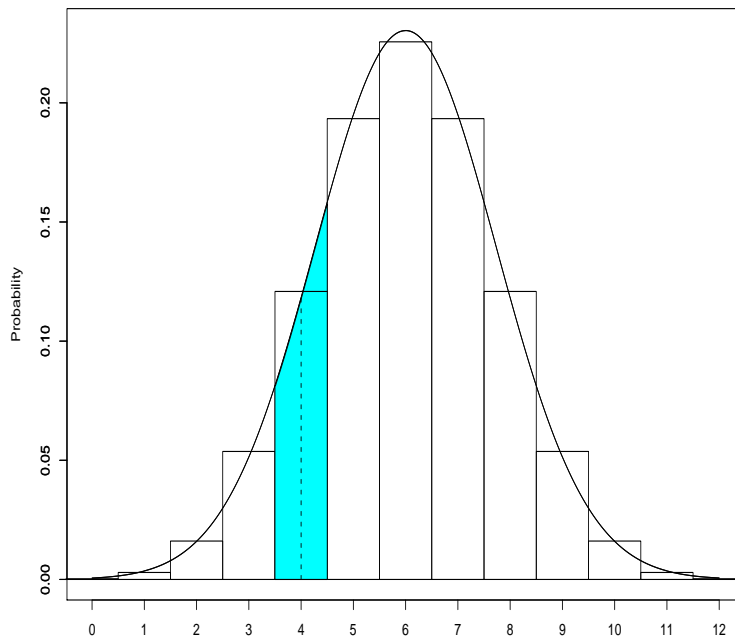


The normal approximation to the binomial

The binomial probability function is not useful for calculating probabilities when the number of trials n is large, as it involves multiplying a potentially very large number $\binom{n}{k}$ with a potentially very small one $p^k(1-p)^{n-k}$. **Fortunately, the normal distribution provides a very good approximation to the Binomial when n is large enough (say $np > 5$ and $n(1-p) > 5$). The approximation involves two steps, one obvious and one not so obvious:**

- (1) **Instead of using the $B(n, p)$ distribution, using the $N(\mu, \sigma^2)$ distribution, where $\mu = np$ and $\sigma^2 = np(1-p)$.**
- (2) **Use a continuity correction to correct for the use of a continuous distribution to approximate a discrete one.**

The need for a continuity correction can be seen in the following situation. 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 goal is to approximate the area of the bar above 4 with an area under the normal

curve. Using the probability corresponding to $P(X = 4)$ is clearly wrong, since that is 0; by using the area between 3.5 and 4.5 (the shaded area above), the area between the bar and curve missed on the left is roughly made up for on the right. **This is the essence of the calculation: whenever the Binomial probability for an integer value k is needed, use the area under the corresponding normal curve for $(k - .5, k + .5)$.**

Example. Let H be the number of heads in 12 flips of a fair coin. What is the probability of observing between 3 and 5 heads, inclusive; that is, $P(3 \leq H \leq 5)$? In this case we can calculate the value exactly: since

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

we have

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

To approximate this we first need the mean

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

and standard deviation

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

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 $\sigma = 1.732$:

$$\begin{aligned} P(3 \leq H \leq 5) &= P(2.5 < H < 5.5) \\ &\approx P\left(\frac{2.5 - 6}{1.732} < Z < \frac{5.5 - 6}{1.732}\right) \\ &= P(-2.02 < Z < -2.29) \\ &= P(Z < -2.29) - P(Z < -2.02) \\ &= .3859 - .0217 = .3642 \end{aligned}$$

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

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

Note that if the question had been about the probability of there being fewer than 150 defective tires, the calculation would have been

$$\begin{aligned} P(T < 150) &= P(T < 149.5) \\ &\approx P\left(Z < \frac{149.5 - 128}{10.85}\right) \\ &= P(Z < 1.98) \\ &= .9761, \end{aligned}$$

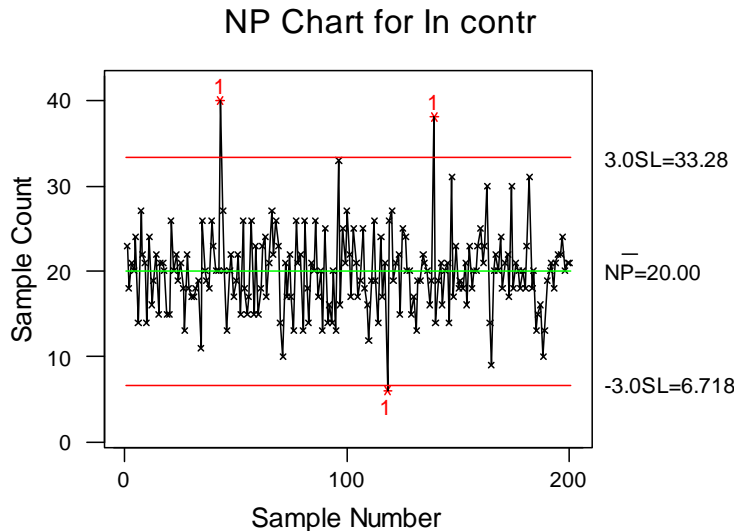
the difference in these two values (.0047) being the estimated probability of exactly 150 tires with defects.

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

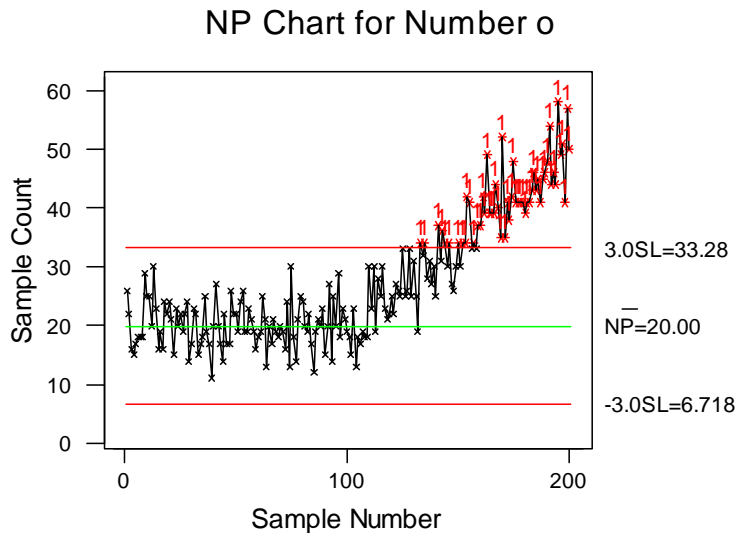
the number of defectives D . We know that $D \sim B(1000, .02)$, so $E(D) = (1000)(.02) = 20$ and $S(D) = \sqrt{(1000)(.02)(.98)} = 4.427$. Thus, D is approximately normally distributed with mean $\mu = 20$ and standard deviation $\sigma = 4.427$. This allows us to assess whether future values of D are 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 101st 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:**.