Info 2950, Lecture 9
27 Feb 2018

Prob Set 3: due Wed night 7 Mar

Prob Set 4: out 8 Mar, due Mon 19 Mar (election simulation)

Midterm: Thu Mar 22, 1:25-2:40, Statler 185
Consider measuring samples of the population for some trait.

Standard red/green color blindness affects roughly 8% of males (the relevant gene that codes for the pigment in the retinal cone cells is on the X chromosome, so is sex-linked, and only .6% of females are affected since that would require two of the variant X chromosomes, .08*.08=.0064).

Suppose that 194 boys in an incoming class of high school students are tested for color blindness, what range of results is expected?
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.

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

def rn(p,n): return sum(rand(n) <= p)
S=100000 # number of simulations
results = [rn(.08,194) for t in xrange(S)]
\[ N = 72, \; q = 0.08 \]
\[ \text{mean} = Nq = 5.76 \]
\[ \text{stdev} = \sqrt{Nq(1-q)} = 2.30 \]
Consider a sample poll of \( n=1000 \) people, of whom \( k=550 \) answered 'Yes' to some question.

Assuming the 1000 people are drawn at random from some very large population, what is the likely percentage of 'Yes' voters in the full population?

The idea is that the fraction \( p=550/1000=.55 \) is sampled from some much larger population whose overall fraction is some unknown value \( q \).

Each sampled person can be considered to be a Bernoulli process with probability \( q \) of success, and therefore the number of 'Yes' voters is normally distributed with mean \( nq \) and standard deviation \( \sqrt{nq(1-q)} \).

\[ e.g. \text{https://terrytao.wordpress.com/2008/10/10/small-samples-and-the-margin-of-error/} \]
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 \( nq + \sigma \).
Consider a sample poll of \( n=1000 \) people, of whom \( k=550 \) answered 'Yes' to some question.

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

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 be off due to finite sample size.

Redo poll many times and average the values of \( p \), gets closer and closer to the underlying probability \( q \).
We can consider either the standard deviation of the number count, or the standard deviation of the percentage.

Since \( p = \frac{k}{n} \), 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: \( \sqrt{n} / n = \sqrt{n} / (\sqrt{n})^2 = 1 / \sqrt{n} \) )
Number count: $\sqrt{1000 \times .55 \times .45} = 550 \pm 15.73$

as Percentage: $\frac{550}{1000} \pm \sqrt{.55 \times .45 / 1000} = .55 \pm .016$

$= 55\% \pm 1.6\%$

(just divide everything by 1000)

here 1.6\% is 1 standard deviation, so the “68\% confidence level”

90\% confidence level would be 1.645 standard deviations, so

$55\% \pm 2.6\%$ (90\% confidence)