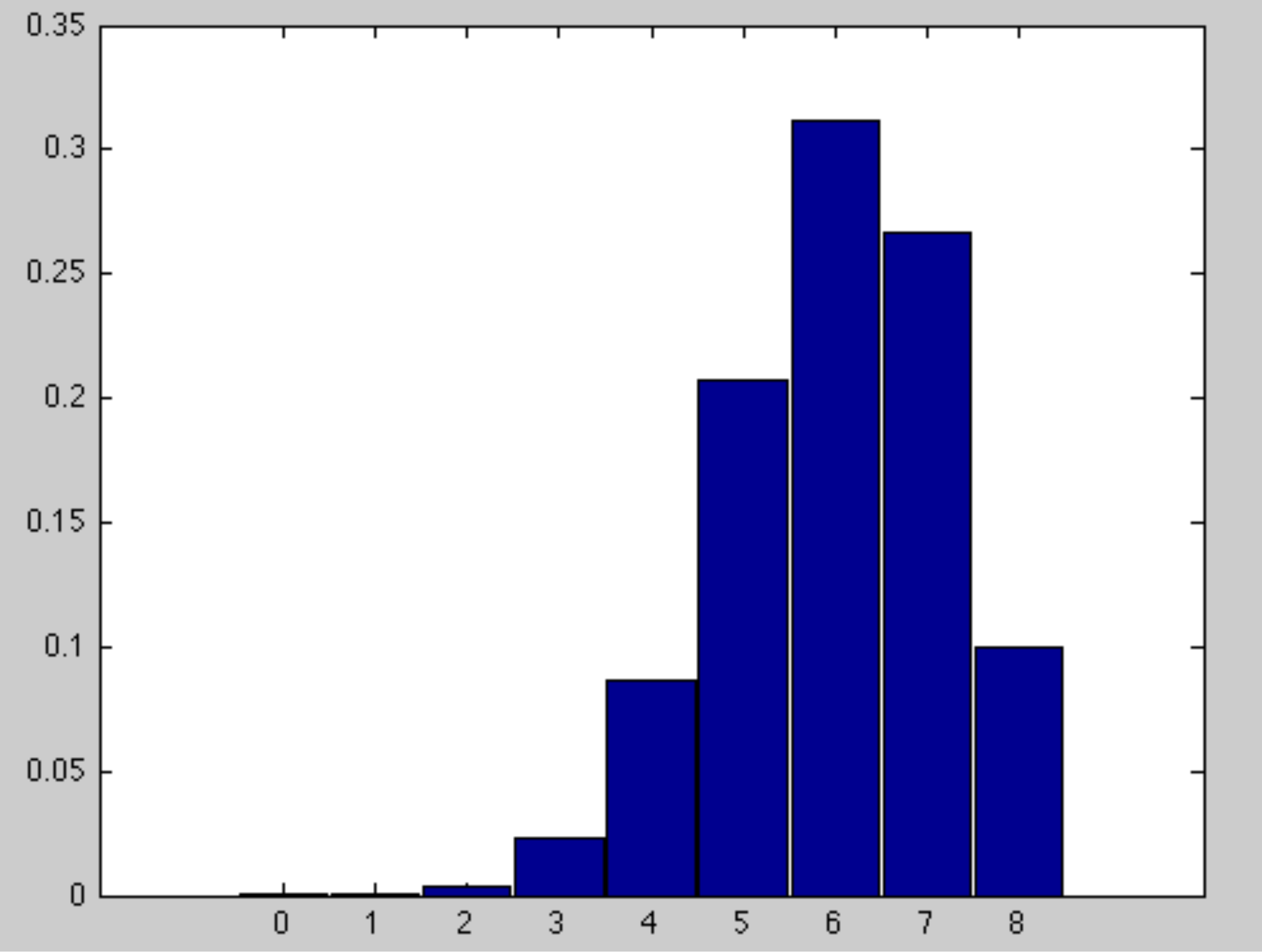
Question 1. 8 tickets are drawn at random with replacement from a box containing 8 tickets — 6 $\boxed{1}$s and 2 $\boxed{0}$s.

What is the probability that we will observe between 5 and 7 $\boxed{1}$ s?

We can calculate the probability using the basic rules of probability and the binomial formula:

$$\begin{aligned}
P\left(\text{between 5 and 7 }\boxed{1}\text{s}\right) &= P\left(5\,\boxed{1}\,s\right) + P\left(6\,\boxed{1}\,s\right) + P\left(7\,\boxed{1}\,s\right) \\
&= \binom{8}{5}\cdot\left(\frac{3}{4}\right)^5\cdot\left(\frac{1}{4}\right)^3 + \binom{8}{6}\cdot\left(\frac{3}{4}\right)^6\cdot\left(\frac{1}{4}\right)^2 \\
&\qquad + \binom{8}{7}\cdot\left(\frac{3}{4}\right)^7\cdot\left(\frac{1}{4}\right)^1 \\
&= 56\cdot\frac{3^5}{4^8} + 28\cdot\frac{3^6}{4^8} + 8\cdot\frac{3^7}{4^8} \\
&= \frac{51516}{65536} \approx 0.786
\end{aligned}$$

(*) We can answer questions about the probability of different outcomes for this experiment by referring to the probability histogram for the number of $\boxed{1}$ s in eight random draws:



Question 2. 80 tickets are drawn at random with replacement from the same box.

What is the probability that we will observe between 55 *and* 65 $\boxed{1}$ *s?*

Answer: We can give a precise answer again using the binomial formula:

$$P\left(\text{ between } 55 \;\text{ and } 65 \boxed{1}\text{s in } 80 \text{ draws}\right)$$
$$= \binom{80}{55}\cdot\left(\frac{3}{4}\right)^{55}\cdot\left(\frac{1}{4}\right)^{25} + \binom{80}{56}\cdot\left(\frac{3}{4}\right)^{56}\cdot\left(\frac{1}{4}\right)^{24} + \cdots$$
$$\cdots + \binom{80}{65}\cdot\left(\frac{3}{4}\right)^{65}\cdot\left(\frac{1}{4}\right)^{15} = \ldots?$$

These days, evaluating expressions like this directly is easy with computers. For example, we can use on-line calculators like the one found here:

http://stattrek.com/online-calculator/binomial.aspx

Answer: 84.55%.

350 years ago this approach was unavailable.

Questions like this led Abraham DeMoivre to discover that ***if the number of draws is large enough***, then the *probability histogram* for the number of $\boxed{1}$s drawn (at random with replacement) from a *zero-one* box is well-approximated by the *normal curve.* In fact, these questions led DeMoivre to *discover* the normal curve.

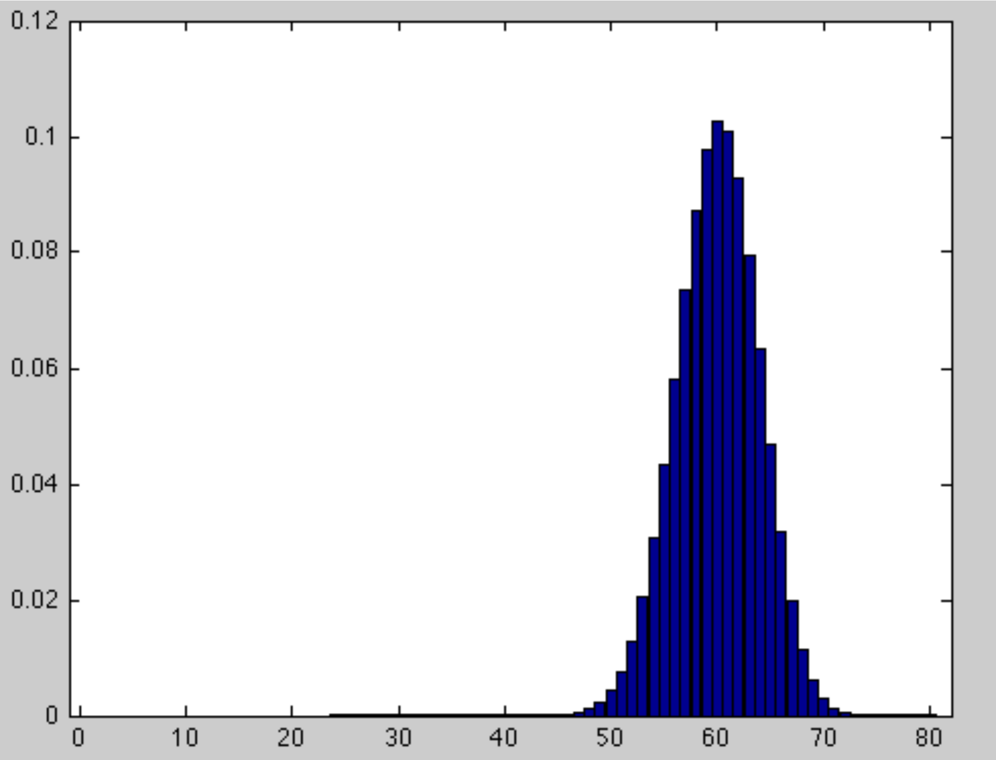
(*) To approximate the probability histogram for the number of $\boxed{1}$s with the normal curve, we need to rescale the histogram to standard units.

(*) To rescale the probability histogram for the number of $\boxed{1}$s to standard units, we use the formula

$$z = \frac{\Big(\text{number of } \boxed{1}\text{s}\Big) - EV\,\Big(\text{number of } \boxed{1}\text{s}\Big)}{SE\,\Big(\text{number of } \boxed{1}\text{s}\Big)}$$

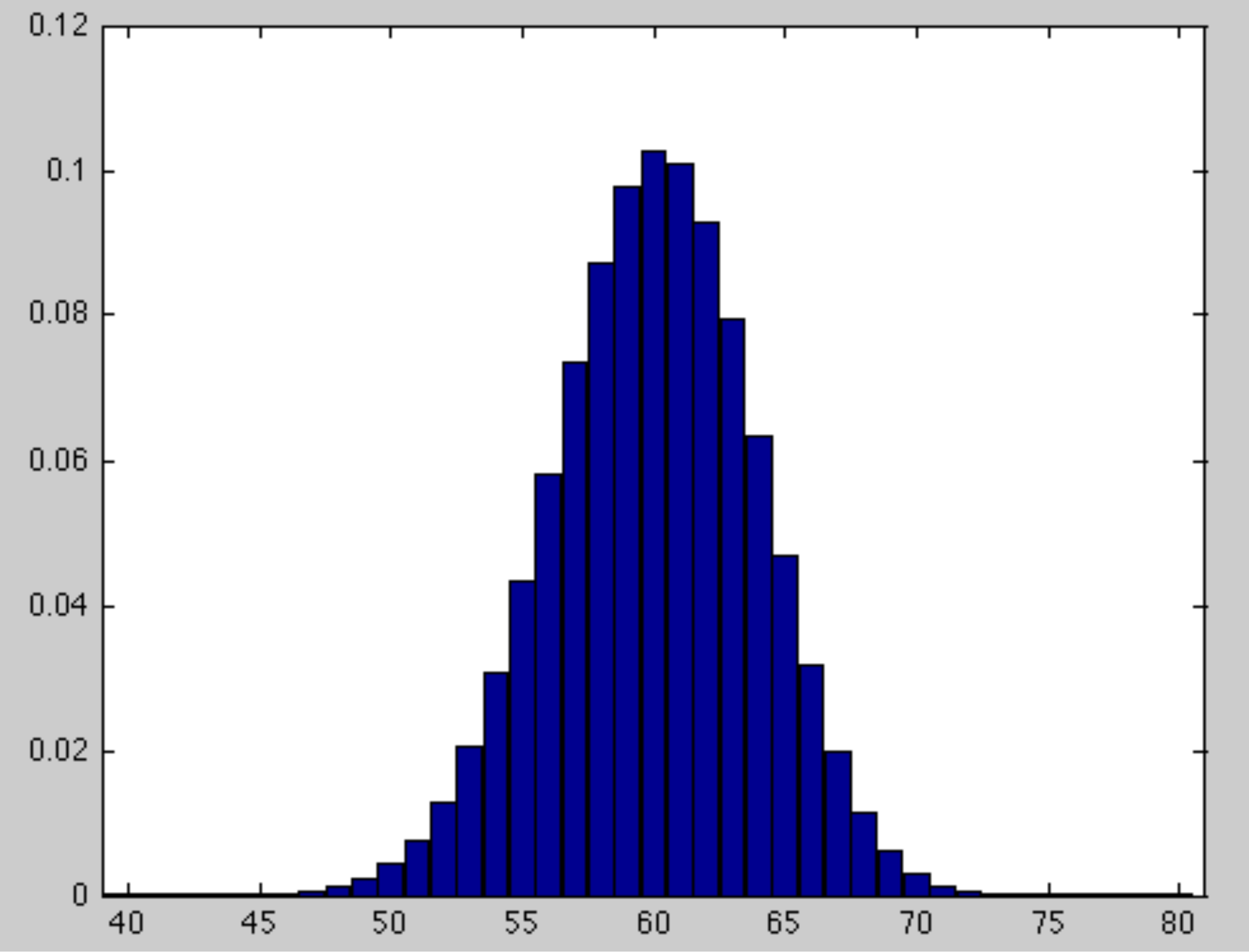
(*) In practice, we don't draw histograms, we use a simple step-by-step process to figure probabilities like this.

Probability histogram for the number of $\boxed{1}$s observed in 80 random draws (with replacement) from $\boxed{\boxed{1}\,\boxed{1}\,\boxed{1}\,\boxed{1}\,\boxed{1}\,\boxed{1}\,\boxed{0}\,\boxed{0}}$:



Observation: This histogram is skewed to the right, but the area under the histogram between 0 and 40 is so small as to be ignorable. The portion of the histogram between 40 and 80 is very symmetric (and normal-looking).

Zooming in to the part of the histogram where $40 \leq$ number of $\boxed{1}$s ≤ 80:



Rescaling to standard units...

First:

(*) The average of the box: $\frac{6}{8} = \frac{3}{4}$.

(*) The SD of the box: use the shortcut for the SD of a *zero-one* box...

$$SD = \sqrt{\frac{3}{4} \cdot \frac{1}{4}} = \frac{\sqrt{3}}{4} \approx 0.433.$$

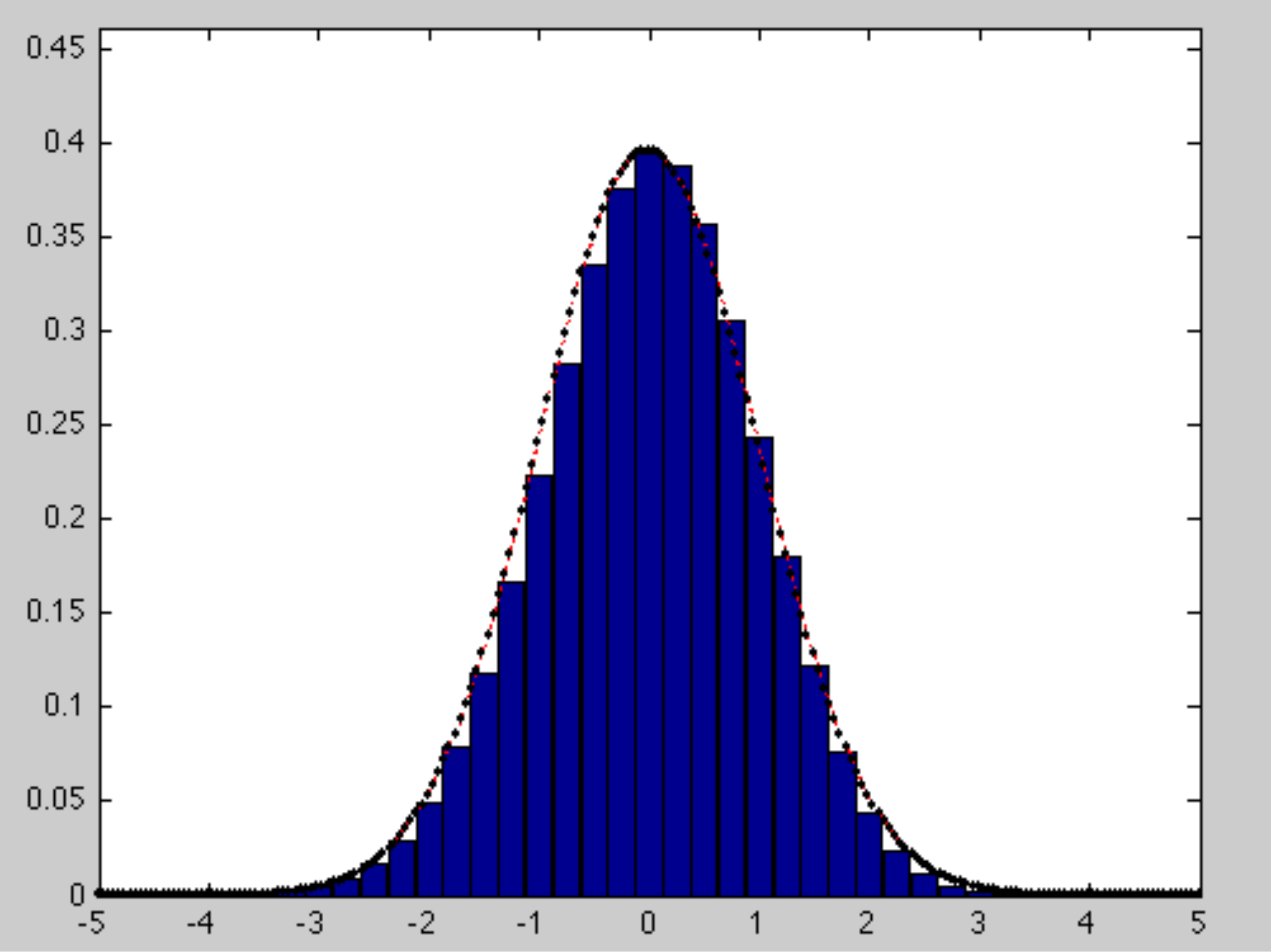
(*) The expected value for the ***number of*** $\boxed{1}$ ***s*** is the same as the expected value for the ***sum of the draws***:

$$EV\left(\text{number of } \boxed{1}\, s\right) = 80 \times \frac{3}{4} = 60.$$

(*) The standard error for the number of $\boxed{1}$s is the same as the standard error for the sum of the draws:

$$SE\left(\text{number of } \boxed{1}\, s\right) = \sqrt{80} \times \frac{\sqrt{3}}{4} \approx 3.873.$$

The rescaled histogram with the normal curve:



To answer Question 2, we can now take a step-by-step approach.

(i) The probability of observing between 55 and 65 $\boxed{1}$s is equal to the area under the original probability histogram between 54.5 and 65.5.

(ii) The area under the original probability histogram between 54.5 and 65.5 is equal to the area under the rescaled histogram between

$$\frac{54.5 - 60}{3.873} \approx -1.42$$

and

$$\frac{65.5 - 60}{3.873} \approx 1.42.$$

(iii) The area under the rescaled histogram between -1.42 and 1.42 is approximately equal to the area under the normal curve between -1.42 and 1.42.

(*) From the normal table, this is about 84.4%.

The Central Limit Theorem for the sum of draws from a box of numbered tickets says:

Suppose that tickets are drawn at random with replacement from a box of numbered tickets. If the number of tickets drawn is ***large enough****, then the probability histogram for the sum of the draws is well approximated by the normal curve, using the standard-units scale for the probability histogram.*

Comments:

1. This is also called the *normal approximation*.

2. The *normal approximation for data* can be explained by this — roughly speaking, many types of data can be modeled by sums of draws from appropriate boxes of tickets.

3. How large the number of draws needs to be for the normal approximation to be accurate depends on the distribution of tickets in the original box (and the desired degree of accuracy).

$\Rightarrow$ The more skewed the distribution of numbers in the original box the larger the number of draws needs to be for the normal approximation to be accurate.

In practical terms, the Central Limit Theorem tells us that:

If tickets are drawn at random with replacement from a box of numbered tickets and if the number of tickets is large enough, then

$$P(A \leq \textit{sum of draws} \leq B) \approx \left\{ \begin{array}{l} \textit{area under the normal curve } \boldsymbol{between} \\[2ex] \dfrac{A - EV(sum)}{SE(sum)} \quad \textit{and} \quad \dfrac{B - EV(sum)}{SE(sum)} \end{array} \right.$$

$$\textit{and} \;\; P(A \leq \textit{sum of draws}) \approx \left\{ \begin{array}{c} \textit{area under the normal curve} \\ \boldsymbol{to\ the\ right\ of} \\[2ex] \dfrac{A - EV(sum)}{SE(sum)} \end{array} \right.$$

$$\textit{and} \;\; P(\textit{sum of draws} \leq B) \approx \left\{ \begin{array}{c} \textit{area under the normal curve} \\ \boldsymbol{to\ the\ left\ of} \\[2ex] \dfrac{B - EV(sum)}{SE(sum)} \end{array} \right.$$

Question 3. A fair die is rolled 400 times, what is the probability that the total number of observed spots is 1450 or more?

(*) '1450 or more' means between 1450 and $2400 = 400 \times 6$ (which is the maximum possible) in this case.

Answer:

- The total number of observed spots in 400 rolls of a fair die is like the sum of 400 random draws with replacement from the box

$$\left\lfloor \boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{4}\,\boxed{5}\,\boxed{6} \right\rfloor.$$

- The average of this box is 3.5, and the SD is approximately 1.7078.
- $EV(sum) = 3.5 \times 400 = 1400$ and $SE(sum) = SD \times \sqrt{400} \approx 34.1565$.
- $(1450 - 1400)/34.1565 \approx 1.46$ and $(2400 - 1400)/34.1565 \approx 2.93$
- By the normal approximation,

$$\begin{aligned} P(2400 \geq \text{sum} \geq 1450) &\approx \text{area under normal curve between 1.46 and 2.93.} \\ &\approx \frac{1}{2}((\text{table entry for 2.93}) - (\text{table entry for 1.46})) \\ &\approx 7.18\% \end{aligned}$$

Question 4. A fair die is rolled 900 times, what is the chance that ⚄ will be observed between 140 and 180 times?

Answer:

- The number of ⚄ in 900 rolls of a fair die is like the sum of 900 random draws, with replacement from the box $\left|\boxed{1}\,\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{0}\right|$.
- The chance that the number of $\boxed{1}$s is between 140 and 180 is equal to the area under the *original* probability histogram between 139.5 and 180.5.

  ⇒ This is sometimes called the *continuity correction*, and applies to problems like this one involving draws from a zero-one box.
- The average of this box is 1/6, and the $SD = \sqrt{\frac{1}{6} \cdot \frac{5}{6}} \approx 0.373$.
- $EV(sum) = \frac{1}{6} \times 900 = 150$ and $SE(sum) = SD \times \sqrt{900} \approx 11.18$.

- $(139.5 - 150)/11.18 \approx -0.94$ and $(180.5 - 150)/11.18 \approx 2.73$.
- By the normal approximation,

$$
\begin{aligned}
P(140 \le & (\text{number of ⚄ in 900 rolls}) \le 180) \\
&= \text{area under original histogram between 139.5 and 180.5} \\
&\approx \text{area under normal curve between } -0.94 \text{ and } 2.73 \\
&= \frac{1}{2}((\text{table entry for } 2.73) + (\text{table entry for } 0.94)) \\
&\approx 82\%
\end{aligned}
$$

Question 5. One hundred draws are made at random with replacement from the box

$$\left\lfloor \boxed{1}\,\overbrace{\boxed{0}\,\boxed{0}\;\ldots\;\boxed{0}}^{99} \right\rfloor .$$

What is the probability that the sum of the draws will be between 0 and 2?

(*) Average of the box: $1/100 = 0.01$ and SD: $\sqrt{(0.01)(0.99)} \approx 0.0995$.

(*) $EV(sum) = 100 \cdot (0.01) = 1$ and $SE(sum) = \sqrt{100} \times SD(box) \approx 0.995$

(*) $(-0.5 - 0.01)/0.995 \approx -0.51$ and $(2.5 - 0.01)/0.995 \approx 2.5$

(*) Normal approximation:

$$\begin{aligned} P(0 \le \text{sum} \le 2) &\approx \text{area under the normal curve between } -0.51 \text{ and } 2.41 \\ &= \frac{1}{2}((\text{table entry for } 0.51)+(\text{table entry for } 2.5)) \approx 68.5\% \end{aligned}$$

I don't think so...

$\Rightarrow$ This box is very skewed. One hundred draws is not enough to invoke the normal approximation in this case! Using the binomial formula:

$$P(0 \le \text{sum} \le 2) = \overbrace{(0.99)^{100}}^{P(\text{sum}=0)} + \overbrace{100 \cdot (0.01) \cdot (0.99)^{99}}^{P(\text{sum}=1)} + \overbrace{\binom{100}{2}(0.01)^2(0.99)^{98}}^{P(\text{sum}=2)}$$

$$\approx 36.6\% + 36.97\% + 18.5\% = 92.07\%.$$

Rule of thumb

Suppose that the proportion of $\boxed{1}$*s in a zero-one box is p (so the proportion of* $\boxed{0}$*s is* $1-p$*). The normal approximation may be used reliably for estimating probabilities for the observed number of* $\boxed{1}$*s in* $\boldsymbol{n}$ *random draws with replacement from this box* ***if*** $\boldsymbol{np > 5}$ ***and*** $\boldsymbol{n(1-p) > 5}$***.***

In the example above, the normal approximation will become somewhat accurate for draws from this box once the number of draws is over 500.

Comment: As the number of draws grows larger, the normal approximation becomes more and more accurate.