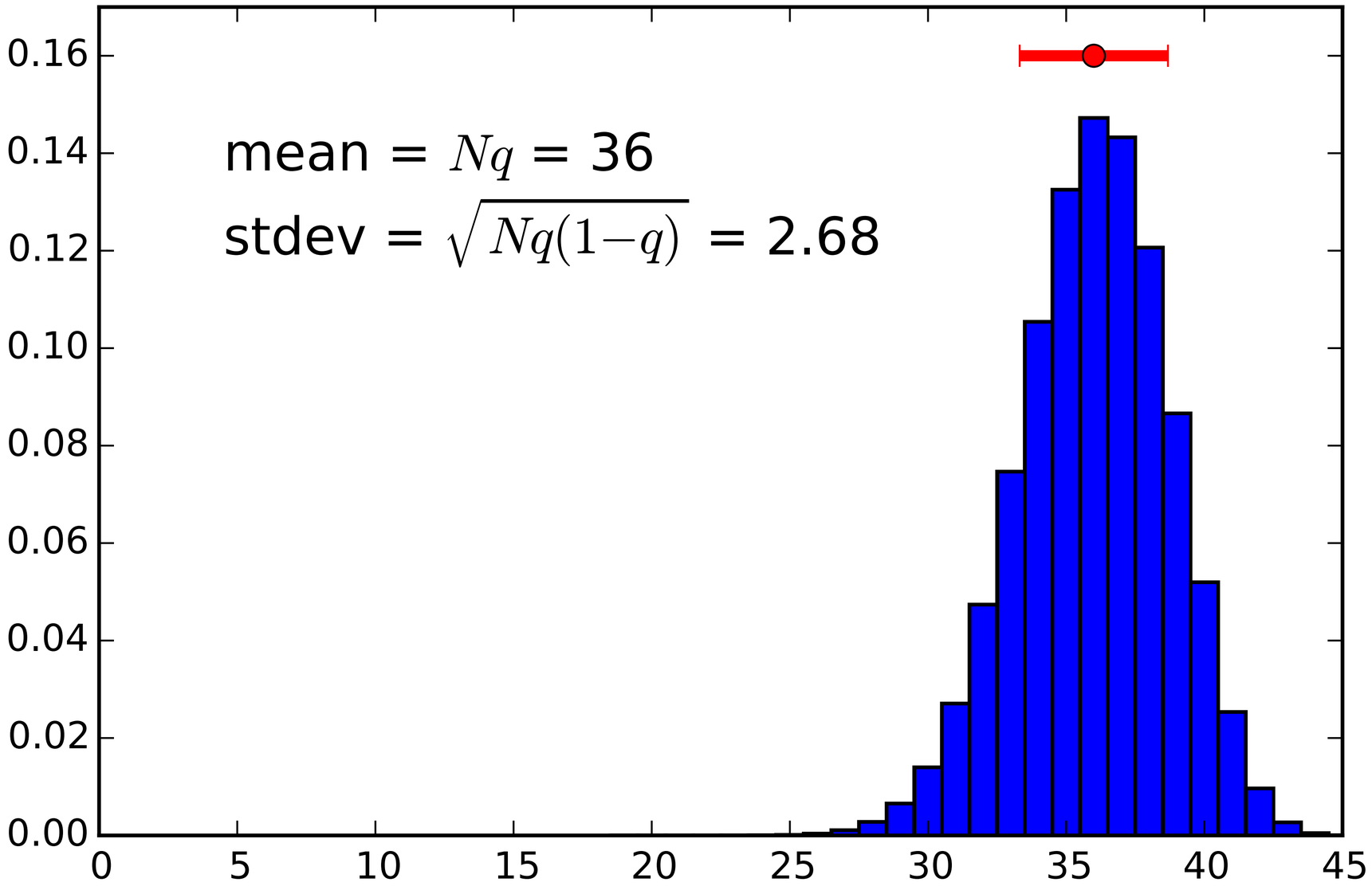
Info 2950, Lecture 12

9 Mar 2017

Prob Set 3: due Fri night 10 Mar

(from lecture 7, 16 Feb)

Midterm poll, N=45, q=4/5



(from lecture 7, 16 Feb)

Suppose that q = m / 132 prefer 23 Mar.

We're random sampling 45 people,
each a Bernoulli trial with probability q of "success"

What statistical variation would we expect for N=45 ?

Estimate q = 36/45 = 4/5, then variance of a trial is (1/5)(4/5), and

V = Nq(1-q) = 45 * (1/5) (4/5) = 36 / 5 , so that $\sigma = \sqrt{V[X]}$ = 2.68

so result for mean Nq is accurate to roughly 36 +/- 2.65 ,
(or as a percentage q = .8 +/- .06 = 80% +/- 6%)

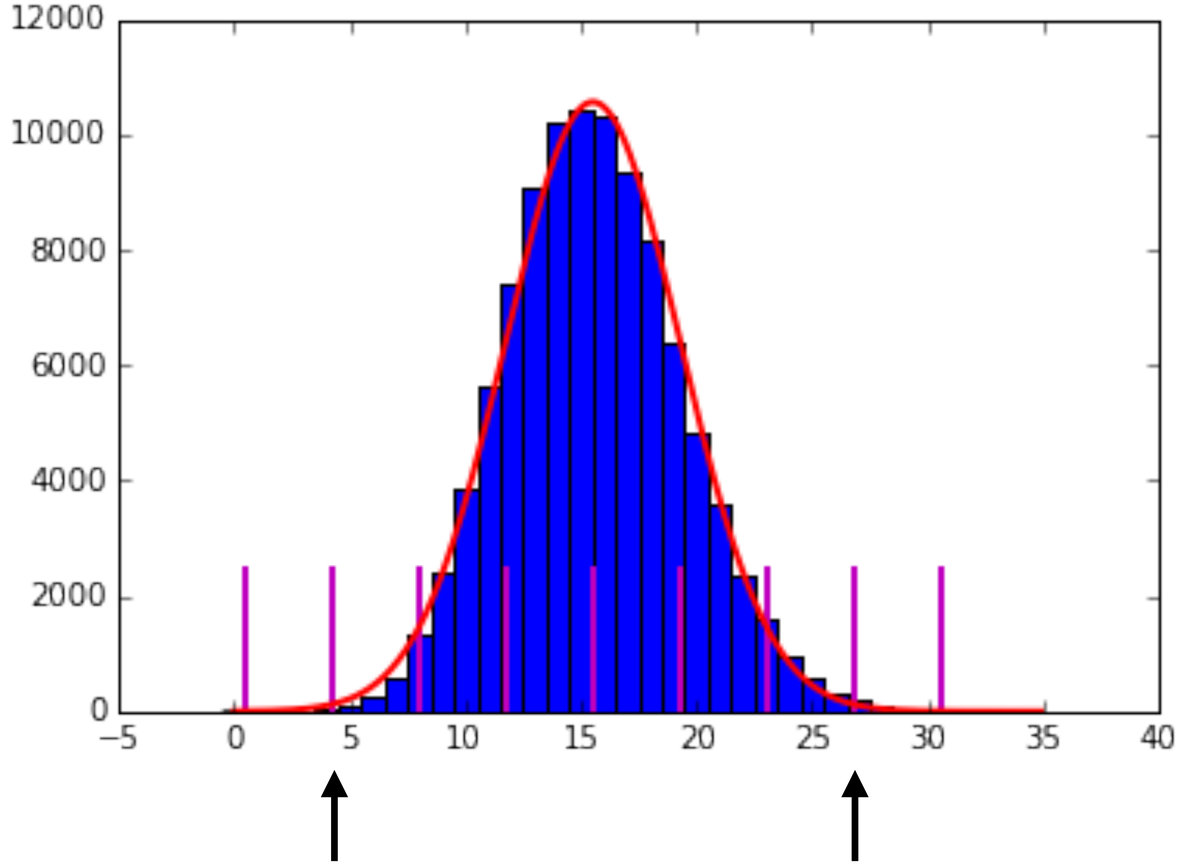
Color blindness test: expect 8% in class with 194 boys

mean = Np= .08*194 = 15.52
std = sqrt(Np(1-p)) = sqrt(.08*.92*194) = 3.78

From the '68-95-99.7' rule, it's very likely (95%) that the results of a single measurement will fall between 15.52 +/- 2*3.78, so from 8 to 23,

and almost certain (99.7%) that they'll fall between 15.52 +/- 3 * 3.78, so from 4 to 27 measured to have color blindness.

```
def rn(p,n): return sum(rand(n) <= p)
S=100000 # number of simulations
results = [rn(.08,194) for t in xrange(S)]
```



Consider a sample poll of $n=1000$ people,
of whom $k=550$ answered 'Yes' to some question.

Again we can consider either the standard deviation of the number count, or the standard deviation of the percentage.

The standard deviation of the number count for a Bernoulli trial consisting of n events each with probability q of success, as derived in class:

$$\sigma=\sqrt{nq(1-q)}$$

If the actual probability is q, then expect nq Yes responses.
If poll redone with different samples of n people taken from same distribution, then 68% of the time expect to get values of k between $nq-\sigma$ and $nq+\sigma$.
If poll done only once and get k yes votes then estimate of the underlying probability is $p=k/n$, but could by off due to finite sample size.
Redo poll many times and average the values of p, gets closer and closer to the underlying probability q.

Since $p=k/n$, the standard deviation of the estimated probability is given by dividing the standard deviation of the number count by n:

$$\sqrt{np(1-p)} \,/\, n \;=\; \sqrt{p(1-p)/n}.$$

Notice that the n is now in the denominator under the square root.

(That is because the standard deviation of the number count had a factor of n in the numerator under the square root, and when we divide by n that gets converted into a factor of n in the denominator under the squareroot: $\sqrt{n} \,/\, n = 1 \,/\, \sqrt{n}$)

Suppose we want to sample a population and determine the fraction of voters p to within +/- .01 accuracy at the 90% confidence level.

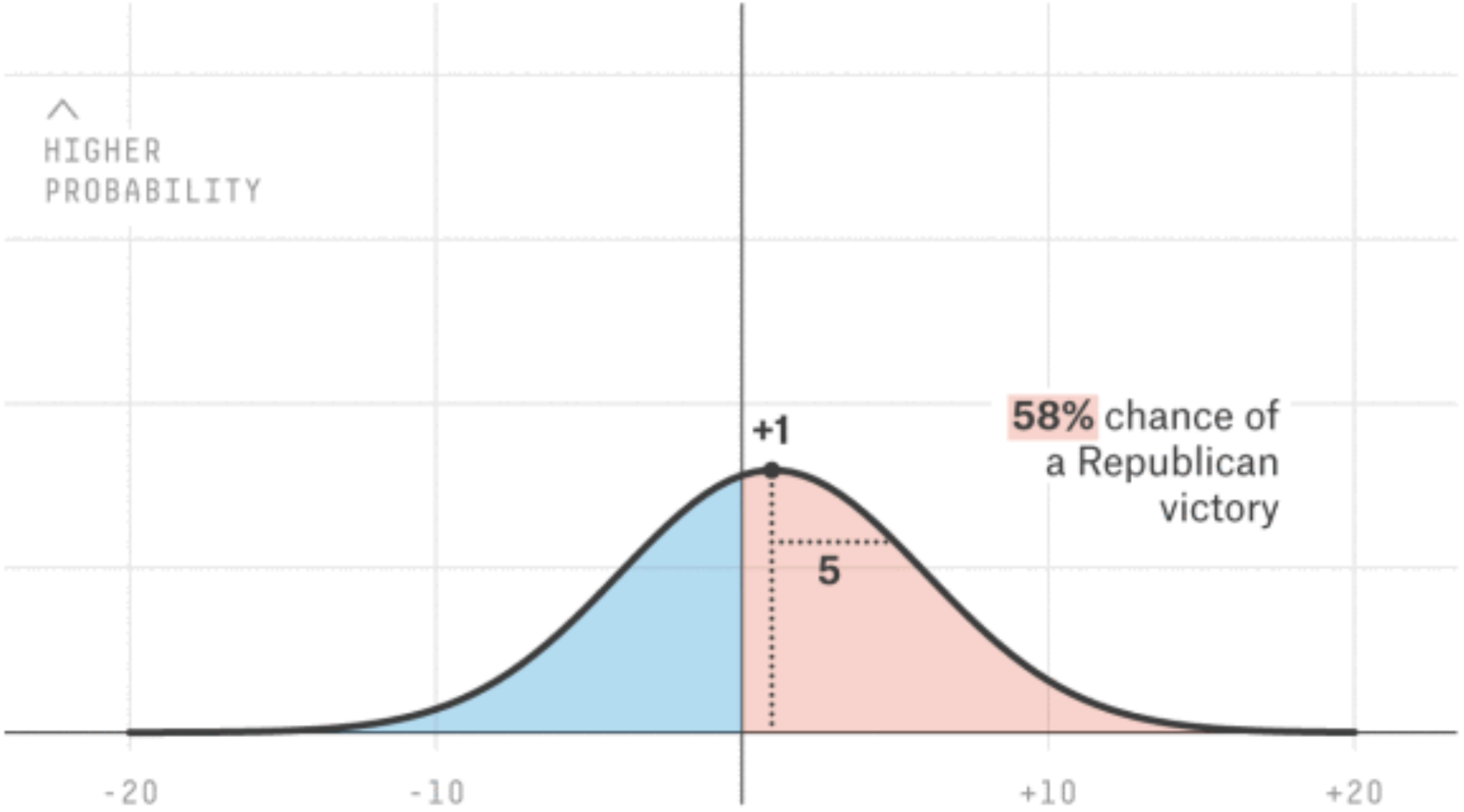
m of N voters say yes, how big does N need to be?
90% confidence level is $m \pm 1.645\sigma$, where $\sigma = \sqrt{Np(1-p)}$.

So we estimate the underlying $p = m/N \pm 1.645 \sqrt{p(1-p)/N}$

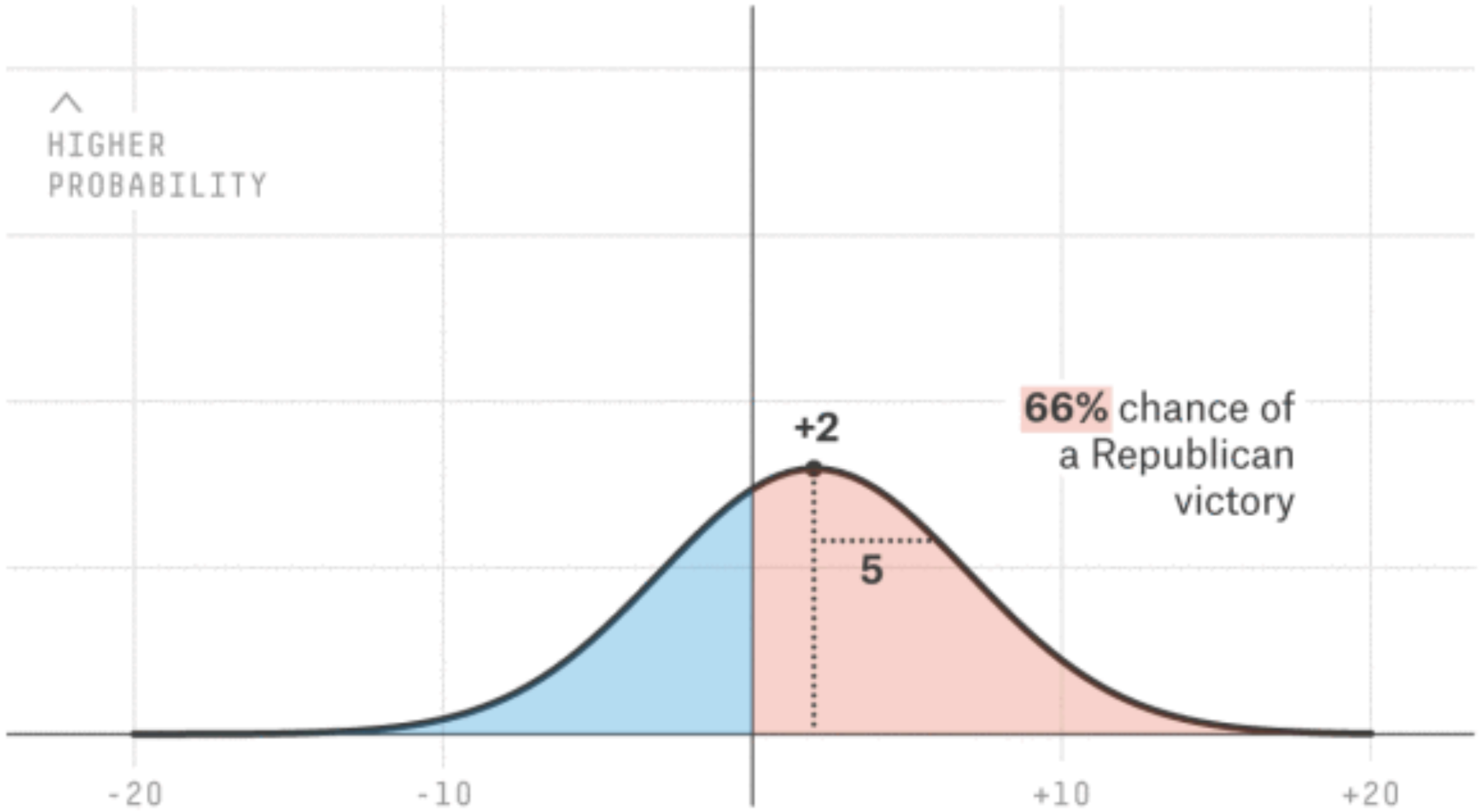
need $1.645 \sqrt{p(1-p) / N} < .01$ to have 90% probability of being within +/- .01

Worst case is p=1/2, so need $\sqrt{N} > 100 * 1.645 / 2$ or $N > 6765$

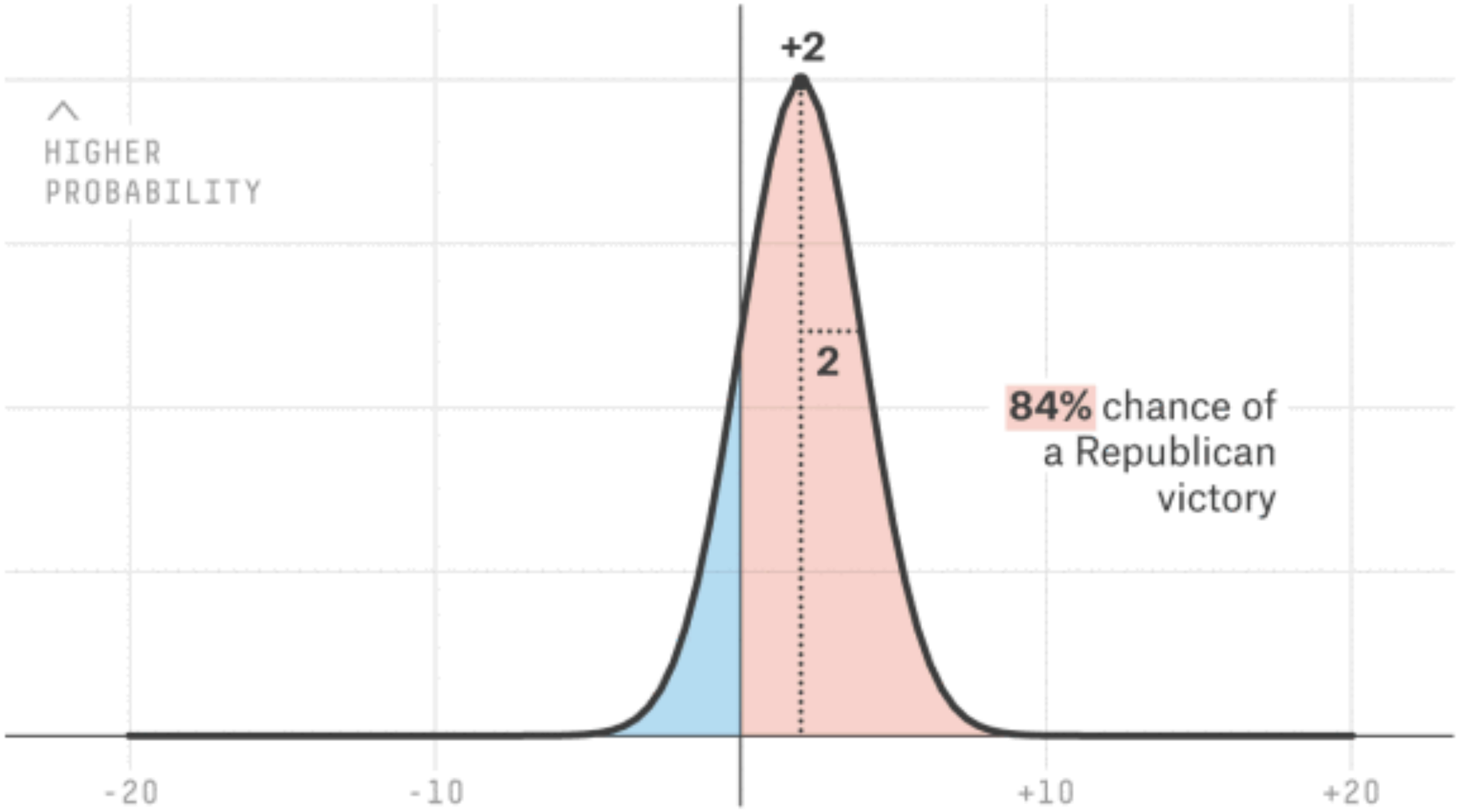
https://fivethirtyeight.com/features/how-the-fivethirtyeight-senate-forecast-model-works/



Suppose poll data gives candidate 51% with standard deviation of 5%,
then p=.51, σ=.05, so .01/.05 SD above mean,
and norm.cdf(.2) = .579 gives candidate a 58% chance of winning



Suppose poll data with same standard deviation of 5% now gives candidate 52% of sample, then p=.52, σ=.05, so .02/.05 SD above mean,
and norm.cdf(.4) = .655 gives candidate a 66% chance of winning



Suppose poll data gives candidate same 52% but now with standard deviation of just 2%, then p=.52, σ=.02, so .02/.02 SD above mean,
and norm.cdf(1) = .841 gives candidate an 84% chance of winning