Normal Distribution

A Logically Developed Introduction

Yousef Ibrahim

August 2023

1 Introduction

The Normal distribution (also known as the Gaussian distribution and the Bell distribution) is a theoretical model (distribution) that describes the possible values of a continuous random variable and how often they occur. A continuous random variable is a variable that can take an infinite number of values between two limits. The normal curve was first derived by the French mathematician Abraham de Moivre in 1733. He was studying the distribution of errors in astronomical measurements. The normal curve was later independently derived by the German mathematician Carl Friedrich Gauss in 1809. Gauss used the normal curve to study the distribution of errors in measurements of physical quantities.

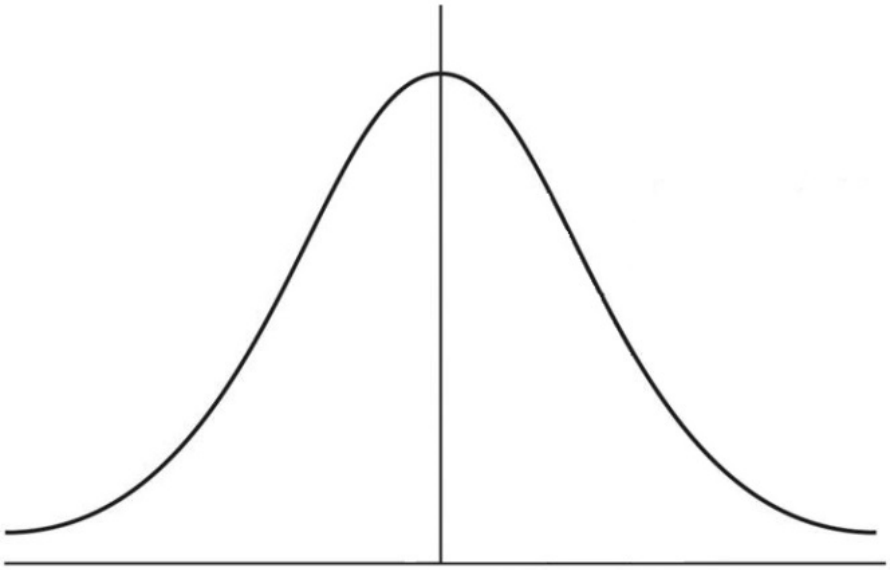
2 What it represents

The normal distribution, as we have previously said, models the distribution of values for a certain continuous random variable, such as height. If we know that a certain continuous random variable is normally distributed, that is, it is well modelled by our theoretical normal distribution, we can plot the Normal curve, which is the curve corresponding to the normal equation, with the equation

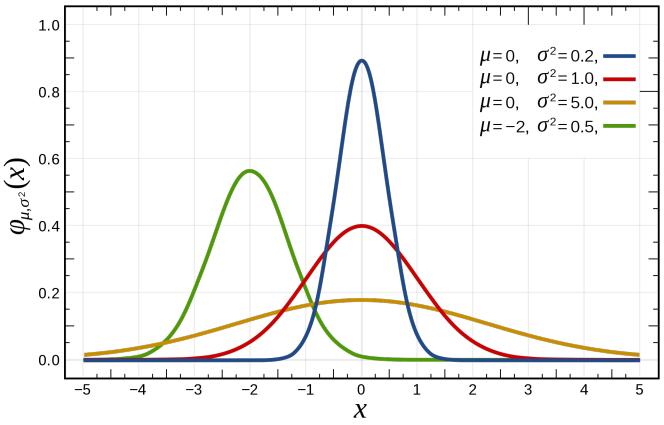
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)}$$

where μ is the mean/median/mode (they are all equal) value of the continuous random variable, and σ is the standard deviation of that continuous random variable. It is important to note that the normal distribution is a symmetric.

Plotting the curve, we get something like this [1]



But wait, the mean and standard deviation of the normal distribution can change! Therefore, the shape of the curve changes depending on the mean and standard deviation.



The mean of the continuous random variable dictates the value on the horizontal axis at which the peak occurs. The standard deviation of the continuous random variable determines the *spread* of values of the continuous random variable; hence, the standard deviation determines the spread of the curve itself.

It is important to know that the area under the curve between two points x_1 and x_2 represents the probability that the random variable takes a value between x_1 and x_2.

[1] Note that we truncate the curve. The actual normal curve extends to positive and negative infinity; it is theoretically possible for the continuous random variable to be *any* value

Before proceeding to the next section, we will introduce the notation.

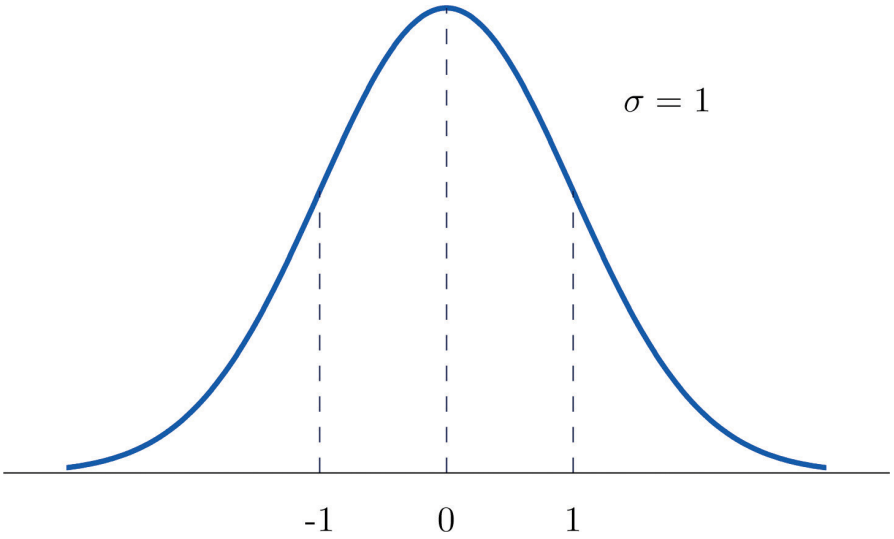
$$X \sim N(\mu, \sigma^2)$$

It reads, "the continuous random variable X is normally distributed with a mean μ and a variance of σ^2."

3 The standard normal curve

It is very difficult to deal with the infinite number of normal curves, as they vary according to the mean and standard deviation. To fix this problem, we *convert* the given normal distribution into the *standard* normal distribution. The standard normal distribution is the normal distribution with mean equal to 0 and standard deviation equal to 1. *Why* we use a standard normal distribution will be made more clear shortly; for now, the standard normal distribution simplifies calculations.

This is the standard normal curve



Now, you may question *why* this works; how can we convert one distribution? We will now introduce a very important property of *all* normal distributions. The area under the curve for all normal distributions between $\mu - k\sigma$ and $\mu + k\sigma$ for all real values of k is constant; this is inherent in the formulation of all normal distributions. The next section will bind everything together.

4 Calculating probabilities

We will now learn how to "convert" normal distributions to the standard normal distribution.

if

$$X \sim N(\mu, \sigma^2)$$

then

$$Z \sim N(0, 1^2)$$

where

$$Z = \frac{X - \mu}{\sigma}$$

Essentially, we are finding out how many standard deviation our X value lies from the mean, μ. Since we know that the probability of a normally distributed random variable being k times above or below the mean is fixed, knowing how many standard deviation above or below the mean our given value is will enable us to find the required probability. We can now deduce that the continuous random variable Z represents the number of standard deviations away from the mean for values of X. Since 50% of values are less than or equal to μ then 50% of values lie 0 or less standard deviations behind μ. The same is true the other way around. Hence the median value of Z is 0. By extension, the mean value of Z is also 0, as we know that Z is normally distributed beforehand.

$$X - \mu$$

finds the difference between the X value and the mean. Dividing by σ, we find out *how many* standard deviations our value is from the mean.

But wait, how do we know the probability of getting a value that is above or below the mean by $k\sigma$ standard deviations or less? This is where the standard normal table come into play. It is basically a table with z-scores (number of standard deviations) and corresponding values of $P(Z \leq z)$. It is customary to denote $P(Z \leq z)$ by $\Phi(z)$. The table is given in the next page. Notice how only positive z values are given. Using the symmetry of the distribution, we can easily find values for negative z scores.

The Normal Distribution Function

The function tabulated below is $\Phi(z)$, defined as $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}t^2} dt$.

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.00	0.5000	0.50	0.6915	1.00	0.8413	1.50	0.9332	2.00	0.9772
0.01	0.5040	0.51	0.6950	1.01	0.8438	1.51	0.9345	2.02	0.9783
0.02	0.5080	0.52	0.6985	1.02	0.8461	1.52	0.9357	2.04	0.9793
0.03	0.5120	0.53	0.7019	1.03	0.8485	1.53	0.9370	2.06	0.9803
0.04	0.5160	0.54	0.7054	1.04	0.8508	1.54	0.9382	2.08	0.9812
0.05	0.5199	0.55	0.7088	1.05	0.8531	1.55	0.9394	2.10	0.9821
0.06	0.5239	0.56	0.7123	1.06	0.8554	1.56	0.9406	2.12	0.9830
0.07	0.5279	0.57	0.7157	1.07	0.8577	1.57	0.9418	2.14	0.9838
0.08	0.5319	0.58	0.7190	1.08	0.8599	1.58	0.9429	2.16	0.9846
0.09	0.5359	0.59	0.7224	1.09	0.8621	1.59	0.9441	2.18	0.9854
0.10	0.5398	0.60	0.7257	1.10	0.8643	1.60	0.9452	2.20	0.9861
0.11	0.5438	0.61	0.7291	1.11	0.8665	1.61	0.9463	2.22	0.9868
0.12	0.5478	0.62	0.7324	1.12	0.8686	1.62	0.9474	2.24	0.9875
0.13	0.5517	0.63	0.7357	1.13	0.8708	1.63	0.9484	2.26	0.9881
0.14	0.5557	0.64	0.7389	1.14	0.8729	1.64	0.9495	2.28	0.9887
0.15	0.5596	0.65	0.7422	1.15	0.8749	1.65	0.9505	2.30	0.9893
0.16	0.5636	0.66	0.7454	1.16	0.8770	1.66	0.9515	2.32	0.9898
0.17	0.5675	0.67	0.7486	1.17	0.8790	1.67	0.9525	2.34	0.9904
0.18	0.5714	0.68	0.7517	1.18	0.8810	1.68	0.9535	2.36	0.9909
0.19	0.5753	0.69	0.7549	1.19	0.8830	1.69	0.9545	2.38	0.9913
0.20	0.5793	0.70	0.7580	1.20	0.8849	1.70	0.9554	2.40	0.9918
0.21	0.5832	0.71	0.7611	1.21	0.8869	1.71	0.9564	2.42	0.9922
0.22	0.5871	0.72	0.7642	1.22	0.8888	1.72	0.9573	2.44	0.9927
0.23	0.5910	0.73	0.7673	1.23	0.8907	1.73	0.9582	2.46	0.9931
0.24	0.5948	0.74	0.7704	1.24	0.8925	1.74	0.9591	2.48	0.9934
0.25	0.5987	0.75	0.7734	1.25	0.8944	1.75	0.9599	2.50	0.9938
0.26	0.6026	0.76	0.7764	1.26	0.8962	1.76	0.9608	2.55	0.9946
0.27	0.6064	0.77	0.7794	1.27	0.8980	1.77	0.9616	2.60	0.9953
0.28	0.6103	0.78	0.7823	1.28	0.8997	1.78	0.9625	2.65	0.9960
0.29	0.6141	0.79	0.7852	1.29	0.9015	1.79	0.9633	2.70	0.9965
0.30	0.6179	0.80	0.7881	1.30	0.9032	1.80	0.9641	2.75	0.9970
0.31	0.6217	0.81	0.7910	1.31	0.9049	1.81	0.9649	2.80	0.9974
0.32	0.6255	0.82	0.7939	1.32	0.9066	1.82	0.9656	2.85	0.9978
0.33	0.6293	0.83	0.7967	1.33	0.9082	1.83	0.9664	2.90	0.9981
0.34	0.6331	0.84	0.7995	1.34	0.9099	1.84	0.9671	2.95	0.9984
0.35	0.6368	0.85	0.8023	1.35	0.9115	1.85	0.9678	3.00	0.9987
0.36	0.6406	0.86	0.8051	1.36	0.9131	1.86	0.9686	3.05	0.9989
0.37	0.6443	0.87	0.8078	1.37	0.9147	1.87	0.9693	3.10	0.9990
0.38	0.6480	0.88	0.8106	1.38	0.9162	1.88	0.9699	3.15	0.9992
0.39	0.6517	0.89	0.8133	1.39	0.9177	1.89	0.9706	3.20	0.9993
0.40	0.6554	0.90	0.8159	1.40	0.9192	1.90	0.9713	3.25	0.9994
0.41	0.6591	0.91	0.8186	1.41	0.9207	1.91	0.9719	3.30	0.9995
0.42	0.6628	0.92	0.8212	1.42	0.9222	1.92	0.9726	3.35	0.9996
0.43	0.6664	0.93	0.8238	1.43	0.9236	1.93	0.9732	3.40	0.9997
0.44	0.6700	0.94	0.8264	1.44	0.9251	1.94	0.9738	3.50	0.9998
0.45	0.6736	0.95	0.8289	1.45	0.9265	1.95	0.9744	3.60	0.9998
0.46	0.6772	0.96	0.8315	1.46	0.9279	1.96	0.9750	3.70	0.9999
0.47	0.6808	0.97	0.8340	1.47	0.9292	1.97	0.9756	3.80	0.9999
0.48	0.6844	0.98	0.8365	1.48	0.9306	1.98	0.9761	3.90	1.0000
0.49	0.6879	0.99	0.8389	1.49	0.9319	1.99	0.9767	4.00	1.0000
0.50	0.6915	1.00	0.8413	1.50	0.9332	2.00	0.9772		

Question: The weights of packages that arrive at a factory are normally distributed with a mean of 18kg and a standard deviation of 5.4 kg

a) Find the probability that a randomly selected package weighs less than 10kg

Solution:

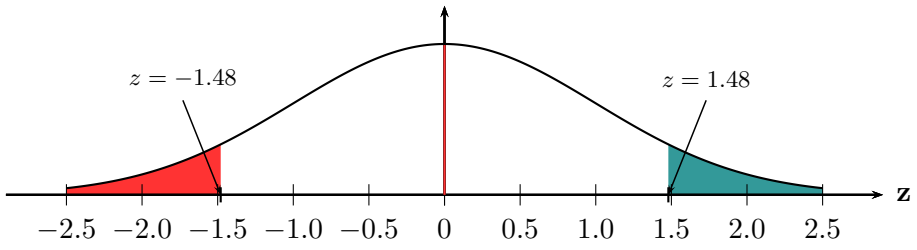
$X \sim N(18, 5.4^2)$ and $Z \sim N(0, 1^2)$ where $Z = \dfrac{X - \mu}{\sigma}$

$$\begin{aligned} P(X \leq 10) &= P\left(Z \leq \frac{10 - 18}{5.4}\right) \\ &= P\left(Z \leq -1.48\right) \\ &= P\left(Z \geq 1.48\right) \end{aligned}$$

Since the total probability is 1

$$\begin{aligned} &= 1 - P\left(Z \leq 1.48\right) \\ &= 1 - 0.9306 \\ &= 0.0694 \end{aligned}$$

We can easily see this if we sketch the normal curve. Note that the use of strict and non-strict inequalities is identical as $P(Z = z) = 0$ is always true; hence, including or excluding a number is not significant. Generally, we can say that



$$P(Z > z) = 1 - P(Z < z)$$

and

$$P(Z > -z) = P(Z < z)$$

The heaviest 15% of packages are moved around the factory by Jemima using a forklift truck.

b) Find the weight, in kg, of the lightest of these packages that Jemima will move.

Solution:

$P(X \geq x) = 15\%$ where x is the weight of the lightest package carried by a forklift.

We must use the following table:

Percentage Points Of The Normal Distribution

The values z in the table are those which a random variable $Z \sim N(0, 1)$ exceeds with probability p; that is, $P(Z > z) = 1 - \Phi(z) = p$.

p	z	p	z
0.5000	0.0000	0.0500	1.6449
0.4000	0.2533	0.0250	1.9600
0.3000	0.5244	0.0100	2.3263
0.2000	0.8416	0.0050	2.5758
0.1500	1.0364	0.0010	3.0902
0.1000	1.2816	0.0005	3.2905

$$P(X \geq x) = P\left(Z \geq \frac{x - 18}{5.4}\right)$$

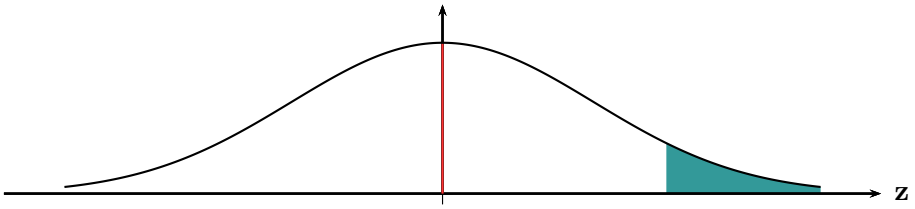
$$P\left(Z \geq \frac{x - 18}{5.4}\right) = 0.15$$

$$\frac{x - 18}{5.4} = 1.0364$$

$$\vdots$$

$$x = 23.6$$

If we wanted to sketch the curve, we would get something like this. This certainly helps in confirming that our z value is indeed above the mean.



One of the packages not moved by Jemima is selected at random.

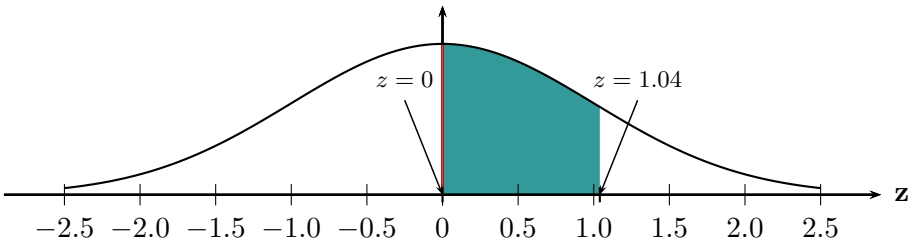
c) Find the probability that it weighs more than 18kg

Solution:

This is a very obvious use of conditional probability. We can reword the question into, "Find the probability that $X > 18$ given that X is *not* moved by Jemima, i.e. given that $X < 23.6$. Hence

$$\begin{aligned} P\left(X > 18 | X < 23.6\right) &= \frac{P(18 < X < 23.6)}{P(X < 23.6)} \\ &= \frac{P(\frac{18-18}{5.4} < Z < \frac{23.6-18}{5.4})}{P(Z < \frac{23.6-18}{5.4})} \\ &= \frac{P\left(0 < Z < 1.037\right)}{P\left(Z < 1.037\right)} \end{aligned}$$

Sketching our curve to get a clearer picture of $P\,(0 < Z < 1.037)$:



We can now clearly see that

$$\begin{aligned} P(0 < Z < 1.04) &= P(Z < 1.04) - P(Z < 0) \\ &= 0.8508 - 0.5 \\ &= 0.351 \end{aligned}$$

and we have found that

$$P(X < 1.04) = 0.8508$$

$$\begin{aligned} \therefore P\,(X > 18 | X < 23.6) &= \frac{0.351}{0.8508} \\ &= 0.413 \end{aligned}$$

A delivery of 4 packages is made to the factory. The weights of the packages are independent

d) Find the probability that exactly 2 of them will be moved by Jemima

Solution:

$$\begin{aligned} P(\text{exactly 2 of the 4 packages will be moved by Jemima}) &= \binom{4}{2} \times 0.15^2 \times 0.85^2 \\ &= 0.975375 \end{aligned}$$

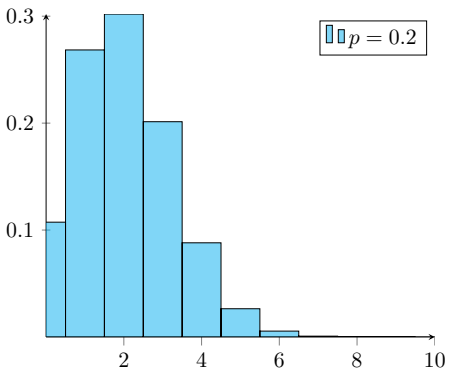
5 Approximating the Binomial distribution with the Normal distribution

This is for the Cambridge Probability and Statistics 1 syllabus *only*.

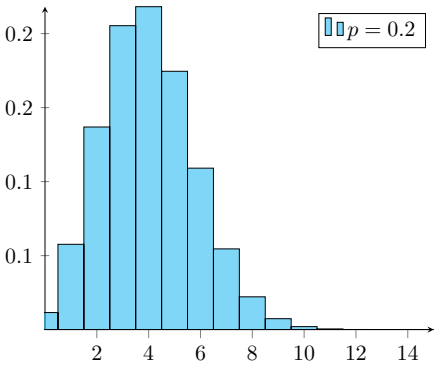
Let us make a table discussing the similarities and differences between the Binomial distribution and the Normal distribution.

Characteristic	Normal distribution	Binomial distribution
Random Variable	Continuous	Discrete
Symmetric	Yes	symmetrical or asymmetrical
Probability of outcomes	Determined by the mean and standard deviation	Determined by the number of trials and the probability of success on each trial

This is an average looking normal distribution. The following plot is for $X \sim B(10, 0.2)$.

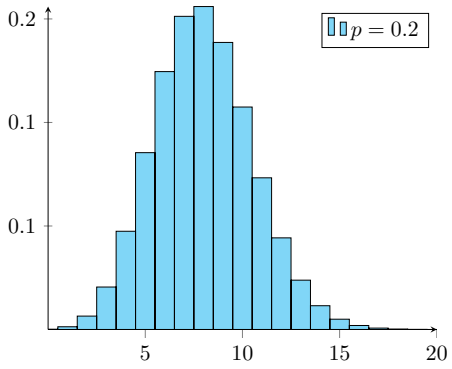


When we increase n to 20, we get something that may look familiar. The following plot shows $X \sim B(20, 0.2)$.



You would be correct if you are seeing a skewed normal distribution.

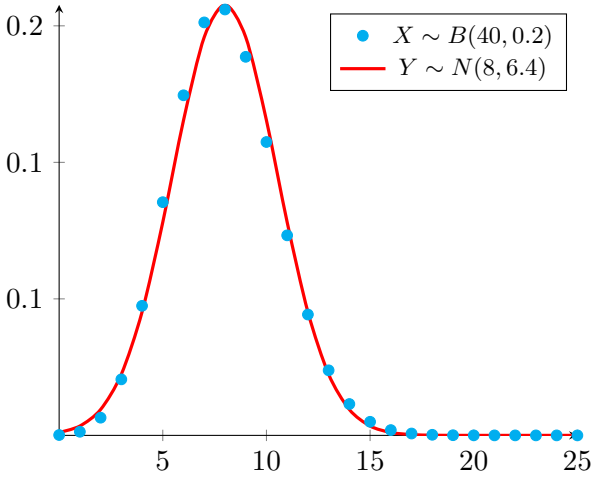
Let us see what happens when $n = 40$. The follow plot shows $X \sim B(40, 0.2)$. We will also superimpose the distribution of a continuous random variable Y, where $Y \sim N(8, 6.4)$



If we superimpose the distribution of the random continuous variable Y, where

$$Y \sim N(8, \sqrt{6.4})$$

, we will see something very interesting (We will only plot points).



We can see that our normal distribution perfectly fits our normal distribution! Therefore, we can say that $X \dot{\sim} N(8, 6.4)$, which reads as "X approximately follows a normal distribution with a mean 8 and standard deviation $\sqrt{6.4}$.[2]

[2] $\dot{\sim}$ is not a standard notation for "approximately distributed"; if you want to use it, you must define it beforehand

Note that when n was small ($n = 10$), our binomial distribution couldn't be modelled well with a normal distribution. Also, when $p = 0.5$, our binomial distribution is symmetrical for any sample size, so a normal distribution could fit in more easily. This brings us to the conditions of modelling a binomial distribution with a normal distribution.

Criteria:

If $X \sim B(n, p)$, then $X \dot{\sim} N(np, np(1-p))$ if and only if n sufficiently large to ensure that both $np > 5$ and $n(1-p) > 5$.

But why would we approximate a binomial distribution with a normal distribution anyways? We will only use it (due to syllabus restrictions) to find probabilities more easily, but we can also produce inferences about the proportion of a population from a sample with a binomial distribution via normal approximation.

However, remember that we are going from a discrete distribution, to a continuous distribution; this creates a problem as there will be some "holes" in the discrete distribution that we must correct. This is done by a *continuity correction*. We will explain this with an example.

Example: Anil is a candidate in an election. He received 40% of the votes. A random sample of 120 voters is chosen. Use an approximation to find the probability that, of the 120 voters, between 36 and 54 inclusive voted for Anil.

Solution:

First we must identify that the base distribution is a Binomial distribution: we have a fixed probability of success (or votes in this context) and independent Bernoulli trials.[3]

Let X represent the number of votes Anil receives in the election, where $X \sim B(120, 0.4)$. If you try doing it using binomial coefficients, it will take ages. Luckily, we can use the normal approximation as $np = 120 \times 0.4 = 48 > 5$ and $n(1-p) = 120 \times 0.6 = 72 > 5$.

$$\mu = np = 48$$

and

$$\sigma^2 = np(1-p) = 48 \times 0.6 = 28.8$$

so

$$X \dot{\sim} N(48, 28.8)$$

which reads, "X approximately follows a normal distribution with a mean 48 and variance 28.8".

[3] A Bernoulli trial is a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment is conducted. It is named after Jacob Bernoulli, a 17^{th}-century Swiss mathematician, who analyzed them in his *Ars Conjectandi*.

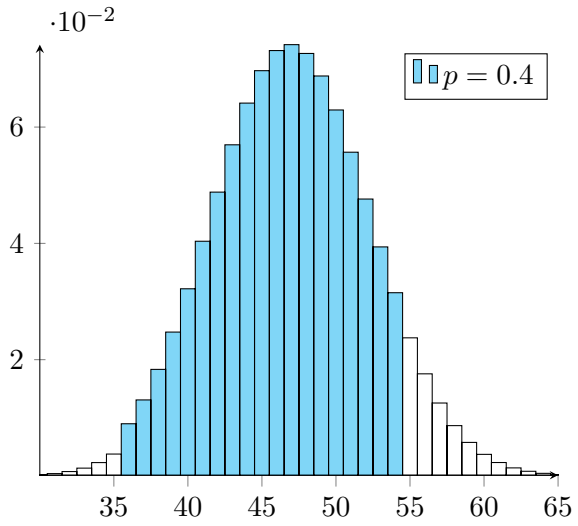
Our required probability is $P\left(36 \leq X \leq 54\right) \approx P\left(\frac{35.5-48}{\sqrt{28.8}} \leq Z \leq \frac{54.5-48}{\sqrt{28.8}}\right)$ where

$$Z \sim N(0,1).$$

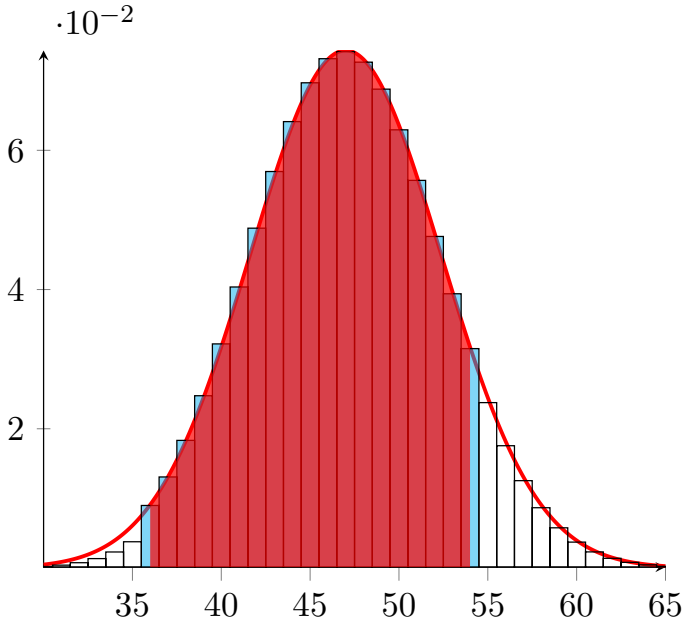
But wait! why aren't we taking $P\left(\frac{36-48}{\sqrt{28.8}} \leq Z \leq \frac{54-48}{\sqrt{28.8}}\right)$.

To explain this, we will make use of plotting.

The exact probability we want is shaded in the following plot of our binomial distribution.

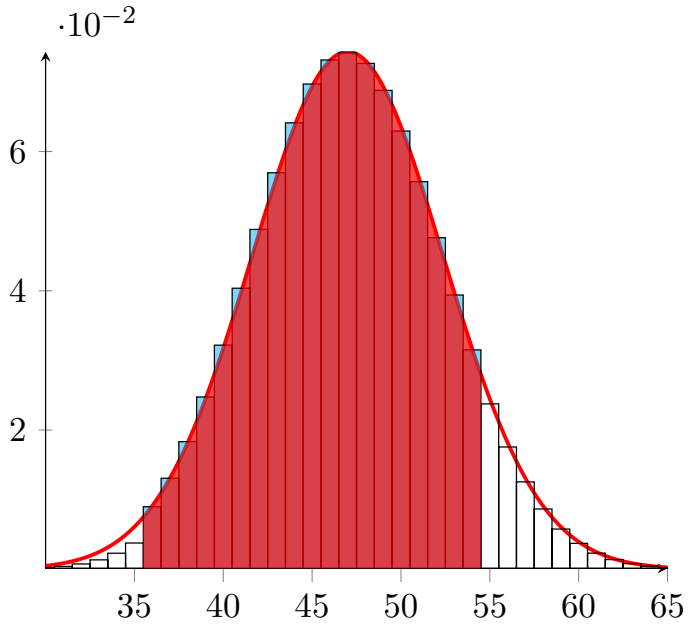


Superimposing the corresponding normal curve and highlighting the expected probability



Note how we have half a bar missing from both ends because the bar 36 in the discrete sense includes all the numbers from 35.5 till 36.5 in the continuous sense. Likewise, the bar 54 in the discrete sense includes all the numbers from 53.5 till 54.5 in the continuous sense.

Taking $P(36 \leq X \leq 54) \approx P\left(\frac{35.5-48}{\sqrt{28.8}} \leq Z \leq \frac{54.5-48}{\sqrt{28.8}}\right)$ instead,



we obtain a better approximation. If we want to formulate it, we subtract 0.5 from the lower bound in the discrete sense ($36 - 0.5 = 35.5$) and add 0.5 to the upper bound in the discrete sense ($54 + 0.5 = 54.5$). To find our answer all we have to do is compute $P(36 \leq X \leq 54) \approx P\left(\frac{35.5-48}{\sqrt{28.8}} \leq Z \leq \frac{54.5-48}{\sqrt{28.8}}\right)$, which we have previously discussed.