Gallup polls published a document called "*How Are Polls Conducted?*" that describes how Gallup selects people to include in its poll and other details. Toward the end of the document there were two exerpts that gave me some pause.

> "*For example, with a sample size of 1,000 national adults (derived using careful random selection procedures), the results are highly likely to be accurate within a margin of error of ±4 percentage points.*"

> "*If Gallup increases a poll sample size to 2,000, the results would then be accurate within ± 2% of the underlying population value, a gain of two percentage points in terms of accuracy, but with a 100% increase in the cost of conducting the survey.*"

As an example, in top-line results of a February 3-16, 2025 poll of 1,004 adults aged 18+ living in all 50 U.S. states and the District of Columbia revealed that 39% of respondents were satisfied with the position of the United States in the world today, compared to 59% who are dissatisfied (2% had no opinion). Gallup reported the ±4% margin of error.

When we provide a "margin of error," we aim to describe the variation in our using the sample proportion $\hat{p}$ as an estimate of p. That is, how much can we expect the sample proportion to differ from the proportion for the entire population?

Often, statisticians and quantitative researchers will report a margin of error that provides 95% confidence. The usual interpretation is that we are 95% confident that the actual proportion that describes the entire population is on the interval created by adding and subtracting the margin of error from the sample proportion. When we say 95% confident, we mean that if we were to conduct the poll repeatedly the interval would contain the actual proportion 95% of the time.

1 Basic Simulation

First, conduct a basic simulation study. Assume that the true probability that someone is satisfied with the position of the United States in the world today is 0.39. Use `rbinom()` to generate 10k polls of the same sample size.

Plot a histogram of the resulting sample proportions with a superimposed density to provide a visual approximation of the sampling distribution for the sample proportion. What do you notice about the shape? What is the range of the middle 95%? You can approximate the margin of error by halving that range (i.e., plus or minus gives the full range). How does this compare to the 4% reported by Gallup?
Now, double the sample size and perform the same computations. Plot a histogram of the resulting sample proportions with a superimposed density to provide a visual approximation of the sampling distribution for the sample proportion. What do you notice about the shape? What is the range of the middle 95%? You can approximate the margin of error by halving that range (i.e., plus or minus gives the full range). How does this compare to the 2% reported by Gallup?

2 Resampling

In the previous section, you needed to make an assumption about the actual population proportion p. Under that assumed value of p, we performed simulations to see what results of the poll should look like under that assumption.

Another option for approximating the sampling distribution for $\hat{p}$ that we have discussed in class is resampling. Create a data frame that contains the data from the Gallup survey and perform resampling.
Plot a histogram of the resulting proportions from resampling with a superimposed density to provide a visual approximation of the sampling distribution for the sample proportion. What do you notice about the shape? What is the range of the middle 95%? You can approximate the margin of error by halving that

range (i.e., plus or minus gives the full range). How does this compare to the 4% reported by Gallup? Note here, we are limited to the data we have – that is, we cannot "double the sample size."

3 Simulation over n and p

For `n` in $\{100, 110, 120, ..., 3000\}$ and `p` in $\{0.01, 0.02, ..., 0.99\}$ perform 10000 simulations as we did in Section 1. For each case, store the half the range of between the 2.5th and 97.5th percentiles. To summarize the results, create a `geom_raster()` plot showing the estimated margin of error as a function of n and p. Provide better guidance for the Gallup readers.
In Figure **??**, we see that the story isn't as simple as Gallup made it out to be. That is, the sample size is only one part of the equation. It is true that an increased sample size reduces the margin of error, but it also depends on p. When p is extreme (close to 0 or 1), the margin of error is smaller as we can not extend beyond the parameter space.

4 Actual Margin of Error Calculation

From Central Limit Theorem, we know that

$$\widehat{p} \sim \mathcal{AG}\left(\mu_{\widehat{p}} = p, \sigma_{\widehat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}\right).$$

We have that

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AG}(\mu_Z = 0, \sigma_Z = 1).$$

We can "pin" Z between the 2.5th and 97.5th percentile:

$$P\left(z_{0.025} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{0.975}\right)$$

and solve the inequality for p.

Solving the Z equation for $(\hat{p} - p)^2$ yields

$$(\hat{P} - p)^2 = Z^2\left(\frac{p(1-p)}{n}\right).$$

We can solve for p, the parameter of interest, as follows.

$$\begin{aligned}
(\hat{P} - p)^2 - Z^2\left(\frac{p(1-p)}{n}\right) &= 0 && \text{(Rewriting)}\\
\hat{P}^2 - 2(\hat{p})(p) + p^2 - \frac{Z^2 p}{n} - \frac{Z^2 p^2}{n} &= 0 && \text{(Expanding)}\\
\left(1 + \frac{Z^2}{n}\right)p^2 + \left(-\frac{Z^2}{n} - 2\hat{P}\right)p + \hat{P}^2 &= 0 && \text{(Rewriting)}
\end{aligned}$$

Noting we have a quadratic with respect to p, we can apply the quadratic formula to find the roots.

$$p = \frac{-\left(-\frac{Z^2}{n} - 2\hat{P}\right) \pm \sqrt{\left(-\frac{Z^2}{n} - 2\hat{P}\right)^2 - 4\left(1 + \frac{Z^2}{n}\right)\left(\hat{P}^2\right)}}{2\left(1 + \frac{Z^2}{n}\right)} \tag{Quadratic Formula}$$

$$= \frac{-\left(-\frac{Z^2}{n} - 2\hat{P}\right)}{2\left(1 + \frac{Z^2}{n}\right)} \pm \frac{\sqrt{\left(-\frac{Z^2}{n} - 2\hat{P}\right)^2 - 4\left(1 + \frac{Z^2}{n}\right)\left(\hat{P}^2\right)}}{2\left(1 + \frac{Z^2}{n}\right)} \tag{Rewriting}$$

Remark: Here, we see something similar to a confidence interval formula, i.e., a point estimator ±, a margin of error.

The point estimator can be simplified as follows. This is called the **Wilson Estimate**.

$$\tilde{P} = \frac{-\left(-\frac{Z^2}{n} - 2\hat{P}\right)}{2\left(1 + \frac{Z^2}{n}\right)} = \frac{\frac{Z^2}{n} + 2\hat{P}}{2\left(1 + \frac{Z^2}{n}\right)} \tag{Distributing}$$

$$= \frac{\frac{Z^2}{n} + 2\left(\frac{X}{n}\right)}{2\left(1 + \frac{Z^2}{n}\right)} \tag{Writing $\hat{P}$ as $\frac{X}{n}$}$$

$$= \frac{\frac{1}{n}\left(Z^2 + 2X\right)}{2\left(1 + \frac{Z^2}{n}\right)} \tag{Factoring}$$

$$= \frac{Z^2 + 2X}{2n\left(1 + \frac{Z^2}{n}\right)} \tag{Rewriting}$$

$$= \frac{\frac{1}{2}Z^2 + X}{n\left(1 + \frac{Z^2}{n}\right)} \tag{Rewriting}$$

$$= \frac{X + \frac{1}{2}Z^2}{n + Z^2} \tag{Distributing}$$

Remark: We can see that the Wilson Estimate is a weighted average between $\hat{P} = \frac{X}{n}$ and $\frac{1}{2}$ where the weights are n and Z^2 respectively. That is, $\hat{P}$ is weighted more heavily as n increases.

The margin of error is simplified as follows.

$$
\begin{aligned}
\frac{\sqrt{\left(-\frac{Z^2}{n}-2\hat{p}\right)^2-4\left(1+\frac{Z^2}{n}\right)\left(\hat{p}^2\right)}}{2\left(1+\frac{Z^2}{n}\right)} &= \frac{\sqrt{\left(-\frac{Z^2}{n}-2\frac{x}{n}\right)^2-4\left(1+\frac{Z^2}{n}\right)\left(\left(\frac{x}{n}\right)^2\right)}}{2\left(1+\frac{Z^2}{n}\right)} && \text{(Writing } \hat{P} \text{ as } \tfrac{X}{n}\text{)}\\
&= \frac{\frac{2}{n}\sqrt{\frac{1}{4}\left(-Z^2-2x\right)^2-\left(1+\frac{Z^2}{n}\right)x^2}}{2\left(1+\frac{Z^2}{n}\right)} && \text{(Factoring)}\\
&= \frac{\sqrt{\frac{1}{4}\left(-Z^2-2x\right)^2-\left(1+\frac{Z^2}{n}\right)x^2}}{n+Z^2} && \text{(Simplifying)}\\
&= \frac{\sqrt{\frac{Z^4}{4}+\frac{4Z^2x}{4}+\frac{4x^2}{4}-x^2-\frac{Z^2X^2}{n}}}{n+Z^2} && \text{(Expanding)}\\
&= \frac{\sqrt{\frac{Z^4}{4}+Z^2X+X^2-X^2-\frac{Z^2X^2}{n}}}{n+Z^2} && \text{(Simplifying)}\\
&= \frac{\sqrt{\frac{Z^4}{4}+Z^2X-\frac{Z^2X^2}{n}}}{n+Z^2} && \text{(Simplifying)}\\
&= Z\frac{\sqrt{\frac{Z^2}{4}+X-\frac{X^2}{n}}}{n+Z^2} && \text{(Factoring)}\\
&= Z\frac{\sqrt{\frac{Z^2}{4}+X\left(1-\frac{X}{n}\right)}}{n+Z^2} && \text{(Factoring)}\\
&= Z\frac{\sqrt{\frac{Z^2}{4}+n\frac{X}{n}\left(1-\frac{X}{n}\right)}}{n+Z^2} && \text{(Rewriting)}\\
&= z_{1-\alpha/2}\frac{\sqrt{n\hat{p}(1-\hat{p})+\frac{z^2_{1-\alpha/2}}{4}}}{n+z^2_{1-\alpha/2}}. && \text{(Writing } \tfrac{X}{n} \text{ as } \hat{P}\text{)}
\end{aligned}
$$

Compute the Wilson margin of error formula for `n` in $\{100, 110, 120, ..., 2000\}$ and `p` in $\{0.01, 0.02, ..., 0.99\}$. To summarize the results, create a `geom_raster()` plot showing the margin of error as a function of n and p. Provide better guidance for the Gallup readers.

5 Optional Challenge: Simulating Bootstrap over n and p

For `n` in $\{100, 110, 120, ..., 3000\}$ and `p` in $\{0.01, 0.02, ..., 0.99\}$ perform 100 simulations of resampling as we did in Section 2. For each case, perform 100 simulations where you generate a random poll result, perform resampling, compute half of the range of the middle 95%, and store it. Finally, store the average of the 100 half-ranges of the middle 95% from the 100 simulations – this will act as the estimated margin of error for the resampling approach.

To summarize the results, create a `geom_raster()` plot showing the estimated margin of error as a function of n and p. Provide better guidance for the Gallup readers.

Note: This code will take a long time to run. I recommend running it on a smaller set of n (e.g., 100, 200, 300, 400), and a smaller grid for p (e.g., 0.05, 0.15,..., 0.95) to make sure it works. Even still, you may want to run it overnight. If you have working code, I will send you the output from the full simulation, which took 65 hours, 38 minutes, and 53 seconds to run.

Added Challenge: You can try to implement a `faster` solution that uses the `foreach` and `doParallel` packages for `R` (Microsoft and Weston, 2022b,a) to perform the computations in parallel.

References

Microsoft and Weston, S. (2022a). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.17.

Microsoft and Weston, S. (2022b). *foreach: Provides Foreach Looping Construct*. R package version 1.5.2.