The Central Limit Theorem

Suppose $\boldsymbol{n}$ tickets are drawn at random with replacement from a box of numbered tickets.

*The **central limit theorem** says that when the probability histogram for the **sum** of the draws is **rescaled to standard units** (using the expected value of the sum and the standard error for the sum) then the rescaled histogram is well-approximated by the normal curve, **as long as the number of draws, n, is large enough**.*

Or, using the '*box-of-sums*' interpretation...

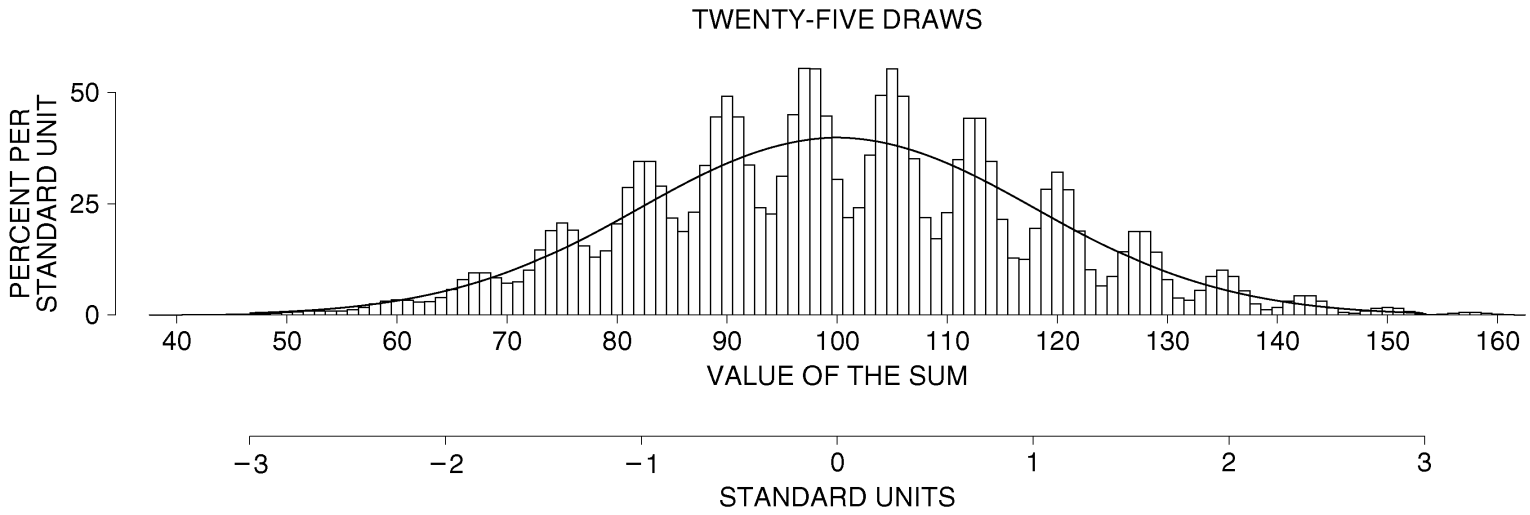
Suppose n tickets are drawn at random with replacement from a box of numbered tickets.

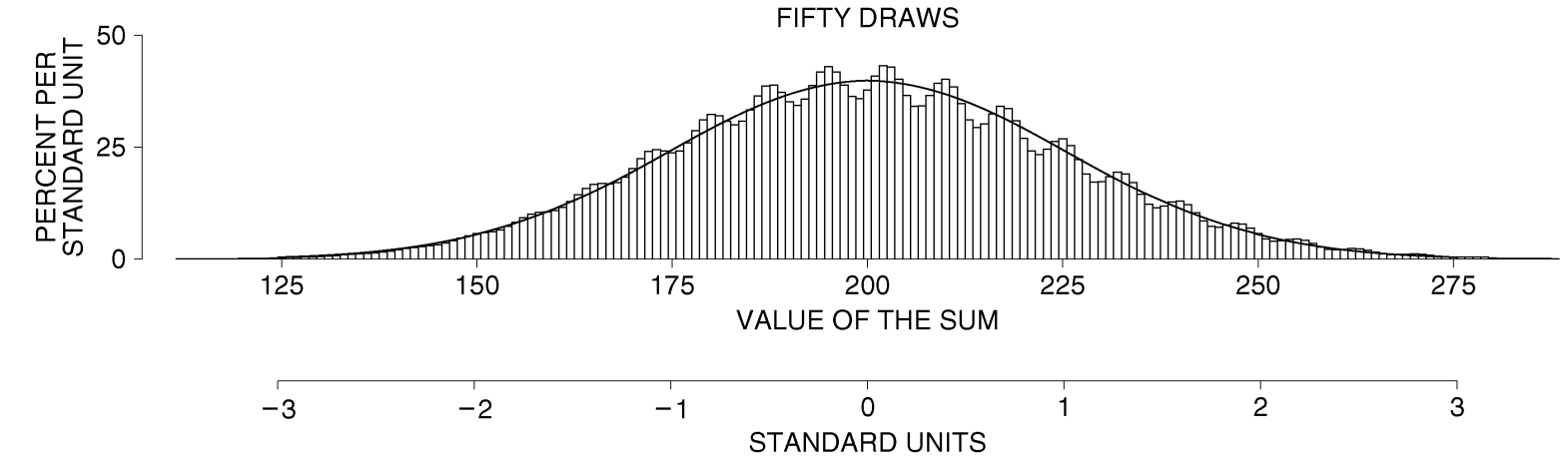
> *The* ***central limit theorem*** *says that when the histogram for the box-of-sums is* ***rescaled to standard units*** *— using the average of the box-of-sums (= expected value of the sum) and the SD of the box-of-sums (= standard error for the sum) — then the rescaled histogram is well-approximated by the normal curve,* ***as long as the number of draws, n, is large enough****.*

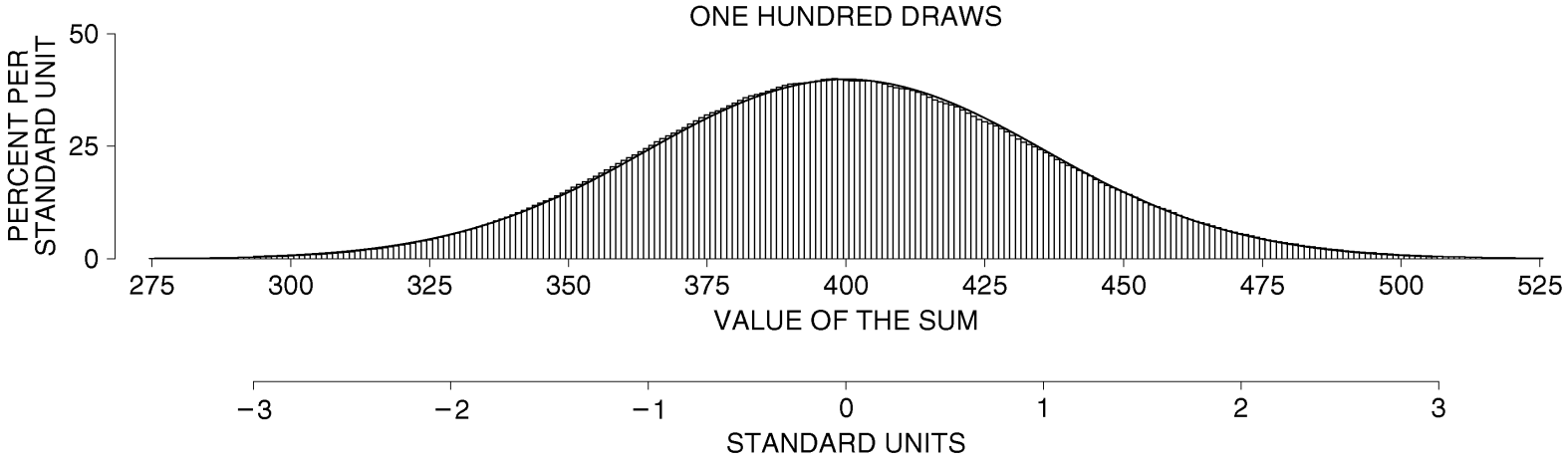
Illustration: Histograms for the box-of-sums corresponding to 25, 50 and 100 draws (made at random, with replacement) from the box

1	2	9

(From page 322 of FPP)




FIFTY DRAWS
PERCENT PER STANDARD UNIT
50
25
0
125
150
175
200
225
250
275
VALUE OF THE SUM
−3
−2
−1
0
1
2
3
STANDARD UNITS


ONE HUNDRED DRAWS
PERCENT PER STANDARD UNIT
50
25
0
275
300
325
350
375
400
425
450
475
500
525
VALUE OF THE SUM
−3
−2
−1
0
1
2
3
STANDARD UNITS

In practical terms, the *Central Limit Theorem* says:

if $\boldsymbol{n}$ *tickets are drawn at random, with replacement from a box of numbered tickets and if* $\boldsymbol{n}$ *is large enough, then*

$$P(\alpha \leq \text{ sum of draws } \leq \beta) \approx \left\{ \text{area under the normal curve between } \frac{\alpha - EV}{SE} \text{ and } \frac{\beta - EV}{SE} \right.$$

where

- $EV = \boldsymbol{n} \cdot$ **(average of box)** *is the* ***expected value*** *of the sum, and*
- $SE = \sqrt{\boldsymbol{n}} \cdot$ **(SD of the box)** *is the* ***standard error*** *for the sum.*

This is called the ***Normal Approximation***.

In particular, if $\boldsymbol{n}$ is large enough, then the normal approximation gives

- $P(EV - SE < \text{ sum } < EV + SE) \approx 68\%$.
- $P(EV - 2SE < \text{ sum } < EV + 2SE) \approx 95.5\%$.
- $P(EV - 3SE < \text{ sum } < EV + 3SE) \approx 99.7\%$.

Question: A fair coin is tossed 10,000 times. What is the probability that we observe Heads between 4900 and 5100 times?

Answer:

- The number of Heads in 10,000 tosses is like the sum of 10,000 random draws (with replacement) from the box $\boxed{\boxed{1}\ \boxed{0}}$.
- The average of this box is 1/2, and the $SD = \sqrt{\frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2}$.
- $EV = \frac{1}{2} \times 10,000 = 5000$ and $SE = SD \times \sqrt{10,000} = 50$.
- By the normal approximation,

$$\begin{aligned} P(4900 \leq &\text{(number of Heads in 10,000 tosses)} \leq 5100) \\ &\approx \text{area under normal curve between } -2 \text{ and } 2 \\ &= 95.5\% \end{aligned}$$

Question: 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 $\boxed{1\ 0\ 0\ 0\ 0\ 0}$.
- The average of this box is 1/6, and the $SD = \sqrt{\frac{1}{6} \cdot \frac{5}{6}} \approx 0.373$.
- $EV = \frac{1}{6} \times 900 = 150$ and $SE = SD \times \sqrt{900} \approx 11.18$.
- By the normal approximation,

$$\begin{aligned} P(140 \leq &\text{(number of ⁘ in 900 rolls)} \leq 180) \\ &\approx \text{area under normal curve between } -0.89 \text{ and } 2.68 \\ &= \frac{1}{2}(T(2.68) + T(0.89)) \approx 81\% \end{aligned}$$

(The notation $T(x)$ means the entry in the FPP normal table corresponding to x standard units.)

The *continuity* correction

Example. A fair coin is tossed 100 times, what is the probability that H is observed between 48 and 54 times?

Answer:

- The number of Heads in 100 tosses is like the sum of 100 random draws (with replacement) from the box $\boxed{\boxed{1}\ \boxed{0}}$.
- The average of this box is 1/2, and the $SD = \sqrt{\frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2}$.
- $EV = \frac{1}{2} \times 100 = 50$ and $SE = SD \times \sqrt{100} = 5$.
- By the normal approximation,

$$\begin{aligned} P(48 \leq &\text{(number of Heads in 100 tosses)} \leq 53) \\ &\approx \text{area under normal curve between } -\frac{2}{5} \text{ and } \frac{3}{5} \\ &= \frac{1}{2}T(0.4) + \frac{1}{2}T(0.6) = 38.115\% \end{aligned}$$

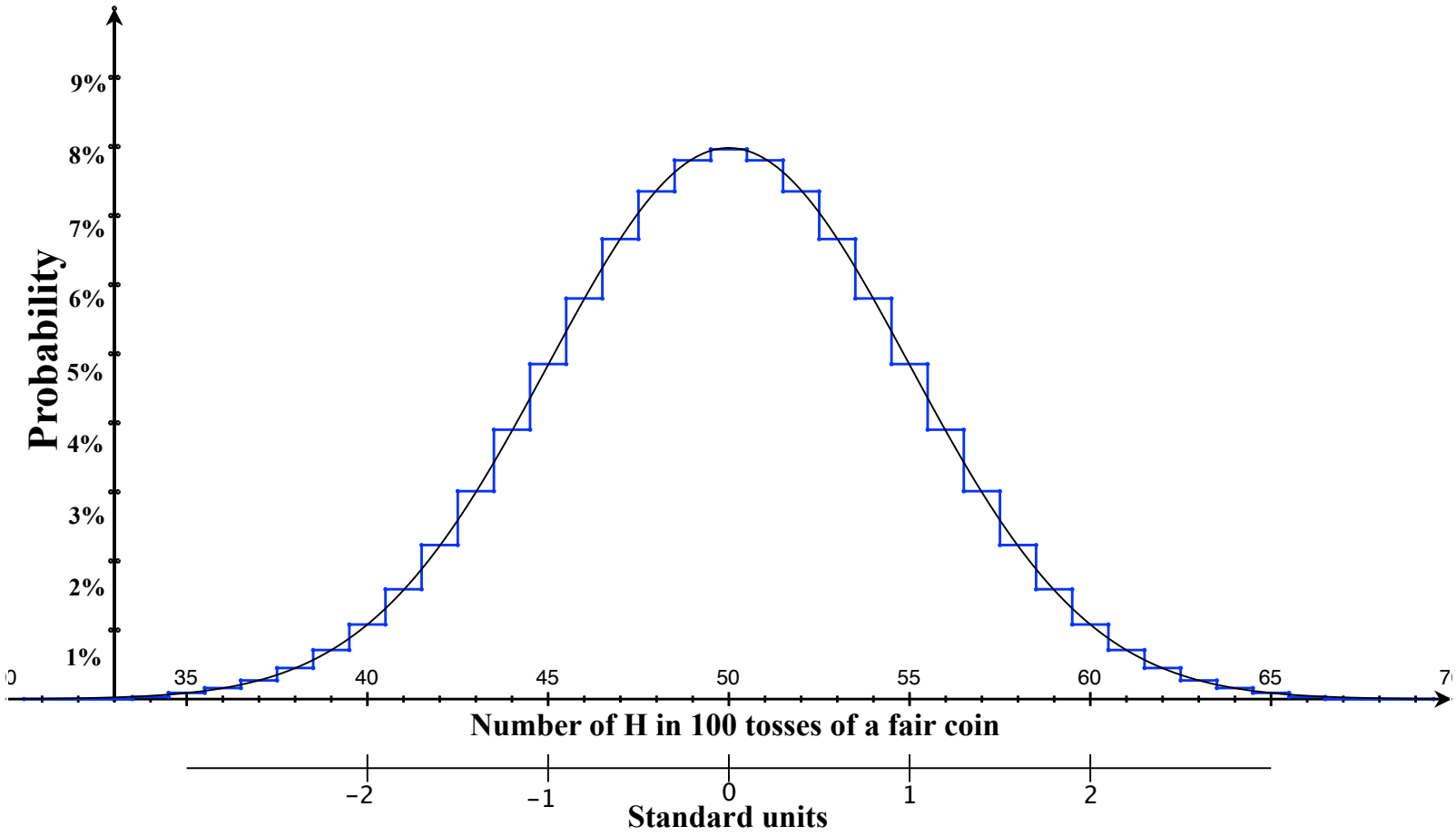
Checking our approximate answer using an on-line calculator (for binomial probabilities):

$$P(48 \leq (\text{number of Heads in 100 tosses}) \leq 53) = 44.93\%.$$

The normal approximation appears to have missed the mark by quite a bit!

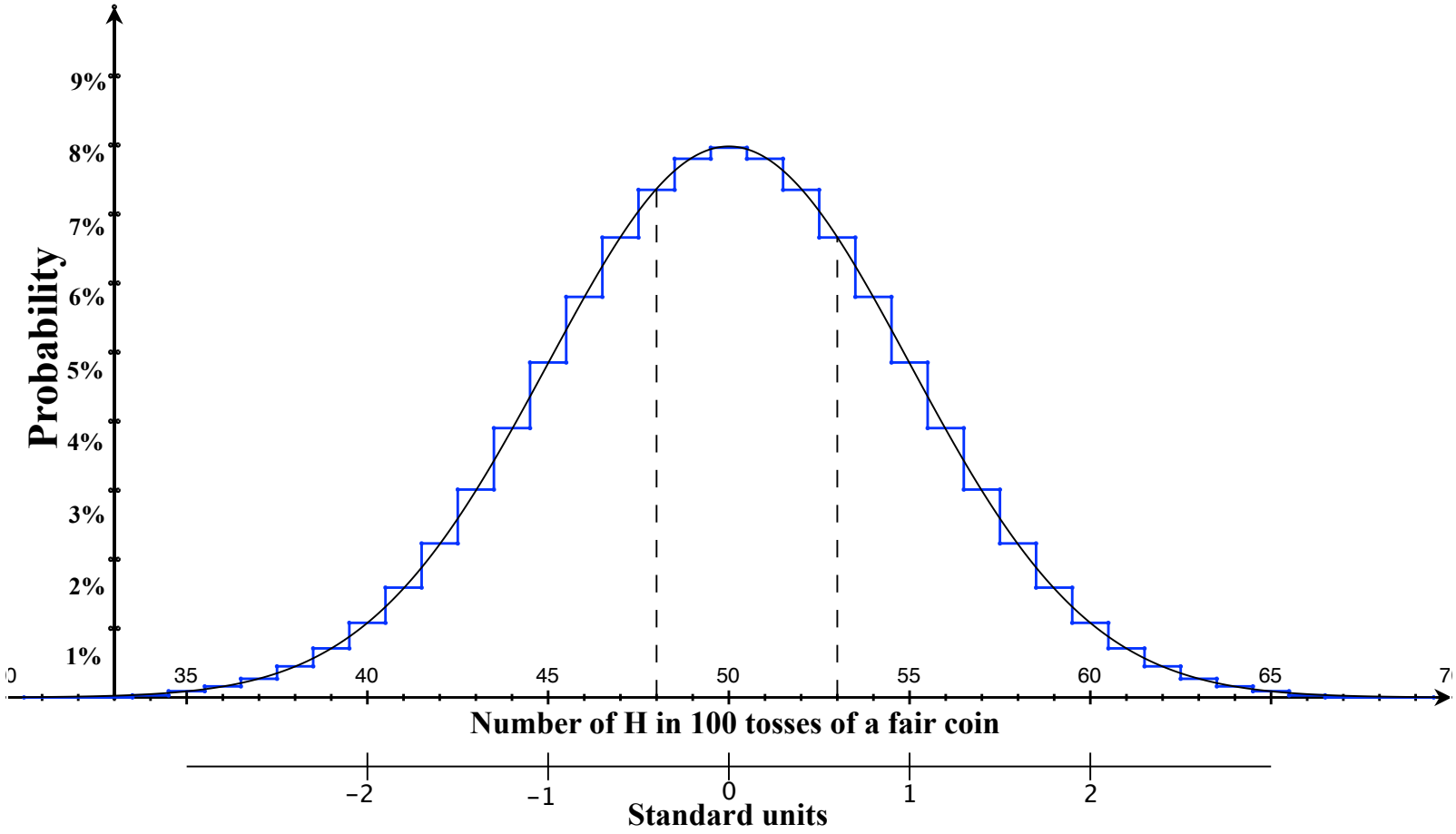
What's wrong?

Answer: The normal approximation is based on the idea that we approximate the area under a probability histogram with the normal curve...

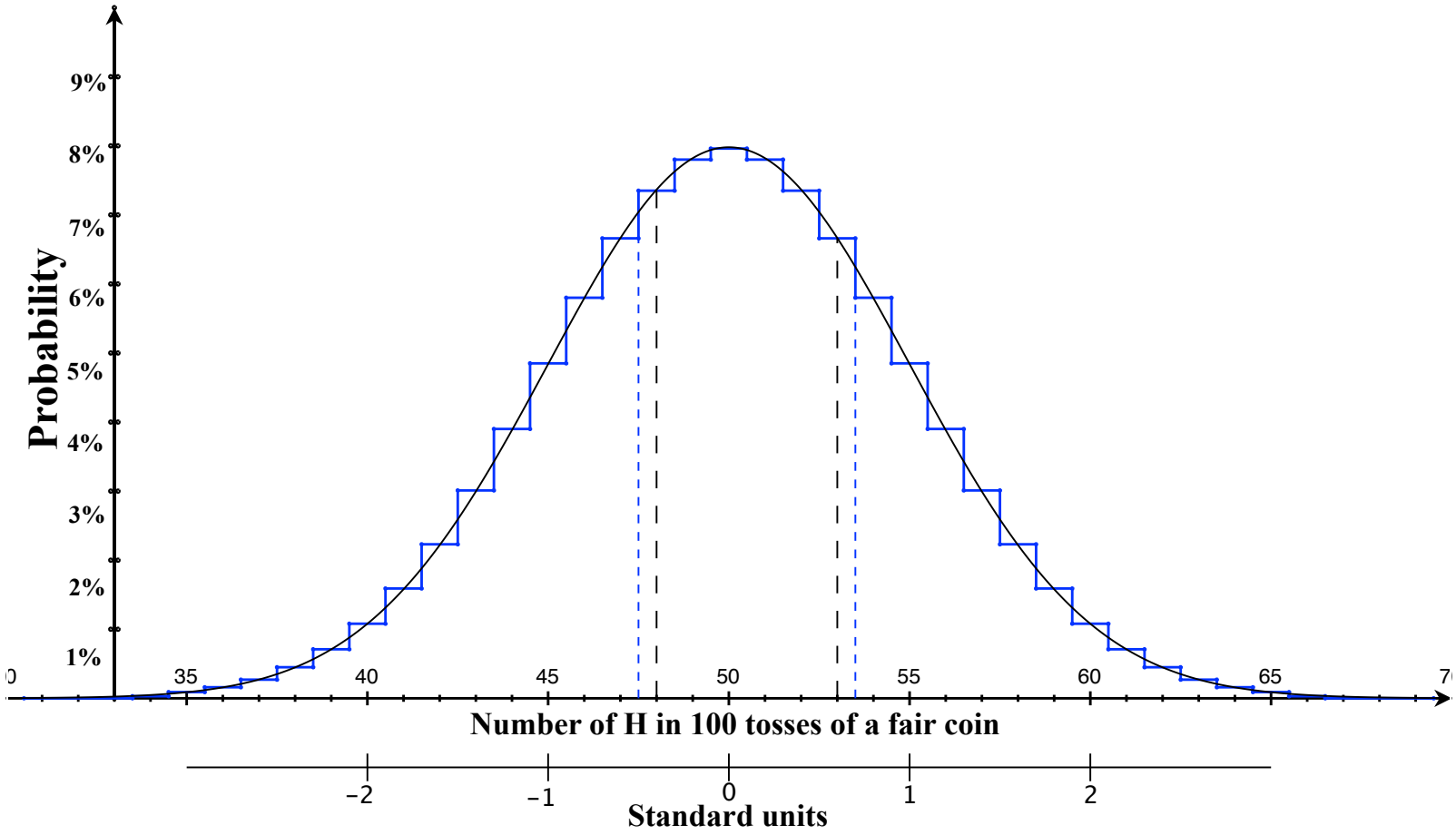


We want to approximate the area of the bars over $48, 49, 50, 51, 52$ and 53...

... and we used the normal curve to approximate the area between the dashed lines below:



But this means that we cut off half of the left-most and right-most bars! The region we should be approximating is between the dashed blue lines below: from 47.5 to 53.5



I.e., the more accurate approximation is

$$P(48 \leq \text{(number of Heads in 100 tosses)} \leq 53)$$

$$\approx \text{area under normal curve between } -\frac{2.5}{5} \text{ and } \frac{3.5}{5}$$

$$= \frac{1}{2}T(0.5) + \frac{1}{2}T(0.7) = 44.95\%$$

More generally, if we draw n tickets (at random with replacement) from a 0-1 box and we want to approximate $P(a \leq \#\text{ of } \boxed{1}\text{s} \leq b)$ we should use the *continuity correction*:

$$P(a \leq \#\text{ of } \boxed{1}\text{s} \leq b) \approx \left\{ \begin{array}{l} \text{area under the normal curve between} \\ \frac{(a-0.5)-EV}{SE} \text{ and } \frac{(b+0.5)-EV}{SE} \end{array} \right.$$

In this case $EV = np$ and $SE = \sqrt{n}\sqrt{p(1-p)}$, where p is the proportion of $\boxed{1}$s in the box.

Comment: The continuity correction becomes less important as the number of draws from the box becomes larger, because the differencs between

$$\frac{(a-0.5)-np}{\sqrt{n}\sqrt{p(1-p)}} \text{ and } \frac{a-np}{\sqrt{n}\sqrt{p(1-p)}}$$

and

$$\frac{(b+0.5)-np}{\sqrt{n}\sqrt{p(1-p)}} \text{ and } \frac{b-np}{\sqrt{n}\sqrt{p(1-p)}}$$

become very small.

Specifically, the differences are

$$\frac{0.5}{\sqrt{n}\sqrt{p(1-p)}}$$

which become smaller and smaller as n gets bigger.