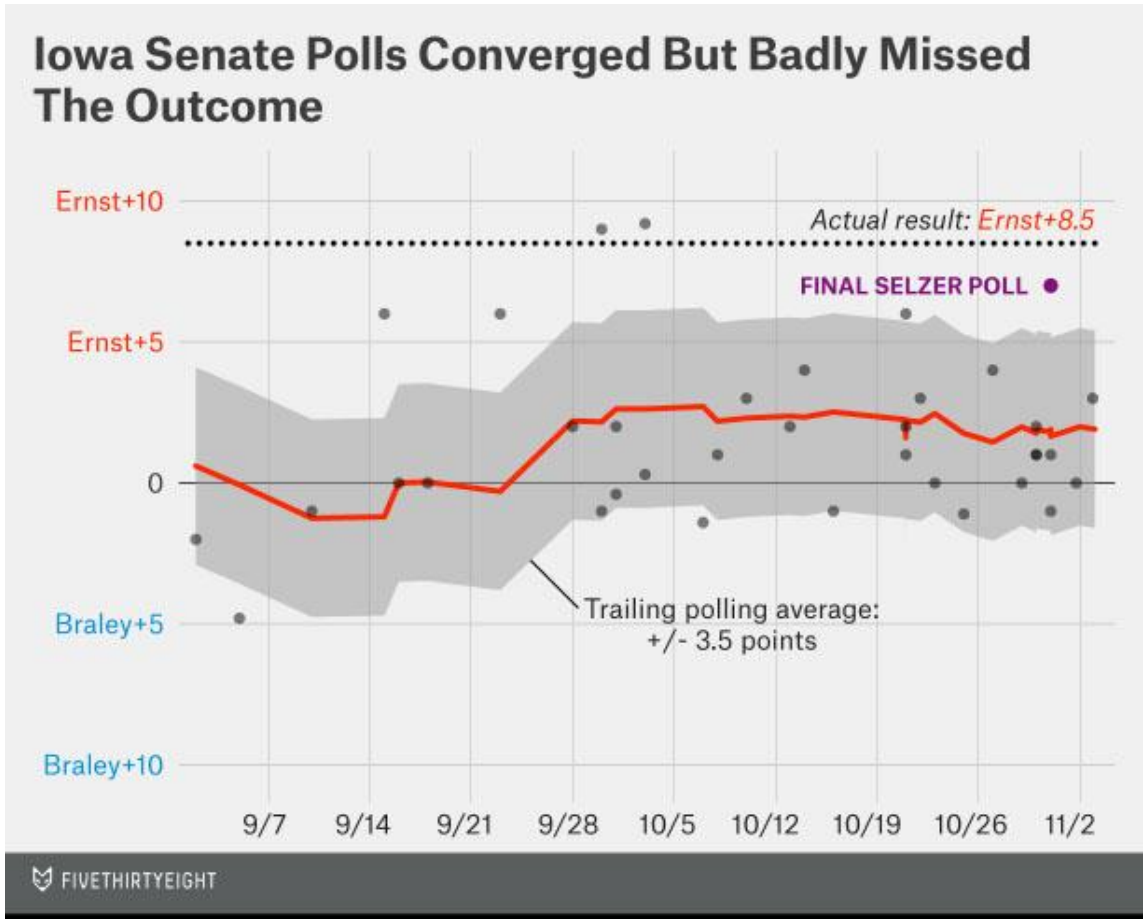
Decoration/Aesthetics

- Titles and footnotes
- Color, weight, etc. of graphical elements
- Axis and legend text
- Grid or guidelines

- Etc. – there tend to be a large number of options at this point

Reconstructed





Programming

- The final program starts with the end result in mind
  - Get the data, convert data types and layout (long vs. wide) as necessary
  - Calculate data values needed
  - Specify the graphics

Get, clean, convert the data

```
data Iowapolls;
        infile "Z:\PUBLIC_web\Stataworkshops\Grammar of Graphics\Iowa Huffingtonpost.csv"
                firstobs=2 dsd dlm=",";
        input Pollster :$46. Dates :$16. Pop $ Ernst Braley Undecided Spread :$10;
        LV = (index(pop, "LV") gt 0);
        if LV; *Just use "likely voter" polls, not "registered voters";
        polldate = input(trim(substr(dates, index(dates, "-")+2))||"/2014", mmddyy10.);
        format polldate mmddyy8.;
        margin = Ernst - Braley;
run;

proc sort;
    by polldate;
    run;
```

Calculate other needed data

```
data Iowapolls2;
            set Iowapolls;
            retain w1 - w51 m1-m51;
            array w {51};
            array m {51};
            w{_n_} = polldate;
            m{_n_} = margin;
            do i = 1 to _n_;
                        if w{i}+21 lt polldate then do;
                                    w{i}=.;
                                    m{i}=.;
                                    end;
                        end;
            trailing21 = mean(of m1-m51);
            do i = 1 to 51;
                        if w{i} ne . then lastdate = w{i};
                        end;
            format lastdate mmddyy8.;
            drop w1 -- i;
            run;
```

Calculate other needed data (cont.)

```
data trailing;
        set Iowapolls2;
        keep trailing21 lastdate;
        by lastdate;
        if last.lastdate then output;
        run;

data Polls;
        merge Iowapolls trailing(rename=(lastdate=polldate));
        by polldate;
        trailing = lag(trailing21);
        lmoe = trailing - 3.5;
        umoe = trailing + 3.5;
        if polldate gt "01Sep2014"d;
        run;
```

Basic graphical specification

```
proc sgplot;
        band x=polldate upper=umoe lower=lmoe;
        series x=polldate y=trailing;
        scatter x=polldate y=margin;
        run;
```

With decoration (setup)

```
proc format;
        value margin -10 = "Braley +10"
                -5 = "Braley +5"
                5 = "Ernst + 5"
                10 = "Ernst + 10";
        run;

title "Iowa Senate Polls Coverged";
title2 "(But Badly Missed the Outcome)";
footnote "data from Huffington Post, analysis after fivethirtyeight.com";
```

With decoration (plot)

```
proc sgplot;
        band x=polldate upper=umoe lower=lmoe / fillattrs=(color=gray)
                legendlabel="+3.5% / -3.5%";
        series x=polldate y=trailing / lineattrs=(color=red)
                legendlabel="Trailing average";
        scatter x=polldate y=margin / markerattrs=(symbol=circlefilled color=black)
                legendlabel="Polling margin";
        refline 8.5 / lineattrs=(pattern=dot thickness=4 color=black)
                label="Actual Result: Ernst +8.5" labelloc=inside;
        xaxis grid;
        yaxis min=-10 max=10 grid label=" ";
        label polldate = "Poll Ending Date";
        format margin margin.;
        run;
```

After all the steps