What is this "Margin of Error"?

On April 23, 2017, The Wall Street Journal reported:

> "Americans are dissatisfied with President Donald Trump as he nears his 100th day in office, with views of his effectiveness and ability to shake up Washington slipping, a new Wall Street Journal/NBC News poll finds.

> "More than half of Americans—some 54%—disapprove of the job Mr. Trump is doing as president, compared with 40% who approve, a 14-point gap. That is a weaker showing than in the Journal/NBC News poll in late February, when disapproval outweighed approval by 4 points."

> [Skipping to the end of the article …] "The Wall Street Journal/NBC News poll was based on nationwide telephone interviews with 900 adults from April 17-20. It has a margin of error of plus or minus 3.27 percentage points, with larger margins of error for subgroups."

Throughout the modern world, every day brings news concerning the latest public opinion polls. At the end of each news report, you'll be told the "margin of error" in the reported estimates.

Every poll is, of course, subject to what's called "sampling error": Evenly if properly run, there's a chance that the poll will, merely due to bad luck, end up with a randomly-chosen sample of individuals which is not perfectly representative of the overall population.

However, using the tools you learned in your "probability" course, we can measure the likelihood of such bad luck. Assume that the poll was not subject to any type of systematic bias (a critical assumption, unfortunately frequently not true in practice). Then there was a 95% chance, when the poll was conducted, that the final estimates would differ from the true numbers which describe the entire population by no more than the reported "margin of error".

In simple terms: The pollster in the WSJ/NBC study reported above, knowing only the results of his/her poll, should be willing to bet $19 against your bet of $1 that Trump's actual level of approval amongst American adults in late April was in the 40% ± 3.27% range. (This range is called "a 95%-confidence interval" for the true percentage.)

The next page extends this language to other types of estimates and to forecasts.

Optional note:

The reported margin of error in the news story refers to the estimated disapproval (54%) and approval (40%) levels separately. How much can we trust the estimate that disapproval in April 2017 led approval by 14 percentage points? Or that disapproval's lead had grown by 10 percentage points over the previous two months?

During our course, we'll see that the margin of error in the estimated "gap" of 14% is roughly 6.6%, and that the margin of error in the 10% estimated change in the gap is roughly 9.2%! The insight is simple: Combining fuzzy numbers increases the resulting fuzz.

The Language of (Statistical) Estimation

Typically, when we're estimating some unknown characteristic of a population, or predicting something, we'll be able to compute (from our sample data) one standard-deviation's-worth of "noise" (i.e., potential error due to the randomness of the sampling process that gave us our data) in our estimate or prediction. And typically, the methods we'll use will lead to normally-distributed uncertainty in our estimate or prediction. Therefore, 95% of the time we'll get an estimate or prediction that differs from the truth (i.e., the true value of that population characteristic, or the actual value ultimately taken by whatever we're predicting) by no more than approximately two standard-deviation's-worth of noise.

One _standard_-deviation's-worth of potential _error_ in an estimate or prediction is called the **standard error of the {quantity being estimated or predicted}**. As we examine the results of a regression analysis, we'll encounter, for example, the standard error of a mean, the standard error of a proportion, the standard error of a prediction, the standard error of an estimated subgroup mean, the standard errors of the estimated coefficients in our regression model, and the standard error of the regression itself.

The **margin of error (at the 95%-confidence level) in our estimate or prediction** will simply be (approximately 2)·(the standard error of that estimate or prediction). Both the appropriate "approximately 2" multiplier (which comes from something called "the t-distribution" with some number of "degrees of freedom") and the appropriate standard error (again, this represents one standard-deviation's-worth of "noise" in our estimation or prediction process) will be computed for us by any modern statistical software.

Finally, a **95%-confidence interval for our estimate or prediction** is

> (the estimate or prediction itself) ± (~2)·(the standard error of the estimate or prediction), or

> (the estimate or prediction itself) ± (the margin of error in the estimate or prediction).

Here's an example from the first dataset we'll analyze in class, consisting of 15 cars sampled from the fleet owned and operated by a municipality. The variables for each car are maintenance Costs over the previous year, Mileage (in 000s) driven over the previous year, Age (in years) at the start of the year, and Make (all the cars were either Fords or Hondas; Make is encoded as Ford = 0, Honda = 1).

	B	Costs	Mileage	Age	Make	G
1	**Univariate statistics**					
2		**Costs**	**Mileage**	**Age**	**Make**	
3	mean	688.866667	16.3733333	1	0.46666667	
4	standard deviation	111.678663	4.34370919	0.84515425	0.51639778	
5	standard error of the mean	28.8353068	1.12154089	0.21821789	0.13333333	
6						
7	minimum	518	8.4	0	0	
8	median	673	16.9	1	0	
9	maximum	861	24.6	2	1	
10	range	343	16.2	2	1	
11						
12	skewness	0.038	-0.214	0.000	0.149	
13	kurtosis	-1.189	-0.068	-1.615	-2.308	
14						
15	number of observations	15				
16						
17	t-statistic for computing					
18	95%-confidence intervals	2.1448				

estimate ± margin of error

95%-confidence interval for mean annual cost per car (across the fleet) $688.87 ± 2.1448·$28.84

95%-confidence interval for mean annual miles driven per car   16,373 ± 2.1448·1,121

95%-confidence interval for mean age of cars in the fleet   1.000 ± 2.1448·0.218

95%-confidence interval for fraction of fleet for which Make = Honda   46.67% ± 2.1448·13.33%

Optional: Connection with the Probability Module

If what appears below feels too technical, please feel free to skip it completely!

Every statistical study begins with a random sampling of data from a population of interest. Any number then computed from the sample data can be thought of as one realization of a random variable which takes different values for different samples.

For example, imagine repeatedly drawing random individuals (say, from the population of EMP alumni who graduated between 2005 and 2010), where each draw is equally likely to yield any one of the individuals in that population. (This is called "simple random sampling with replacement.") We'll then average the annual incomes of all of the sampled individuals, and use that as an estimate of the true population mean, μ (the mean annual income of EMP alumni 5 to 10 years after graduation). Call the computed sample mean $\bar{x}$ (a number); this is a realization of the random variable $\bar{X}$, which varies from one sample to the next.

The individual random draws $X_1, X_2, …, X_n$ are independent, identically-distributed random variables, and $\bar{X}$ is their average. Therefore, from your probability course:

(1) $E[\bar{X}] = \mu$, (2) $StdDev(\bar{X}) = \sigma / \sqrt{n}$, and (3) $\bar{X}$ is approximately normally distributed

(where σ is the population standard deviation, and approximate normality follows from the Central Limit Theorem).

If $\bar{X}$ were precisely normally distributed, there would be a 95% chance that its realized value $\bar{x}$ differs from its expected value (the true population mean μ) by no more than $\pm 1.96 \cdot \sigma / \sqrt{n}$.

We don't know σ, of course. But we can use the sample standard deviation s as an estimate of σ, and therefore we can be 95%-confident that the interval

$$\bar{x} \pm (\sim 2) \cdot s / \sqrt{n}$$

contains the true population mean μ. This is called a **95%-confidence interval** for μ. (The "approximately 2" multiplier is a bit larger than 1.96, to cover for our use of s instead of σ.)

Similarly, when estimating or predicting *anything* statistically, a 95%-confidence interval will be

(the estimate or prediction)
$\pm (\sim 2) \cdot$ **(one standard-deviation's-worth of random variability in the estimate or prediction).**

That standard-deviation's-worth of "noise" in our estimate or prediction is called the "standard error" of whatever we're estimating or predicting. On the previous page, note that each "standard error of the mean" is simply the sample standard deviation, divided by $\sqrt{15}$, the square-root of the sample size.

In regression studies, computing standard errors typically requires much more work than using the simple $s / \sqrt{n}$ formula for the standard error of the mean. That's why we leave the calculation to the computer.