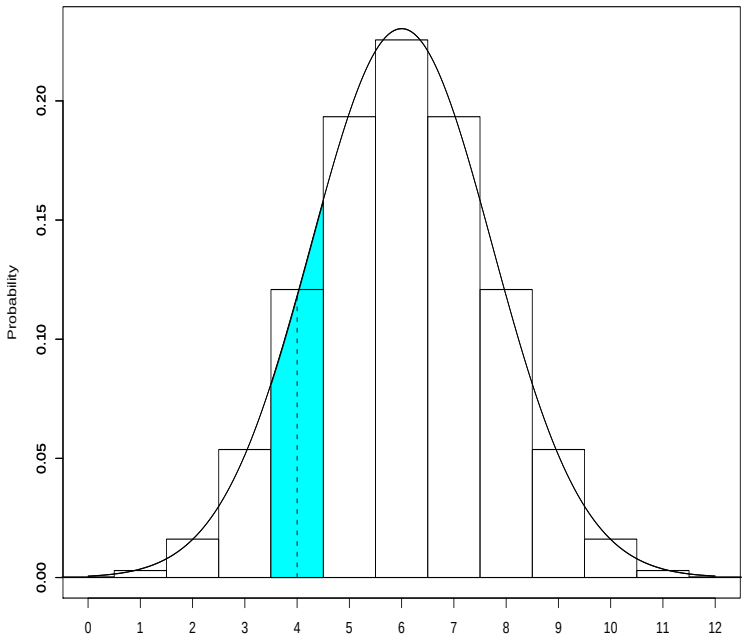
The normal approximation to the binomial

In order for a continuous distribution (like the normal) to be used to approximate a discrete one (like the binomial), a **continuity correction** should be used. There are two major reasons to employ such a correction.

First, recall that a discrete random variable can only take on only specified values, whereas a continuous random variable used to approximate it can take on any values whatsoever within an interval around those specified values. Hence, when using the normal distribution to approximate the binomial, more accurate approximations are likely to be obtained if a continuity correction is used.

Second, recall that with a continuous distribution (such as the normal), the probability of obtaining a *particular* value of a random variable is zero. On the other hand, when the normal approximation is used to approximate a discrete distribution, a continuity correction can be employed so that we can approximate the probability of a specific value of the discrete distribution.

Consider an experiment where we toss a fair coin 12 times and observe the number of heads. Suppose we want to compute the probability of obtaining *exactly* 4 heads. Whereas a discrete random variable can have only a specified value (such as 4), a continuous random variable used to approximate it could take on any values within an interval around that specified value, as demonstrated in this figure:



The continuity correction requires adding or subtracting .5 from the value or values of the discrete random variable X as needed. Hence to use the normal distribution to approximate the probability of obtaining *exactly* 4 heads (i.e., $X = 4$), we would find the area under the normal curve from $X = 3.5$ to $X = 4.5$, the lower and upper boundaries of 4. Moreover, to determine the approximate probability of observing *at least* 4 heads, we would find the area under the normal curve from $X = 3.5$ and above since, on a continuum, 3.5 is the lower boundary of X. Similarly, to determine the approximate probability of observing *at most* 4 heads, we would find the area under the normal curve from $X = 4.5$ and below since, on a continuum, 4.5 is the upper boundary of X.

When using the normal distribution to approximate discrete probability distribution functions, we see that semantics become important. To determine the approximate probability of observing *fewer* than 4 heads, we would find the area under the normal curve from 3.5 and below; to determine the approximate probability of observing *at most* 4 heads, we would find the area under the normal curve from 4.5 and below, since the latter event includes the value $X = 4$.

Now consider the binomial distribution in particular. Let H be the number of heads in 12 flips of a fair coin. We know that the probability of observing exactly k heads in 12

flips is

$$P(H = k) = \binom{12}{k}(.5)^k(.5)^{12-k}, \qquad k = 0, 1, 2, \ldots, 12.$$

We also know that the mean of this binomial distribution is given by

$$\mu = np = (12)(.5) = 6,$$

while the standard deviation is given by

$$\sigma = \sqrt{np(1-p)} = \sqrt{(12)(.5)(.5)} = 1.732.$$

Say we are interested in the probability of observing between 3 and 5 heads, inclusive; that is, $P(3 \leq H \leq 5)$. We can calculate this exactly, of course:

$$P(3 \leq H \leq 5) = P(H = 3) + P(H = 4) + P(H = 5) = .05371 + .12085 + .19336 = .36792$$

What about the normal approximation to this value? First, we should check whether the normal approximation is likely to work very well in this case. A useful rule of thumb is that the normal approximation should work well enough if both np and $n(1-p)$ are greater than 5. For this example, both equal 6, so we're about at the limit of usefulness of the approximation.

Back to the question at hand. Since H is a binomial random variable, the following statement (based on the continuity correction) is **exactly** correct:

$$P(3 \leq H \leq 5) = P(2.5 < H < 5.5).$$

Note that this statement is **not** an approximation — it is **exactly correct**! The reason for this is that we are adding the events $2.5 < H < 3$ and $5 < H < 5.5$ to get from the left side of the equation to the right side of the equation, but **for the binomial random variable, these events have probability zero**. The continuity correction is **not** where the approximation comes in; that comes when we approximate H using a normal distribution with mean $\mu = 6$ and standard deviation $\sigma = 1.732$:

$$\begin{aligned} P(3 \leq H \leq 5) &= P(2.5 < H < 5.5) \\ &\approx P\left(\frac{2.5-6}{1.732} < Z < \frac{5.5-6}{1.732}\right) \\ &= P(-2.02 < Z < -.29) \\ &= P(Z < -.29) - P(Z < -2.02) \\ &= .3859 - .0217 = .3642 \end{aligned}$$

Note that the approximation is only off by about 1%, which is pretty good for such a small sample size.

Example. Suppose that a sample of $n = 1,600$ tires of the same type are obtained at random from an ongoing production process in which 8% of all such tires produced are defective. What is the probability that in such a sample not more than 150 tires will be defective?

Answer. We approximate the $B(1600, .08)$ random variable T with a normal, with mean $(1600)(.08) = 128$ and standard deviation $\sqrt{(1600)(.08)(.92)} = 10.85$. The probability calculation is thus

$$\begin{aligned} P(T \leq 150) &= P(T < 150.5) \\ &\approx P\left(Z < \frac{150.5 - 128}{10.85}\right) \\ &= P(Z < 2.07) \\ &= .9808 \end{aligned}$$

Example. Based on past experience, 7% of all luncheon vouchers are in error. If a random sample of 400 vouchers is selected, what is the approximate probability that

(a) exactly 25 are in error?

(b) fewer than 25 are in error?

(c) between 20 and 25 (inclusive) are in error?

Answer. We approximate the $B(400, .07)$ random variable V with a normal, with mean $(400)(.07) = 28$ and standard deviation $\sqrt{(400)(.07)(.93)} = 5.103$. The probability calculations are thus

(a)

$$\begin{aligned} P(V = 25) &= P(24.5 < V < 25.5) \\ &\approx P\left(\frac{24.5 - 28}{5.103} < Z < \frac{25.5 - 28}{5.103}\right) \\ &= P(-.69 < Z < -.49) \\ &= .3121 - .2451 = .0670 \end{aligned}$$

(b)

$$\begin{aligned} P(V < 25) &= P(V < 24.5) \\ &\approx P\left(Z < \frac{24.5 - 28}{5.103}\right) \\ &= P(Z < -.69) \\ &= .2451 \end{aligned}$$

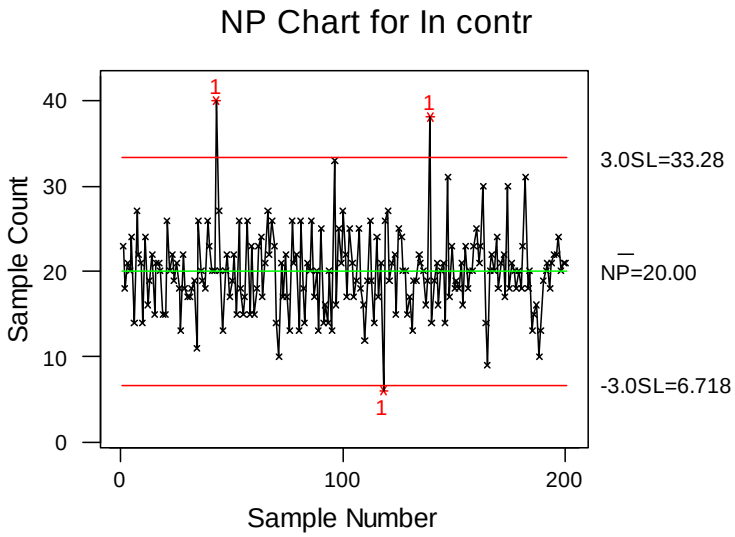
(c)

$$\begin{aligned} P(20 \leq V \leq 25) &= P(19.5 < V < 25.5) \\ &\approx P\left(\frac{19.5 - 28}{5.103} < Z < \frac{25.5 - 28}{5.103}\right) \\ &= P(-1.67 < Z < -.49) \\ &= .3121 - .0475 = .2646 \end{aligned}$$

The normal approximation to the binomial is the underlying principle to an important tool in statistical quality control, the **Np chart**. Say we have an assembly line that turns out thousands of units per day. Periodically (daily, say), we sample n items from the assembly line, and count up the number of defective items, D. What distribution does D have? $B(n, p)$, of course, with p being the probability that a particular item is defective. Thus, examination of D allows us to see if the probability of a defective item is changing over time; that is, the process is getting *out of control*.

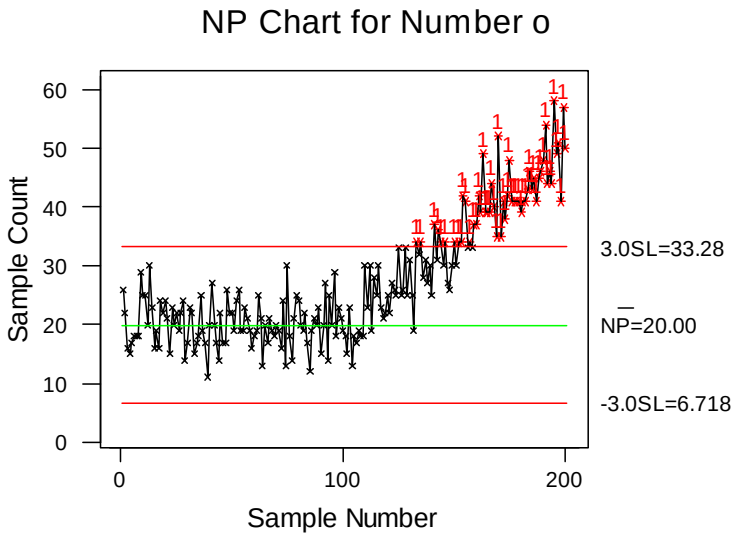
Consider the following situation. Our assembly line has been running for a while, and based on this historical data, we've seen that the probability of an individual item being defective is .02 (this would come from sampling and getting the empirical frequency of defectives). Our online quality control system samples 1000 items per day, and counts up the number of defectives D. We know that $D \sim B(1000, .02)$, so $E(D) = (1000)(.02) = 20$ and $S(D) = \sqrt{(1000)(.02)(.98)} = 4.427$. Thus, D is approximately normally distributed with mean $\mu = 20$ and standard deviation $\sigma = 4.427$. This allows us to assess whether future values of D are unusual, by seeing whether they get "too far" from 20.

Here is an example of an Np chart.



The chart consists of the values for the process plotted with three lines: the expected number of defects in the center, and two control limits at $\mu \pm 3\sigma$ (the number 3 is arbitrary, but standard). Since D is roughly normally distributed, we know that the probability of D being outside the control limits, assuming that p is staying at .02, is $P(|Z| > 3) = .0026$. So in this case, where there are 200 days worth of data, we're not surprised to see a day or two outside the limits. There are actually three, but this is just random fluctuation; we know that because the process "settles down" to its correct value immediately. In fact, these data do come from a stable process.

Now consider this chart:



This chart is very different. About 100 days into the sample, the process starts to go "out of control." Eventually the D values move outside the control limits, and the process should be stopped and corrected. In fact, for these data, starting with the 101^{st} observation, I increased p steadily by .0003 per day.

Control charts also can be constructed based on other statistics, such as means and standard deviations. They are an integral part of the idea of *kaizen*, or continuous improvement, that has revolutionized manufacturing around the world in recent years.

MINITAB commands

To obtain an Np chart, click on `Stat` $\mapsto$ `Control Charts` $\mapsto$ `NP`. Enter the variable with the binomial counts in it next to `Variable:`, and insert the number of binomial trials each observation refers to next to `Subgroup size:`. If there is a value of p that is known from historical data to correspond to the process being in control, enter that value next to `Historical p:`.

