Lecture 14: One distribution to rule them all. Part 1. Normal approximation to the binomial distribution

CSCI2244-Randomness and Computation

April 9, 2019

1 Some pictures

Figure 1 shows the PMFs of the random variables $S_{n,p}$ for $n = 40, 80, 100$, and $p = 0.5, 0.35, 0.5.$ where $S_{n,p}$ denotes the number of heads on n tosses of a coin with heads probability p. These PMFs are given by the binomial probability distribution

$$P(S_{n,p} = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The three random variables of course all have different expected values (20, 28 and 50, respectively) so the PMFs are nonzero on different parts of the number line. By our previous calculations, the standard deviation of $S_{n,p}$ is $\sqrt{np(1-p)}$, so these three random variables have standard deviations 3.16, 4.27 and 5, respectively.

While these are discrete distributions, you can imagine making a smooth curve by connecting the tops of the stems in each of the three plots. Since the stems in this case are 1 unit apart, and the sum of the heights of the stems is 1, the area under each of the three imaginary curves is 1, and so they represent the graphs of densities.

We transform these random variables. In general, if X is a random variable with mean $E(X) = \mu$, and m is a constant, then $X - m$ is a random variable with the same variance as X, and with mean $E(X - m) = E(X) - E(m) = \mu - m.$ In particular, $X - \mu$ has mean 0. The effect on the graph of the PMF of X (or, in the continuous case, the PDF) is to shift the graph a distance μ to the left. Otherwise the graph is the same.

In Figure 2, $S_{n,p}$ is transformed in this way to $S_{n,p} - np$, so that all three random variables have expected value 0.

Further, if X is a random variable with standard deviation σ, , then the standard deviation of the random variable X/s is σ/s. Thus if we start with X having mean μ and standard deviation σ, then we can apply both these transformations in succession and get a random variable $\frac{(X-\mu)}{\sigma}$ with mean 0 and standard deviation 1. The net effect is to shift by a distance of μ and to shrink horizontally by a factor of σ. Figure 3 shows the effect of these transformations, displaying the PMFs of

$$\frac{S_{n,P} - np}{\sqrt{np(1-p)}}.$$

In the last step, we destroyed the property that the area enclosed by each stem plot is 1: when we compressed the graph by a factor of σ, we divided this area by σ. To restore the property, we multiply the height of each graph by the standard deviation $\sigma = \sqrt{np(1-p)}$. Figure 4 shows the remarkable result— all the points appear to lie on the same smooth curve, which we have traced. What is this curve?

It was drawn by plotting the graph of

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2},$$

and the crucial result illustrated by these pictures is that this shape closely approximates the binomial distribution. In other words, this famous ‘bell curve’ represents a continuous probability density that is a kind of limiting case of the binomial distributions as n grows large.

2 The normal distribution

The function ϕ is called the *standard normal density*. ‘Standard’ here means that it has mean 0 and standard deviation 1. It’s not the least bit obvious that this is a density: you would need to show that

$$\int_{-\infty}^{\infty} e^{-x^2/2}dx = \sqrt{2\pi}.$$

The proof of this is not hard, but it cannot be obtained by usual integration techniques, because there is no closed form expression for an antiderivative of $e^{-x^2/2}$.

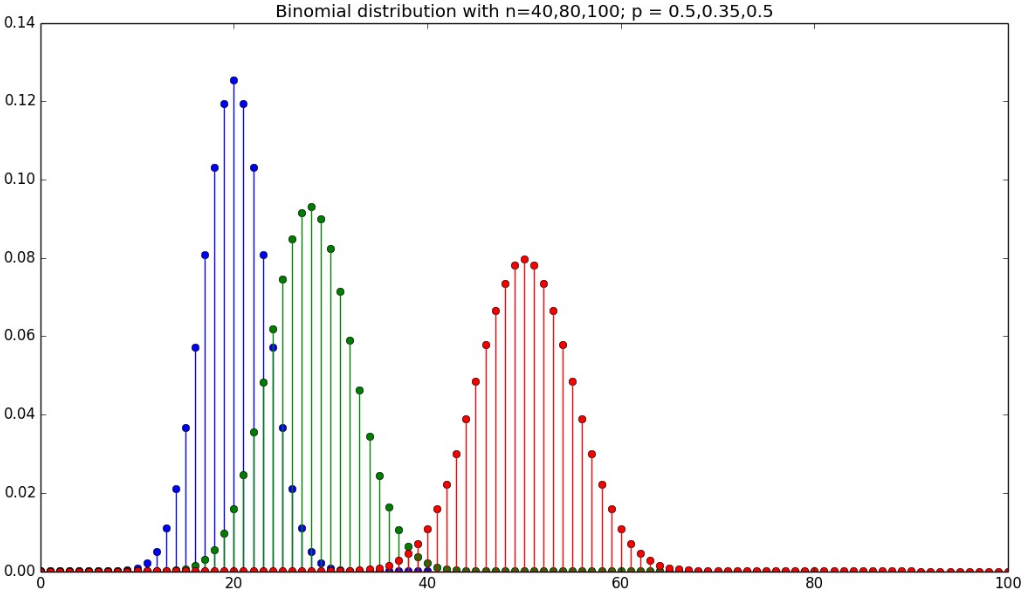

Figure 1: PMFs of binomial distribution with $n = 40, 80, 100$ and $p = 0.5, 0.35, 0.5$

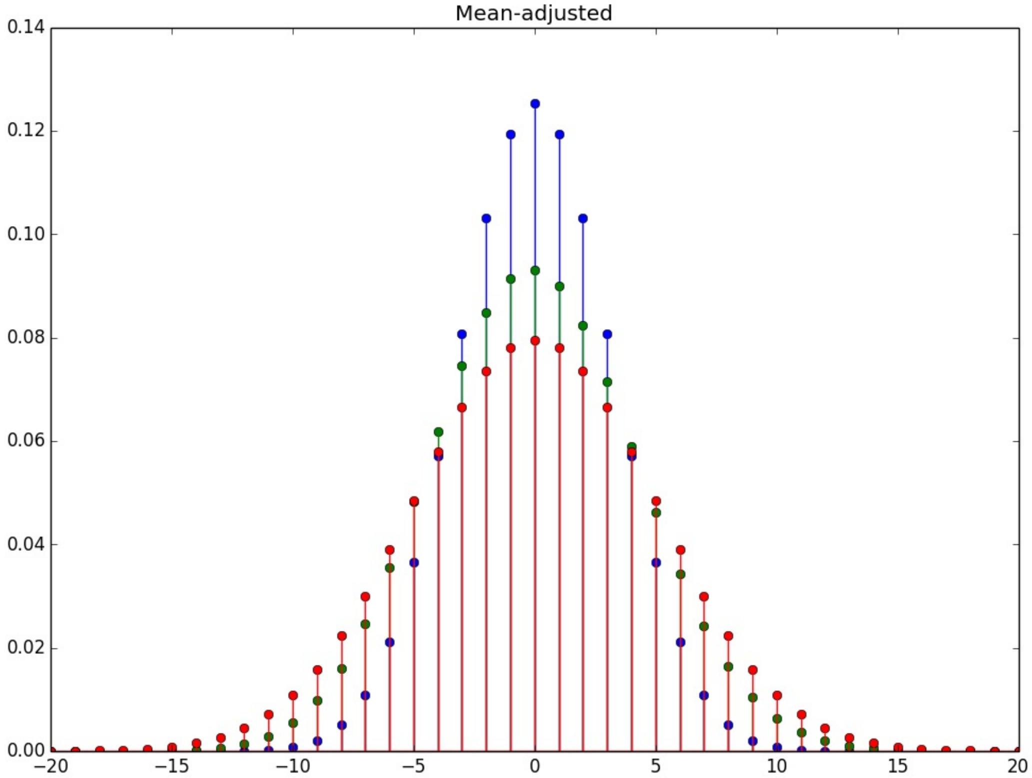

Figure 2: The same distributions shifted to all have mean 0...

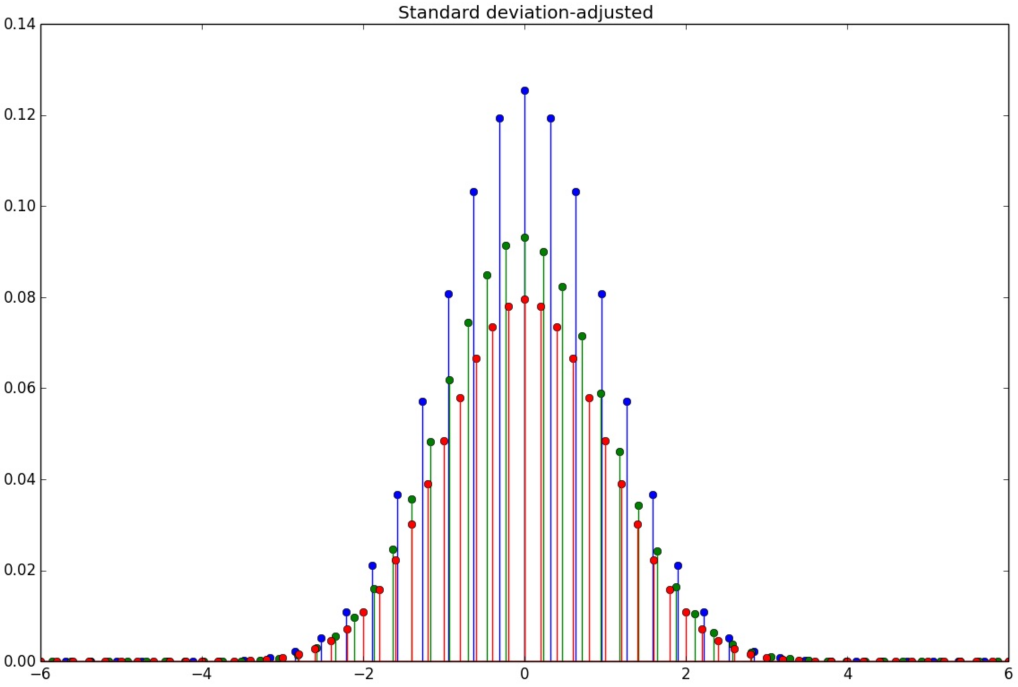

Figure 3: ...and scaled horizontally to have standard deviation 1

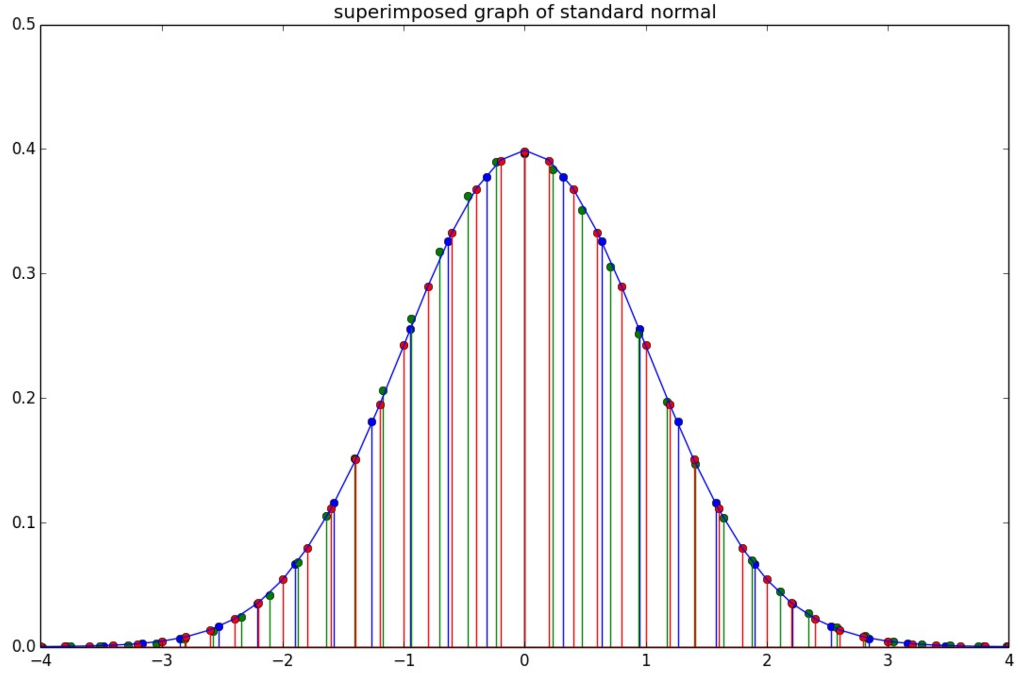

Figure 4: The previous figure stretched vertically by the standard deviation so that all the total area under each graph is 1, superimposed on the graph of $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

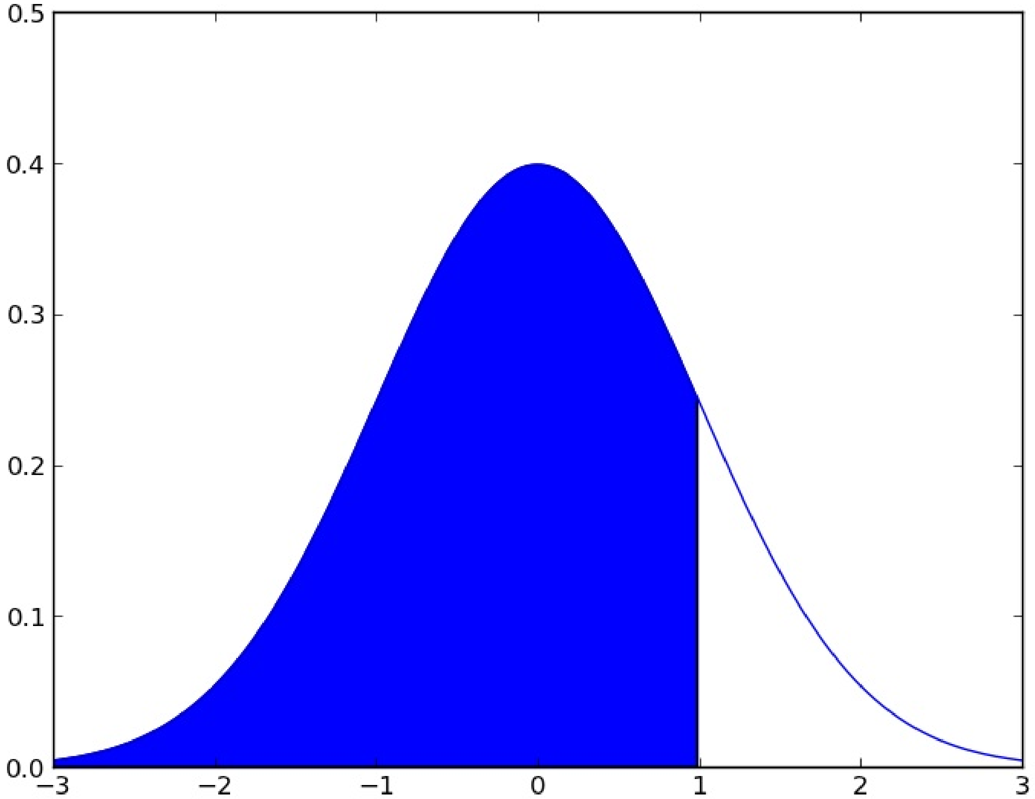

Figure 5: The standard normal density $\phi(x)$: the shaded area is $\Phi(1)$, the probability that the standard normal random variable has value less than 1.

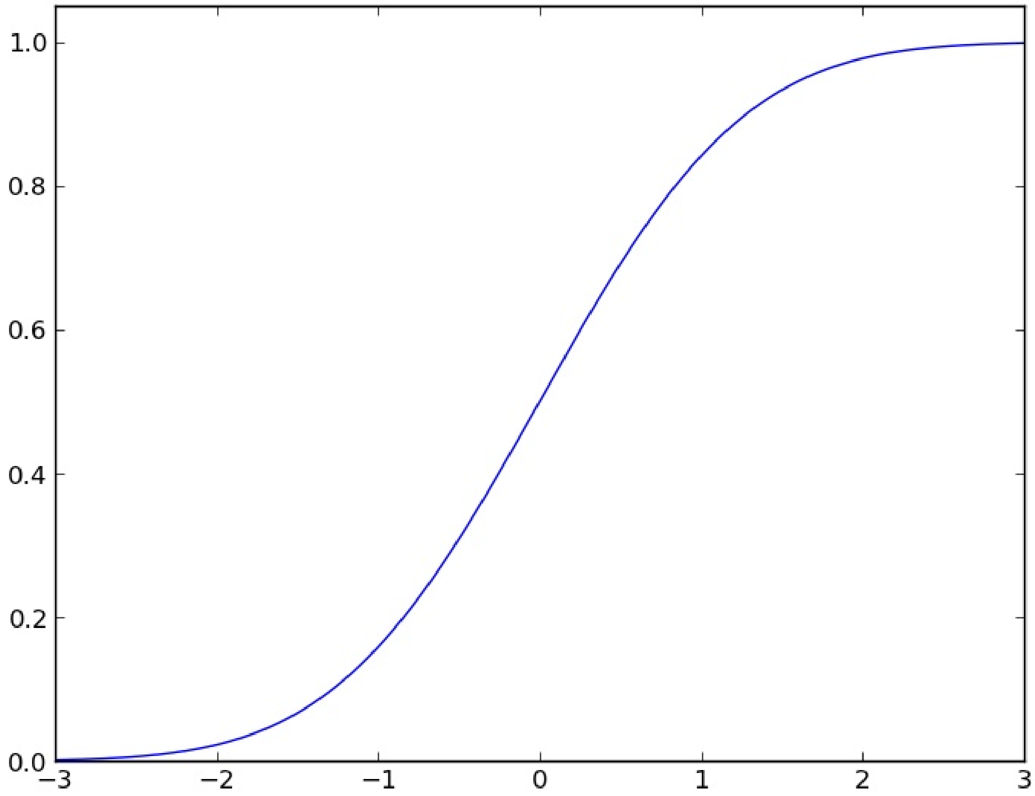

Figure 6: The cumulative normal density $\Phi(x)$.

The corresponding cumulative distribution function is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Since we cannot evaluate $\Phi(x)$ analytically, it has to be approximated numerically. You can compute $\Phi(x)$ in Python with

```
from scipy.stats import norm
norm.cdf(x)
```

You can similarly compute the inverse of the cumulative distribution function with `norm.ppf`.

If you drop the 'standard' part, there are infinitely many normal densities, one for each choice of mean and standard deviation. The normal density with mean μ and standard deviation σ is

$$\frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-x^2/2\sigma^2}.$$

3 Normal approximation to binomial distribution

Our observations above illustrate an important fact: the binomial distribution, adjusted to have mean 0 and standard deviation 1, is closely approximated by the normal distribution, especially as n gets larger.

The general principle is that

$$P(a \leq \frac{S_{n,p} - np}{\sqrt{np(1-p)}} \leq b) \approx \Phi(b) - \Phi(a).$$

In the appendix (which is non-required reading) there is something like an explanation of *why* this is true. Below we discuss what you can do with this fact.

Example. We use the normal approximation to the binomial distribution to estimate the probability that a fair coin tossed 100 times comes up heads between 45 and 55 times, inclusive. That is, we want to evaluate

$$P(45 \leq S_{100,0.5} \leq 55).$$

Let us first note an issue. Since $S_{n,p}$ is a discrete random variable that only takes integer values, the probability above is identical to

$$P(44.5 \leq S_{100,0.5} \leq 55.5)$$

and even

$$P(44 < S_{100,0.5} < 56),$$

but these will give us three different answers if we use them as the basis for calculating the normal approximation. It turns out that we get the best results by going exactly halfway between the integer values. We then have

$$\begin{aligned} P(44.5 \leq S_{100,0.5} \leq 55.5) &= P(\frac{44.5-50}{\sqrt{100\times 0.25}} \leq \frac{S_{100,p}-50}{\sqrt{100\times 0.25}} \leq \frac{55.5-50}{\sqrt{100\times 0.25}}) \\ &= P(-1.1 \leq \frac{S_{100,p}-50}{\sqrt{100\times 0.25}} \leq 1.1) \\ &\approx \Phi(1.1)-\Phi(-1.1) \\ &= 0.728668. \end{aligned}$$

Because of the symmetry in the event, we could have also evaluated this as

$$1 - 2\cdot\Phi(-1.1).$$

For purposes of comparison, we actually can compute the exact value in this case. It is

$$2^{-100}\sum_{j=45}^{55}\binom{100}{j} \approx 0.728747,$$

so our approximation is accurate to four decimal digits. Incidentally, had we used 45 and 55 as the initial bounds, we would have gotten the approximation

$$\Phi(1)-\Phi(-1) = 0.682689,$$

which gives only one decimal digit of accuracy!

If we wanted to know the probability that we get more than 55 heads, we can use the result of the above calculation, subtracting from 1 and dividing by 2. (Or, what is the same thing, taking $\Phi(-1.1)$.) The result is 0.13566.

Let us ask more generally, what is the probability of getting between 45% and 55% heads on N tosses of the coin, for large N? If we repeat the above calculation, the lower limit -1.1 in the approximation gets replaced by

$$(0.45N - \frac{1}{2N} - N/2)/(\sqrt(N)/2)$$

so the probability is approximately

$$1 - 2\Phi(-0.1\sqrt{N} - \frac{1}{N\sqrt{N}}).$$

Since $-0.1\sqrt{N} - \frac{1}{N\sqrt{N}}$ tends to $-\infty$ as N gets large, the probability approaches 1. We already knew this, thanks to the law of large numbers, but now we can estimate it very accurately. For example, with $N = 1000$ we get 0.9986, which agrees with the exact answer to 4 decimal places.

Example. This example concerns election polling. A large number N of people have voted in an election in which there are two candidates, A and B. We sample S voters and determine the number k of these S voters who voted for A. The value

$$\bar{p} = k/S$$

serves as the estimate of the proportion p of voters in the entire population who voted for A.

How large should we make S? Let's say we have a target accuracy—we would like our estimate to be within 3% of the actual vote, so that if, say, 54% of the total population voted for A, then $\bar{p}$ would be between 0.51 and 0.57. That is, we would like to be sure that

$$|\bar{p} - p| \leq 0.03.$$

Of course, there is no way we can be absolutely sure of this without sampling almost all the voters in the population. Instead, noting that $\bar{p}$ is a random variable, we will try to choose k such that

$$P(|\bar{p} - p| \leq 0.03) \geq 0.95.$$

The threshold 95% is frequently used in such calculations. This is referred to as the confidence level, and 3% as the margin of error.

Since the sampling is done without replacement, k has a hypergeometric distribution, which depends on both N and S. But, as we've observed before, since N is very large, k is for all practical purposes a binomial random variable that depends only on S and p. Thus

$$\frac{k - Sp}{\sqrt{Sp(1-p)}}$$

is well approximated by a standard normal random variable. We have

$$\begin{aligned}\frac{k - Sp}{\sqrt{Sp(1-p)}} &= \frac{S}{\sqrt{Sp(1-p)}} \cdot \left(\frac{k}{S} - p\right) \\ &= \sqrt{\frac{S}{p(1-p)}} \cdot (\bar{p} - p).\end{aligned}$$

So this last expression is a random variable with approximately the standard normal distribution. We thus have

$$\begin{aligned}P(|\bar{p} - p| \leq 0.03) &= P(-0.03 \leq \bar{p} - p \leq 0.03) \\ &= P\left(-0.03\sqrt{\frac{S}{p(1-p)}} \leq \sqrt{\frac{S}{p(1-p)}} \cdot (\bar{p} - p) \leq 0.03\sqrt{\frac{S}{p(1-p)}}\right) \\ &\approx \Phi\left(0.03\sqrt{\frac{S}{p(1-p)}}\right) - \Phi\left(-0.03\sqrt{\frac{S}{p(1-p)}}\right) \\ &= 1 - 2\Phi\left(-0.03\sqrt{\frac{S}{p(1-p)}}\right).\end{aligned}$$

Thus we need to solve

$$1 - 2\Phi\left(-0.03\sqrt{\frac{S}{p(1-p)}}\right) \geq 0.95,$$

that is,

$$\Phi\left(-0.03\sqrt{\frac{S}{p(1-p)}}\right) \leq 0.025.$$

We can could use the inverse normal cdf for this. This result is

$$0.03\sqrt{\frac{S}{p(1-p)}} \approx 1.96.$$

The usual practice is to call this 2, and to treat ± 2 standard deviations as the '95% confidence interval'. In fact this gives 95.4% confidence.

So now we need to solve

$$0.03\sqrt{\frac{S}{p(1-p)}} \geq 2.$$

We don't know what p is, but if $0 \leq p \leq 1$, then $p(1-p) \leq \frac{1}{4}$. Thus it is sufficient to solve

$$0.03\sqrt{4S} \geq 2,$$

which gives $S \approx 1111$. So say, candidate A receives 54% of the total vote and you poll 1200 voters. Approximately 95% of the time you perform this experiment, you will find that between 51% and 57% of the voters in your sample voted for candidate A.

4 Nonrequired reading for the mathematically curious: The standard normal density really is a density and it really is standard

While we don't have a simple closed formula for the integral of e^{-x^2}, we can nonetheless evaluate the improper integral

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx.$$

We'll sneak up on the answer by finding the *volume* under the surface obtained when you rotate the graph of e^{-x^2} about the y-axis. The shape is shown below. In (x, y, z)-coordinates, the surface is the graph of the equation

$$z = e^{-(x^2+y^2)}.$$

There are two different approaches for computing this volume: We can take cross sections for each fixed value of x and add up (that is, integrate over) the areas of all these cross sections. Let's call the area of the cross section C_x. Then the volume is

$$V = \int_{-\infty}^{\infty} C_x dx.$$

Now the cross section is just the graph of $e^{-(x^2+y^2)}$ where x is fixed and y varies, so its area is

$$C_x = \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dy = e^{-x^2} \int_{-\infty}^{\infty} e^{-y^2} dy = e^{-x^2} \cdot I.$$

Thus

$$V = I \cdot \int_{-\infty}^{\infty} e^{-x^2} dx = I^2.$$

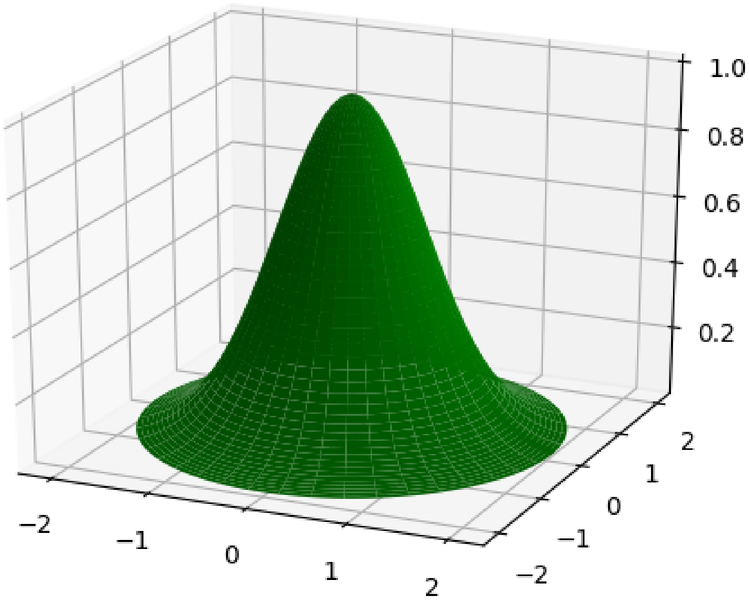

Figure 7: What you get when you rotate the graph of e^{-x^2} about the vertical axis.

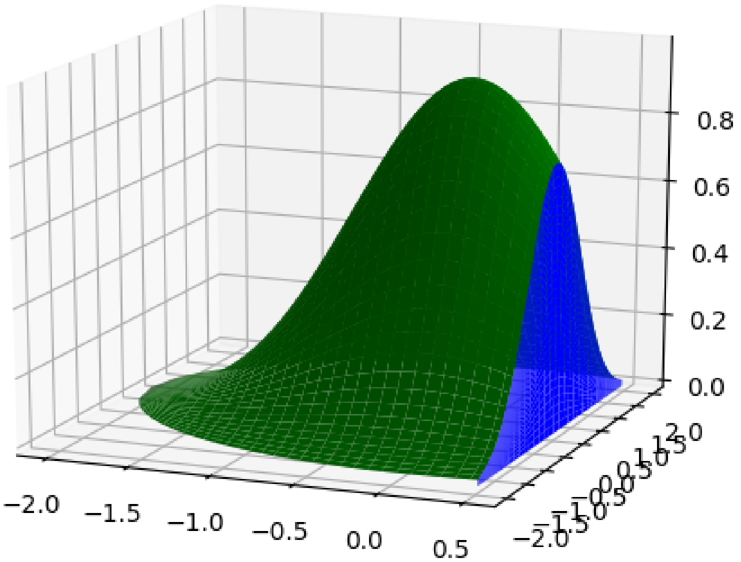

Figure 8: A planar cross-section obtained by fixing the x-coordinate

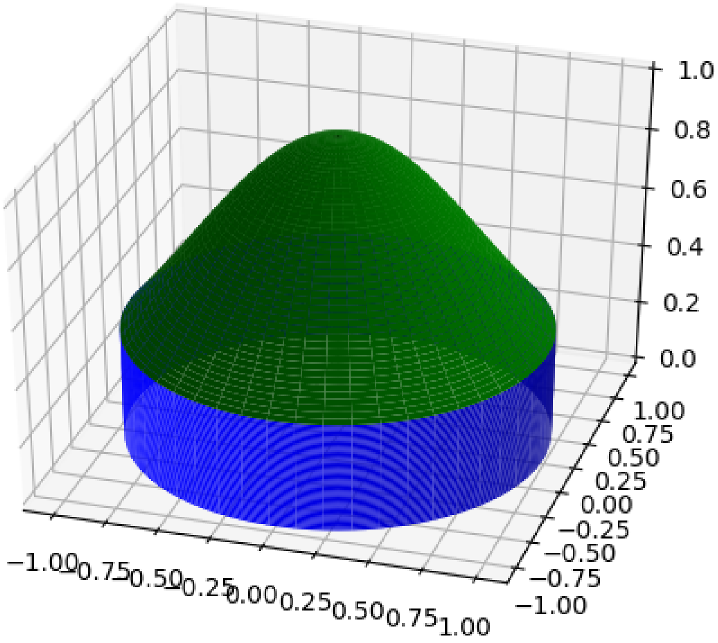

Figure 9: A cylindrical cross-section obtained by fixing the radius of rotation

The other way to obtain the volume is to integrate over the lateral areas of all the cylindrical cross-sections you get by fixing the distance $r > 0$ from the origin in the (x, y)-plane–this is a technique for finding the volumes of solids of rotation that you may have learned in calculus. The area C_r of the cylinder is $2\pi re^{-r^2}$ and thus

$$\begin{aligned} V &= \int_0^\infty 2\pi re^{-r^2} dr \\ &= -\pi \cdot e^{-r^2}\Big|_0^\infty \\ &= -\pi \cdot (0-1) \\ &= \pi. \end{aligned}$$

So $I^2 = \pi$ and thus $I = \sqrt{\pi}$.

We find

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$$

by making a change of variables $u = \frac{x}{\sqrt{2}}$ and applying the above result. Thus

$$\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

integrates to 1: The standard normal density really is a density.

What is the variance? Since $E(X) = 0$, the variance is just

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{\frac{-x^2}{2}} dx.$$

We integrate by parts, setting $u = x$, $dv = xe^{\frac{-x^2}{2}} dx$, so $du = dx$, $v = -e^{\frac{-x^2}{2}}$. The formula for integration by parts gives

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \cdot \left(xe^{\frac{-x^2}{2}}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx \right) &= \\ \frac{1}{\sqrt{2\pi}} (0 + \sqrt{2\pi}) &= 1. \end{aligned}$$

So the standard normal density really is ‘standard’, in the sense that its standard deviation is 1.