Control Charts for Attributes: Fraction Non-Conforming

Dr. Austin Brown

Kennesaw State University

Introduction

- In the last couple of classes, we have talked about monitoring processes where the quality characteristic is inherently numeric/continuous using the Shewhart $\bar{x}$ and R/s charts or the CUSUM/EWMA charts.

- However, in some instances, the way that we have decided to operationally measure sample data is by dichotomizing it into either a conforming or non-conforming category.

Introduction

- There are certainly manufacturing applications of this, but let's think of a non-manufacturing example. One big metric that university presidents and provosts monitor carefully is the undergraduate 5-year graduation rate.

- This rate is looking at the number of students in an incoming freshman cohort who graduate within five years of enrollment. So what is our outcome variable we're most interested in here?

Introduction

- A student either did graduate within five years (conforms) or did not graduate within five years (non-conforming).

- Obviously, this is not a continuous or normally distributed variable, and thus, using the Shewhart $\bar{X}$ or CUSUM/EMWA charts would be inappropriate. So what do we do? A chart specifically designed for this type of situation has to be used.

The p Chart

- We know that for a given student, there is a true probability, p, that they graduate within five years.

- This implies that there is also a true probability that a random student does not graduate within five years, and we'll call that $1 - p$.

- If we let Y be an indicator variable which assumes a value of 1 if the student graduates in five years and 0 if they don't, then we can create a probability distribution:

$$P[Y = y] = p^y(1 - p)^{1-y}, \quad y = \{0, 1\}$$

The p Chart

▶ You may recall from math stats that Y is a Bernoulli random variable. In our example, if we took a random sample of n students (i.e., a sample of n-independent Bernoulli trials), then the probability that a subset of them, say x, graduated within five years follows the Binomial distribution:

$$P[X = x] = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = \{0, 1, \ldots, n\}$$

The p Chart

- We also know that when we don't know p and have to estimate it, that we typically use a maximum likelihood estimator for that purpose (we'll denote it $\hat{p}$).

- MLEs have the nice characteristic that their sampling distributions are normally distributed (most of the time) and that is the case here, too.

$$\hat{p} \overset{.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

The p Chart

▶ All of this leads to the development of the p chart, which is a Shewhart-style chart.

▶ The p chart functions much like the $\bar{x}$ chart: we take random samples from our process, calculate a plotting statistic, and compare that statistic to control limits.

The p Chart

▶ Our plotting statistic, as you may have guessed, is the sample proportion. We take random samples of size n and calculate the proportion of the sample which <u>DO NOT</u> conform.

▶ For example, if we took a random sample of 100 students five years after they enrolled at KSU and 54 did not graduate (transferred, dropped out, etc.), then the sample proportion would be:

$$\hat{p} = \frac{54}{100} = 0.54$$

The p Chart

▶ Okay, so what about control limits? In the case where p is known, and assuming we're using 3 σ limits:

$$UCL = p + 3\sqrt{\frac{p(1-p)}{n}}$$

$$LCL = p - 3\sqrt{\frac{p(1-p)}{n}}$$

The p Chart

▶ Now, one thing you may have observed is that in some cases, our $LCL < 0$ (depending on the value of p and n).

▶ In such cases, it is customary to set the LCL to zero and operate with only an UCL.

▶ Here, you have to be a little bit mindful about the patterns plotting on the chart as systematic patterns may indicate an out-of-control process.

The p Chart

▶ Like with the $\bar{x}$ and R/s charts, in many instances we plainly won't know p and will have to estimate it using Phase I data.

▶ In this case, we will take m samples of size n and calculate a $\hat{p}_i$ for each of the m samples. Then our estimate of p will be:

$$\bar{p} = \frac{1}{m} \sum_{i=1}^{m} \hat{p}_i$$

The *p* Chart

▶ Now, our Phase I control limits are:

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

▶ Much like with the $\bar{x}$ chart, we will plot our individual $\hat{p}_i$'s against these calculated limits to ensure the process is in statistical control, and if not, search for an assignable cause and if found, correct it, omit the OOC point(s), and recalculate the limits.

The *p* Chart

- Let's go through an example. Let's say we are monitoring freshman retention at KSU. If a new student enrolls in classes for the next term, the more likely they are to stay at the university. We want to monitor the proportion of new students in a given term who do not enroll in classes for the next term.

- Suppose we take samples of size $n = 50$ from each of the last 30 spring and fall terms. These data are located in the "Enrollment" Excel spreadsheet in D2L.

The p Chart

▶ Now, let's do another example using the Orange Juice Cans data. Here, we are manufacturing the cardboard cans used to store frozen orange juice.

▶ We visually inspect the cans for obvious defects that would cause leaking. We take $m = 30$ samples of size $n = 50$ and the data are contained in the Orange Juice Cans Excel spreadsheet (sheet 1) in D2L.

The *p* Chart

- ▶ Now, practically, there are decisions we have to make with regards to the design and implementation of the *p* chart.

- ▶ Namely, we have to choose the sample size, *n*, the widths of the control limits, *L*, and the sampling interval (time between samples).

The p Chart

▶ Regarding the sampling interval, we ought to take samples at a rate that makes sense for our process.

▶ If we're manufacturing Pringles chips, it wouldn't make much sense to sample one can per day because the production rate is quite fast.

▶ In today's world, it's not unreasonable to think we could sample 100% of the units. In this case, we have to decide how to best group our observations together in order to get $\hat{p}$'s (hourly, shift-wise, etc.).

The p Chart

▶ Sample size selection can be a bit more nuanced. For instance, if our $p/\bar{p}$ is small and so is n, then every point that contains a non-conforming point would be considered OOC, which is probably not reasonable.

▶ For example, suppose $p = 0.01$ and $n = 8$. This means:

$$UCL = 0.01 + 3\sqrt{\frac{0.01(0.99)}{8}} = 0.1155$$

$$LCL = 0.01 - 3\sqrt{\frac{0.01(0.99)}{8}} = -0.0.0955 \rightarrow 0$$

The p Chart

▶ So if we take a sample and have 1 nonconforming unit, $1/8 = 0.125 > UCL$.

▶ To get around this, we want to choose our sample size such that we have a high probability of obtaining at least one nonconformity.

$$P[X \geq 1] = \sum_{x=1}^{n} \frac{n!}{x!(n-x)!}(0.01)^x(0.99)^{n-x}$$

The p Chart

▶ We know from probability theory that:

$$P[X \geq 1] = 1 - P[X < 1] = 1 - P[X = 0]$$

▶ Thus:

$$P[X = 0] = \frac{n!}{0!(n-0)!}(0.01)^0(0.99)^n = (0.99)^n$$

The p Chart

▶ So, if we wanted to fix $P[X \geq 1]$ such that it equaled some probability like 0.95 we can arrive at our sample size with some straightforward algebra:

$$0.95 = 1 - (0.99)^n$$

$$0.05 = (0.99)^n$$

$$\implies n = \frac{\ln(0.05)}{\ln(0.99)} \approx 299$$

The *p* Chart

▶ One thing that we have very briefly touched on so far is that depending on the nature of our process, we may calculate a lower control limit that is less than 0 which is obviously an impossible observation.

▶ Like in our last example, the actual calculated $LCL = -0.0.0955$.

▶ What if we wanted to detect a downward shift? We'd need a sample size large enough to ensure we have a positive LCL.

The p Chart

$$LCL = p - 3\sqrt{\frac{p(1-p)}{n}} > 0$$

$$\implies p > 3\sqrt{\frac{p(1-p)}{n}}$$

$$\implies n > \frac{9(1-p)}{p}$$

The p Chart

- ▶ Like the $\bar{x}$ and R/s charts, there may be instances where we do not have equal sample sizes at each time point.

- ▶ Fortunately, we can handle this situation without much trouble, but it makes our Phase I and Phase II control limit calculations slightly different.

- ▶ We'll still take m samples, but each sample has a variable sample size n_i.

The p Chart

▶ If we let the number of nonconforming units taken at sample i be denoted as D_i, then:

$$\bar{p} = \frac{\sum_{i=1}^{m} D_i}{\sum_{i=1}^{m} n_i}$$

▶ And the control limits will be nearly identical to what we had previously, but with our sample size being indexed by i:

$$\bar{p} \pm 3\sqrt{\frac{\bar{p}(1 - \bar{p})}{n_i}}$$

The *p* Chart

▶ So let's go through an example. Let's say we work in the purchasing department of some company. Each week, we issue purchase orders for various supplies our company needs. Clearly, this won't be constant from week to week.

▶ We want to create a control chart monitoring the proportion of purchase orders which contain errors. Obviously, if our POs have errors, this impedes expediency on getting supplies, which can detrimentally effect other aspects of our organization.

▶ Our Phase I data are contained in the PO Excel spreadsheet in D2L.

The p Chart

▶ Okay, so in Phase I, how can we know how to set up our control chart to be efficient at detecting shifts away from target?

▶ Just like for $\bar{x}$ charts, we can use OC curves and average run length.

▶ When using OC curves, recall that we are primarily interested in estimating the probability of making a type II error, β.

The p Chart

$$\beta = P[LCL < \hat{p} < UCL | p = p_1]$$

$$\beta = P[\hat{p} < UCL] - P[\hat{p} \leq LCL]$$

$$\beta = P[D < nUCL] - P[D \leq nLCL]$$

The p Chart

▶ Remember, D follows a binomial distribution so we won't be able to calculate exact probabilities. We'll have to round to approximate β.

▶ Let's use the original control limits from the Orange Juice Can example to calculate β.

The *p* Chart

▶ Since the *p* chart assumes that samples are mutually independent, then the in-control and out-of-control ARL's can be calculated the same way as we learned previously.

$$ARL_0 = \frac{1}{\alpha}$$

$$ARL_1 = \frac{1}{1 - \beta}$$

▶ So in the orange juice example, let's figure out what the ARL_0 is as well as the ARL_1 when the fraction nonconforming has shifted from 0.2081 to 0.1108.

Final Remarks

▶ As shown, when our quality characteristic is a proportion, the *p* chart can be used as an effective tool for monitoring, especially when we're mostly interested in detecting large shifts effectively.

▶ However, the *p* chart has two main drawbacks. First, it doesn't take into account the type of nonconformity or the number of nonconformities a single inspected unit possesses.

▶ Second, the fraction nonconforming may be a less intuitive metric than the raw count of nonconformities in a single batch. When there are any difficulties in interpretation, this can impede wide adoption of a quality program.