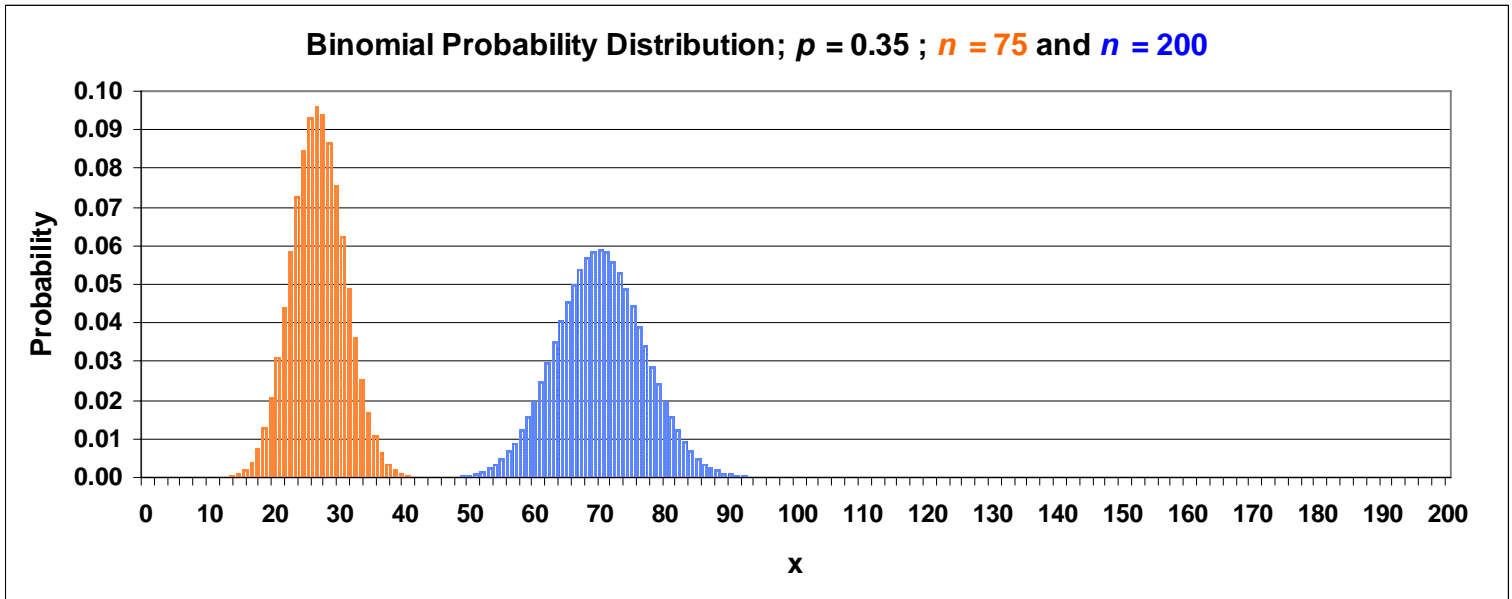


For a binomial distribution if n the number of trials is very large and the probability of a success, p , remains constant, the mean np and the standard deviation $\sigma_x = \sqrt{np(1-p)}$ both increase as n increases. If $n = 75$ and $p = 0.35$, the mean is 26.5 and the standard deviation is 4.13. If p stays constant but n is increased to 200, the center of the distribution shifts right to a mean of 70.0. The distribution also “spreads out” with a larger standard deviation of 6.75 and the probabilities for each category with the class width $\Delta x = 1$ become smaller. This is shown in the figure below.



To analyze the behavior as n becomes large it would be easier if the center and spread stayed constant. To achieve this

end we transform from x to z scores. Recall that $z = \frac{x - \mu_x}{\sigma_x}$, so that $\mu_z = \left\langle \frac{x - \mu_x}{\sigma_x} \right\rangle = \frac{1}{\sigma_x} (\langle x \rangle - \mu_x) = 0$ and

$$\sigma_z^2 = \left\langle \left[\frac{x - \mu_x}{\sigma_x} - \mu_z \right]^2 \right\rangle = \left\langle \left[\frac{x - \mu_x}{\sigma_x} \right]^2 \right\rangle = \frac{\langle (x - \mu_x)^2 \rangle}{\sigma_x^2} = \frac{\sigma_x^2}{\sigma_x^2} = 1. \text{ Thus, using } z \text{ scores the distribution stays centered at } 0$$

with a constant spread of 1. As n increases more categories are “squeezed” into the same space around $z = 0$, so that the distribution starts to approach a continuum. To analyze this limit as $n \rightarrow \infty$, consider the probability density which is the height of the rectangle for a given value of z in the histogram of the distribution. The total area under the

histogram represents the probability of the sample space and so must be one. $1 = \sum_{x=0}^n P(x) = \sum_{x=0}^n P(x) \Delta x$, where $\Delta x = 1$

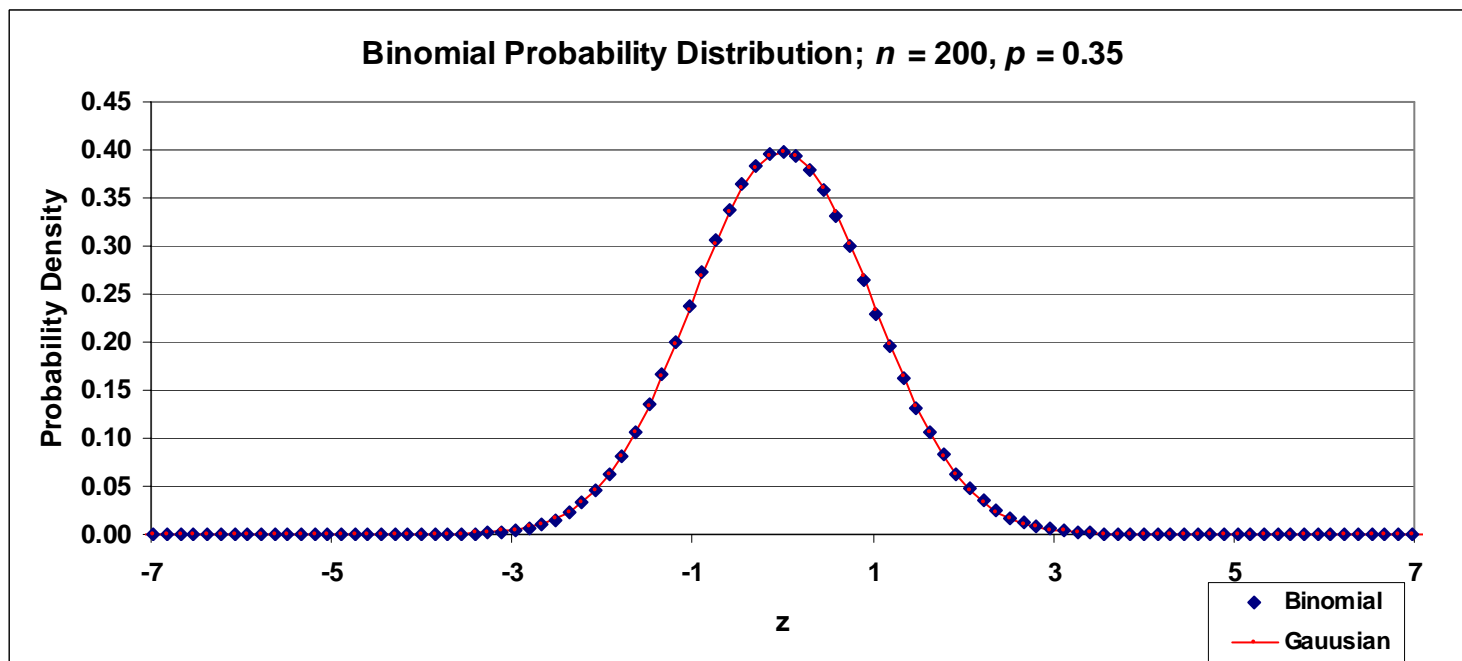
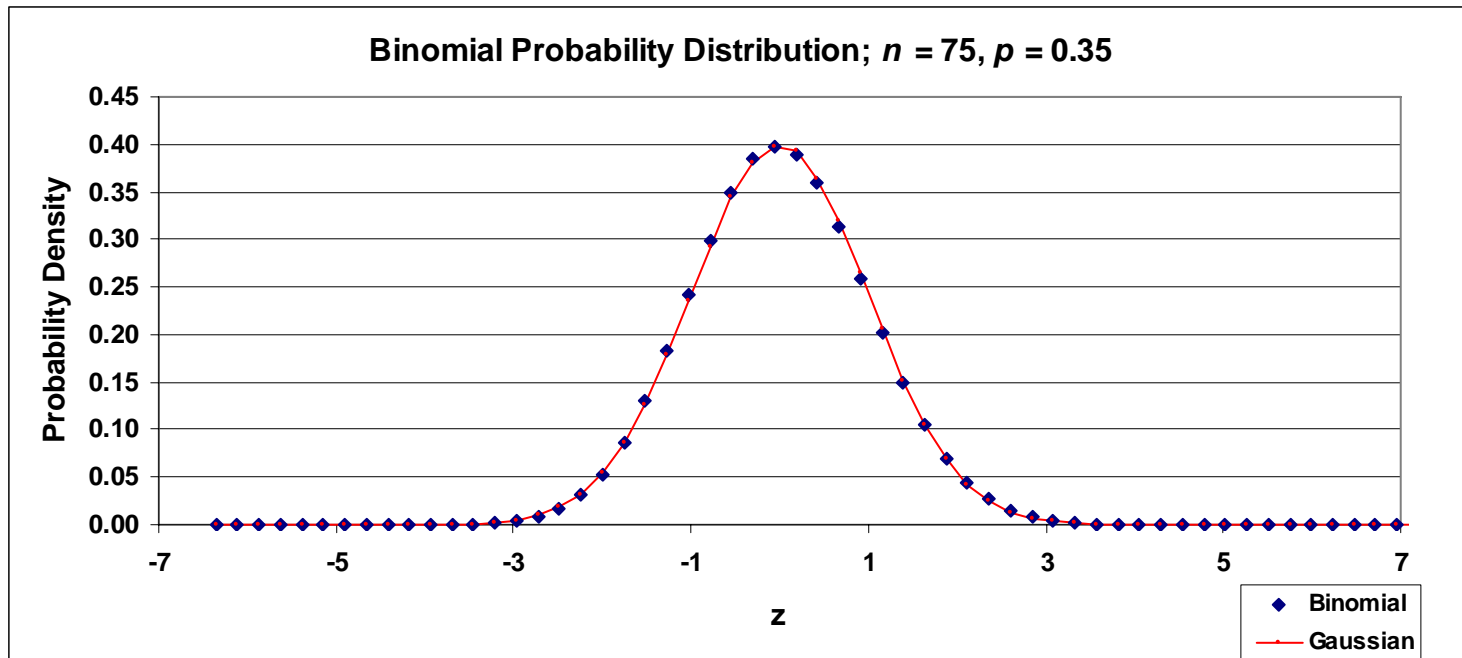
is the difference between consecutive values of x and is the width of each of the rectangles in the probability histogram. The area under the probability histogram of z must also equal one. The width of the rectangles in the z score histogram is the difference between consecutive z values and is given by

$$\Delta z = \Delta \left(\frac{x - \mu_x}{\sigma_x} \right) = \frac{\Delta x}{\sigma_x} = \frac{\Delta x}{\sqrt{np(1-p)}} = \frac{1}{\sqrt{np(1-p)}}.$$

So that the total area under the histogram remain 1, we define the probability density function $f(z)$ as

$$f(z) = \frac{P(x)}{\Delta z} = P(x) \sqrt{np(1-p)} = \sqrt{np(1-p)} \binom{n}{x} p^x (1-p)^{n-x}, \text{ where } x = \sigma_x z + \mu_x.$$

Then $\sum_z f(z)\Delta z = \sum_{x=0}^n P(x)\Delta x = 1.$



Now, the difference between two consecutive probabilities is given by

$$\begin{aligned} \Delta P &= P(x+1) - P(x) = \binom{n}{x+1} p^{x+1} (1-p)^{n-x-1} - \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} p^x (1-p)^{n-x} \left[\binom{n-x}{x+1} \left(\frac{p}{1-p} \right) - 1 \right] = \binom{n}{x} p^x (1-p)^{n-x} \left[\frac{np-xp}{(x+1)(1-p)} - \frac{x-xp+1-p}{(x+1)(1-p)} \right] \\ &= \binom{n}{x} p^x (1-p)^{n-x} \left[\frac{np-xp-x+xp-(1-p)}{(x+1)(1-p)} \right] = \binom{n}{x} p^x (1-p)^{n-x} \left[\frac{(np-x)-(1-p)}{(x+1)(1-p)} \right] \end{aligned}$$

So that the difference in the probability density function between two consecutive z scores is given by

$$\begin{aligned} \Delta f &= \Delta P \sqrt{np(1-p)} = \sqrt{np(1-p)} \binom{n}{x} p^x (1-p)^{n-x} \left[\frac{-z\sqrt{np(1-p)} - (1-p)}{(np + z\sqrt{np(1-p)} + 1)(1-p)} \right] \\ &= f(z) \left[\frac{-z\sqrt{np(1-p)} - (1-p)}{(np + z\sqrt{np(1-p)} + 1)(1-p)} \right] = -\frac{z\sqrt{np(1-p)}}{np(1-p)} f(z) \left[\frac{1 + \frac{1-p}{z\sqrt{np(1-p)}}}{1 + z\sqrt{\frac{(1-p)}{np}} + \frac{1}{np}} \right] \\ &= -\frac{z}{\sqrt{np(1-p)}} f(z) \left[\frac{1 + \frac{1-p}{z\sqrt{np(1-p)}}}{1 + z\sqrt{\frac{(1-p)}{np}} + \frac{1}{np}} \right] \end{aligned}$$

Thus, since $\Delta z = \Delta \left(\frac{x - \mu_x}{\sigma_x} \right) = \frac{1}{\sqrt{np(1-p)}}$,

$$\begin{aligned} \frac{\Delta f}{\Delta z} &= \frac{\Delta f}{1} \div \frac{1}{\sqrt{np(1-p)}} = \Delta f \sqrt{np(1-p)} = -\frac{z\sqrt{np(1-p)}}{\sqrt{np(1-p)}} f(z) \left[\frac{1 + \frac{1-p}{z\sqrt{np(1-p)}}}{1 + z\sqrt{\frac{(1-p)}{np}} + \frac{1}{np}} \right] \\ &= -zf(z) \left[\frac{1 + \frac{1-p}{z\sqrt{np(1-p)}}}{1 + z\sqrt{\frac{(1-p)}{np}} + \frac{1}{np}} \right] \end{aligned}$$

Now, as $n \rightarrow \infty$, $\Delta z \rightarrow 0$, so $\lim_{\Delta z \rightarrow 0} \frac{\Delta f}{\Delta z} = \lim_{n \rightarrow \infty} -zf(z) \left[\frac{1 + \frac{1-p}{z\sqrt{np(1-p)}}}{1 + z\sqrt{\frac{(1-p)}{np}} + \frac{1}{np}} \right] = -zf(z) \left[\frac{1+0}{1+0+0} \right]$

From the definition of a derivative, we have that $\frac{df}{dz} = \lim_{\Delta z \rightarrow 0} \frac{\Delta f}{\Delta z} = -zf(z)$. This differential equation describes the probability density function of the standard normal distribution. Separating variables gives

$$\frac{df}{f} = -zdz$$

$$\int \frac{df}{f} = -\int z dz$$

$$\ln(f) = -\frac{z^2}{2} + C$$

$$f(z) = \exp\left(-\frac{z^2}{2} + C\right) = e^C e^{-z^2/2} = A e^{-z^2/2}$$

The constant A is chosen so that the total area under the probability density curve is 1. A is determined by the following trick using a transformation to polar coordinates.

$$\begin{aligned} \int_{-\infty}^{\infty} A e^{-z^2/2} dx &= A \int_{-\infty}^{\infty} e^{-z^2/2} dx = A \sqrt{\left(\int_{-\infty}^{\infty} e^{-x^2/2} dx\right) \left(\int_{-\infty}^{\infty} e^{-y^2/2} dy\right)} \\ &= A \sqrt{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dy dx} = A \sqrt{\int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta} \\ &= A \sqrt{\left(\int_0^{2\pi} d\theta\right) \left(\int_0^{\infty} e^{-r^2/2} r dr\right)} = A \sqrt{(2\pi) \left(-e^{-r^2/2}\bigg|_0^{\infty}\right)} \\ &= A \sqrt{(2\pi)(-0 - -1)} = A \sqrt{2\pi} \end{aligned}$$

So $A = \frac{1}{\sqrt{2\pi}}$ and the Normal or Gaussian Probability density function is given by $f(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$.