S

OPTIONAL EXTRAS

(all more or less related to control-charting)

NB If you are viewing this file on-screen then you may find that a few of the diagrams are somewhat distorted. To the best of my very limited knowledge about such matters, the nature and extent of this problem seems to depend on the size of the screen and/or the version of the Adobe Acrobat Reader that you are using, However, you should nevertheless see everything as intended if you print the file.

If you have more expert knowledge than I do about such problems and can offer me some help on how to overcome them, do please contact me on at henryneave@sky.com.

OPTIONAL EXTRAS

(all more or less related to control-charting)

CONTENTS

NOTE

(This note would have been referred to elsewhere as an "Out-of-hours" note; however, of course, *all* of the Optional Extras section is "out-of-hours"!)

If you are intending to study Part A: "Using Control Charts on Funnel Experiment data" then on pages 6 and 7 you will need to draw run-charts and then control charts of your data for the first two Rules of the Funnel. Your data for Rule 1 are on Day 3 page 47 *[WB 40]*, and your data for Rule 2 are on Day 3 page 44 *[WB 38–39]*. You may therefore find it convenient to make a separate copy of those pages in order to save you from lots of page-turning while you are constructing those charts.

INTRODUCTION

Yes: this material *is* all "optional extra". Your understanding of the main material of *12 Days to Deming* does *not* depend on your reading any of this. So then the obvious question is why *should* you bother to read any of it?! The only reason right now, as you read this page for the first time, would be that one or more of the topics in the list of Contents that you've just seen have struck you as being of possible interest to you. If not then, indeed, don't bother!

But, in addition to the main title "Optional Extras", there were some words in brackets. Yes, these Optional Extras *are* all connected with control-charting—although some of what's in that list of Contents does not appear to fit that description. In particular, what has "A crash-course in conventional Statistics!" got to do with control-charting? I'll get onto that very soon.

On Days 2 and 3 and during parts of both of the projects, the emphasis was on understanding variation and why that is useful, and the role it plays in the improvement of processes of all kinds, especially management processes. That main material naturally included some initial work on constructing and interpreting control charts, with the section "Control Chart + Brain" on Day 3 pages 26–29 being particularly important for helping you to interpret them. But, as with valuable tools in any area of work, there is always much more experience to be gained on how to use them wisely and to best advantage over and above merely following the basic guidance in the instruction leaflet. So this optional extra material focuses not on *why* control charts should be used—which I believe you now know already—but *how to use them wisely*.

So there can be little doubt as to why Parts A and B are here. Firstly, following the coverage of understanding variation and control charts in the morning of Day 3, you spent the afternoon working with the Funnel Experiment. But unfortunately there was insufficient time available there to gain more experience by using control charts on data (both yours and mine) from that experiment. So Part A here fills that gap.

Secondly, the coverage of control charts during the main course included their use only with what I called "one-at-a-time" data, because that is all that is available in most areas of application. However, there are some areas, particularly (but not only) in manufacturing, where "a-few-at-a-time" data are easy to obtain: so Part B provides some introductory details and guidance on that further development of the technique.

However, as already mentioned, on looking further down the list of Contents there is little doubt that the title of Part D catches your eye! Why on Earth should "A crash-course in conventional Statistics!" be here? For doubtless you will have noticed in the main course material that both Dr Deming and I, amongst others, have been at some pains to emphasise the *irrelevance* of conventional Statistics to the understanding of variation as developed by Walter Shewhart and so enthusiastically adopted by Dr Deming that I sometimes refer to it as the "launchpad" of his vitally important life's work.

But don't misinterpret what I have said about conventional Statistics. It is a fascinating subject, and it is a very *useful* subject in many areas. Hopefully, if you have no background in conventional Statistics, what we have developed in this course regarding the understanding of variation and the use of control charts will have seemed pretty sensible. But that might not have been so true if you *have* had some background in conventional Statistics since some conflicts between the two approaches may well have then become apparent to you. Let me refer ahead to the start of the very final part of these Optional Extras:

> "When introducing control charts (for one-at-a-time data) to delegates at my seminars, reactions were usually very positive, even from those who started out by saying such things as 'I can't do Statistics' or, worse still, telling me in advance that they hated the subject! A little while later there were instead expressions of relief, even surprise, when they discovered how straightforward the technique is, how relatively simple are the calculations involved and, before long, how they were able to interpret what the charts were telling them."

But there were exceptions. On Day 1 page 6 I said:

> "Over the nearly 20 years of my seminars on Dr Deming's teaching I rarely suffered from any 'difficult' delegates. The few that I had could be divided into two types. One type were very senior managers, the other type were those with some qualification in Statistics."

To tell you the truth, the latter were often the more difficult type. That may sound rather flippant, but it isn't. It can be very serious. If it happens to you then I want to help you to deal with it. For *you* may not have any qualification in Statistics. So are the people in your organisation likely to believe you or the one who *is* qualified in the subject?

If you had a good teacher on conventional Statistics then, wholly unlike those delegates referred to at the bottom of the previous page, you may have become very enthusiastic about that version of the subject. I know, for I was one who *was* fortunate enough to have such an excellent teacher. He was the late Dr Clive Granger who was eventually awarded the Nobel Memorial Prize in Economic Sciences in 2003 and was knighted in 2005. Had it not been for the way he taught the subject, I might have joined the ranks of those who "couldn't do Statistics" and who "hated the subject"! As it was, I decided to specialise in Statistics, had Clive as the supervisor of my PhD research, and subsequently became the first full-time Lecturer in Statistics in the University of Nottingham's Department of Mathematics. *But* ... recall some of what I said about my first exposure to the Red Beads Experiment on Day 2 and to some of Deming's other teaching on the more statistical aspects of his work. Where were the probabilities, where was the normal distribution, why that "rough-and-ready" guidance about "3σ" coming from Shewhart rather than, for example, "3.09σ" which has a really nice probability interpretation? Fortunately, *very* fortunately, I also had two very excellent teachers in the shape of Drs Deming and Wheeler to help me through all that. But relatively few others have had all that supreme good fortune.

So, whether or not you have a background in conventional Statistics, there you have many clues as to why much of the remaining content has been included within these Optional Extras. If, like me, you had a good grounding in conventional Statistics, I hope that what you will find here, starting at Part C, will help you through the problems that I had. And if you haven't, but there are others around you in your organisation who do have such a background, they are likely to be quite an obstacle to your progress with what I may call "the real thing". The "crash-course" will help you to understand what their view of Statistics is all about and why they don't think much of what you are trying to tell them (and *vice-versa)*! That understanding will help you to communicate with them. If you understand the way that they think, and why, they are likely to be more willing to listen to you. So there is more material here to help you to then make progress. I would point in particular to (a) some writing about "statistical studies" from Deming which I describe in Part C and to (b) some computer simulation studies that are in Part E. Those simulation studies are actually set wholly within the conventional statistician's understanding of the subject; nevertheless, they prove conclusively that some of the main arguments which are often voiced about control charts by the conventional statistician are patently and completely wrong.

Finally, Part F is a "Technical Section" which is likely to be of more interest to mathematically-inclined students: a number of results are verified here that are simply quoted and then used either elsewhere in the course or in these Optional Extras, and there is also some additional discussion on the practical problem which is faced by anyone who starts using control charts: how many data to use when computing your control limits.

So please just pick and choose what, if anything, you read here in these Optional Extras. Or just browse through them and see if anything catches your eye. Maybe there will be nothing—if so, that's fine: just stick to the main course. But these Optional Extras will, of course, all still be here if and when the time comes later when you suspect that some of them *might* be useful to you after all.

PART A: USING CONTROL CHARTS ON FUNNEL EXPERIMENT DATA

—an extension of Major Activity 3–h

1. Introduction

If you have read my discussion of the Funnel Experiment Major Activity 3–h on Appendix pages 15–18, you will remember that I studied two simulated sets of Funnel Experiment data that had been obtained using a computer program which I'd written in order to demonstrate the Funnel Experiment in my seminars. You, of course, might also like to reproduce that experiment by using a spreadsheet or writing a computer program if you are talented in that way—and, if you are able, I recommend that you do so. I personally found it extremely instructive to play around with my program in order to become familiar with the experiment and with the learning that it helps to develop.

In case that might be possible, let me tell you what I did in my seminars. As you will remember, Dr Nelson's original version of the experiment is rather tricky to carry out for real except in a small group that has plenty of time to spend on it. And my "one-dimensional" version of the experiment that you used during Major Activity 3–h is also really only convenient for at most a small group rather than being in a form suitable for presentation in front of a larger audience. So that is the main reason I initially wrote a computer program to demonstrate the Funnel Experiment in my seminars. It was presented on-screen in front of the delegates, and represented a bird's-eye view of the table in Lloyd Nelson's original version of the experiment. A little icon indicated the current position of the funnel above the table and, at the click of a button, effectively a marble was dropped through the funnel and its final resting position was shown. At the following click of the button, the funnel was moved to its next position according to whichever Rule was being demonstrated. And so on. So this program enabled each of the four Rules to be initially carried out slowly step by step, then slightly faster, and then as fast as we liked. This therefore allowed the delegates to see and understand precisely what the Rules are before proceeding to examine their effects. All the previous resting positions of the marble were retained on-screen in order to study the long-term patterns and interpretations.

The prime purpose of Day 3 as a whole is best summarised by the short title of Don Wheeler's excellent little book: *Understanding Variation*. So, looking forward to the time when you are actively working on interpreting real data and improving processes, etc, a really important aspect of Day 3 is your becoming familiar with the use of control charts in order to help you to do just that—understand variation—and know what is sensible to do as a consequence of that understanding. From that viewpoint, a one-dimensional version of the Funnel Experiment with my approach of using a couple of dice or something similar to simulate the variation is actually more fruitful than Dr Nelson's original two-dimensional version in the sense that the data we obtain can indeed be immediately analysed on control charts. So, assuming you are now familiar with the Rules from your work in Major Activity 3–h, I would definitely recommend that you develop a *one*-dimensional version if you are willing and able to develop a computer program or devise some spreadsheet method to represent the experiment. However, that's for the future! For now, let's content ourselves by working with the data you generated during the Major Activity and with the data from my two computer simulations. Having spent some considerable time in the morning of Day 3 on becoming familiar with control charts, it would of course have been logical to construct control charts of all the Funnel Experiment data in the afternoon. But time would not permit. That is why I have decided to begin these Optional Extras by making up for that forced omission. Depending on how much time you would like to spend on this, there are various possibilities available.

If either you never got round to reading that material which begins on Appendix page 15, or it's been quite a while since you did so, I suggest you first read through that as it will help to put you in the picture and remind you of matters which will be useful as you revisit the Funnel Experiment here.

As you'll recall, in Major Activity 3–h you summarised your data from Rules 1 and 2 of the Funnel on histograms and your data from Rules 3 and 4 on run charts. Histograms would have been almost meaningless with Rules 3 and 4 since those processes were hopelessly out of control. However, as it turned out, histograms were extremely useful in comparing Rules 1 and 2. Rule 1 is the straightforward in-control process, and a histogram can often provide some additional useful information about the output from a stable process. Further, if you hadn't constructed and compared the histograms for those two sets of data, it would have been quite tricky to figure out what Rule 2 was producing. Run charts and, even more so, control charts can often tell you the most important things to know about a process's behaviour, but sometimes the histogram can shed extra light on the subject.

So, what are those "various possibilities" of what you might do next? The most challenging approach would be to throw you in at the deep end and simply suggest you construct control charts of your data for all four Rules, and see how you get on. When you get to Rules 3 and 4, starting on page 8, you will already have your run charts available in the main text or the Workbook, but with Rules 1 and 2 you will need to begin by drawing the run charts here. Similarly as with Rules 3 and 4 in the main text, I have provided here your graph-paper for Rules 1 and 2 with the first five points of the run charts already drawn in. So firstly complete those run charts using your data from Day 3 page 47 *[WB 40]* which came from Rule 1 (the *Second* Strategy in the Ford case) and page 44 *[WB 38–39]*, i.e. Rule 2 (Ford's *First* Strategy). Then move on to producing your control charts: refer back to Technical Aid 6 on Day 3 page 16 if you need to. Make some notes on what you learn and also on any problems that you encounter. When constructing your own control charts, you will, as always, need to decide on your baseline, i.e. how many data to use for computing the control limits (see Technical Aid 8 on Day 3 page 17). As a quick reminder, in that Technical Aid I suggested a baseline of between 10 and 15 observations—or less if you're in a hurry to get started! There is much more detailed discussion on the length of the baseline at the start of the Technical Section in these Optional Extras on pages 71–74.

Having worked on control charts for all four Rules using your own data, then move on to my control charts and discussion, starting on page 9, for the two simulations that I worked with on Appendix pages 15–18. Here I've found it useful to extend the control charts to 50 points rather than the 40 shown there—the extra length permits a few effects to be demonstrated more clearly. You might be able to expand on the notes you will have made earlier, and maybe the discussion here will shed light on any problems you found.

A possibly more appealing approach might be to reverse the suggestions I've just made. That is, *first* read the material on my two simulations beginning on page 9. Then, in particular and if it appeals to you, you could then use the same approach with your own data from Activity 3–h as I have done there by producing both "short" and "long" control charts: that can be quite instructive.

If you are short of time then you could, of course, take the really easy way out: just read my material here for the time being, and then return to try out control charts with your Funnel Experiment data on some later occasion!

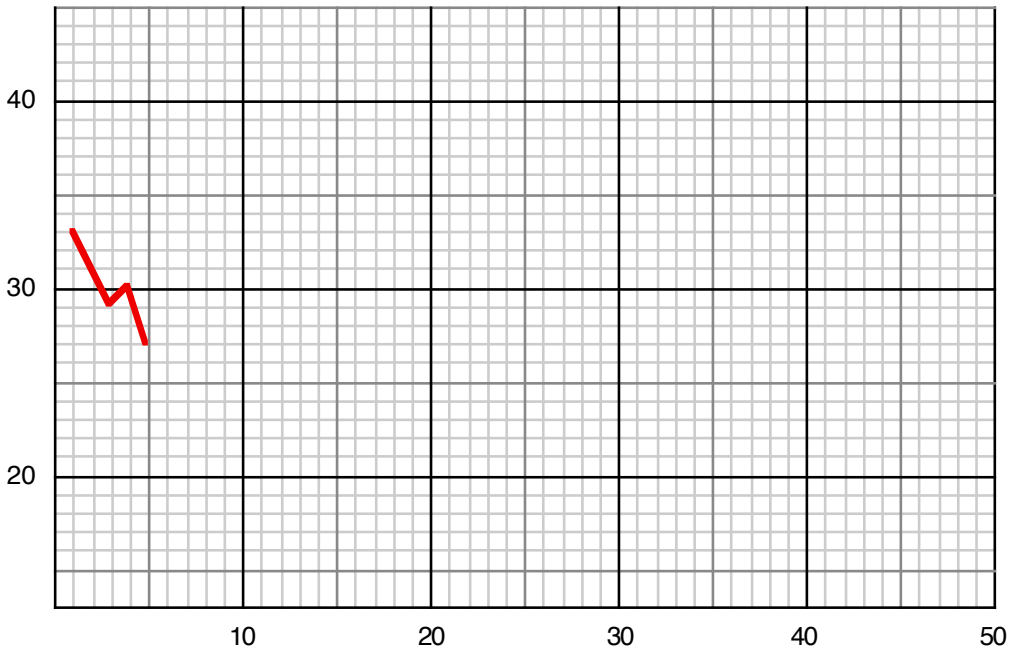
Pages 5–7 are also on Workbook pages 249–251.

2. Rules 1 and 2 of the Funnel

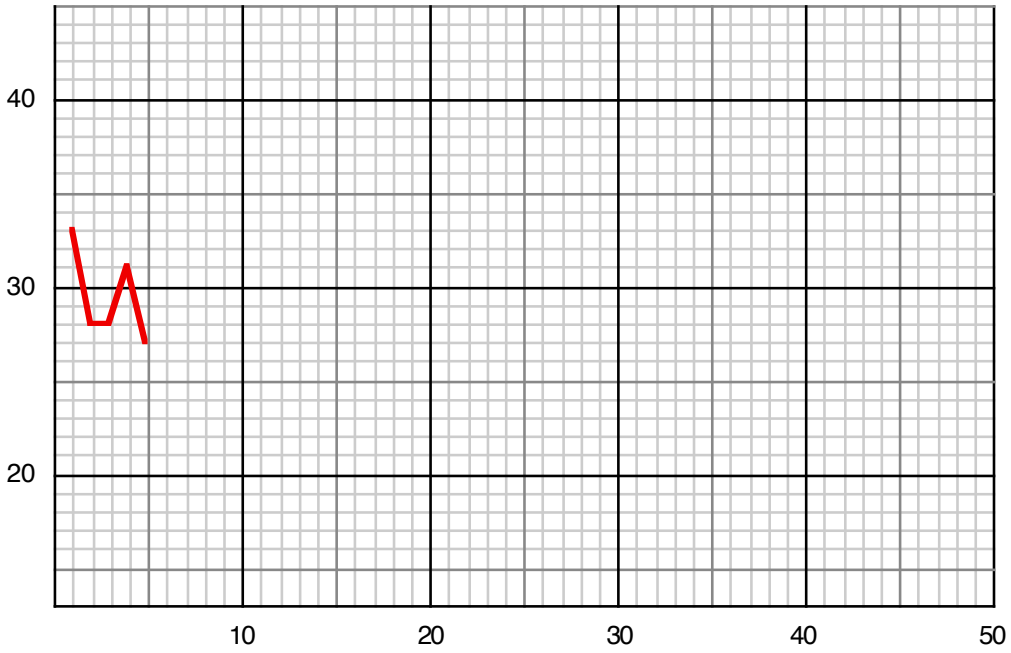
So now go ahead with Rules 1 and 2 using the data you generated on Day 3. Again, your Rule 1 (Ford's Second Strategy) data are on Day 3 page 47 *[WB 40]* and the data from Rule 2 (Ford's First Strategy) are on Day 3 page 44 *[WB 38–39]*.

Seeing that you have not drawn the run charts of these data previously, I suggest it would be a good idea for you to develop the charts for Rules 1 and 2 "live". That is to say, reflect the more usual and better practice of first (a) drawing the run chart over your chosen baseline *but no further*; then (b) computing the positions of the Central Line and the control limits from those data; (c) inserting these three lines on the graph-paper throughout the baseline and then somewhat further into the future if all seems to be well at that stage (i.e. if it currently appears feasible that the process is in control); and finally (again if all seems well) (d) continuing the chart *one point at a time*. Imagine that you were seeing these data for the first time, so that you don't know beforehand what you learned when generating them and constructing their histograms back on Day 3. There is room for your computations and whatever notes you care to make under the graph-paper on the next two pages. In both cases, describe what you feel the control chart is telling you *as and when you are developing it*: a kind of brief running commentary.

Rule 1



Rule 2



3. Rules 3 and 4 of the Funnel

The "motivation" for the various Rules is discussed in *DemDim* Chapter 5, so there is no need for me to say much about that here. In brief, recall that, as we've indicated previously, Rule 3 is at first sight a rather innocuous-looking variant of Rule 2, while Rule 4 concentrates on trying to minimise average short-term variation. The latter is, of course, an interesting mixture of good and bad. It is good to reduce variation, but is it wise to do so only in the short term? Let's carry out similar procedures as previously but now for Rules 3 and 4, and see what happens.

Seeing that you have already drawn the run charts for your Rules 3 and 4 data (Day 3 pages 51 and 55 *[WB 43 and 45]* respectively, in each case preceded by the relevant data), I'm not going to suggest that you now start again! But you can at least *pretend* that you are going through the same procedure you have just been following with Rules 1 and 2, i.e. developing the control charts "live": you will, of course, find some substantial differences compared with what happened in them!

So, in both cases, compute the positions of the Central Line and control limits from the data over your chosen baseline and draw them in on your run chart. Now, it's not impossible that in either or both cases you could actually get one or more signals (points outside the control limits) even during the baseline. This is more likely with Rule 4 than with Rule 3. But, even if that doesn't happen, imagine you haven't seen the rest of the run chart (try covering it up for the time being!) and see if you would already have any different thoughts compared with what you had at the same stage in your "running commentaries" on Rules 1 and 2. Seeing that, of course, you already know what actually happened with these processes, that pretence might be quite difficult! But, with Rule 3, are you already seeing the first signs of the zig-zag effect that becomes the overwhelming feature of that rule sooner or later? Or, with Rule 4, are you already seeing some indication of the "wandering" nature of that process? It might be a good idea for you to briefly look back at Day 3 page 19 and remind yourself of the control charts on the left-hand side of that page. If you recall, those processes were mainly fairly happily in control, although there was some doubt regarding a possible seasonal effect in the chart at the bottom of that page. But, generally, the question to ask is whether or not your charts look noticeably different from those charts on the left of Day 3 page 19 over the baseline. And then gradually uncover the rest of the chart: in both cases, describe what you feel the chart is telling you, and *how soon* it is doing so.

Obviously, I don't know how your data turned out, so I can't describe what you will or won't see. So it's over to you now to discover what happens. When you return, move onto the next section to study the control charts and my discussions on them for the two sets of computer-generated Funnel Experiment data that I introduced on Appendix page 15.

4. Control charts for the computer-generated data

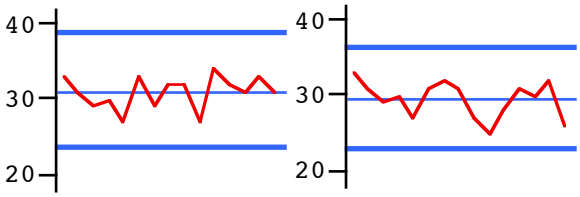
I hope you will have found it helpful to gain that additional experience of constructing and interpreting control charts. However, to be fair, Funnel Experiment data are not the best kinds of data to impress you of the control chart's usefulness! The reason stems back to something I said on Day 3 page 18: "The control chart becomes really valuable when it is *unclear* as to whether or not the run chart is indicating there are some special causes—which is the more usual situation". However, as we have seen, the run charts that result from the Funnel Experiment *are* mainly pretty easy to interpret! The Rule 1 run chart will of course have generally appeared very stable; and, unless you were very unlucky with your throws of the dice, inserting the control limits should then have produced control charts reminiscent of the various control charts that you have seen of stable processes such as those from the Red Beads Experiment and the other in-control processes on the left of Day 3 page 19. And, almost certainly, you didn't really need control limits when dealing with Rules 3 and 4 to convince you that those processes were unstable!

But let's fill in a little more detail. Upgrading a run chart to a control chart by inserting control limits results in two important gains. First, it enables you to have much greater confidence in your judgment as to whether the process is or is not in statistical control, whereas with many run charts such judgment is little more than guesswork. And second, it often allows you to make your judgment *earlier* than if you were only using a run chart. Both of these features are extremely important advantages in practice.

So let's now examine control charts produced by the two runs of the Funnel Experiment that were demonstrated in the Appendix. It will be useful to look at both the complete control charts and also how the charts appear when the control limits are first drawn in—I'll refer to the latter as the "short" control charts. That, of course, occurs as soon as the baseline data have been recorded: I have used a baseline length of 15 in the following charts. What might we learn and what might we predict at that early stage? I'll deal with the charts from the two simulations for one Rule at a time.

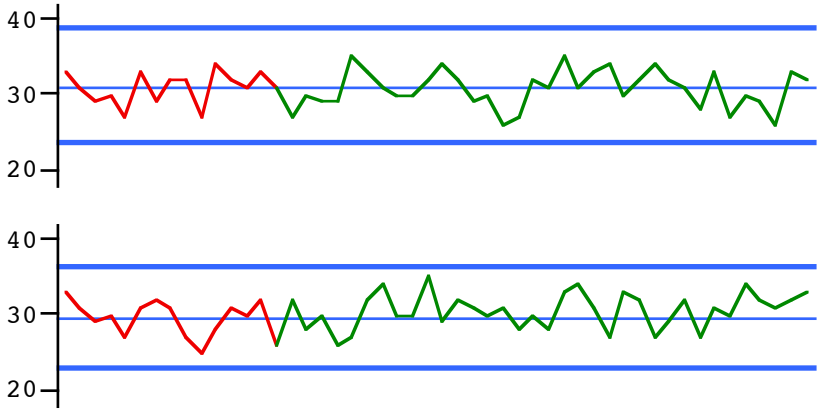
Rule 1

As expected, the short versions of the control charts for Rule 1 hold no surprises for us.



As we would have hoped, the Central Lines and the control limits are quite similar in the two simulations—although, of course, they're not in *exactly* the same places. How could they have been? They're almost bound to be different when using different sets of data from a process, however stable the process may be. This is analogous to the situation in conventional Statistics when drawing samples from the same distribution or "population": the sample means and sample standard deviations (see page 18 and onward in Part B of these Optional Extras) will always differ, except for a very rare fluke.

In the long versions of the control charts I've coloured in green the sections following the baseline, i.e. after the positions of the control limits have been calculated, drawn in and then extended into the future. And again, as expected, we get two charts strongly reminiscent of the control charts representing the stable processes on Day 3 page 19.



Rule 2

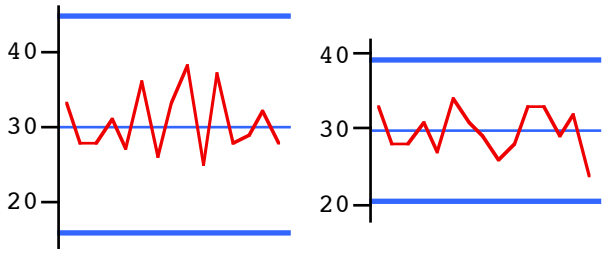
Rule 2 is the case where, as we know both from the Ford example and from your own work in the Major Activity, the situation is undesirable compared with Rule 1, in particular suffering from approximately 40% greater variation. However, going by Bill Scherkenbach's account of the Ford example, it appears that the people involved there had no history of using Rule 1—it would seem that the process was actually *set up* as Rule 2, i.e. using the automatic compensation equipment. Perhaps a convincing salesperson for the equipment was around when the process was being designed! Also, as far as we know, they were not analysing their data even on a run chart, let alone a control chart: the implication from those histograms is that they were simply judging quality in terms of conformance or non-conformance to specifications (a *poor* method of judgment investigated on Day 7). And, with the compensation equipment in operation, almost all of the shaft diameters *were* within specifications—as seen in the first histogram on Day 3 page 5.

> Mind you, the histogram indicates that quite a few were uncomfortably close to the edges of the specifications. In fact (as observed on Day 3 page 9), if you count carefully, that histogram appears to show only 49 diameters rather than the 50 that were claimed, so maybe one had just slipped over the edge and—shall we say?—vanished!

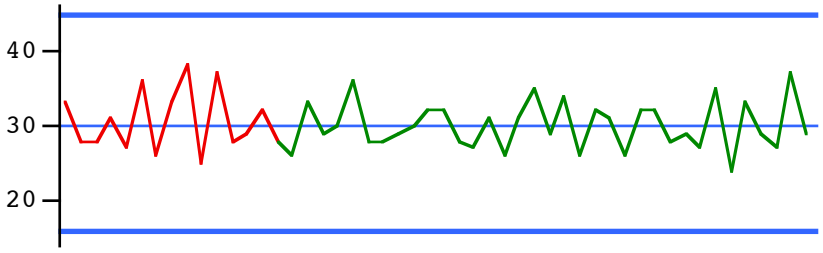
But suppose they had been using a run chart. Would they have noticed anything was amiss? Without a Rule 1 run chart with which to compare, the answer is probably No. (Presumably, had a comparison with Rule 1 been available, particularly with histograms, they would have noticed Rule 2's larger variation along with the fact that *none* of Rule 1's diameters were close to the edges of the specifications.) So what else might have been seen? I did point out in the Appendix that Rule 2's run charts are relatively "jagged", but it would probably have taken an experienced eye to notice something of that nature.

So how about control charts of the Rule 2 data in our two simulations? Do they appear at all different in nature from control charts of genuinely in-control processes? Let's see.

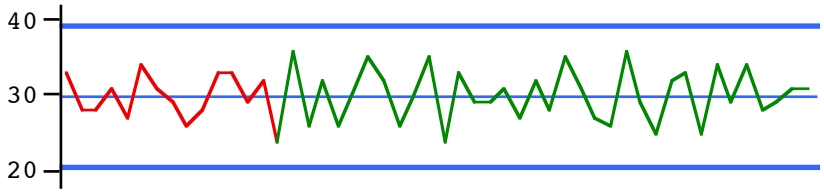
Again let's look first at the short control charts of the two simulations. The second one does not appear to have anything much to tell us, but I suggest the first one does. There are two features which, compared with the control charts of the six stable processes on Day 3 page 19, look rather odd. First, there is the pronounced zig-zag in the central part of the baseline. It's rare to see anything like that in the pictures on Day 3 page 19. But also note the relatively large amount of "white space" between the graph itself (i.e. the run chart) and the control limits, especially between the graph and the lower control limit. That's the same effect as was seen in the illustration on Day 3 page 24—the effect often referred to as "hugging the Central Line", i.e. where none of the points are *anywhere near* the control limits. We have previously described an in-control process as one where (almost) all the points are "comfortably contained" between the control limits; but "hugging the Central Line" is where they are far *too* comfortably contained between the limits! Hugging the Central Line is not good—you should be very suspicious of it!



And, as becomes very clear, in the first simulation of the experiment this effect continues all the way through the complete control chart. Even in the second simulation (see the chart at the top of the next page) where the effect is less pronounced, note that, throughout all 50 values, not one gets at all close to



either limit—so I suggest that that might be considered as at least *slightly* suspicious! But I'd say the clear impression in the first case is that the control limits are simply "wrong": they should be closer together (so yet again compare with those six in-control charts on Day 3 page 19). And, having made that observation, maybe it would seem that those limits in this second simulation should at least be a *little* closer together.



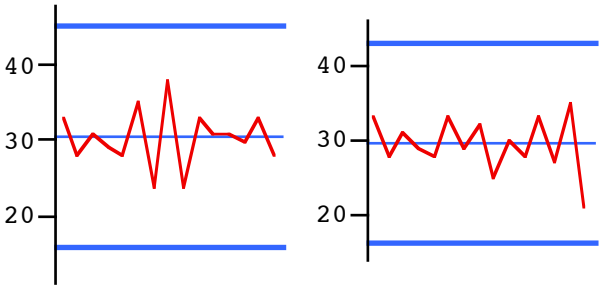
But, with our knowledge of what Rule 2 is, isn't that precisely what we would expect? Rule 2's compensation mechanism virtually ensures that particularly high values are followed by particularly low values, and *vice-versa*—i.e. the zig-zag effect already observed. But recall how the control limits are computed—the distance between them is simply proportional to the average moving range $\overline{MR}$. Clearly, Rule 2's zig-zag effect *increases* the moving ranges (the differences between adjacent values) compared with what would be expected if that compensation scheme were not operating. The control limits are indeed "wrong" in the sense that they now have no chance of reflecting the *actual* variation, precisely because of that zig-zag effect.

Now, as we have seen, the hugging-the-Central-Line effect will be more apparent with some sets of Rule 2 data than others: it's a matter of luck! But, at least, we now know that the control chart stands a reasonable chance of indicating a problem with Rule 2, *even when there isn't a Rule 1 version with which to compare it*. That is much less true with both the run chart and the histogram.

(Please move on to the next page for Rule 3 and then Rule 4.)

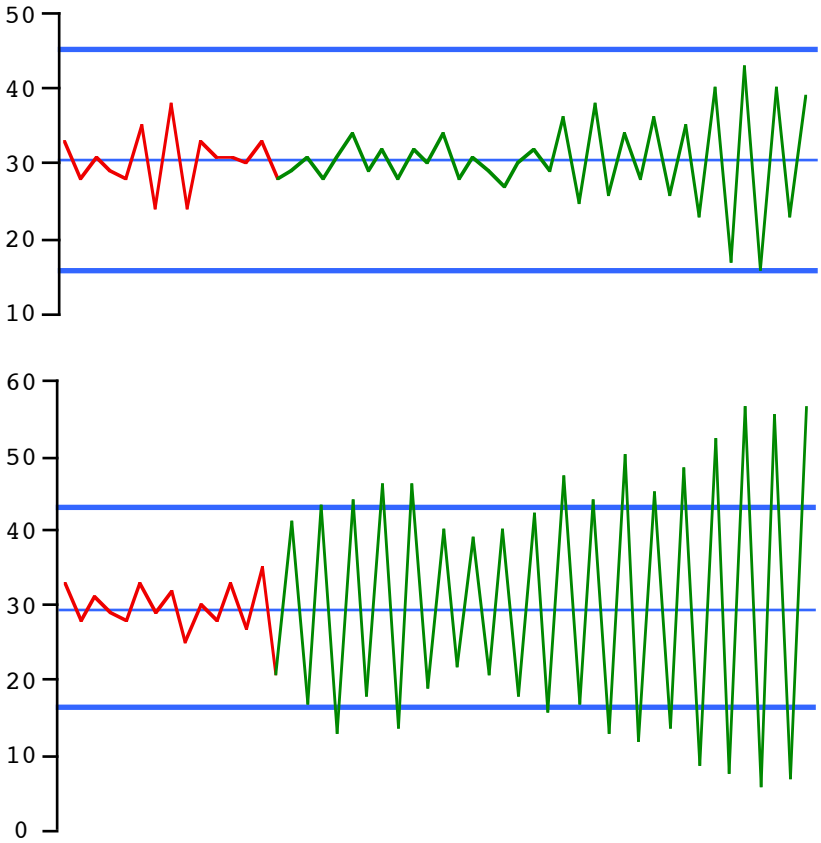
Rule 3

Since you have already seen the run charts of both Rules 3 and 4, you may well suspect there is not much further description necessary in either case. You'd be right! There are just a couple of points worth mentioning but, by and large, the run charts told the stories more than adequately.

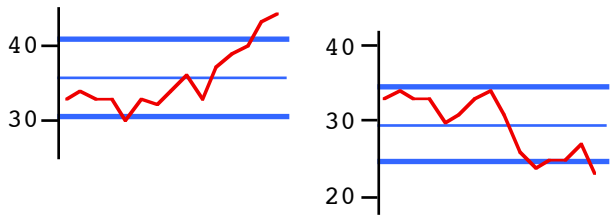


One aspect immediately noticeable about Rule 3's short control charts is their similarity to those of Rule 2 on page 10. A moment's reflection will show why. Recall that the only difference between the two Rules is that in Rule 2 the funnel is moved relative to *its current position* while in Rule 3 it is moved relative to *the target of 30*. So while the funnel, and hence the marble, both stay fairly close to the target, the **Outcome**s (i.e. positions of the marble) will be quite similar in the two cases. And that is what we see in these short control charts.

Sooner or later, however, the marble will finish up rather further away from the target than previously which will, in Rule 3's case, position the funnel at that same larger distance *on the other side of the target*. And then the very severe zig-zags are likely to really get moving! It is possible that, with some lucky throws of the dice, they may die out for a while—as does indeed happen during the first simulation alongside following those early zig-zags within the baseline). But be sure: they will always return and in time will become absurdly large as is demonstrated here in the second simulation:

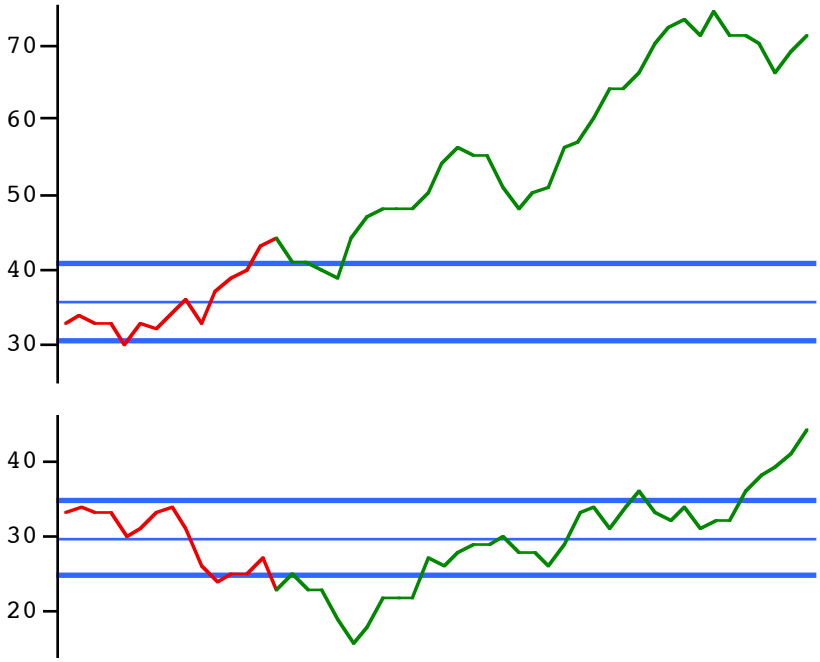


Rule 4



Whereas Rule 3's short control charts did not give much warning of the horrors to come, Rule 4's short charts immediately show conclusive evidence of severe problems. They would have done so even if we had been using a shorter baseline than recommended in my general guidance. The reason lies in the very motivation for Rule 4: to reduce *short-term* variation. And remember that it is precisely the short-term variation which the moving ranges (and hence $\overline{MR}$ itself) measure, with the immediate effect here of substantially *narrowing* the gap between the control limits. That reduced short-term variation is, of course, a delusion: in reality, this process is not capable of such low variation. So, if you were trying Rule 4 after you had tried Rule 2 (or indeed Rule 1), you might be quite excited when you compute Rule 4's limits! But not as soon as you insert them on the run chart and see how the graph behaves in relation to them. The "wandering" effect that was already discussed on Appendix page 19 causes points well outside the control limits to start arriving thick and fast—probably, as in both simulations here, even within the baseline itself or if not then very soon afterwards.

So if you obtained either of those short control charts in practice then there'd have been no point in extending their control limits into the future. The purpose of control limits is to help you to notice when the process goes out of control. But here you already *know* it's out of control, and so hopefully instead you would be immediately trying to find out why! However, just for completeness (and perhaps amusement!), here alongside are the long control charts!



Let's now summarise and develop what we have learned in these last few pages. Note that we haven't really been studying the four processes in terms of control charts—we've been doing the reverse! The truth is that we had already learned most of what there is to know about these processes from your own work in the Major Activity and from the two sets of simulated data in the Appendix. We have instead been investigating what the control charts look like when each of the four Rules is in operation. And that's useful: the way one often studies suggested procedures or methods is to let them run under known conditions and see what happens. We now know that if an unjustified compensation effect is in operation then the control chart will have unusually wide control limits, possibly resulting in the hugging-the-Central-Line phenomenon and/or (in the case of Rule 3) wild zig-zags which eventually go outside even those over-wide control limits. Conversely, we also know now that if adjacent values in the data are unnaturally close together—as they are bound to be with Rule 4—then the control limits will be correspondingly close together: so much so that we are likely to get points outside those limits even within the baseline, let alone subsequently. Also, of course, the "wandering" effect will be all-too-obvious.

The fact that the control chart shows these features so clearly is particularly notable since these are kinds of data which it was not really designed to work with! As you know, the basic idea of the control chart is that the formula for deriving the control limits is designed to indicate the scale of the process's common-cause variation—even, in many circumstances, when the process is already *out of control* (which is itself, of course, quite an achievement). But how else could the control limits enable the chart to detect special causes? When the process is in control, we've seen both here and with the collection of six processes on Day 3 how well the control limits illustrate the extent of the common-cause variation. But how can they still manage to do that when the data from which they are computed are disturbed by special causes?

That question was largely addressed in the "How do we compute those control limits—and why?" section beginning on Day 3 page 13. There we argued how moving ranges can often succeed pretty well in meeting that challenge, certainly compared with the conventional statistician's standard deviation (about which, if you wish, you can read more in Part B of these Optional Extras). However, in the middle paragraph of Day 3 page 15, I warned you that the Funnel Experiment reveals some serious exceptions where even the moving-range method is wholly unable to perform as just described. For to say that moving ranges can reflect the scale of the process's common-cause variation *even if computed from data recorded when the process is out of control* does depend on an implicit assumption. This assumption is that it's *mostly* true that the two adjacent values contributing to a moving range are themselves still free to be "typical" values from the process. This is not unusual with many special causes that occur in practice. But it's clearly not the case with Rules 2, 3 and 4. In Rules 2 and 3, if one value is high then it is likely that the next will be low. Conversely, in Rule 4 any two adjacent values are bound to be relatively close to each other. Both effects clearly destroy the ability of moving ranges (and hence $\overline{MR}$ which is used in the computation of the limits) to guide us as to the size of common-cause variation. Data in which any one value considerably influences what the next value is are said to be "autocorrelated"; in the case of Rules 2 and 3 they are said to be *negatively* autocorrelated, and in the case of Rule 4 they are *positively* autocorrelated.

So now you know that, if you see either the zig-zag effect or the "wandering" effect in a control chart, you might immediately have a suggestion as to what might be happening in the process being studied. But beware: don't jump to conclusions without further thought. Although unjustified compensation causes the zig-zag effect, it's not the *only* possible cause of that effect. As a stupidly-simple example, let's suppose you are recording the outside temperature every 12 hours: at midday and at midnight. Yes, I think that would produce a pretty impressive zig-zag!

Less trivially, suppose the data alternately measure efficiency or productivity etc of both a day-shift and a night-shift. There could be many reasons for a zig-zag here: effectively, you are likely to have *two* processes in operation rather than just the one. The night-shift might be disadvantaged because of poor lighting or an inefficient heating system. Alternatively, the day-shift may be disadvantaged since the computer

system runs rather slowly during the daytime because everybody is using it—whereas if relatively few staff are working the night-shift then the computer may be operating like greased lightening! The crucial practical point is that, as soon as the zig-zag is seen on the control chart, discussions can immediately start on the reason(s) for it—and, because the control chart is such a straightforward statistical tool (not encumbered by what some might consider to be the usual mathematical mumbo-jumbo of conventional statistical techniques), *everyone* can be involved in the discussion—not just the "experts". The importance of this with, say, process improvement teams is beyond price. Similarly, the control chart can serve very effectively as a *communication language* within and between departments in an organisation, between different levels of management, and even between organisations.

There are also lots of circumstances which will lead to positively autocorrelated data. Financial data are a good case in point. Stock market indices, inflation figures, exchange rates, etc are bound by their very nature to be highly positively autocorrelated since normally it's the case that any figure is relatively close to the previous one compared with the variation as a whole. Does that mean control charts can have no useful role to play in studying such processes? No. But agreed, there's no point in simply plotting data whose main characteristics we already know and which will almost surely drown out anything else that might be of interest. One very simple but often effective ploy is instead to chart the *changes* day-to-day (or whatever time-interval is relevant)—these changes are, of course, the same as the moving ranges except for being recorded as positive or negative according as whether they go up or down. Since this simple manoeuvre almost completely extinguishes the autocorrelation effect, such a chart is now quite likely to be able to discover other special causes that may be affecting the process.

An aspect I would like to emphasise regarding this discussion is that we haven't been involved here with an exercise in Mathematics but rather an exercise in common sense. And that is a valuable contrast between (a) using control charts and (b) using many of the conventional statistical approaches.

Of two books by Dr Wheeler that I particularly recommend, some such matters are touched upon in *Understanding Variation* but are dealt with more comprehensively in *Making Sense of Data*.

PART B: A-FEW-AT-A-TIME DATA

1. Introduction

I first mentioned "a-few-at-a-time" data a long while ago: in the discussion on the First Paradox (Appendix page 2). So it may have been quite some time since you saw it, and therefore I'll reproduce it here:

> "Another great value of Don *[Wheeler]*'s work is that, whereas it used to be the case that control charts were generally considered to require small *samples* of data to be available at each time-point represented on the chart, he popularised an excellent and simple method of how to construct control charts when only *one* value is obtainable at any particular time. Samples are often easily obtainable in manufacturing processes, but single values are all you can get with the majority of other types of process. Since most people only have 'one-at-a-time' data available in their processes, that type of data is all I consider in the main material of this course, in the Springboard article, and in the case studies covered in both *ST* and *EST*. However, in the latter little books, I do also briefly describe charts that are suitable for 'a-few-at-a-time' data."

Let me emphasise a few basics. By "a-few-at-a-time" data I am implying that we are now dealing with what were referred to above as "samples" rather than just single values. The implication of "at-a-time" is that the data concerned *are recorded pretty much at the same time and under the same conditions*. The adjective "*random*" is often seen, and a "random sample" effectively implies that, in addition to "recorded pretty much at the same time and under the same conditions", the data in the sample are otherwise not related to each other in any way. You may think that I am being rather fussy with my details here, but it turns out that such details are important regarding how these data will be regarded and used when being analysed on control charts. In fact, rather than just "samples" or "random samples", a different word has traditionally been used to refer to them in the control-charting context, namely "subgroups".

Why "subgroups"? Actually, I can't recall ever having seen that word very precisely described! But presumably the idea is that there is some *group* of observations that were recorded "pretty much at the same time and under the same conditions" but that we only have a few of those observations available to use.

2. Calculations on a subgroup

> If you want to move straight on to finding out how to construct control charts for subgrouped data (without bothering to find out *why* the details are what they are), you are welcome to skip this section for now and move straight on to Section 3 (page 20). However, if later you decide to embark upon Part D: the "crash-course in conventional Statistics!", you will need to read some of this section at that time.

For illustration, here is a *subgroup* of size 5:

1.4 1.2 2.1 1.8 1.2

> Note that there's no reason why two or more values in the data shouldn't be equal to each other (depending on the precision being used, one place of decimals in this case) as has happened here with 1.2.

In the literature it is common to refer to individual values in the subgroup by the letter X and to the size of the subgroup by n, so that here $n = 5$. When used with control charts, n is usually no larger than this.

As in the main text of Days 2 and 3, one of the first things we think of doing with a set of numbers such as these is to calculate their average. Now, although the way that we have previously computed an average (i.e. add up the numbers and then divide that total by how many numbers there are) is by far the most com-

mon method, it isn't the *only* way used by statisticians. For example, an alternative approach is introduced on page 84 in the Technical Section along with a little discussion on why that alternative might be preferable in some contexts. So, to avoid ambiguity, statisticians normally use a different term to refer to the type of average with which everybody is familiar: this is the *mean* or (to give it its full name) the *arithmetic mean*. Therefore, with this terminology, we can now say that the *mean* of our subgroup is

$$\bar{X} = (1.4 + 1.2 + 2.1 + 1.8 + 1.2) \div 5 = 7.7 \div 5 = 1.54$$

since $\bar{X}$ is the notation specifically reserved by statisticians for "the *mean* of the values being represented by X". The notation $\bar{X}$ ("X-bar") isn't used for *any* other type of average.

Also as previously, we are likely to be interested in some way of measuring the *variation* of the numbers in our subgroup. On Day 3 we became familiar with the idea of using *moving ranges* for this purpose because we were then focused on examining the way the data varied *over time*. But, in a subgroup, we are now dealing with numbers "recorded pretty much at the same time": so moving ranges are irrelevant in this context. This inevitably brings us back to the conventional statistician's favourite measure of variation, namely the *standard deviation*. This has occasionally been mentioned in the main text, notably at the bottom of Day 3 page 13 (which it would be useful for you to briefly refer back to), although it has never actually been *defined* previously—and now you'll soon be discovering why!

The definition starts out sensibly enough. The general idea is that, roughly speaking, the standard deviation indicates the *typical size of gap* between the n values of X and their mean $\bar{X}$. This is a perfectly reasonable approach to measuring the variation in a subgroup: the larger the gaps, the larger the variation; and the smaller the gaps, the smaller the variation. The trouble is that statisticians do not tackle this "perfectly reasonable approach" in what many people would consider to be a "perfectly reasonable" way!

Let's be sure about what I mean by "gaps" here: they are the *distances* between the individual numbers and the mean. So check that, with our subgroup, the gaps are, in turn, 0.14, 0.34, 0.56, 0.26 and 0.34. (0.14 is the distance between 1.4 and the mean 1.54, 0.34 is the distance between 1.2 and 1.54, and so on.) Note that I am carefully avoiding the use of words like "difference" or "deviation" since those words are generally understood to mean the value of $X - \bar{X}$. The latter is, of course, a *negative* value if X is smaller than $\bar{X}$ whereas gaps or distances are understood to always be positive (or zero). So the differences or deviations are –0.14, –0.34, 0.56, 0.26 and –0.34 respectively whereas the gaps are as listed above.

Now, surely the obvious way to get a measure of the variation in the subgroup would be to simply calculate the average (mean) gap or distance. That's to say:

$$(0.14 + 0.34 + 0.56 + 0.26 + 0.34) \div 5 = 1.64 \div 5 = 0.328\,.$$

> You might immediately notice a very good reason *why* we're using "gaps" or "distances" in preference to "differences" or "deviations": if we were to include the latter's minus signs then their sum would be zero—as would be the case with *any* set of data. Try it for yourself if you don't believe me!

This measure (using gaps) is occasionally seen in the literature, glorying in the name Mean Absolute Deviation—or MAD for short! (In Mathematics, the effect of the word "Absolute" is to get rid of those minus signs.) However, the experts rarely use the MAD. Instead, the traditional and almost universal method is to

(a) *square* the gaps (or the deviations or differences since squaring them gets rid of all minus signs!),

(b) add up the resulting "squared gaps",

(c) divide that sum of the "squared gaps" by $n-1$, and finally

(d) take the *square root* of the result.

And there you have the long-awaited definition of the standard deviation of a sample (or subgroup). *Now* you can probably appreciate why I have avoided showing it to you previously!

Before I discuss it and try to give it some rhyme and reason, let's go through the arithmetic with our illustrative subgroup, just to make sure of what's involved. I'll lay it out in those same four steps: if you wish, check it with your calculator.

(a) $0.14^2 = 0.0196$, $0.34^2 = 0.1156$, $0.56^2 = 0.3136$, $0.26^2 = 0.0676$, $0.34^2 = 0.1156$;

(b) Adding up the above gives 0.6320;

(c) $n - 1 = 5 - 1 = 4$, so dividing 0.632 by 4 gives 0.158; and

(d) the square root of $0.158 = \sqrt{0.158} = 0.397$.

For a long while I couldn't understand *why* this method was in common use! Sure, the standard deviation is *some* kind of way of producing something like an average or typical gap: but why so complicated and why do Mathematical Statisticians prefer it to the MAD? And why on Earth divide by $n-1$ in Step (c) rather than the more obvious n? The answers eventually became clear to me as I learned more about Mathematical Statistics theory.

The answer to the first pair of questions is that, quite simply, despite being easier and quicker for you and me to calculate, the MAD turns out to be very awkward to use in algebra and other mathematical derivations, whereas it is much easier to develop nice mathematics with the standard deviation. As a matter of fact, nice mathematics is even easier to carry out without even bothering with Step (d), i.e. not bothering to take the square root of the result in Step (c). The result in Step (c) is called the *variance* and, if you turn the pages of any Mathematical Statistics textbook, you will actually find the variance being mentioned far more often than the standard deviation.

Years later, when I first came across the Taguchi Loss Function (studied on Day 7), I found rather more justification for concentrating on the variance than purely mathematical convenience. What we learn with the Taguchi Loss Function is that the *square* of the gap between a figure and its middle or optimum value actually has greater practical significance than the straightforward gap. So I, at least, became rather more comfortable than previously about concentrating on variances (using squares of gaps) rather than on just the straightforward gaps themselves. However, I feel pretty sure that convenience for the Mathematical Statisticians is the more likely main reason for the common use of the variance rather than the MAD!

Actually, the good news is that, for the purpose of constructing and using control charts for a-few-at-a-time data, i.e. using *subgroups* rather than one-at-a-time data, you don't *need* to know the answers to those questions I've just raised! So why have I bothered to mention standard deviations at all? The reason is that some of the details used in constructing these and other types of control charts do involve the standard deviation (or variance) in the background theory. If you are content to simply accept the details that I tell you as being the truth rather than knowing anything about that background, then fine! Much of the content in this optional material is not essential for you to know: I'm simply providing it for those who are curious about such things. So you are most welcome to pick and choose what you bother with. But, in particular, for your benefit if you are not really mathematically inclined, questions which need answers here in terms of anything at all substantial in terms of Mathematics have been postponed to the final (Technical) section—therefore that section is even more optional than the rest of these Optional Extras!

There is still more good news to come! The almost universally accepted way of computing control limits for control charts based on a-few-at-a-time data, i.e. data which come to us in subgroups, doesn't involve our computing *either* the standard deviation *or* the MAD! Instead it uses something much quicker and easier than either of them, namely the *range* of the subgroup. The range of a subgroup is simply defined as the highest value in the subgroup minus the lowest value in the subgroup. One typically computes control

limits using data from somewhere between, say, 8 and 15 subgroups. So, even if you have a scientific calculator which has the standard deviation programmed in, you'd still be doing lots more button-pressing in order to evaluate it than you'd need in order to compute the ranges of those subgroups. Indeed, depending on the precision of the data being recorded, you might well be able to write down the ranges without using a calculator at all. But you will still need a calculator to produce the control limits themselves.

However, as mentioned earlier, the theory underlying where we place the control limits relates to standard deviations rather than to any other type of measure of variation. Now, since the standard deviation is *some* kind of average or typical gap between the values in the data and their mean, it is obvious that the range (largest value minus smallest value) will be *greater* than the standard deviation. We shall therefore need to divide it by some *conversion factor* to reduce it to a number which is on the same scale as the standard deviation—so that, in fact, it could then be regarded as an *estimate* of the standard deviation. The conversion factor involved is h, shown in the following table for subgroup sizes $n = 2$ to 6.

n	2	3	4	5	6
h	1.128	1.693	2.059	2.326	2.534

The technical details as to how this table is derived will be left (as you would expect!) to the Technical Section (page 83). With our illustrative subgroup of size 5, the largest and smallest values are respectively 2.1 and 1.2, so that the range, which we'll denote by R, is $2.1 - 1.2 = 0.9$. With $n = 5$, the converted value of R (converted in order to make its value comparable with the standard deviation) is thus $0.9 \div 2.326 = 0.387$. The actual standard deviation of this subgroup was found on page 19 to be 0.397. Of course, we could not expect the converted R to be *exactly* equal to the standard deviation since it uses less detailed information about what's in the subgroup. In fact, what the conversion method does (under the conditions assumed in the theory of the method) is to produce values that are equal to the standard deviation "on the average".

> You may recall that the value of the MAD for this subgroup was noticeably *less* than what the standard deviation turned out to be: 0.328 compared with 0.397. Interestingly, similar theory produces a conversion method which requires the MAD to be *multiplied by* 1.253 in order to obtain a value comparable with the standard deviation. This gives $0.328 \times 1.253 = 0.411$—which is again (of course) not *equal* to the standard deviation but is nevertheless considerably closer to it.

3. Control charts for subgrouped data

Most of the time in the previous section was spent on considering how to measure *variation* in subgrouped data. In case you skipped that section, I'll summarise it in just a single sentence as follows. Although the underlying ideas about measuring variation are based on the standard deviation—the statistician's favourite measure, as I have previously described it—practical work normally uses something much simpler: R, the subgroup's *range*, i.e. the distance between the subgroup's smallest and largest values.

In this section we'll introduce the type of control charts that are almost universally employed for studying subgrouped data. There are other possibilities but, from the practical point of view, I do not think they are generally worth bothering with. However, particularly because of using ranges, it will be necessary to consult tables of so-called "control-chart constants" in order to compute the control limits. Such control limits are in accord with Shewhart's "3σ-limits" as referenced by Dr Deming in his quotation reproduced on Appendix page 4. (That is followed by some discussion which, in particular, indicates why Mathematical Statisticians tend not to like the method!) As in the previous section, technical details about these control-chart constants are contained in the Technical Section at the end of these Optional Extras. Here we shall simply concentrate on how to use them.

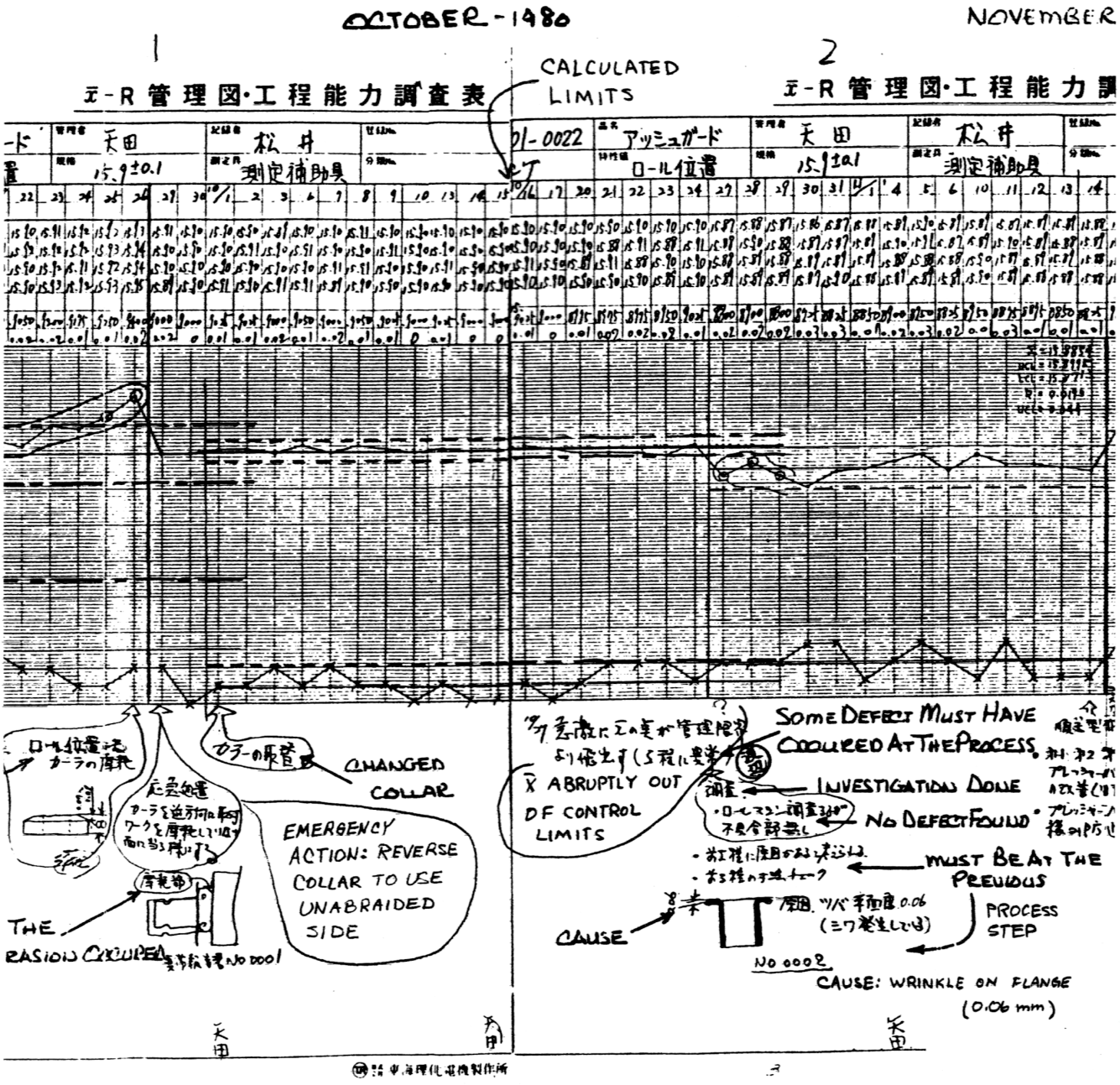
With subgrouped data it is usual to construct not just one but *two* control charts, one for the subgroup *means* (the term introduced in the previous section for what we had previously simply referred to as the "average"): the $\bar{X}$- chart, and one for the subgroup *ranges*: the R - chart. Thus we have one chart focused on whether or not the process *average* is stable and one chart focused on whether the amount of process *variation* is stable. This second chart, specifically studying whether the *variation* in the subgrouped data over time is or is not stable, has now become possible because of having "a few" data available at each time-point rather than just one. Since it is usual practice to deal with these charts as a pair rather than separately, they are sometimes referred to in the singular: the $\bar{X}$- R chart.

Before covering the details, it would be useful to take a look at part of the rather famous hand-drawn $\bar{X}$- R chart of which we have already seen something as the fifth of the "Six Processes" briefly mentioned on Day 3 page 21 and then described in some detail on Day 3 pages 32–33. Since those pages were "effectively 'extra-curricular'", I'll repeat some of their content here. On the next page there is a short portion of the chart which in 1982 some Ford Motor Company personnel brought back from the Tokai Rika Company in Japan (for, at that time, anything of this nature was totally new to them). Something else that was wholly new to them was that such charts were being constructed, used and interpreted not by statistical "experts" but by the personnel on the factory floor. The writing is not very clear, but you will be able to see the daily data in subgroups of size 4 written above the graph-paper and the very active notes and comments below the graph-paper with some translations written in. The $\bar{X}$- chart is drawn in the top half of the graph-paper and the R - chart is at the bottom of the graph-paper. You will also easily see, beginning on 27 October, clear signals below the Lower Control Limit on the $\bar{X}$- chart indicating that the process had suddenly gone out of control. Also note the efforts made to find the reason and also that the control limits (on both parts of the chart) were accordingly recomputed. It is also worth observing that, although it was the $\bar{X}$- chart which had the signals, the opportunity was also taken to update the R - chart since that also soon started getting points above its previous Upper Control Limit. Although ancient, this *Japanese Control Chart* is extremely interesting to study, and an excellent presentation about it is provided in Chapter 7 of *Understanding Statistical Process Control* (Third Edition) by Don Wheeler and the late David Chambers.

So let's see how to construct an $\bar{X}$- R chart. As with the control charts for single values in the main text, we will need to use data collected over a few time-points (the *baseline*). There's no "rule" governing how many time-points to use but I would suggest that, with n as large as 4 or 5, 10 would generally be ample. The Central Line of the $\bar{X}$- chart will naturally be the mean of the subgroup means, for which the traditional very logical notation is $\bar{\bar{x}}$. But *how far* above and below the Central Line should the control limits be? The answer is a multiple of the mean range $\bar{R}$, that multiple being H which is provided in the following table:

n	2	3	4	5	6
H	1.880	1.023	0.729	0.577	0.483

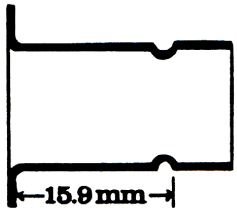
And how about the second of the pair of control charts, the R-chart, on which we plot subgroup ranges? Obviously enough, the Central Line here will be the mean range $\bar{R}$. And, very conveniently, the control limits are also multiples of $\bar{R}$ — except that, strictly following Shewhart's guidance on 3σ-limits, there is no Lower Control Limit using small subgroup sizes! Why might this be sensible? Simply because, for small values of n such as in the brief table of control-chart constants that I have shown above, his guidance leads to *negative* values for the LCL: and, clearly, no range can be negative. Interpreting this in practice, this implies that there are some circumstances in which it is quite possible for a zero subgroup range to occur: i.e. all the values in the subgroup happen to be equal to each other — whether or not the process is in statistical control. Clearly, this depends on the amount of precision with which the data-values are being recorded. But the fact that such circumstances do exist in practice surely implies that it would be rather inconvenient if a zero range always *had* to be below the Lower Control Limit; so the fact that Shewhart's guidance doesn't provide one is not so peculiar after all — indeed, it's rather fortunate!



The tables and notation being included here are extracts from the table of control-chart constants on page 46 of *EST*, my book of *Elementary Statistics Tables*. Those tables cover a wider variety of situations than the most common ones that I am including here, including subgroup sizes larger than $n = 6$. And then Lower Control Limits can and do exist. Using the notation in that *EST* table, the constants to be used for the Lower and Upper Control Limits on the R-chart are denoted h_1 and h_2. So all that concerns us here is that the Upper Control Limit on the R-chart is computed as $h_2 \bar{R}$ where h_2 is provided in the following table:

n	2	3	4	5	6
h_2	3.267	2.575	2.282	2.114	2.004

To put a little flesh on the bones, it would be a good idea for you to compute these various control limits and see them in action. The easiest way to do that is to use some data from that piece of the *Japanese Control Chart* reproduced on the previous page. It would therefore be useful for you to know a little more about some of the background to this chart and also to this particular part of it. So below I have reproduced some short extracts from Don Wheeler's write-up. There is plenty to learn from these extracts, not only about the use of the control chart but also about the management and working environment—they are *not* unrelated! As you can see, the part of the chart being illustrated begins on Monday 22 September 1980 and runs through to Friday 14 November, and that is the period covered by the description below. I won't repeat here any of the details already mentioned on page 21, so you might like to remind yourself of what I wrote there before moving on to the following extracts.

> "As the Ford group was touring the Tokai Rika plant, they observed eight production workers 'engaged in active discussion' around this Average and Range Chart. To the people from Ford, it seemed that something must be wrong with the process represented by the chart, so they asked about it. They expected there to be an internal production problem, or an assembly plant problem, or a problem of too many rejects. However, they were told that this was simply a routine review of an ongoing process and, in fact, the process was currently being operated predictably and was well within the specifications. *[In case you know about such things, the Process Capability Index was measured as about 2.25, indicating that virtually everything being produced was within the middle half of the specification range and that the process's performance was even superior to so-called six-sigma quality. Not bad for 1980!]*
>
> The process represented by this chart is the fabrication of a cigar lighter shell. The dimension tracked by the chart is the distance between the flange and the detent, as shown. The target value for this dimension is 15.90 mm, and the specified tolerance is ±0.10 mm. The measurements shown on the chart were made with a snap gauge and were recorded to the nearest 0.01 mm. Based on production data given later, about 17,000 pieces were being produced each day.
>
> 
>
>
> Points outside the limits are noted on September 25 and 26, 1980. Having noted that exceptional variation was present, they looked for the Assignable *[Special]* Cause. The notes at the bottom of the chart document these efforts. 'Abrasion on the positioning collar' is identified as the Assignable Cause for the process excursion noted in late September, 1980. In addition to writing down the Assignable Cause on the chart they took action—the very next day the process average shifted back to the target of 15.90 mm. Again, a note on the chart tells what was done.
>
> As a temporary solution, a worker turned the worn collar over to use the back side. Two days later a new collar was installed. This incident displays a desire on the part of the Tokai Rika personnel to operate at the target. The process was in no danger of producing nonconforming product, yet they took the trouble to fix it so that it would stay centred on the target value of 15.90 mm. Moreover, just as the shift on September 29 shows the desire of the workers to operate at the target value, the replacement of the collar on October 1 shows the support of the management for this policy.
>
> Why do the operators and their supervisors want to operate right at the target value when they have such *[relatively]* wide specifications? Isn't this excessive? Would it not be cheaper to let the process run until the process average was above 15.95 mm? While it might be cheaper for this one operation, it would eventually prove to be more expensive for the company. The definition of World Class Quality is 'On Target with Minimum Variance'. This example shows how this concept is put into practice. *[Don then continues with some discussion involving the Taguchi Loss Function, studied on Day 7 of our course. The following two paragraphs now cover October 1980, and I'll then ask you to compute the control limits used on the chart during October.]*
>
> Following the installation of the new collar on October 1, data were collected for recalculating the limits. The process stayed within these new limits until October 27. At that time the product

dimension suddenly shifted downward. The fact that it was a sudden change in the process was a clue to the nature of the problem, and as such was noted at the bottom of the chart.

The search for the Assignable Cause led back to the preceding step, a blanking operation. When a problem was found, it was checked to see if it corresponded to the indications given by the process behaviour chart *[control chart]*. Since this problem involved the repair of a die, the fix was postponed until the weekend of November 15 and 16."

The rest of this page and all of page 25 are also on Workbook pages 252–253.

Now, returning to the chart on page 22, you will see the note at the top pointing out that the new control limits were computed after the 15 October subgroup had been recorded. To save you from having to try to decipher the writing above the chart, here are the data that were used for the computation:

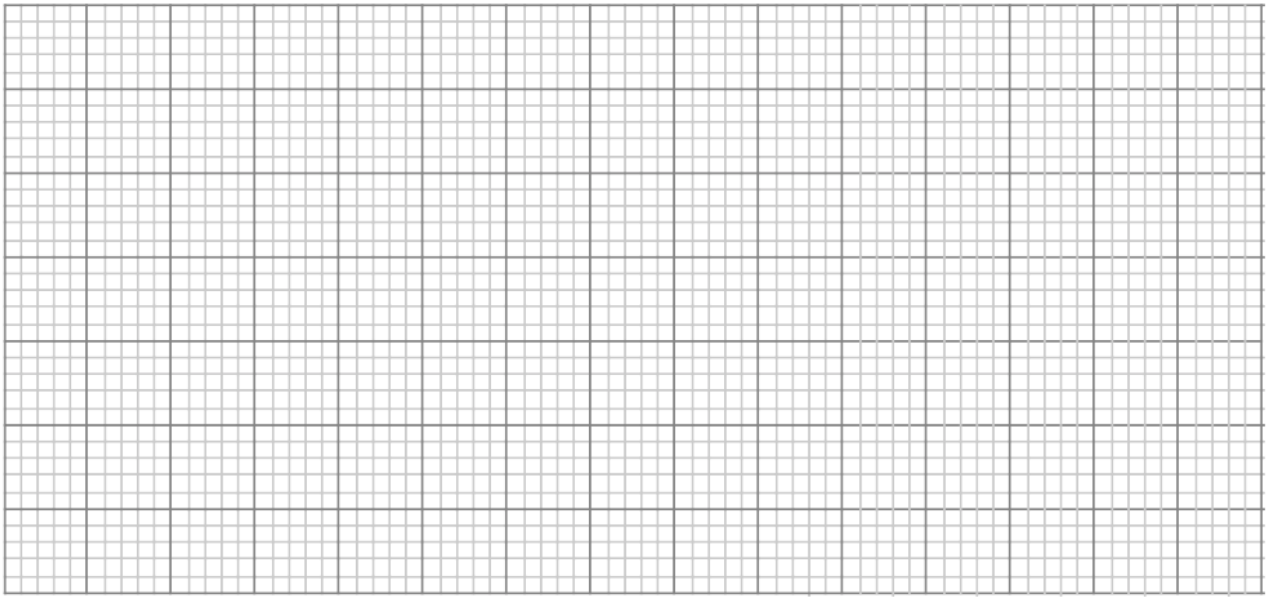
	$^{10}/_{1}$	2	3	6	7	8	9	10	13	14	15
	15.90	15.90	15.89	15.90	15.90	15.91	15.90	15.90	15.90	15.90	15.90
	15.90	15.91	15.90	15.91	15.90	15.90	15.91	15.90	15.90	15.90	15.90
	15.90	15.90	15.90	15.90	15.91	15.91	15.90	15.90	15.91	15.90	15.90
	15.91	15.90	15.91	15.91	15.89	15.90	15.90	15.90	15.90	15.90	15.90
$\bar{X}$											
R											

So, go ahead and write down the values of $\bar{X}$ and R for each subgroup—not very difficult with these data! And yes, you will soon see a few of those zero ranges that I mentioned on page 21. Next compute the values of $\bar{\bar{X}}$ and $\bar{R}$. Finally, compute the control limits for both parts of the chart as described earlier and check that they agree (approximately) with what the Japanese workers had drawn as reproduced here on page 22. My answers are on page 27 if you need them—but first try it yourself here:

As you have seen on page 22, all remained well until the out-of-control signals which began on 27 October. You have read at the top of this page what then happened. After the repair had been carried out during the weekend of 15–16 November and some subsequent data had been collected, new control limits were drawn on the chart from 17 November onward. Meanwhile, someone had checked through the data from 27 October to 14 November and had computed control limits for that period using that whole set of data. If you would like to try out the computations one more time, here are the data recorded over that period (I don't know the reasons for the apparent six-day week followed by a three-day week):

	27	28	29	30	31	11/1	4	5	6	10	11	12	13	14
	15.87	15.88	15.87	15.86	15.87	15.88	15.89	15.90	15.89	15.89	15.87	15.89	15.89	15.88
	15.88	15.90	15.88	15.87	15.87	15.89	15.90	15.91	15.87	15.89	15.90	15.89	15.88	15.89
	15.88	15.89	15.88	15.89	15.89	15.89	15.88	15.88	15.88	15.90	15.89	15.89	15.89	15.88
	15.89	15.89	15.89	15.87	15.90	15.88	15.89	15.89	15.89	15.90	15.89	15.88	15.88	15.88
$\bar{X}$														
R														

If you'd like to examine this period by drawing the $\bar{X}$-R chart on the graph-paper below, you will see that the process remained in control during this time, albeit with the lower mean and with increased variation. However, you will also see that, despite the problem that had been discovered, it is highly unlikely that any piece manufactured even during this period was anywhere near going outside the specifications of 15.90 mm ± 0.10 mm. This is, of course, an illustration of the value of having improved the process well beyond the minimum that had appeared to be necessary: despite the current perturbation, everything produced *still* remained "fit for purpose" (as conformance to specifications is often described).



4. Discussion

One of the expressions that is in vogue these days is "doing more with less". What better illustration could there be of that aim than this Japanese Control Chart? Firstly, only a tiny amount of data was used: just four readings out of around 17,000 units per day. Secondly, although maybe the accuracy of the readings (to the nearest one-hundredth of a millimetre) might have been quite impressive over 35 years ago, nevertheless it may appear to be rather crude considering the variability of the measurements involved: they were usually only varying between, say, 15.87 mm and 15.93 mm, and often over an even narrower interval than that. But just look at what putting those "rather crude" data on control charts still enabled the production workers to learn and do!

There is one aspect of the Japanese Control Chart that I must advise you *not* to copy unless you already have really considerable knowledge and understanding of your process. That was the way in which the subgroups were formed. To quote from Wheeler and Chambers page 155, "The four pieces for the daily subgroup were drawn ... at 10.00 am, 11.00 am, 2.00 pm, and 4.00 pm". That is hardly consistent with my introductory remarks about subgroups on page 17: " ... the data concerned *are recorded pretty much at the same time and under the same conditions*"! Immediately following that quotation about when the readings were taken, there is some discussion on this very point. Translating some of the Japanese writing under the chart for August 1980, the Japanese already had evidence that this schedule for recording the data was such that "the present measurement method can detect process change". The discussion then continued with these wise words:

> "As long as a process behaviour chart *[control chart]* is capable of detecting exceptional variation, it is sensitive enough to use ... . There is no need to increase sensitivity by increasing subgroup size. Furthermore, one subgroup per day had proven to be adequate since this process was one that usually would change slowly, over a period of days. For these reasons, the subgroup size and the subgroup frequency were not changed."

But let me repeat my above warning: "I must advise you *not* to copy *[such a method of data-collection]* unless you already have really considerable knowledge and understanding of your process." I learned that lesson from an incident which occurred some time before I had even met Dr Deming. Since this happened well over 30 years ago, I cannot recall all the details, but I can recall what was most important. I was visiting a plant that was manufacturing long rolls of some rubbery material which was being used in a paper-manufacturing process. The people there had become puzzled because both their $\bar{X}$-chart and R-chart were showing extraordinary "hugging the Central Line" effects—much more extreme than you saw when studying the control charts for Rule 2 of the Funnel on page 10 of these Optional Extras. The reason became clear when I asked them how they were forming their subgroups. They told me that they were measuring the thickness of the material at both edges, and in the middle of the roll, and then halfway in-between those measurements, producing subgroups of size $n = 5$. Unfortunately, when I examined the data, it turned out that there was a relatively substantial difference between the thickness of the roll at its edges compared with most of the material away from the edges! That difference far exceeded the natural variation at any one of the five positions, and so resulted in the range of any such five measurements being vastly greater than what would have been obtained from the natural variation alone. You can soon visualise the effect that *that* had on the control limits on both parts of their $\bar{X}$-R chart. Those good people would have been far better off by just taking single measurements *either* at an edge *or* in the middle and using the ordinary one-at-a-time chart!

My purpose in introducing you to the Japanese Control Chart has been twofold. Firstly, of course, it contains very suitable data for you to practise with, both on the computations and with the interpretations. But secondly, I hope the short extract you have seen here will have whetted your appetite for seeing more of it!

I have already mentioned the superb coverage of the Japanese Control Chart in the book by Don Wheeler and David Chambers cited on page 21. The whole 20-month chart is contained (in sections) within that chapter along with excellent discussion. Further, a slightly updated version of Don's original video of *A Japanese Control Chart* is also available from SPC Press (www.spcpress.com), either by download or on DVD. I particularly recommend the book if you are interested in learning much more on control charts than I have included either in these Optional Extras or in the main course. It is quite expensive, but the whole book contains extremely interesting and useful material. If you have a local library, see if they can find it for you!

Computed control limits

With the data on page 24 *[WB 250]* the control limits for the $\bar{X}$-chart are at 15.895 mm and 15.908 mm, while the (upper) limit on the R-chart is at 0.0207 mm.

With the data on page 25 *[WB 251]* the control limits for the $\bar{X}$-chart are at 15.871 mm and 15.899 mm, while the (upper) limit on the R-chart is at 0.0440 mm.

PART C: NE'ER THE TWAIN SHALL MEET?

1. Introduction

Well, the twain *may* meet—but there's often a serious problem when they do: neither can understand what the other is talking about! The "twain" to whom I am referring are students from what we might call the Deming/Shewhart school of Statistics on the one hand and from the "conventional" or "traditional" or "mathematical" school of Statistics on the other. As I said as early as Day 1 page 6, and repeated here in the Introduction on page 2:

> "Over the nearly 20 years of my seminars on Dr Deming's teaching I rarely suffered from any 'difficult' delegates. The few that I had could be divided into two types. One type were very senior managers; the other type were those with some qualification in Statistics."

On page 2 I then pointed out that

> "... the latter were often the more difficult type. That may sound rather flippant, but it isn't. It can be very serious. If it happens to you then I want to help you to deal with it. For *you* may not have any qualification in Statistics. So are the people in your organisation likely to believe you or the one who *is* qualified in the subject?".

For example, the "conventional" student has been taught that both the theory and "validity" of control charts depend upon traditional Mathematical Statistical fodder such as probability theory, the normal distribution, the Central Limit Theorem, and hypothesis testing. The Deming/Shewhart student may never even have *heard* of such things because, truth to tell, neither the "validity" nor even the basic ideas behind control-chart methodology depend on any of them—at least, that is, according to both Dr Walter Shewhart (who, as you know, was the subject's creator) and his famous protégé, Dr W Edwards Deming (and, by now, you know quite a lot about him as well!). I'd say they were both fairly safe sources of wisdom. However, if you *are* interested in what those things are about, you will find some explanation and discussion on them in Part D of these Optional Extras—and that will be useful if you ever become one of the twain that *do* meet! Why? Well, let's see.

You, the Deming/Shewhart student, might well like to convince the conventional Statistics student that those supposed mathematical underpinnings are irrelevant. But how can you, if you know not what they are, let alone why the conventional student deems them to be so essential? Yet you almost certainly *need* to be able to communicate such arguments, else your organisation will continue to be held back by misconceptions taught to them by the statistical "expert", misconceptions that are very likely to result in over-restrictive use of control charts and often *fear* of their use. As Dr Deming pointedly expressed it (*Out of the Crisis* page 286[335]), the mathematical concepts mentioned above "are misleading and derail effective study and use of control charts". I most certainly could not have expressed it better myself.

So some of the material that follows attempts to enable you (a) to understand *how* the conventional statistician thinks, and *why* he thinks that way, and (b) to thus be able to communicate with him, and then (c) to have at least a sporting chance of helping him see the error of his ways.

One big problem that exists between the two schools is that there are likely to be incompatibilities of perceived *purpose*, and therefore of use and interpretation, of control charts. The Deming/Shewhart student recognises control charts as a valuable guide to appropriate action for *improvement*. The conventional student usually regards them merely as a monitoring device to provide an early warning of something going wrong, so as to trigger timely corrective action. But that is not *improvement*—it is, at best, *maintenance of the status quo*.

That is clearly a matter touching upon the very management philosophy and approach of the organisation in which the control charts are being used—thus it is strongly related to the main material of this course. The concentration here in this optional material is largely restricted to merely technical issues. But students from the two schools often find that, even on technical matters, they *still* cannot understand each other.

The clear rejection by both Shewhart and Deming of the relevance of the conventional statistician's "life-blood" of probability calculations and normal distributions in this context have already been evidenced by quotations from both of them that you have seen in the main text. But they surely bear repeating here.

Firstly, there was the extract from *Out of the Crisis* page 286[pages 334–335] that I have just mentioned, and I make no apology for repeating part of the final sentence:

> "It would ... be wrong to attach any particular figure to the probability that a statistical signal for detection of a special cause could be wrong, or that the chart could fail to send a signal when a special cause exists. The reason is that no process, except in artificial demonstrations by use of random numbers, is steady, unwavering.
>
> It is true that some books on the statistical control of quality and many training manuals for teaching control charts show a graph of the normal curve and proportions of area thereunder. Such tables and charts are misleading and derail effective study and use of control charts."

Then there was this extract from page 12 of Shewhart's 1939 book:

> "Some of the earliest attempts to characterise a state of statistical control were inspired by the belief that there existed a special form of frequency function f and it was early argued that the normal law characterised such a state. When the normal law was found to be inadequate, then generalised functional forms were tried. Today, however, all hopes of finding a unique functional form f are blasted."

And finally there was this passionate language from Deming when he was speaking to some senior executives in France in 1989 (recorded in *Profound Knowledge*, BDA Booklet A6 page 4 and recently already mentioned here on page 20 of these Optional Extras):

> "How can we aim for minimum economic loss? It is *nothing* to do with probabilities of the two kinds of mistakes. No, no, no, no: not at all. What we need is an operational definition of when to look for a special cause, and when not to. That is, a rule which guides us when to search in order to identify and remove a special cause, and when not to. It is not a matter of probability. It is nothing at all to do with how many errors we make on average in 500 trials or 1,000 trials. No, no, no—it can't be done that way. We need an operational definition of when to act, and which way to act. Shewhart provided us with a communicable operational definition: the control chart using 3σ-limits. Shewhart contrived and published the rules in 1924—65 years ago. Nobody has done a better job since."

2. The essence of the argument

Two types of "statistical studies"

Assuming you do not have much or any background in conventional Statistics, this section will (like the first section) contain some words, terms and phrases with which you are not familiar. But please read it nonetheless! My purpose here is to provide a broad introductory description of all of the rest of this optional extra material so that you can get an advance sense of the shape of things to come. This includes (on page 33) a typical syllabus for an introductory course on conventional Statistics which obviously will indeed include some terms with which you are unfamiliar. But, after reading these current two pages, you will have a good idea of *why* I am introducing them to you. They are mostly "part and parcel" of what the conventional statistician *is* familiar with, and so then you will have a chance of discussing things with him. That will be even more the case if you undertake the "crash-course in conventional Statistics!" that I offer you in Part D.

But even the conventional statistician will probably not be familiar with some or all of what I shall now introduce—so you're on a level playing-field for the time being! At least, that was my experience with what follows. Maybe a couple of years before I first met Dr Deming, I tried to read some of what he had written about the subject of Statistics. And almost immediately I was faced with terms such as "analytic studies", "enumerative studies" and "frames". All totally new to me.

Now, at that time, I had already been a Lecturer in Statistics in the University of Nottingham's Department of Mathematics for over 15 years. So I suppose I had already become somewhat set in my ways and in my understanding. For it seemed from what I was reading that Deming was claiming all I had so far learned and had therefore so far been teaching others, was "merely" concerned with "enumerative" studies—whereas what was really needed in order to be useful in the real world was "analytic" studies. Surely that couldn't be right, *could it?*

But a lot of other stuff that Deming had written did appeal to me, so I rather set aside that business about the two types of statistical studies and read about other things instead. Nevertheless, some time later when the British Deming Association began its work, one of the first things I did was to set up a study group to examine Deming's writings about Statistics to see if that group could shed some light on those puzzling matters. I was extremely fortunate to have the late Professor David Kerridge as leader of that study group, and light eventually began to dawn under his patient and wise guidance. David became a great supporter and helper and friend in the years that followed.

Again assuming that you do not have much background in conventional Statistics, you probably won't have the mental blocks that I had and will therefore be able to get the gist of what Deming was writing about much more quickly than I did.

What do dictionaries tell us about those puzzling words? Actually, I discovered that neither dictionaries on the internet nor in print seem very helpful with the adjective "enumerative". All I could find was either the noun "enumeration" or the verb "enumerate" with "enumerative *adj*" then appearing merely as an appendage without definition. The verb "enumerate" is typically described as "To count off or name one by one; list" and the noun "enumeration" as "The act of enumerating" or "A detailed list of items". Is that really *all* I had been teaching all those years?!

Maybe I'd have better luck with "analytic". A well-known dictionary on the internet produces: "Generally speaking, 'analytic' refers to 'having the ability to analyse' or 'division into elements or principles'." OK: maybe that makes a bit of sense. How about my favourite hard-copy dictionary—what did I find there? "Of, pertaining to, or based on analysis; showing an ability to analyse and reason from a perception of the

parts and interrelationships of a subject; skilled in analysis." Hmm, I found a glimmering there, but not a lot.

How about "enumerative *study*" or "analytic *study*"? No: I drew a blank on both of those.

So I'd better try some brief explanations in my own words! "Analytic studies" are indeed what we have been involved with, particularly in the early days of *12 Days to Deming*, primarily enabled by the use of control charts. Let me try a description in just a single sentence: *The purpose of analytic studies is to enhance knowledge and understanding of processes, for prediction into the future, and to provide guidance for improvement.* You will observe that we could just as well replace "analytic studies" by "control charts" in that sentence. Looking back to the early years of my career-life, I must indeed confess that that sentence does *not* describe what I was then teaching. I had actually come across a version of control charts in a textbook quite early on while teaching in America, and thought the topic interesting enough to devote, say, half of a 50-minute lecture to it in the introductory Statistics course that I gave back here in the UK, but I certainly cannot claim that that sentence describes what I taught during those few minutes. (No, I shall *not* embarrass myself by describing to you what I *did* cover in them!)

How about a brief explanation of "enumerative studies" in my own words? As we shall see later, an introductory course on conventional Statistics often begins by talking about drawing a sample (or preferably a "*random* sample") from a "population". A "population" is some collection of "things"—possibly people but often not. In the Red Beads Experiment, the population is a collection of 4,000 beads. A "sample" from the population is, of course, a selection of some of those "things" from the population—we are familiar with samples consisting of 50 beads obtained using the "paddle". If the term "*random* sample" is used, what does that imply? It actually implies something very specific: namely, that each and every possible sample (of the specified size) is *equally likely* to be drawn as any other. Using the sample, the output from an "enumerative study" is then an attempted description of what is in that population—or, rather, in that part of the population which is available for sampling, and that's what Deming meant by the word "frame". A census is a good example of an enumerative study using an extremely large (though *not* random) sample.

Note that simply attempting to describe what is in the population (or frame) does not include any intent to explain *why* the population (frame) contains what it does nor what anything related to it might become or deliver in the future. As Deming would say, it contains no "temporal spread". So, in an enumerative study, there is no reason to consider matters such as whether or not a state of statistical control exists—indeed, the times at which the data are taken are often not even noted—whereas, of course, that is top priority in analytic studies. As I understand it, *this* is the essential difference between those two types of statistical studies—and, as I believe you will appreciate, that's a *big* difference. There may well be more, but at least I hope this will give you a reasonable start if you ever do decide to delve into such matters.

If books and courses on Statistics were to make clear—maybe not using the same terms, but at least indicating the purpose and limitations—that what they include is designed *only* for enumerative rather than for analytic studies, presumably they would not do much harm. However, if one looks at the examples illustrated in, say, chapters on histograms in introductory Statistics texts, even in the more "practical" ones, it is often the case that they are actually involved not with "populations" but with *processes*, i.e. with their data being generated over time and with a strong likelihood that time-dependence is important. Thus, not only are the purpose and limitations of enumerative studies not *made clear*—it would appear that they are not even *recognised* by many authors and teachers.

A typical syllabus for an introductory Statistics course

To help you understand something of the conventional statistician's mindset, Part D of these Optional Extras will introduce you to some content of a typical introductory Statistics course. But, to prepare you

for that pleasure, here is a possible syllabus for such a course. Again it will, of course, contain some terms that are unfamiliar to you—but then that is likely to be true of a syllabus for a course on *any* topic with which you are not already familiar.

- Summarising "raw data" through pictures such as histograms and the calculation of "sample statistics" such as the sample mean $\bar{X}$ and the sample standard deviation s and/or the sample variance s^2 *[a "sample statistic" is anything which can be computed from the data in the sample].*
- Probability, particularly as the long-term proportion of occurrences of an event and using symmetry considerations (as with coins, dice and playing cards).
- The natural link between the above two topics, i.e. that if one takes an ever-larger random sample from a population then the corresponding histogram (appropriately scaled) gets ever-closer to a similar pictorial representation of the probabilities of the possible outcomes—thus leading to the ideas of the "true mean" μ and the "true standard deviation" σ of a probability distribution being respectively the long-term values of $\bar{X}$ and s as the sample size $n \longrightarrow \infty$ ("tends to infinity").
- Discrete and continuous probability distributions: particularly the binomial and normal distribution respectively.
- Properties of the normal distribution, including the Central Limit Theorem.
- Statistical inference: in particular, confidence intervals and hypothesis tests (tests of significance), and how assumptions of normality and/or the use of the Central Limit Theorem enable these to be placed on an appealing mathematical footing.

There: that's going to be fun, isn't it?

The conventional statistician's view of control charts, ...

Here is an abbreviated version of Deming's second quotation on page 30:

> "How can we aim for minimum economic loss? It is *nothing* to do with probabilities of the two kinds of mistakes. No, no, no, no: not at all. ... It is not a matter of probability. It is nothing at all to do with how many errors we make on average in 500 trials or 1,000 trials. No, no, no—it can't be done that way."

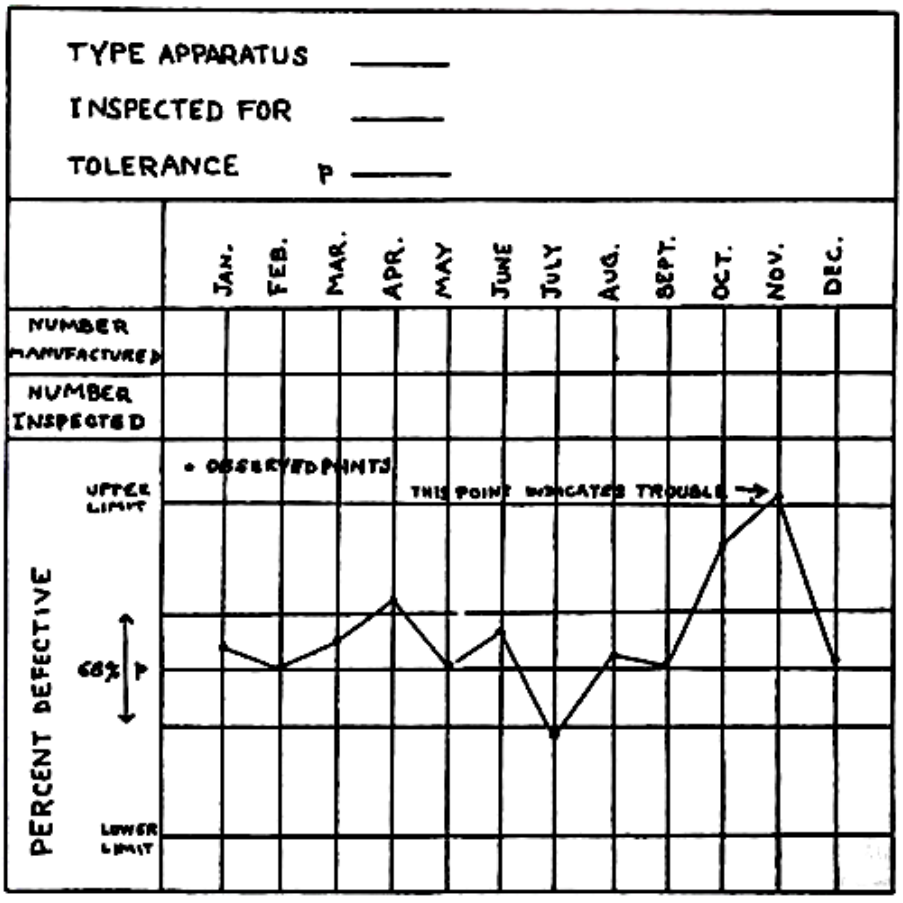
As you will see later, Deming was alluding here to the conventional statistician attempting to treat control charts as if they were hypothesis tests. A hypothesis test results in the rejection or acceptance of a so-called "null hypothesis" H_0. This decision is made according respectively to whether an appropriate sample statistic, the "test statistic", lies inside or outside some region of values defined by one or two "critical values": this region is therefore called the "critical region". Moreover, the "significance level" of the test is defined to be the probability of the test statistic *wrongly* rejecting H_0, i.e. rejecting H_0 (because its value falls inside the critical region) when H_0 is in fact true. So that's the direct connection with what Deming was talking about above. Assumptions about the data being normally distributed, or equivalently the Central Limit Theorem, allow particularly easy and convenient derivation of the critical value(s) for any desired significance level.

If the conventional statistician, trained in this way, subsequently comes across a control chart—by far the most important tool for an analytic study—it is rather easy to see why he is immediately inclined to regard it as a kind of glorified hypothesis test with H_0 representing "in statistical control", and why he thinks that normality has such an important part to play. But, as we have seen, the traditional Statistics course such as that whose syllabus is summarised above is based upon foundations only pertaining to *enumerative* studies

—in particular, ignoring any questions about behaviour possibly changing over time. (There are a few areas in conventional Statistics that study behaviour changing over time in some *specified well-defined* manner. But that's a very different matter, and actually probabilities and the normal distribution again often dominate the theory in such areas.) The conventional statistician's interpretation of a control chart as a glorified hypothesis test is thus wholly without foundation in practice. As a matter of fact, even in enumerative studies, aspects of hypothesis testing stand on rather thin ice because, in most practical applications, H_0 is never, or almost never, true! So I'm rather glad that "in statistical control" is *not* an appropriate H_0!

The conventional statistician will hardly give a friendly hearing to the suggestion that the foundations upon which he has built his beliefs, career and reputation might be inappropriate in the "real world" as opposed to in the Mathematics classroom. So, is there anything that can be done? Yes, *fortunately and remarkably*, there is.

But, before that, it is worth making the intriguing and salutary point that Shewhart himself started out thinking that the subject *could* be developed from the conventional Statistics viewpoint. What is often referred to as "the very first control chart" (pictured below, dating from 1924) shows some guidelines placed at just one standard deviation either side of the Central Line, with the indication "68% *p*" written against them. That probability, expressed as a percentage, is derived from the normal distribution, as you will be able to confirm when you reach page 49.



Yet, despite this, Shewhart was open-minded enough to eventually see the error of this mode of thinking. His statement on page 30 (sandwiched between the two quotations from Dr Deming) was clearly autobiographical!

... and what can be done about it

Substantially, the conventional statistician's case for requiring normality to make control charts and their control limits "valid" exists on two main fronts. Such a statistician believes that normality is needed:

- because control-chart constants that are used in computation of control limits, such as the 2.66 with which we became familiar on Day 3, are derived from normal distribution theory (as indeed they are); and
- so that a *probability interpretation* can be given to control limits: specifically, the claim is often made that, under normality, there is a probability of 0.0027 (i.e. 0.27%) that any particular data-point falls outside Shewhart's 3σ-limits (note that 0.27% = 2 × 0.135% when you look at page 49) if the process is in statistical control.

Now again, any acceptance of Shewhart's and Deming's teachings immediately leads to the denial of the conventional statistician's claims in both of these respects. The "*fortunate and remarkable*" facts alluded to on the previous page are however that, even if we *ignore* what Shewhart and Deming said about both normality and probability interpretations (recall, in particular, page 30), the above claims are *still* demonstrably wrong! In other words, we are able to wade right into the conventional statistician's camp, onto ground which both Shewhart and Deming believed to be without foundation yet which the conventional statistician needs to have faith in (for all that he has learned is built upon it), and to talk to him in language which he both understands and accepts (even if we don't), and to *still* produce evidence which disproves his beliefs!!

That evidence is demonstrated in Part E on pages 65–70.

PART D: A CRASH-COURSE IN CONVENTIONAL STATISTICS!

Histograms and sample statistics

We're starting off on quite familiar ground, particularly if you have read Part B of these Optional Extras, and so this first section is very short.

From the practical viewpoint, one could simply express the prime purpose of Statistics as being "data analysis". So let's suppose we have available for analysis a random sample of data (in the sense described on page 17) taken from some "population" of values. Alternatively, our data might consist of the results from repeated trials of some experiment, operation or procedure, etc. What can the data tell us about that population or other source from which they've been drawn? To get some idea about that, the first thing you might well do is to construct a histogram of those data, just as you have seen and done on Day 3. You will see many more histograms in this crash-course.

Secondly, the term "sample statistics" simply refers to any quantities that can be computed from the data. We have already discussed some sample statistics in the "Calculations on a subgroup" section in Part B (beginning on page 17) although I didn't use that term there. So again I need say very little here. However, at the time I did give you the option of skipping that section. If you accepted that option then I'm afraid I must now ask you to go back to it, for quite a lot that I would otherwise have had to include here is all there. You won't have to read the whole section, but you will need to read the first two pages: you can stop once you've read the paragraph in the middle of page 19 which introduces the "variance".

Probability

The conventional statistician regards Probability as the branch of Mathematics on which the subject of Statistics is based.

The idea of the *probability* of an *event* occurring in some particular situation is often described in terms of the *long-term proportion of occurrences* of that event, implying (conceptually at least) that it is possible to repeat the situation being envisaged under the same conditions *ad infinitum*.

> That is, of course, a far more exacting notion of stability than is considered in the control-charting context: and so, when necessary to avoid ambiguity, I shall refer to the situation now being described as that of "*exact* stability".

Many examples of probability calculations, both at the introductory stage and later, make use of considerations of *symmetry*, which is where all the various possible outcomes are regarded as being *equally likely* to occur. Common illustrations in the introductory texts are the two sides of a coin, the six faces of a die, and the 52 playing cards in a "well-shuffled" deck.

Linking it all together

Now comes a master-stroke in the introductory conventional Statistics course: to combine the ideas presented above into a unified theory so as to create a basis of probability for results obtained from sampling.

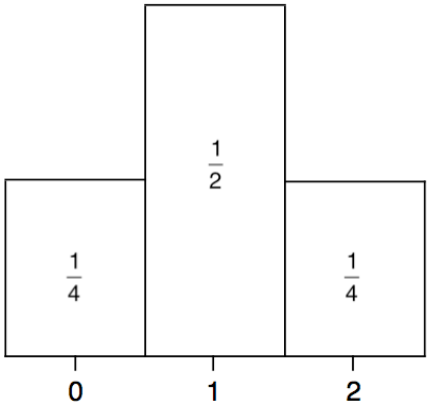
Let's demonstrate by means of a very simple example: the number of Heads obtained when two coins are tossed. Assuming that the coins obey the mathematical ideal of an exactly 50–50 chance of Head or Tail every time they are tossed, it follows that, when both are tossed:

Probability of no Heads	=	$\frac{1}{4}$
Probability of 1 Head and 1 Tail	=	$\frac{1}{2}$
Probability of 2 Heads	=	$\frac{1}{4}$

A typical way of verifying this (if you need one) is as follows. Suppose we label the coins as A and B. Then there are *four* possible outcomes, all equally likely. They are:

Coin A:	Head	Head	Tail	Tail
Coin B:	Head	Tail	Head	Tail

In other words, each of these four possibilities occurs a quarter of the time in the long run, i.e. each one of the four possibilities has probability $\frac{1}{4}$. This easily translates into the above probabilities since the event "1 Head and 1 Tail" corresponds to *two* of the four equally-likely possibilities. We can draw a picture of these three probabilities where, similarly to a histogram, heights or areas are proportional to the probabilities:



Note that these three probabilities add up to 1, as of course is bound to be true in any situation when adding up the probabilities of all possible mutually exclusive outcomes. So, in terms of such a picture of probabilities (whether the situation being described is simple like this or far more complex), the *total area* contained in any such picture must be 1.

Now suppose we toss the two coins several times, keeping track of how often we get no Heads, one Head, and two Heads. Here are some typical results. After tossing the coins ten times, we might have:

0 Heads:	3 times
1 Head:	6 times
2 Heads:	once

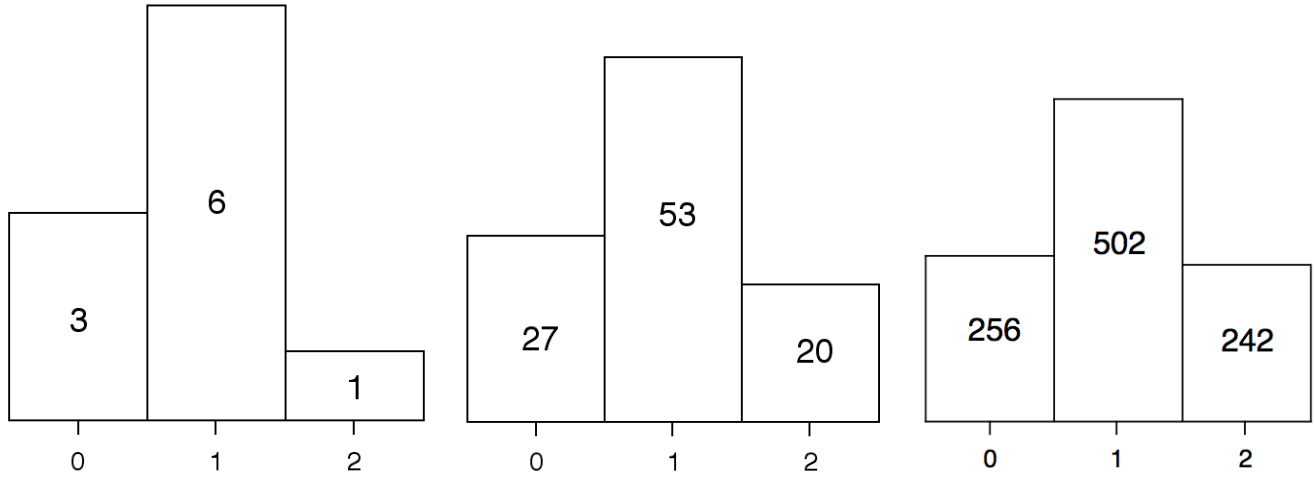
Then, after 100 tosses, we could have:

0 Heads:	27 times
1 Head:	53 times
2 Heads:	20 times

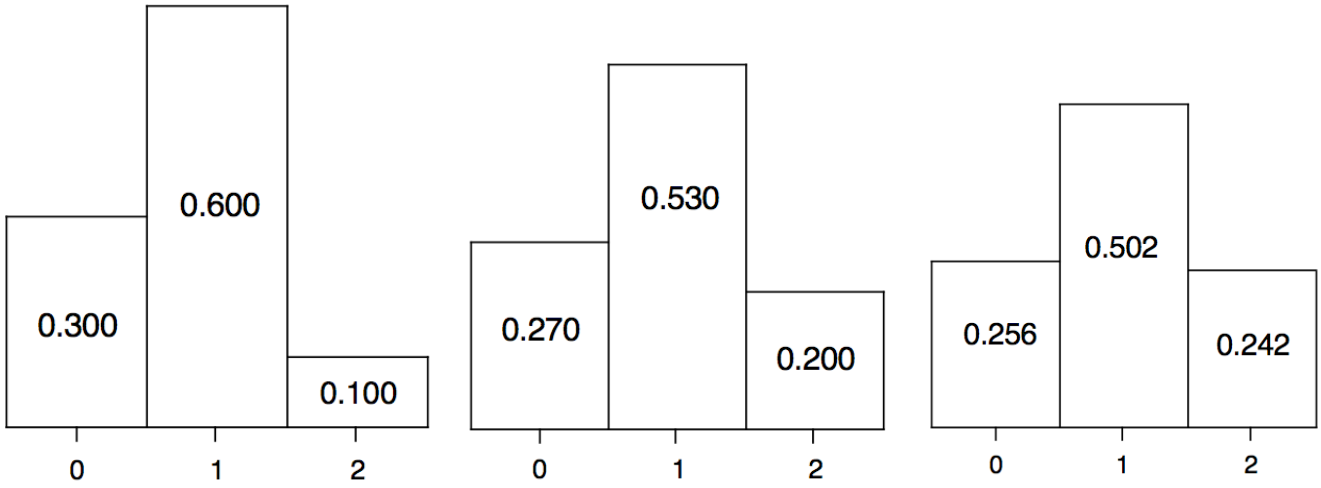
And after 1,000 tosses, the results might be:

0 Heads:	256 times
1 Head:	502 times
2 Heads:	242 times

Here are the histograms corresponding to those three sets of data (with vertical scales adjusted so that the pictures are comparable with each other):



However, following what we just observed about the *total area* contained within pictures of probabilities, it is more useful to indicate the *proportions* of times each possibility occurred, as that will then immediately enable us to see approximations or estimates of the *probabilities* of the outcomes. In the following pictures, the proportions are expressed in terms of decimals rather than fractions as that will make it easier to see what is happening numerically as well as pictorially. Yet another advantage of using proportions is that we then have no need to worry about adjusting the vertical scales: the areas now automatically add up to 1.



Notice how, with sample size $n = 10$, the picture is noticeably different from the picture of probabilities on the previous page but that, by the time we reach $n = 1{,}000$, the shape has become very similar to that picture of probabilities. So the histogram of the data gets closer and closer to the picture of the actual probabilities as the amount of sampling increases. This is no surprise: it is simply a natural consequence of considering probabilities as "long-term proportions of occurrences".

That link between the picture of the probabilities and the histograms as the amount of sampling increases has its parallel with the sample statistics. Let's consider the sample mean $\bar{X}$. Our sample of size 10 consisted of 3 zeros, 6 ones and 1 two. Clearly, adding up these ten numbers gives us 8, and so $\bar{X} = 8 \div 10 = 0.8$. With the sample of size 100 we had 27 zeros, 53 ones and 20 twos. Adding up these 100 numbers gives 93, and so $\bar{X} = 93 \div 100 = 0.93$.

Before proceeding to the final case, let's note something that will be important a little later. This is that, in the two cases so far considered, the calculations can be written respectively as

$$\bar{X} = 0 \times \tfrac{3}{10} + 1 \times \tfrac{6}{10} + 2 \times \tfrac{1}{10} \quad \text{and} \quad \bar{X} = 0 \times \tfrac{27}{100} + 1 \times \tfrac{53}{100} + 2 \times \tfrac{20}{100}.$$

You can easily verify that, in the final case, the 1,000 numbers add up to 986 but, for the moment, building on what I have just pointed out, let's simply write $\bar{X}$ as

$$\bar{X} = 0 \times \tfrac{256}{1000} + 1 \times \tfrac{502}{1000} + 2 \times \tfrac{242}{1000}.$$

Probability distributions, particularly the binomial distribution

Portrayal of the probabilities of all possible individual outcomes in the situation being considered, either as a picture like that on page 38 or just the numbers on which that picture is based, leads us to the idea of a *probability distribution*. That coin-tossing case of

Probability of no Heads	=	$\frac{1}{4}$
Probability of 1 Head and 1 Tail	=	$\frac{1}{2}$
Probability of 2 Heads	=	$\frac{1}{4}$

is a simple example of an important type of probability distribution known as the *binomial* distribution. Some other types of probability distributions are often included in an introductory course, glorying in such names as Poisson, geometric, hypergeometric, uniform, exponential and, as you already know, normal.

The above link between probability distributions and histograms immediately raises the idea of a *probability distribution* having its own *mean* and *standard deviation*. What we have seen is that, as the number of data represented in the histogram increases, that picture becomes more and more like the picture of the probability distribution. Correspondingly, the *probability distribution*'s mean and standard deviation can be described as the "long-term" values of the *histogram*'s mean and standard deviation as the sample size becomes ever-larger. These "long-term" values of the mean and standard deviation are almost universally denoted respectively by the Greek letters μ (pronounced "mu") and σ (you know how this one is pronounced: "sigma"). The mathematician talks of *defining* μ and σ (and σ^2) as the *limiting values* of the histogram's mean and standard deviation (and variance) as the sample size *tends to infinity*, using the symbolism:

$$\bar{X} \to \mu \quad \text{as} \quad n \to \infty \quad \text{and}$$

$$s \to \sigma \quad (\text{or equivalently } s^2 \to \sigma^2) \quad \text{as} \quad n \to \infty.$$

Such symbolism often looks pretty scary to newcomers but, as you can see, its purpose is simply to provide a mathematical "shorthand" to prevent mathematical derivations and proofs becoming inconveniently wordy and cumbersome.

Let's take another look at the calculations of the sample mean $\bar{X}$ at the end of the previous section. You saw that we finished up writing them like this:

$$\bar{X} = 0 \times \tfrac{3}{10} + 1 \times \tfrac{6}{10} + 2 \times \tfrac{1}{10} = 0.800,$$
$$\bar{X} = 0 \times \tfrac{27}{100} + 1 \times \tfrac{53}{100} + 2 \times \tfrac{20}{100} = 0.930,$$
$$\bar{X} = 0 \times \tfrac{256}{1000} + 1 \times \tfrac{502}{1000} + 2 \times \tfrac{242}{1000} = 0.986.$$

Can you see what's happening? We are multiplying each possible value by the proportion of times that it occurs. But it is those very *proportions* that get closer and closer to the *probabilities* of those values occur-

ring. So imagine we toss the coins 10,000 times, 100,000 times, 1,000,000 times and so on. Surely our calculation will get closer and closer to:

$$\bar{X} \;=\; 0\times\tfrac{1}{4} \;+\; 1\times\tfrac{1}{2} \;+\; 2\times\tfrac{1}{4} \;=\; 1,$$

i.e. it's the sum of the values multiplied by the *probabilities* that those values occur. Ah, so to compute μ we don't need to do all that sampling after all! We can simply calculate it directly from the *probabilities*.

Let's develop the same theme to find the *standard deviation* σ or the *variance* σ^2 of this simple binomial distribution. So now we want to find the long-term value of s or of s^2 as $n \longrightarrow \infty$. I suggest we copy the mathematicians here and focus on the variance: it will avoid having square root signs all over the place. We can compute the variance first and then simply take its square root to immediately get the standard deviation.

If we think of this in terms of tossing the coins thousands or millions of times, it all looks rather messy! E.g, suppose we look at the situation where we have the above results from tossing the two coins 1,000 times. Recall what s^2 is: look back at Steps (b) and (c) of the four-step procedure on pages 18–19. At that stage our value of $\bar{X}$ is 0.986, and so "the sum of the 'squared gaps'" $\div$ ($n-1$) would be:

$$s^2 = \tfrac{1}{999}\Big\{(0-0.986)^2\times 256 \;+\; (1-0.986)^2\times 502 \;+\; (2-0.986)^2\times 242\Big\}$$

or, if you like,

$$s^2 = \;(0-0.986)^2\times\tfrac{256}{999} \;+\; (1-0.986)^2\times\tfrac{502}{999} \;+\; (2-0.986)^2\times\tfrac{242}{999}\,.$$

No: I don't intend to compute that unpleasant sum! We don't need to. Let's see what happens as $n \longrightarrow \infty$. First, as we saw above, $\bar{X}$ gets closer and closer to the distribution's mean, i.e. to $\mu = 1$. And, secondly, the fractions (the proportions of occurrences) get closer and closer to the probabilities. So, in the limit (as the mathematicians would say), that unpleasant expression simply becomes:

$$\sigma^2 \;=\; (0-1)^2\times\tfrac{1}{4} \;+\; (1-1)^2\times\tfrac{1}{2} \;+\; (2-1)^2\times\tfrac{1}{4}\,.$$

That's better—much easier arithmetic! It simply boils down to $\sigma^2 = \frac{1}{2}$. So then taking the square root gives us $\sigma = 1\div\sqrt{2}$ which is equal to 0.707.

Finally for this section, let's take a look at another example of a binomial distribution. Incidentally, as you may have realised, "bi–nomial" implies "two names". A binomial distribution always concerns situations or experiments or trials, etc where the possible outcomes can simply be regarded as of just *two* types, let's say S and F, and we are interested in the probabilities that S occurs x times (and F occurs the remaining $n-x$ times) in n trials for all the possible values of x. I've used S and F there because teachers often talk in terms of "Successes" and "Failures" in this context.

Let's suppose we throw three dice (or, what comes to the same thing) one die three times. What is the probability of there being no sixes, or 1 six, or 2 sixes, or 3 sixes? No need to bother with sampling and histograms now: let's head straight for the probabilities. We are, of course, assuming that the dice are "fair" so that, in particular, the probability of a six occurring when a die is thrown is $\frac{1}{6}$.

Since each of the three dice produces a six with probability $\frac{1}{6}$, the probability of all three dice showing a six is surely $\frac{1}{6}\times\frac{1}{6}\times\frac{1}{6} = \frac{1}{216}$. Almost as easily, the probability that *no* die finishes up as a six is $\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6} = \frac{125}{216}$.

But what about the probability of there being exactly *one* six out of three? Here it's easier to think of one die being thrown three times. There are three different ways of getting exactly one six: *either* the first throw produces the six and the other two throws do not produce a six, *or* the single six occurs on the second

throw, *or* it occurs on the third throw. Using the S and F notation, we could therefore have either SFF or FSF or FFS. So, with the probability of S being $\frac{1}{6}$ and of F being $\frac{5}{6}$ and using the same kind of multiplication technique as above, the probability of SFF is $\frac{1}{6} \times \frac{5}{6} \times \frac{5}{6} = \frac{25}{216}$. The same result will be true of both FSF and of FFS since you'll still be multiplying the same fractions together: they'll just be in a different order. Then the probability of there being exactly one six is therefore $3 \times \frac{25}{216} = \frac{75}{216}$.

Finally, what is the probability of there being exactly two sixes? We *could* produce a similar argument as in the previous paragraph, but there is a neater way. Since all the four probabilities (of 0, 1, 2 and 3 sixes) must add up to 1, the probability of two sixes is equal to (1 minus the three probabilities we have just computed), i.e. $1 - \frac{1}{216} - \frac{125}{216} - \frac{75}{216} = \frac{15}{216}$. Job done!

So we have finished up with

Probability of 0 sixes $= \frac{125}{216}$	Probability of 1 six $= \frac{75}{216}$
Probability of 2 sixes $= \frac{15}{216}$	Probability of 3 sixes $= \frac{1}{216}$.

For practice, you might like to verify that the mean μ of this binomial distribution is equal to $\frac{1}{2}$ and its variance σ^2 is equal to $\frac{5}{12}$ (and so, taking the square root, $\sigma = 0.6455$).

A more organised way of dealing with *all* binomial distributions is presented in the Technical Section on pages 85–89.

I appreciate that, at this stage, the standard deviation σ of a probability distribution may not mean a great deal to you! But its importance will become more apparent as we now move on to the next section where we introduce the famous "normal" distribution.

In Statistics books and courses you will often see the term "random variable", and I shall also use this term from time to time later in this material. A "random variable" just means a variable (like the number of Heads or the number of sixes in the recent examples) *whose behaviour is governed by a probability distribution*. The adjective "random" is used to distinguish it from variables in algebra and elsewhere which do not have any such connection with probability.

Continuous probability distributions, particularly the normal distribution

Probability distributions divide themselves into two types depending on what kind of data they represent. So far, we have only been considering what we might call "count data" since the values are obtained by counting something, e.g. the number of Heads when coins are tossed or the number of sixes when dice are thrown (or we could also be considering the number of red beads in the paddle). Count data are a common type of so-called "discrete" data, where "discrete" indicates that all possible values of the data are at a "discrete" distance from each other. Thus, indeed, so far we have only been working with *discrete* probability distributions, of which the *binomial* distribution is a particularly important example. The other type of data is "continuous" data: these usually arise from a *measurement* operation as opposed to a *counting* operation. Obvious examples are lengths, weights, times, etc. In such cases the relevant probability distributions are, as you would expect, called "continuous" probability distributions.

So, when we have continuous data, can we copy what we did with discrete data such as when we were examining the number of Heads when two coins are tossed? That is, can we collect various amounts of data, form histograms of those data as we did on page 39, and see those histograms approaching a picture of a probability distribution such as the one we saw on page 38? The answer is partly Yes and partly No. The devil, as people say, is in the detail. There are both some important similarities and some important differences in the continuous case compared with the discrete case.

For example, in the discrete case our approach was to obtain approximations for the probabilities of each of the various possible values as the proportion of times that each value occurs in quite a large amount of sampling. But in the continuous case (would you believe?!) it actually doesn't even *make sense* to talk about the probability of any particular value occurring! Let's see why.

Suppose that you are taking readings of your body temperature once a day. You might just be taking very rough readings as a simple check that nothing untoward is happening to you. Let me interpret "very rough readings" as simply noting your temperature in degrees Celsius to the nearest integer (whole number). All being well, you will probably get 37°C most of the time, with 36°C some of the time, 38°C now and again, and anything else pretty rarely. But what does "37°C" really mean in this situation? To repeat, you're recording the temperature "to the nearest integer (whole number)". Therefore "37°C" actually implies that your temperature lies somewhere in the interval between 36.5°C and 37.5°C. So then it's quite reasonable to "get 37°C most of the time".

However, it is more usual to read body temperatures to one place of decimals. So in that case, on the occasions when you record exactly 37°C (which, to one place of decimals, we should now write as 37.0°C), this actually implies that your temperature is somewhere between 36.95°C and 37.05°C. And if you had a more expensive thermometer that can read temperatures accurately to *two* places of decimals then, of course, "exactly 37°C" should now be written as 37.00°C and would imply that the temperature is somewhere in the very narrow interval from 36.995°C to 37.005°C. Those three interpretations of "exactly 37°C" would obviously yield very different probabilities. Thus, as I've indicated, to consider the probability that the temperature ***is*** 37°C doesn't really make sense—so it wouldn't make much sense to try to estimate it! In fact, as you can see, the greater the precision, the smaller is the interval surrounding whatever number we record, and so the smaller the probability of recording that particular number. And it's not just a *little* smaller: you can probably see that the probability of recording exactly "37°C" reduces by something of the order of 90% for *each* extra decimal place! The obvious but possibly worrying conclusion is that the probability of recording "exactly 37°C" (or any other precise value) rather rapidly heads toward 0 as we improve the precision of our measurements! That doesn't mean it's *impossible*, but it does mean that's it's something which would only happen just "once in a blue moon", as the saying has it.

On the other hand, although this demonstrates what we *can't* do in the continuous case, it also shows us what we *can* do. That is to consider probabilities of the measurement *lying in any specified interval*. And indeed, that is what is essentially always done when we have a continuous probability distribution.

This is therefore all quite opposite to the discrete case in which the probability distribution actually *consists* of evaluating the probabilities of each and every possible value (or showing a picture of those probabilities). So what form does the "probability distribution" take in the continuous case? What happens if we collect data and form histograms as we did before? One thing I will warn you about in advance: you would need to collect a lot more data than in the discrete case. That takes it rather outside the range of what you could envisage doing manually. If you had had the time and the patience, you *could* have tossed those two coins 1,000 times to get results like you saw on pages 38–39. With similar time and patience you could similarly have thrown three dice 1,000 times and counted the number of sixes.

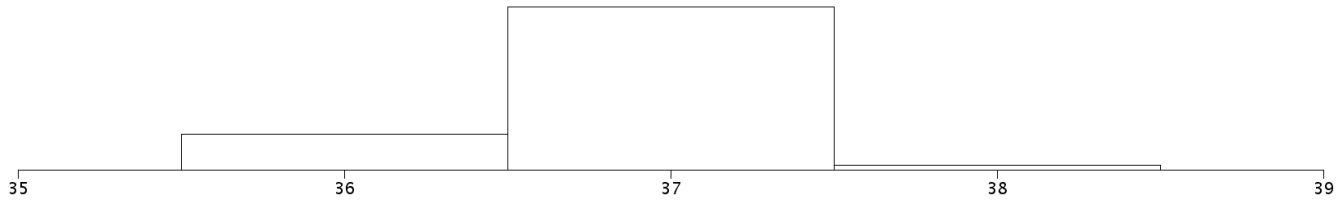
Going by the results for tossing the two coins, it looked as if a sample size of 1,000 was quite sufficient to get a fairly accurate picture of the probability distribution. You'd have probably found the same had you thrown three dice 1,000 times. But, as you'll soon see, the amount of data needed in the continuous case to get a close approximation to the probability distribution is of a different order of magnitude. However, we can get there eventually.

Since manual sampling is now effectively out of the question, we'll need to resort to computer simulations (and it won't be the last time in these Optional Extras). For many decades there have been well-known and

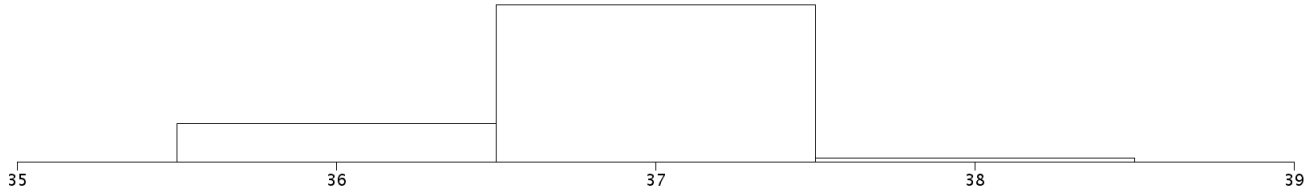
well-tested methods available for generating data from *any* probability distribution (discrete or continuous). This current section is focused on what I often describe as "the statistician's favourite" distribution (and understandably so): the *normal* distribution. So, for illustration, let's use the above case of measuring body temperature; we'll suppose that the temperature is normally distributed, and see what happens.

Let's start as suggested above by measuring the temperature to the nearest degree Celsius. As I said, you would usually get 36ºC, 37ºC or 38ºC (unless you are suffering from a fever or some other medical condition). You *could* get an occasional 35ºC or 39ºC, but that *would* only be very occasionally—unless you have a problem.

So, following on from the previous illustrations, let's first consider a sample of size 1,000. (We'll have to forget the situation previously suggested of daily readings since 1,000 is already close to three years of data, and we're going to need a lot more than that!). The computer simulation that I wrote produced the following histogram (I've used much wider boxes than previously because of what will soon develop):

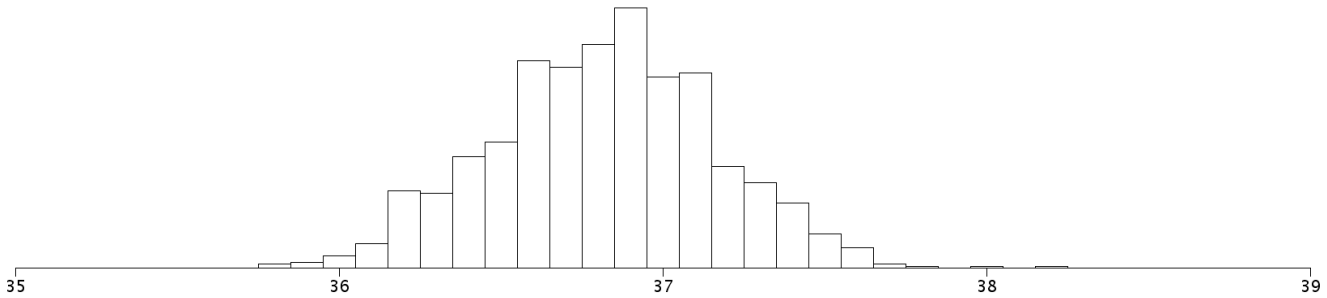


Now, of course, this picture is somewhat reminiscent of the example of tossing two coins, except we must remember that the 36, 37 and 38 are no longer *counts* but *measurements* rounded to the nearest integer. Nevertheless, it would be quite reasonable to deduce from previous work that, to this rather crude level of precision of "the nearest integer", this picture gives a quite reasonable approximation to the probabilities of observing those integer values. But, just to be sure, I then got the program to generate a sample of size 10,000 readings and draw the resulting histogram. Here it is:

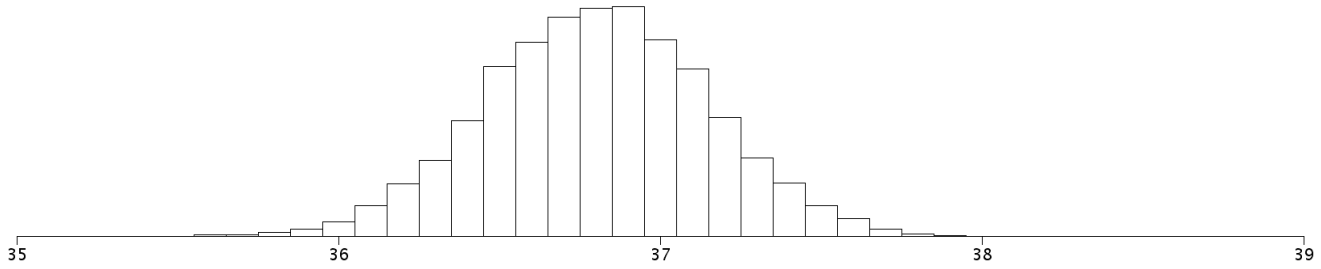


If you look very closely, you will see that this isn't *quite* the same as before, although it is pretty similar. If you're interested, the exact proportions of 36, 37 and 38 with sample size 1,000 were respectively 0.176, 0.801 and 0.023, and with sample size 10,000 were 0.1928, 0.7855 and 0.0217.

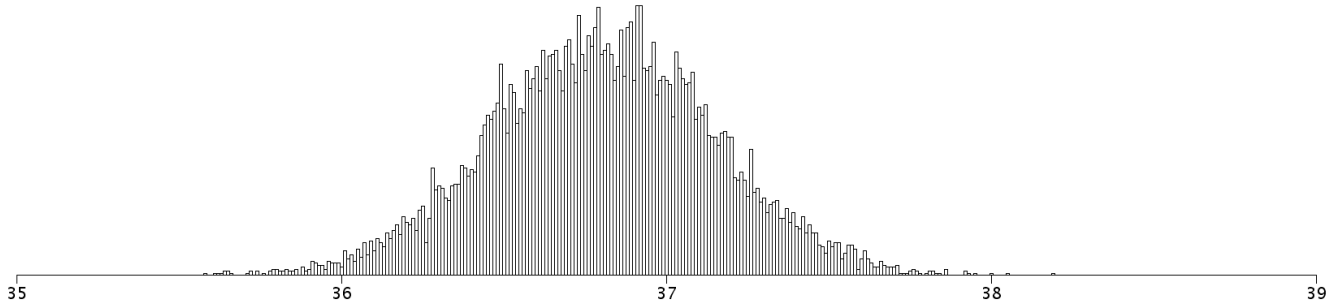
But, obviously, this does not give us much of a clue about what would happen if we recorded the temperatures to a more sensible level of precision. So let's move on to recording them to one place of decimals. Here's the histogram for those first 1,000 observations when measured to one decimal place:



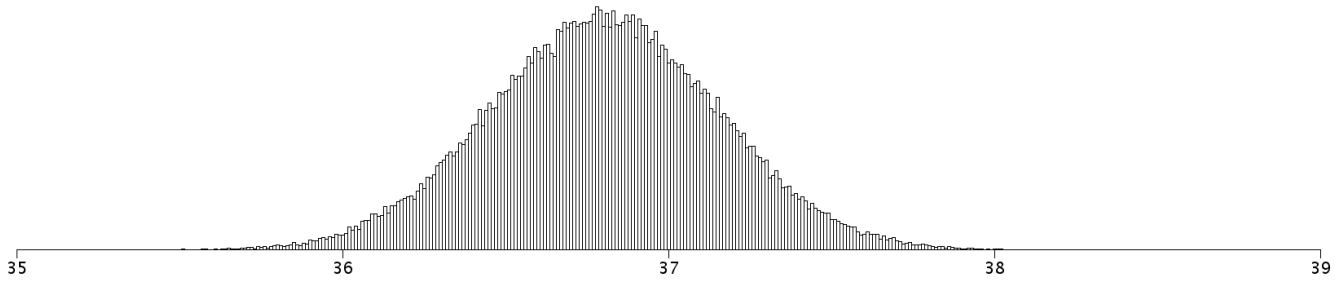
OK, that gives us more of an idea, but this picture is rather ragged. So let's try 10,000 observations; this is what I got:



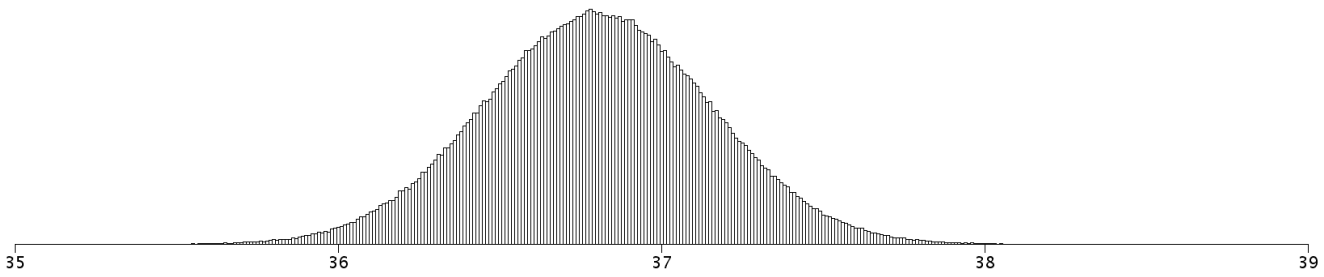
Good: that's a better picture—now we have some reasonable idea of what's going on. Thus encouraged, let's try recording those same 10,000 temperatures to *two* decimal places:



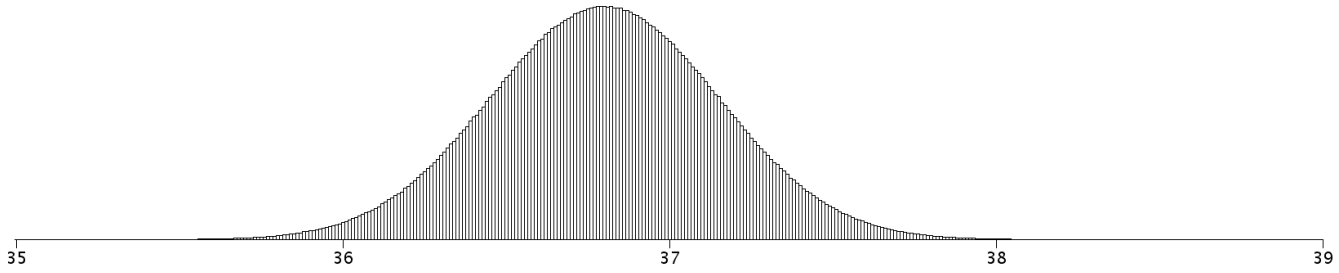
Although this somewhat matches what we were seeing before, we're again back to a rather ragged picture. So let's give the computer a little more work to do and try a sample size 100,000:



Less ragged, but computer time is cheap these days, so let's go up to 1,000,000 observations:



Still just *a little* wobbly, especially at the top, so we'll go for one more picture. With 10,000,000 as our sample size we get:

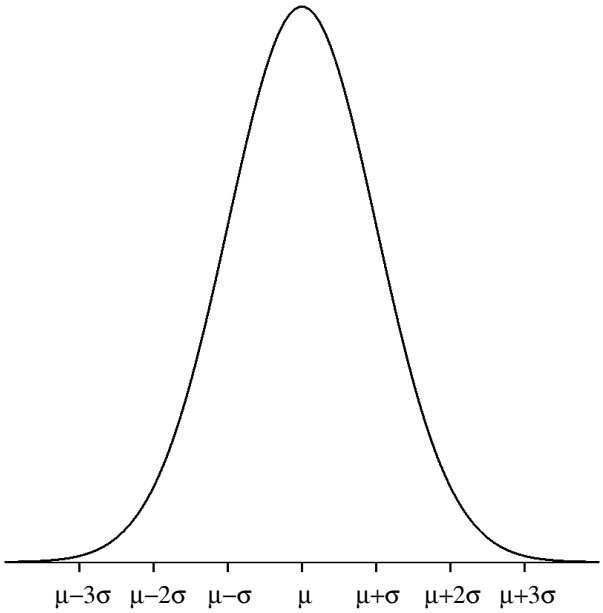


I think that's good enough—if you've ever seen a picture of a normal distribution then I think you'll recognise it!

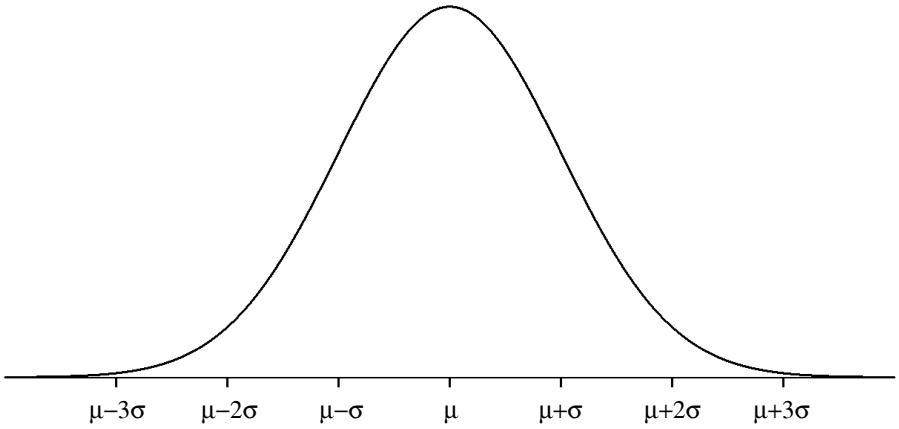
But *which* in the whole family of normal distributions is it? It so happens that there is one, and *only* one, normal distribution for each feasible choice of the mean μ and standard deviation σ. (This characteristic is actually *not* shared by all families of probability distributions.) As you would expect, μ specifies where the distribution is centred. And, as we see in the pictures on the next page, σ defines the *shape* of the distribution: if σ is small then the distribution is tall and thin, whereas if σ is large then the distribution is wider and flatter. Thus, if you have different normal distributions all with the *same* σ, they all have the same *shape* but are shifted sideways from each other.

The normal distribution with mean μ and standard deviation σ is often denoted by $N(\mu,\sigma^2)$—again showing the conventional statistician's preference for referring to the variance σ^2 rather than the standard deviation σ. In particular, the normal distribution with mean 0 and standard deviation = variance = 1, i.e. N(0,1), is known as the *standard* normal distribution.

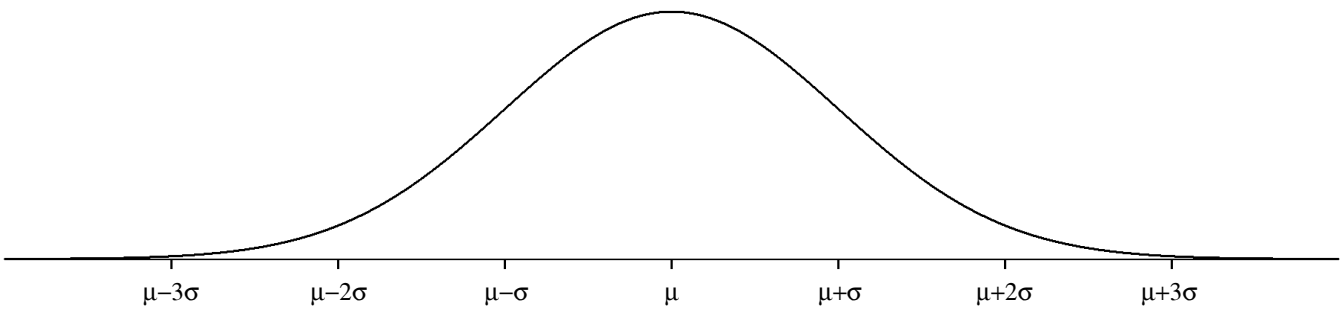
Smallish σ



Larger σ



Still larger σ



Besides seeing the *shape* of the distribution developing from the histograms as the sample size increases and with the boxes becoming correspondingly narrower, you can also get reasonable approximations to μ and σ as the sample size increases. The computer program which produced the histograms also computed the values of $\bar{X}$ and s each time. The values obtained for all the sample sizes illustrated in this and the previous section were as follows:

Sample size	$\bar{X}$	s
10	36.33460	0.38805
100	36.77109	0.36109
1,000	36.84587	0.34558
10,000	36.80593	0.34372
100,000	36.80136	0.35047
1,000,000	36.79796	0.35034
10,000,000	36.80027	0.35015

As you might guess from examining those figures, I was in fact generating data from $N(36.8, 0.35^2)$.

Because of its appealing symmetrical shape, the normal distribution is often referred to as having a *bell-shaped* curve. It has further appeal to the Mathematical Statistician for two main reasons:

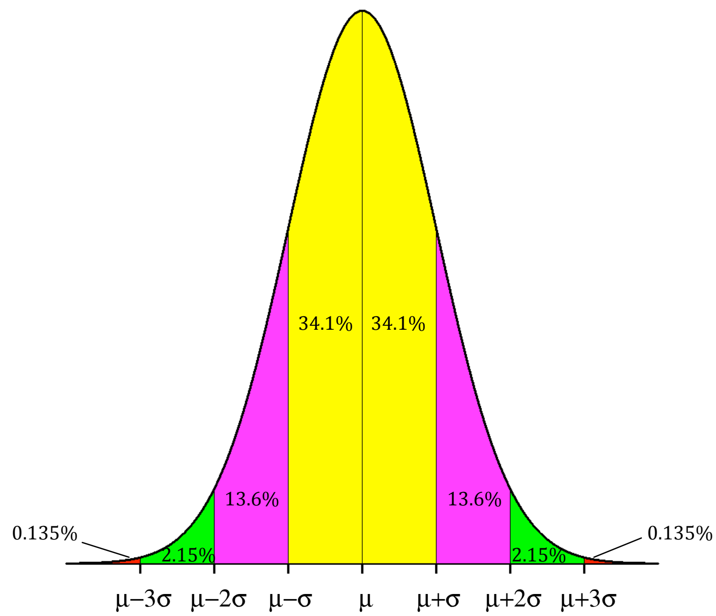
- In practice, data from many sources are mainly clustered around their average (mean) and occur less frequently further away from their mean—which is nicely illustrated by this bell shape; and
- there turn out to be a whole host of pleasant mathematical theorems and results relevant to normal distributions, but to no others. We shall see one of the most famous creations in the next section: the Central Limit Theorem.

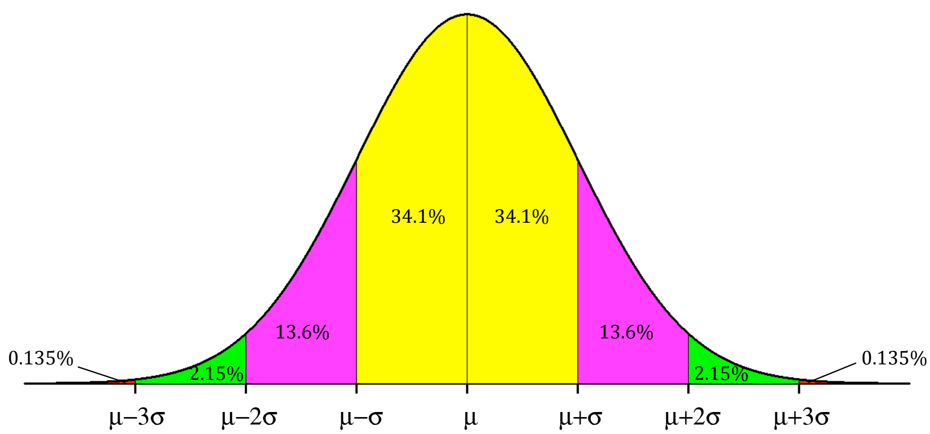
However, before moving on to that section, let's get into just a little detail about finding probabilities when we have a normal distribution. So far we have established the fact that (as is the case with all continuous distributions) there is no point in trying to consider probabilities of getting individual values (because all such probabilities are zero!), so that instead we must concentrate on finding the probability that a normally distributed random variable lies within any specified *interval*. (Incidentally, in that respect, we must slightly extend the idea of an "interval" here to include "one-sided" intervals, i.e. the probability that the observed value is *at least* some number or the value is *at most* some number.) But how can we do all this?

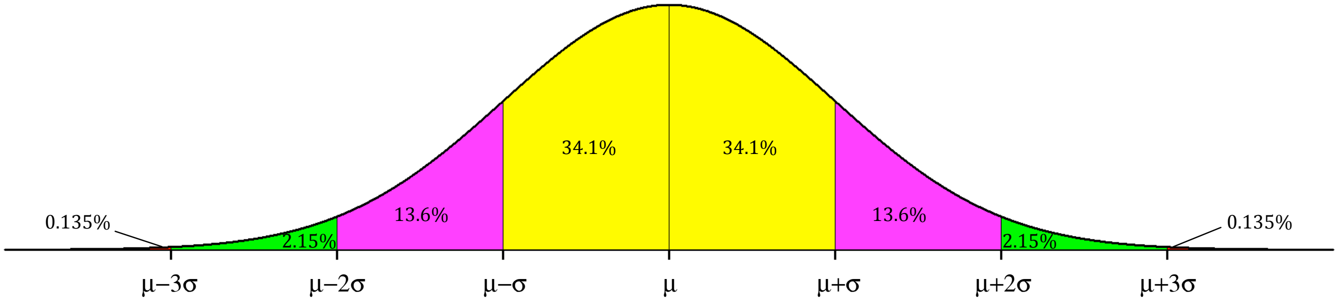
We know how we could do it *approximately* by the method we have already demonstrated in this and the previous section: get lots of data from that normal distribution and find the *proportion* of those data that lie within any desired interval. But how can we do it without going through all that? The answer actually lies in one of Dr Deming's quotations that we saw on page 30. He spoke of a "graph of the normal curve and proportions of area thereunder". We have already noted that any pictures of probability distributions or of the approximating pictures of histograms with *proportions* in the boxes (rather than *frequencies* of occurrence) contain a total area of 1. So the "proportions of area thereunder" are, in fact, *probabilities*. In this sense, the normal distribution has an extremely appealing and convenient feature (a feature which is not wholly unique amongst probability distributions but is nevertheless pretty rare): those "proportions of area thereunder" apply *irrespective of the particular values of* μ and σ.

On the next page there are the same three "graphs of the normal curve" as on page 47 but now with some "proportions of area thereunder" inserted. So, as an example, you can immediately read off that, for *any* values of μ and σ, the probability of the normal random variable lying in the interval between $\mu - 2\sigma$ and $\mu + 2\sigma$ (often expressed as "lying within two standard deviations of the mean") is:

$$2 \times (34.1\% + 13.6\%) \;=\; 2 \times 47.7\% \;=\; 95.4\% \;\text{ or }\; 0.954\,.$$


0.135%
2.15%
13.6%
34.1%
34.1%
13.6%
2.15%
0.135%
μ−3σ
μ−2σ
μ−σ
μ
μ+σ
μ+2σ
μ+3σ


0.135%
2.15%
13.6%
34.1%
34.1%
13.6%
2.15%
0.135%
μ−3σ
μ−2σ
μ−σ
μ
μ+σ
μ+2σ
μ+3σ


0.135%
2.15%
13.6%
34.1%
34.1%
13.6%
2.15%
0.135%
μ−3σ
μ−2σ
μ−σ
μ
μ+σ
μ+2σ
μ+3σ

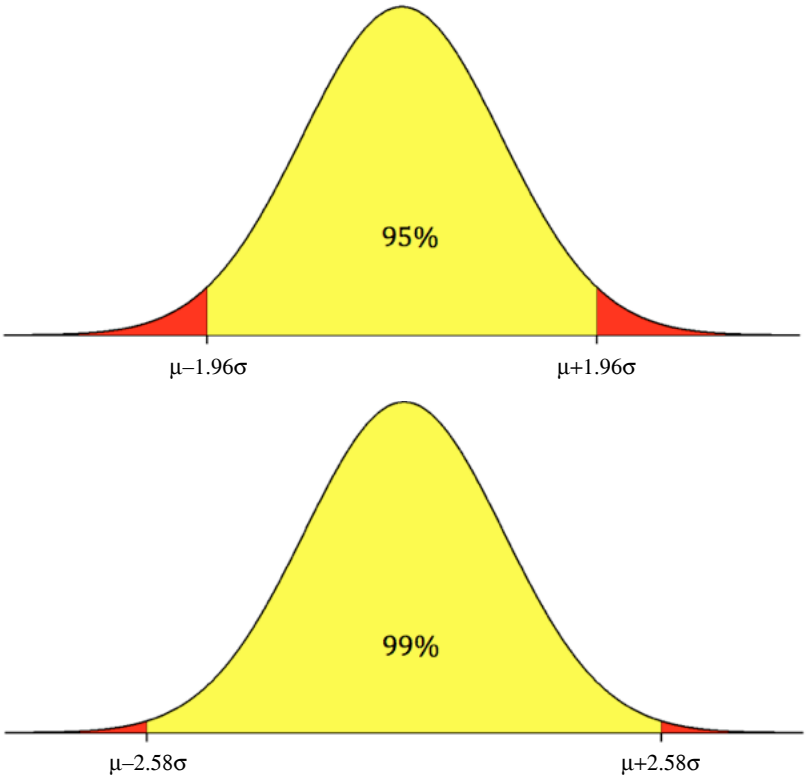
As is clearly seen in those pictures, and as is in any case obvious from the preceding development, the curve which illustrates a continuous probability distribution is relatively high in regions of high probability and relatively low in regions of low probability. The curve is therefore usually referred to as the *probability density function*, or *pdf* for short. (A relatively ancient expression for the pdf was *frequency function*—and this is what Shewhart was referring to in his quotation on page 30.)

On pages 90–91 in the Technical Section I shall describe how the probability of any normal random variable lying within any specific interval can be quickly found using widely-available tables of the normal distribution.

However, there are a couple of particular probabilities that are frequently used in applications of the normal distribution to popular techniques of so-called "statistical inference". Those two probabilities are illustrated below as "proportions of area thereunder". The two diagrams show that, with a normal distribution,

- a random variable has a 95% probability (coloured yellow) of lying within 1.96 standard deviations of its mean μ, 2.5% probability (left-hand tail coloured red) of being less than $\mu - 1.96\sigma$, and 2.5% probability (right-hand tail coloured red) of being greater than $\mu + 1.96\sigma$; and
- a random variable has a 99% probability (coloured yellow) of lying within 2.58 standard deviations of its mean μ, 0.5% probability (left-hand tail coloured red) of being less than $\mu - 2.58\sigma$, and 0.5% probability (right-hand tail coloured red) of being greater than $\mu + 2.58\sigma$.

On pages 54 and 55 I briefly describe the two most commonly-used methods of statistical inference, and the way in which these particular figures apply to them.



Incidentally, if I were to show you a similar diagram with 99.8% in the yellow region and thus 0.1% in each of the two red tails then this would correspond to replacing the 1.96 or 2.58 with 3.09—which explains the remark made about "3.09σ" in the middle of page 2 in these Optional Extras.

The Central Limit Theorem

The Central Limit Theorem is a truly remarkable result which, within the context of conventional Statistics, does indeed place the normal distribution in a position of truly unique importance. Descriptively, it can be summarised as follows:

> Irrespective of *which* probability distribution is being sampled, for sufficiently large samples the sample mean $\bar{X}$ is almost exactly normally distributed. (In theory, there are *some* probability distributions that are exceptions to this statement but, in practice, they are rare.)

There is one respect in which this statement can be confusing. Let's suppose the data are being drawn from a distribution having mean μ and standard deviation σ. As the sample size increases, the distribution of $\bar{X}$ keeps changing. In particular, since we know that $\bar{X}$ gets closer and closer to μ as the sample size increases, the distribution of $\bar{X}$ must keep getting narrower and narrower and therefore taller and taller (since, as always, the total area underneath it has to be equal to 1). So eventually it will get pretty difficult to see *what* kind of shape the distribution has other than it is very tall and very thin! To overcome this problem, the Central Limit Theorem is often expressed as follows:

> If $\bar{X}$ is the mean of a random sample of size n taken from a population having mean μ and variance σ^2 (i.e. standard deviation σ) then
>
> $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
>
> is a random variable whose distribution approaches that of the standard normal distribution N(0,1) as $n \to \infty$. (Z can be referred to as a "standardised" version or form of $\bar{X}$.)

The fact that the random variable Z in this statement has a mean of 0 is fairly obvious: the mean value of $\bar{X}$ is, of course, equal to μ and so the mean value of $(\bar{X} - \mu)$ must surely be 0, as must any multiple of it.

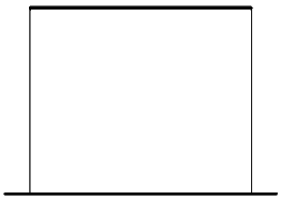
The expression in the denominator of Z is the *standard deviation* of $\bar{X}$: this will be proved in the Technical Section on page 80. That expression is obviously consistent with a fact which we know to be true, i.e. that, as the sample size increases, the variability of $\bar{X}$ keeps getting smaller (so that it approximates the value of μ better and better).

As you might expect from the fact that σ is a measure of variability, *dividing* a random variable by its standard deviation always changes its standard deviation to 1. So the facts that the mean and variance of Z are respectively 0 and 1 are *bound* to be true. The new and remarkable information from the Central Limit Theorem is that the distribution of Z almost always becomes closer and closer to the standard *normal* distribution as the sample size increases (irrespective of what kind of distribution the sample is being drawn from—either discrete or continuous), rather than to any other distribution which has mean zero and standard deviation 1. I implied above that there exist some rare and rather peculiar distributions for which this is not true, but I think that you're unlikely to ever meet one in the "real world"!

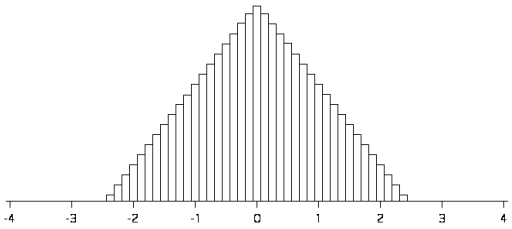
Of course, usually our sample sizes n are rather small compared with "$n \to \infty$"! But what makes the Central Limit Theorem even more attractive is that, almost always, the movement toward normality of $\bar{X}$'s distribution already becomes apparent with very reasonable sample sizes: the distribution is usually effectively indistinguishable from normal for n no larger than 20 or 30 at most, and is often very close to normal for much smaller n. As evidence for this, it's time for some more computer simulations! As a rather remarkable fact, even Shewhart showed the results of some simulations to verify this attractive feature. You notice that I do not say "*computer* simulations" there: Shewhart published the results in his 1931 book—

and that was rather a long while before computers were around. If you are interested in how Shewhart carried out his simulations without a computer, I refer you to pages 182–183 of his book.

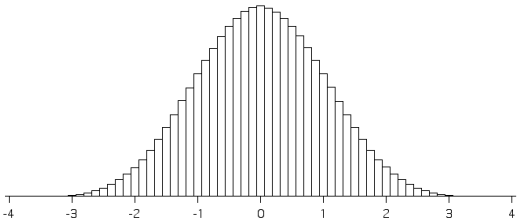
For the results of the computer simulations shown here, I generated samples of various sizes from three continuous probability distributions. The first was one of the two that Shewhart used: a uniform distribution whose pdf is illustrated alongside. The reason for the name is obvious: the probability is spread uniformly over an interval. As a good example, the values you get by pressing the "RAND" key on your calculator have a uniform distribution over the interval from 0 to 1.



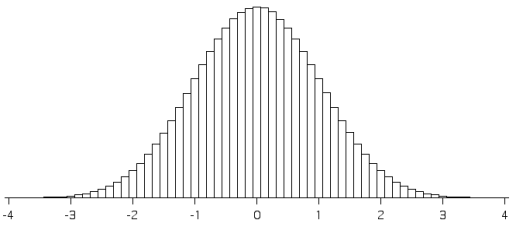
Firstly, I generated samples of size just $n = 2$ from a uniform distribution: here is the interesting histogram of values of the "standardised" version of $\bar{X}$ (i.e. the "Z" in the formal statement of the Central Limit Theorem). In this and all the simulations that follow, the samples were generated 10,000,000 times. This is in contrast to Shewhart's 4,000 times—but even those few must have taken him quite a while!



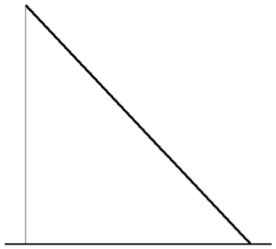
I then moved on to samples of size $n = 4$. Here is the histogram that was obtained. I confess that this one surprised even me—it's incredibly like N(0,1) already:



It hardly seems worth going any further, but here is the histogram with $n = 10$:

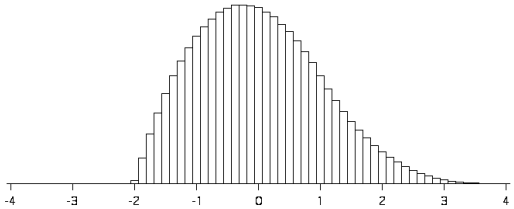


The feature of the uniform distribution which really helps the Central Limit Theorem's effect to become clear so quickly is the fact that, like the normal distribution itself, it is *symmetric*. So presumably that is why Shewhart then moved on to a triangular distribution whose pdf is illustrated here and is, of course, nothing like symmetric:

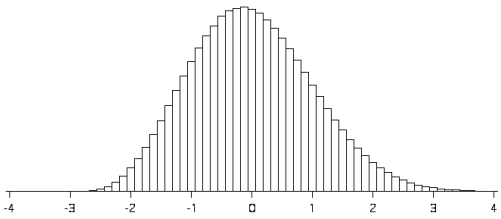


Let's see what happens with samples of size 2 now:

This is again quite a remarkable change from the distribution with which we started.

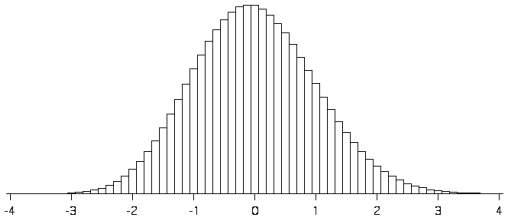


So let's try $n = 4$:

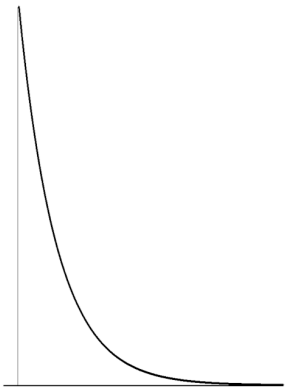


If you look closely, you will see that this is still slightly lop-sided, thus still showing the effect of the non-symmetry of the triangular distribution with which we started. But the Central Limit Theorem effect is remarkably evident already.

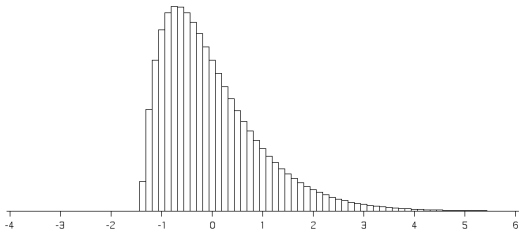
And so we move on to $n = 10$. Still not perfect, but it's almost there.



Some similar simulation work is reported in Chapter 4 of Wheeler and Chambers' *Understanding Statistical Process Control.* Three further distributions are illustrated there, but I'll use just one of them here: the one that is, by far, the most severe test of the five for the Central Limit Theorem. This is the exponential distribution, whose pdf is shown alongside. It's a distribution which is used as a model in many situations including failure analysis and queuing theory, but its main interest here is as one of the most *un*symmetric distributions imaginable.

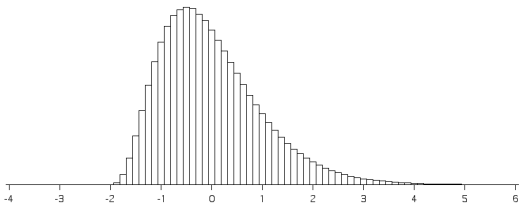


So let's see what shape of distribution the standardised version of $\bar{X}$ has with $n = 2$. Here it is:

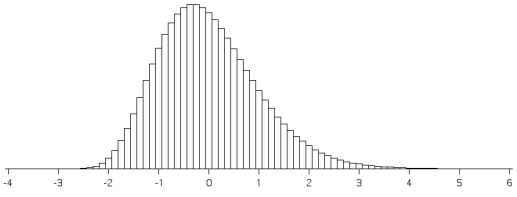


In this histogram It is fairly easy to see the combined influence of both the original shape of the exponential distribution and of an early stage of the Central Limit Theorem. But there's a long way to go.

With $n = 4$ we have:

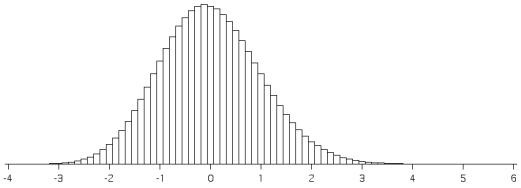


And with $n = 10$ we have:



Well, it *is* getting there but, unsurprisingly, not very quickly.

So, in this case, let's give the computer some real work to do and go right up to $n = 100$:



I think that just about does it. Yes, the Central Limit Theorem even works with something as unfriendly as the exponential distribution!

After all this, it's hardly surprising that, in developing their theory as well as their tools and techniques, conventional statisticians have been very happy to base them on assumptions of normality, justifying this via the Central Limit Theorem in the case of large samples, or directly *requiring* the normality assumption in the case of small samples—for another really nice feature is that if the population being sampled is already normally distributed then $\bar{X}$ itself is also *exactly* normally distributed, however small be the sample size.

Finally then in this crash-course, we come to two of the conventional statistician's favourite applications of the above ideas, often regarded as the most important aspects of so-called "statistical inference": confidence intervals and hypothesis tests.

Confidence intervals

If one wants to estimate the mean μ of the population from which data are being drawn, it is pretty obvious that we should use the sample mean $\bar{X}$ as the estimator. Yes, but that's not very useful unless we have some idea of *how close* $\bar{X}$ is likely to be to μ.

Assuming that $\bar{X}$ is near enough to being normally distributed, so that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is near enough to being standard normal, then, as we have seen on page 50, there is about a 95% probability of Z lying between –1.96 and +1.96. A little algebra enables this range to be expressed as

$$\bar{X} - 1.96\,\sigma/\sqrt{n} \leqslant \mu \leqslant \bar{X} + 1.96\,\sigma/\sqrt{n}\,.$$

If the value of σ is known then the first and last parts of this relationship can be evaluated, thus providing an interval within which we "are 95% confident that μ lies". This interval is referred to as a "95% confidence interval" for μ. However, of course, if the value of μ isn't known then it's rather likely that σ is also not known! So it is quite common practice to compute the sample standard deviation s and throw that into the relationship instead, presumably hoping it's "near enough" to σ. Alternatively, if n is judged to be too small to do that, the conventional statistician conveniently assumes that the data-values themselves are normally distributed, in which case

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has what is known as a "Student t" distribution, tables of which can then be used to find a number to insert into the relationship in place of the 1.96. ("Student" was the pen-name of an English statistician whose real name was W S Gosset.)

I expect you will immediately appreciate (referring again to page 50) that a "99% confidence interval" can be obtained by replacing 1.96 by 2.58 throughout this discussion.

Hypothesis tests (also known as significance tests or tests of significance)

Finally, suppose we want to use the value of $\bar{X}$ to formally "test the hypothesis" that μ has some proposed value: for sake of argument, say $\mu = 37$. "$\mu = 37$" is then usually referred to as the *null* hypothesis and is denoted H_0. How can we formally test for the truth or otherwise of H_0?

The standard approach is as follows. It starts by tentatively *assuming* that H_0 is true. It would then follow, of course, that

$$Z = \frac{\bar{X} - 37}{\sigma/\sqrt{n}}$$

has the standard normal distribution, at least approximately, under the same conditions as previously. Also as before, it is unlikely that the value of σ is known and so the sample standard deviation *s* is usually substituted instead, especially if the sample size *n* is large. It is easy to see how the resulting test statistic *Z* behaves. If the assumption that H_0 is true holds then $\bar{X}$ should be reasonably close to 37 so that *Z* will be relatively small (positive or negative). But if H_0 is not true, particularly if it is *seriously* untrue—i.e. μ is *very* different from 37—then $\bar{X}$ will reflect that very different value of μ, leading to a relatively large (positive or negative) value of *Z*.

But how large is "relatively large"? The same figures as before provide the answer. For example, if H_0 is true then we know there is (exactly or approximately) 95% probability that *Z* lies between –1.96 and +1.96. This interval is therefore often referred to as the "acceptance region", thus leading to formally accepting H_0 if *Z* falls within it. All other values, i.e. all values outside the interval –1.96 to +1.96, thus form a corresponding "rejection region" such that if *Z* is found to have any such value then the formal decision is to "reject H_0". This "rejection region" is usually referred to as the *"5% critical region"*, and the operation is referred to as carrying out the test at the *"5% significance level*". Thus the "significance level" is defined as the probability that the test *wrongly* rejects H_0, i.e. rejects H_0 when H_0 is in fact true. Note that the *smaller* the significance level at which H_0 can be rejected, the *stronger* is the evidence for so doing.

It should also be noted that there is a considerable non-symmetry in such a testing procedure. If we consider continuous distributions in particular then, while strong evidence may be found for rejecting H_0, one can *never* find evidence of any kind for believing that H_0 is actually *precisely true*: all one can truthfully do is to *not reject* H_0. As soon as I realised this long ago, I never subsequently used phrases like "accept H_0" or "acceptance region" since I regard them to be misleading because of their sounding more positive than is appropriate. I guess it's rather like a person in a law-court being judged as either "guilty" or "not guilty": the latter verdict does *not* imply that innocence has been proved. The difference with this analogy is that it is nevertheless *possible* for innocence to be proved in a law-court, whereas in a hypothesis test involving a continuous distribution, it can *never* be proved that H_0 is true.

There are two further aspects of confidence intervals that can also be carried over into hypothesis testing. First, if *n* is small but normality is assumed then a replacement figure for 1.96 can be found from tables of the Student *t* distribution. And secondly, the above hypothesis test can be carried out at the 1% significance level by replacing 1.96 by 2.58 throughout.

PART E. IS THERE *ANYTHING* NORMAL ABOUT CONTROL CHARTS?

Not a lot! Firstly, as you now know and unlike most techniques in traditional Statistics, control charts are specifically designed for what Deming referred to as "analytic studies", i.e. their prime purpose is to provide help toward an improved future rather than simply describing characteristics of the current (or past) state. Thus in that sense they are indeed not "normal" compared with most statistical methods. Further, their "validity" does not depend upon data being normally distributed—or being describable in terms of *any* probability distribution: "the reason is that no process ... is steady, unwavering" (page 30).

But what about those control-chart constants like 2.66? Now, some nice mathematics *is* needed in order to derive them. And it is true that they are derived using the model of a normal distribution. But the fact is that the mathematician *needs* some such model or assumption or else he cannot produce *any* nice mathematics! So we *either* do without ever having such numbers *or* we have to allow some model or assumption in order to get them. And to use the normal distribution for this purpose is particularly convenient for the mathematician. The meaningful question in practice is not whether such numbers are "valid": it's whether or not they are found to be *useful*. As we saw on Day 1 page 8 and quoting from the creator of the control chart rather than from his famous student, their "validity" does not come from the fact that they have been derived using "a fine ancestry of highbrow statistical theorems". On page 18 of Shewhart's 1931 book, his next sentence was: "Such justification must come from empirical evidence that it works." Experience of well over three-quarters of a century is clear: it works.

1. Back to basics

You have seen several instances in the course where I've made what might have appeared to be derogatory comments about "conventional" or "traditional" statisticians. Actually the problem is, of course, not with the statisticians themselves but with the way that they have been taught. By now, if you have read Parts C and D, you know something about that. And then, as so often the case, one of Dr Deming's famous and perceptive one-liners comes into my mind; in this case it's: "How would they know?".

The basic problem is the teaching of Statistics as if it were merely a branch of Mathematics. Or rather: doing that but not clarifying the *limitations* of so doing when attempts are subsequently made to apply the consequences of that teaching to the real world: in real situations, with real data, with real processes, in real circumstances.

Both learning and teaching Statistics as if it were merely a branch of Mathematics can be very convincing. I should know: I've done plenty of both in my lifetime. Mathematics *is* very convincing. Of course it is: in essence, Mathematics is simply (but not necessarily easily!) an exercise in logic. Mathematical logic consists of arguments like

> "If this is true and if that is true then here's something else which is true."

These *are* statements of absolute truth—and that's very comforting! So *of course* Mathematics is convincing: you keep learning what are unarguably new truths! What follows the "then" in that statement *is* a new truth—*as long as what follows the "if"s are true.* And there's the problem in moving to real-world applications.

In Mathematical Statistics you will find loads of logical progressions such as:

> "*If* we have one or more normal distributions and *if* we can draw random samples from those distributions then ... "

except that the "*if*"s are not usually emphasised as I have just done. Where do the "*if*"s come from? In a subject like Mathematical Statistics they are quite likely to be *partly* motivated by things that might be fondly *hoped* to approximately happen in real life. But they are *definitely* motivated by what makes the mathematics possible to do! You will often come across histograms whose shapes look roughly like a normal distribution, implying that (if the source from which the data are taken was exactly stable) the distribution from which they come is something like a normal distribution. Certainly you may believe that tossing coins or throwing dice—or selecting 50 beads out of a container of 4,000 beads using a paddle—produces something like a random sample. But Mathematics is not do-able with "something like"s. Mathematics needs "exactly"s. And so what follows the "*if*"s in the mathematical argument are *not* "something like"s, for no progress with the mathematics can then be made. Instead, at best, they are *idealisations* of what might be regarded as roughly happening in practice which, if those idealisations are *assumed* to be true, permit and enable the mathematical argument to proceed.

As I said back on page 29 (which in turn recalled something on Day 1 page 6), delegates having some qualification in Statistics could be something of a problem in my seminars. Initially I did not have the wit to describe the obstacles to transferring mathematically-obtained results into statistical practice in the way that I have just expressed them above. So what could I do instead to attempt to convince those statisticians? Eventually I hit upon some things that worked. If ever you find yourself confronted by mathematically-educated statisticians, I hope that both the above discussion and the rest of what follows in this part of the Optional Extras will prove helpful to you if you ever find yourself faced with any similar kind of awkward situation.

Now, it might be that you became so excited by some of what you were learning in Part D's "crash-course", such as the famous and amazing Central Limit Theorem and the simplicity and elegance of forming confidence intervals, that you may have forgotten why I have included Part D in this material! It was to help you to understand how the conventional statistician thinks and why he thinks that way. Thus you now have some means of communication with him which might well not have existed previously.

So, thinking back to the very beginning of the crash-course (page 37), let's get back to reality.

We began with what I described there as "quite familiar ground": the histogram. And, of course, to an extent, it was. But when the histogram was introduced early in our *12 Days* course, the approach and discussions were rather different from what is usually met in an introductory conventional course, as typified by the crash-course. The "common ground" was indeed that the histogram is a particular way of producing a "picture" representing a collection of data. But a prime difference relates to the kind and likely source of the data being illustrated.

With our Shewhart/Deming background we are very likely to presume that the data being illustrated in a histogram come from some *process*: we regard the understanding and improvement of *processes* to be our main aim and purpose of collecting data. But that is not the language you normally see and hear in the introductory conventional course—there was nothing of that in Part D's crash-course. Instead, in the crash-course, the source of the data illustrated in a histogram was expressed more in terms of what would soon be met in the conventional course: what I have previously referred to as the "life-blood" of the conventional statistician, namely probability and probability distributions. This is also the reason you are relatively unlikely to see the *run chart* at an early stage in the conventional course, despite its simplicity and its usefulness: run charts do not fit into this mould. Yes, conventional Statistics can eventually come up with some methods of analysing run charts: but nothing that would suit the early stages of an introductory course and nothing which is anywhere near as straightforward and effective as control charts. Recall the fundamental problem of the histogram. It was this: if there was any time-dependence in the behaviour of the process (and, to put it mildly, there usually is!), the histogram ignores it. It was there in the original

data, ready to be seen on a run chart or control chart and thus to be useful to us. But *it is rendered invisible by the histogram.*

As we have seen in the crash-course and would soon be seen in the introductory conventional Statistics course, one common way of describing the data is in terms of something like "a random sample from a population". Another common way is as the results obtained by carrying out repeated trials of an experiment or of some operation. (Although sounding rather different, these two ways often turn out to be pretty much equivalent to each other.) But surely it is sensible to carefully examine these two descriptions of the source of the data—which is usually *not* done in the conventional course. So let's do it.

First, what does "a random sample from a population" imply? A "population" is some collection of "things"—possibly people but often not. In the Red Beads Experiment, the population is a collection of 4,000 beads. A "sample" from the population is, of course, a selection of some of those "things" from the population—we are familiar with samples consisting of 50 beads obtained using the "paddle". So what does a "*random* sample" imply? It implies something very specific: namely, that each and every possible different sample (of the specified size) is *equally likely* to be drawn as any other.

In the case of the Red Beads Experiment, that's a *lot* of different possibilities! According to my reckoning, there are exactly

30,645,728,733,196,872,985,716,792,733,231,423,451,265,770,337,904,251,108,650,805,313,511,833,016,584,223,503,281,029,862,021,442,852,832,482,389,462,720

different possible selections of 50 beads from the population of 4,000 beads. (If anybody thinks differently then do please get in touch with me—I'm still waiting!) And "random sampling" means that each and every one of those selections is equally likely to be drawn as any other. That's a pretty stringent requirement! It would require remarkably consistent workmanship in the production of the beads and, I suggest, a rather more refined sampling mechanism than that wooden or plastic paddle!

The introductory course then swiftly moves on to Probability—no wonder since, as I've said before, the conventional statistician regards Probability as the branch of Mathematics on which the whole subject of Statistics is based.

As we have seen in Part D, the idea of the *probability* of an *event* occurring in some particular situation is often described in terms of the *long-term proportion of occurrences* of that event, implying (conceptually at least) that it is possible to repeat the situation being envisaged *ad infinitum*. Moreover, it also implies that the likelihood of occurrence of the event of interest remains wholly unchanged throughout all that rather long time. So again we have the implication that whatever is being considered stays "steady, unwavering" forever, the property which I refer to as "exact stability". As mentioned in the crash-course, exact stability is a far more exacting notion of stability than that which is considered in the control-charting context; whereas the latter is an entirely practical proposition, the former is pretty fanciful! But usually such doubts about this major assumption are hardly even mentioned, so again the student is effectively being "brainwashed" to ignore time-dependence.

As already implied when referring to random sampling, many examples of probability calculations, both at the introductory stage and later, make use of considerations of *symmetry*, which is where the various possible outcomes are all regarded as *equally likely* to occur. As previously mentioned, common examples are the two sides of a coin, the six faces of a die, or the 52 playing cards in a "well-shuffled deck". The difficulty, or rather the impossibility, of creating such symmetry in practice is also rarely mentioned, so yet again one might suggest that there is some unfortunate brainwashing going on: effectively that the world can be described in terms of mathematical idealisations that are unattainable in practice. Recall the evidence which Deming himself provides (*Out of the Crisis* page 300[351–352]) which we saw on Appendix page 11. Over the years he tried four different paddles, two of them for large numbers of experiments. "Paddle No. 1, used for 30 years, shows an average of 11.3" whereas "the cumulated average for paddle

No. 2 over many experiments in the past has settled down to 9.4 red beads per lot of 50." Deming carried out the Experiment on Red Beads a very large number of times. Random sampling and assumptions of symmetry would, without any real doubt, instead have produced long-term averages pretty close to 10.

Now, don't get me wrong. I'm not trying to imply that, because the Mathematician's assumptions are rarely if ever attainable in practice, Mathematics is useless in practice. Of course not. I often quote a statement made by the famous British statistician Professor George Box (whose huge Statistics Department at the University of Wisconsin I was privileged to work in during 1967–68, very near the beginning of my career). His very astute observation was that "All models are wrong—but some are useful"; I rather wish that George had also appended a few qualifying words such as "in some circumstances". By making those idealising assumptions, the Mathematician can produce all sorts of results that would be quite impossible to derive otherwise. And many of them *are* highly useful in practice. My "grumble" is that the student is not reminded often enough, if at all, that the proof of those results *does* strictly depend on those idealising assumptions. Thus, when trying to use those results in practice, the student is likely to be insufficiently wary about whether practical situations of interest might be *too far removed* from the idealising assumptions for those results to hold sufficiently closely to be useful.

So what should be done? Let me complete Shewhart's quotation to which I alluded in the second paragraph of page 57 (from page 18 of his 1931 book):

> "... the fact that the criterion which we happen to use has a fine ancestry of highbrow statistical theorems does not justify its use. Such justification must come from empirical evidence that it works. As the practical engineer might say, the proof of the pudding is in the eating."

The conventional statistician has sometimes produced some claims about control charts, claims that indeed have a very fine ancestry. But where is their empirical evidence: where is the proof of the pudding? On the other hand, I *shall* produce some empirical evidence, but my empirical evidence will confirm that what the conventional statistician is often heard to say about control charts *actually doesn't work*!

2. Two simulation studies

If you decided to work through the crash-course (Part D) then it may be quite a while since you read Part C. I shall therefore start by reproducing the final section of Part C from page 35 since here I shall then continue directly on from that point.

> **... and what can be done about it**
>
> Substantially, the conventional statistician's case for requiring normality to make control charts and their control limits "valid" exists on two main fronts. Such a statistician believes that normality is needed:
>
> - because control-chart constants that are used in computation of control limits, such as the 2.66 with which we became familiar on Day 3, are derived from normal distribution theory (as indeed they are); and
> - so that *a probability interpretation* can be given to control limits: specifically, the claim is often made that, under normality, there is a probability of 0.0027 (i.e. 0.27%) that any particular data-point falls outside Shewhart's 3σ-limits (note that 0.27% = 2 × 0.135% when you look at page 49) if the process is in statistical control.
>
> Now again, any acceptance of Shewhart's and Deming's teachings immediately leads to the denial of the conventional statistician's claims in both of these respects. The "*fortunate and remarkable*" facts alluded to *[on page 34]* are however that, even if we *ignore* what Shewhart and Deming said about both normality and probability interpretations (recall, in particular, page 30), the above claims are *still* demonstrably wrong! In other words, we are able to wade right into the conventional statistician's camp, onto ground which both Shewhart and Deming believed to be without foundation yet which the conventional statistician needs to have faith in (for all that he has learned is built upon it), and to talk to him in language which he both understands and accepts (even if we don't), and to *still* produce evidence which disproves his beliefs!!
>
> That evidence is demonstrated in Part E on pages 65–70.

A conclusion which must surely be drawn is that, if you hear (as one does) a conventional statistician proclaiming something to the effect that, for control charts to be "valid", the data must be normally distributed, the said statistician really does not know what he's talking about.

The two particular issues raised above both stem from the fact that *some* help from Mathematics is necessary in order to develop the details of where control limits should be placed. We have the guidance from Shewhart about "3σ" but that isn't sufficient to provide exact details. So how can Mathematics help to provide those details? As already emphasised: only by making idealised assumptions about where the data will come from. Unsurprisingly, the Mathematical Statistician likes to assume the data come from a normal distribution. That *is* the assumption made (plus exact stability, random sampling, and the rest), combined with Shewhart's "3σ"-guidance, in order to obtain the 2.66 for producing the control limits using moving ranges. The same is true of all the values of h, H and h_2 on pages 20–22. The manner in which the values of these constants are derived is described in the Technical Section on page 83.

So the first issue which arises is that, because these control-chart constants depend upon the assumption of normality, the conventional statistician claims that the data *need* to come from a normal distribution in order for the control chart to be "valid". Now, you and I know that, if that were really true, the control charts we use (and any others that could ever be devised) would *never* be "valid"! With real process data we don't even believe in exact stability, i.e. that they come from *some* fixed probability distribution—real

processes are *never* completely unchanging over time—let alone having a normal distribution in particular. So, isn't "Are they *useful* in practice?" the important question to ask about control charts? Well, nearly a century of their use would seem to answer that question! The second issue is that the Mathematical Statistician needs to express his conclusions in terms of probabilities—he feels he hasn't done a "proper job" otherwise. But you and I know that, again with processes never being completely unchanging over time, it cannot ever be possible to express *anything* about them in terms of exact probabilities.

Now, this matter of processes never being wholly unchanging, i.e. not being exactly stable, is a really hard nut to crack with Mathematical Statisticians. That's not surprising for, as you now know, almost all of their methods *depend* on exact stability (probability, probability distributions, and the like). Hence, when I was faced with such debates, I regarded that as a brick wall upon which I could make no impression, at least for the time being. So I asked myself whether there could be *some* way that (in order for them to listen to me) I could *assume* exact stability but *still* show that what they were saying with regard to the two issues just outlined was simply untrue. There was. I devised two computer simulations, one for each issue. I no longer have my original programs (it was a long while ago!) but fortunately I still have some of the results that they produced. I have therefore recently rewritten those programs from scratch and have found (with relief!) that these new programs are producing output that is entirely similar to the old results. I can therefore now proceed to the rest of this section with confidence!

So firstly let's take the issue that, because the constants we use to find our control limits have been computed by mathematicians after they assume the data come from normal distributions, our data *have* to come from normal distributions in order for the control limits to be "valid". For this I developed an argument based on one-at-a-time data. As you know, there is one "magic number" needed to construct control charts in this case: 2.66. It will be seen on page 65 that there is a direct relationship between the 2.66 and the conversion factor *h* which was tabulated on page 20. At that stage I was discussing the issue that, when working with a-few-at-a-time data, variation is measured using ranges rather than with sample standard deviations. I pointed out that, since the standard deviation is a kind of average or typical gap between the values in the data and their mean, it is obvious that the range (largest value minus smallest value) will be *greater* than the standard deviation. Therefore if one wanted to change the range into a measure of variation which is on the same scale as σ then it would have to be divided by some conversion factor: that's the conversion factor which is denoted by *h* and which, of course, depends upon the subgroup size. With one-at-a-time data the argument is similar: the only difference is that *moving* ranges are used to measure variation. But moving ranges are effectively ranges of subgroups of size 2, and so then the relevant conversion factor is the value of *h* for $n = 2$, and that is 1.28. A sensible estimator of σ (the standard deviation of the assumed normal distribution) is thus obtained by dividing the mean moving range $\overline{MR}$ by 1.28.

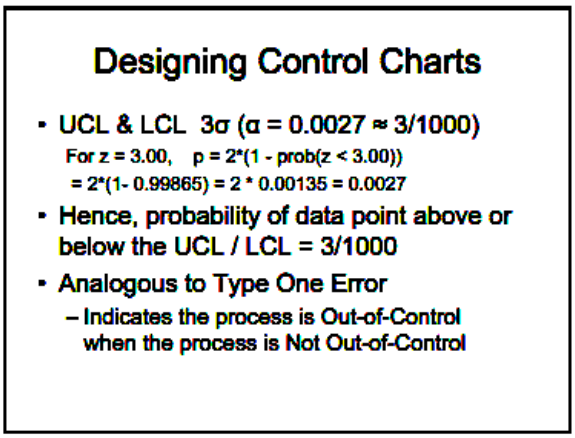
I thought it would first be interesting to go along with the conventional statistician by generating data from a normal distribution and then examining *how good* $\overline{MR} \div 1.28$ turned out to be as an estimator of σ.

Having learned something about that, there was an obvious way to examine the claim that our data *have* to come from a normal distribution in order for the control limits to be "valid". That was to modify the program so as to generate data from some other probability distributions (chosen to be clearly very different from normal distributions) and examine how good the estimator $\overline{MR} \div 1.28$ turned out to be as an estimator of σ with them. Now presumably, if there were any virtue in the notion that normality is *necessary* for the standard method of constructing control charts for one-at-a-time data to be "valid" (or, at least, *useful*), the performance of this estimate should be "good" under normality while it should be "bad" otherwise. Was it? Both the full details and the results of this simulation study are contained in Section 3 (pages 65–67).

Now let's move on to the second computer simulation. This was to examine the claim made by some Mathematical Statisticians that, if the data to be plotted on a control chart come from a normal distribution, they can provide probability interpretations of what the control chart does—just as they can in the case of

other statistical techniques with which they are much more familiar. There is one particular probability interpretation that is often seen. It is the claim that, if we have an exactly stable process with the data coming from a normal distribution, the use of Shewhart's "3σ" control limits on an $\bar{X}$-chart results in there being just a 0.0027 probability that a value of $\bar{X}$ falls outside the control limits.

Here are a couple of examples of such a claim that I found on the internet. I cannot recall exactly where—again, it was a long while ago—and, in any case, I would not want to encourage you to chase up such sources! Firstly, this is a copy of part of some material supposedly telling us "where do typical control chart signals come from":



And, secondly, here is an extract from something titled "Designing Control Charts Control Region":

> "If we observe a signal even though the process has not changed, we have made a *Type I error* (α). This error leads to inefficiency since we will react to a signal but not find any actual cause, the process having not actually changed. By convention, the probability of a Type I error (α) is specified as 0.0027 (0.27%). This results in the control limits trapping 99.73% of the statistic that is being plotted on the control chart.
>
> Note: 99.73% equates to ±3 standard deviations from the process average, if the data being plotted is normally distributed."

You have previously seen how that 0.0027 arises. To remind you, it comes from observing in the diagrams on page 49 that there's a 0.00135 probability in each of the two "tails" of a normal distribution beyond the $\pm 3\sigma$ points. So it is indeed true that the probability of a normally-distributed random variable being more than 3 standard deviations away from its mean is 0.0027.

But, hold on. In order to be able to communicate with the conventional statistician, we have already gone a long, long way. We have gone so far as to *ignore* Deming's crucial observation that "no process ... is steady, unwavering" so as allow him the possibility (which he *needs* to assume) that a process *can* produce data from the *same* probability distribution hour after hour, day after day, week after week—and moreover that that probability distribution could be a normal distribution. Despite not believing this for a host of reasons, we are prepared to *pretend* that it could be true (at least for the time being).

> For interest, I asked the delegates at several of my seminars whether any of their processes behaved like that. Nobody ever said "Yes". Some even greeted the suggestion with considerable mirth!

But pretending it is feasible to actually know *which* normal distribution we have is surely a step too far even for the conventional statistician! However, that *is* what is needed for that probability of 0.0027 to be "valid" (there—we can use some of his own language!).

If we politely point out that fact, he will go along with it (unless he is of extraordinarily closed mind). He's quite used to that kind of problem in his own field, so I think he will accept that it *is* a necessary evil that, in the absence of divine inspiration about the values of μ and σ, we shall have to be content with *estimates* of them computed from the data. Never mind, the control chart has been around for the best part of a century and has been seen to be pretty useful during that time, in spite of this little difficulty. So presumably the inaccuracy caused by having to estimate μ and σ rather than knowing their true values can't be much. After all, he's often done the same kind of thing elsewhere. Therefore that probability computation of 0.0027 should surely be at least *approximately* correct.

There's only one way I know of finding out—yes, another computer simulation. This one was quite easy to write. All that was needed was to (a) generate a-few-at-a-time data from a normal distribution for various subgroup sizes and baseline lengths (the number of time-points used in the calculations), then (b) compute the control limits for an $\bar{X}$-chart in the usual way as described in Part B (pages 21–22), and (c) in each case, compute the probability of a subsequent value of $\bar{X}$ falling outside those control limits.

Obviously, except for the very occasional fluke, every set of data will produce a different $\bar{\bar{X}}$ and a different $\bar{R}$, and thus a different probability. These probabilities can then be collected into a histogram, and the histogram can be investigated to see if there is any sense in claiming that the probability of a point falling outside the limits ***is*** 0.0027.

Full details of this simulation study and the results it produced are provided in Section 4 (pages 68–70).

3. "Control-chart constants depend on normally-distributed data, so unless your data are normally distributed your control chart isn't valid."

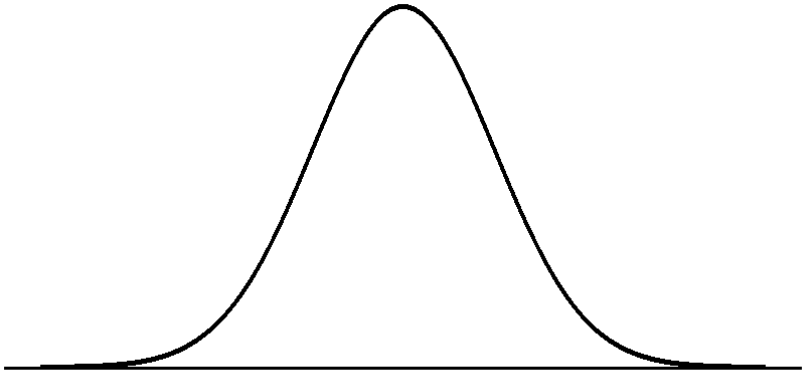
As observed on page 57, the familiar 2.66 does depend on the assumption that we have normally distributed data (in conjunction with Shewhart's 3σ-guidance). The dependence arises since the conversion factors *h* tabulated on page 20 depend on that assumption. With moving ranges effectively being subgroups of size two, the relevant value of *h* is 1.128 so that $\overline{MR} \div 1.128$ becomes a "valid" estimator of σ under that normality assumption. The 2.66 then arises as the result of dividing 3 by 1.128 since this makes the resulting control limits consistent with Shewhart's 3σ-guidance. By "valid" in this context is the implication that the estimator is equal to σ "on the average" in the long term, making it a so-called "unbiased" estimator of σ; there is more discussion on this matter on page 82 in the Technical Section. However, being correct on the average in the long term doesn't really say much about how close to σ we expect it to be in the short term, i.e. on the occasions when we actually use it! So my initial aim when writing this first of the two computer simulations was to find out how this estimator actually behaves in practice.

I therefore wrote a program to generate a large number of series of data from a normal distribution having σ = 1, and in each case calculate $\overline{MR} \div 1.128$. I chose to use baselines (series-lengths) of 12 and to generate 10 million such series. The program then drew a histogram of those 10 million estimates. That histogram is shown at the top left of page 67 and, as you can see, the values of the estimator varied from below 0.5 to above 2.0. That may well strike you as rather wider variation around the true value of σ = 1 than you might like. However, to substantially reduce that variation would require a much longer baseline. There are many arguments as to why that's undesirable. Some reasons are discussed in detail in Section 1 of Part F (pages 71–74). A further reason which is not raised there is that, unless data are arriving thick and fast (impossible with many processes), there will a considerable delay before the control chart can begin to be used. As yet a further objection, there's a law of decreasing returns in operation here: for example, to halve the width of this histogram would require *quadrupling* the baseline length to 48—wholly unsatisfactory for the numerous reasons just referred to. However, that whole issue is another matter. We are simply concerned at the moment with the Mathematical Statistician's claim that, because that 2.66 figure depends on the assumption of a normal distribution, this method is "not valid" if the data do not fit that assumption.

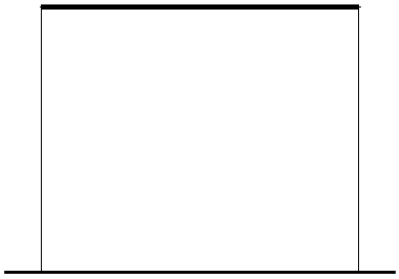
There's an obvious way to examine this (without raising the tricky matter of exact stability not being feasible!). And that is to modify the program so as to generate data from distributions other than the normal distribution and take a look at those resulting histograms. I used the same three distributions as in the simulation study on the Central Limit Theorem (pages 51–54): the uniform distribution, the unsymmetric triangular distribution and the exponential distribution, each being in a version with σ = 1 for ease of comparison with the initial normal version. You've seen pictures of these distributions on pages 52–53 in Part D but, for your convenience, they are illustrated again on the next page along with the normal distribution for comparison. The shapes of these three distributions are so different from each other (as well as from the normal distribution itself) that examining the behaviour of $\overline{MR} \div 1.128$ as an estimator of σ in such a variety of cases would appear to be a suitably severe test of its usefulness.

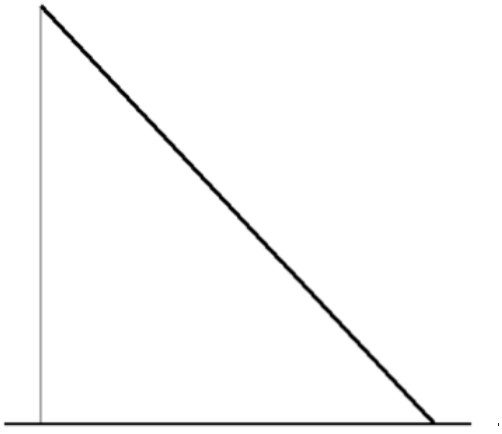
The resulting histograms are shown on page 67. Now, presumably if there *were* any truth in the notion that normality is *necessary* for the standard method of constructing control charts for one-at-a-time data to be useful, the top left histogram for the normal distribution should stand out as representing "good" behaviour while the others should be clearly "bad" in comparison. But, to put it mildly, such evidence is not easy to see! Despite its lack of symmetry, the triangular case appears to be virtually identical to the normal case while the uniform case is just a bit superior to the normal case! The exponential distribution's histogram is a little wider but, to be honest and seeing how incredibly different that shape of distribution is from the normal distribution, I was pleasantly surprised to see how *similar* to the others its histogram turned out to be. The overall conclusion from this simulation exercise surely has to be that the claim of data needing to be normally distributed in order to use the 2.66 $\overline{MR}$ method for constructing control charts lacks credibility.

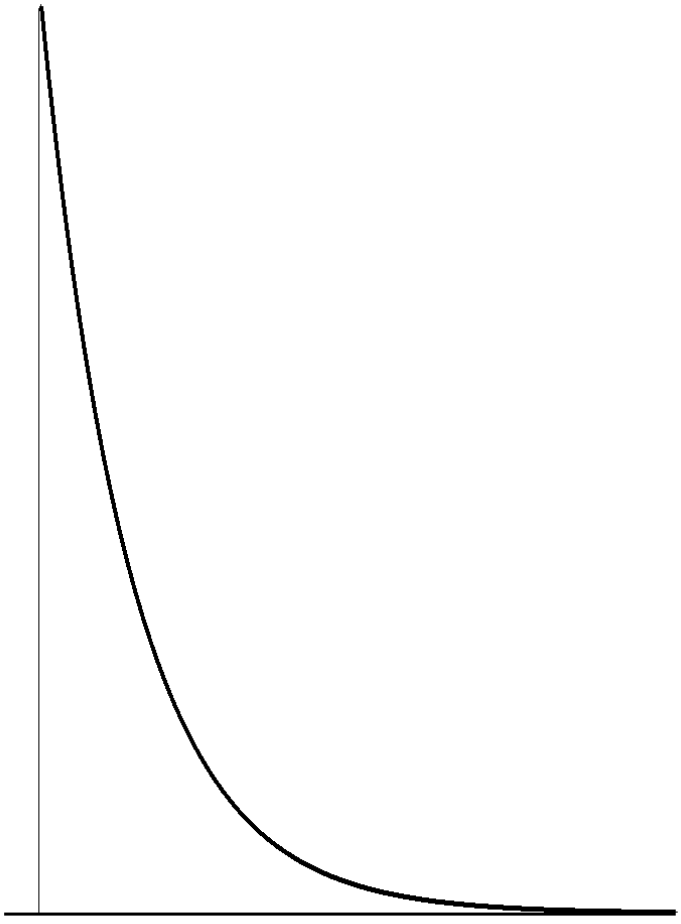
NORMAL



UNIFORM

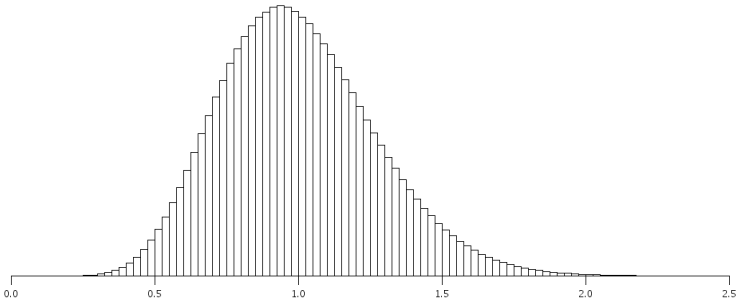




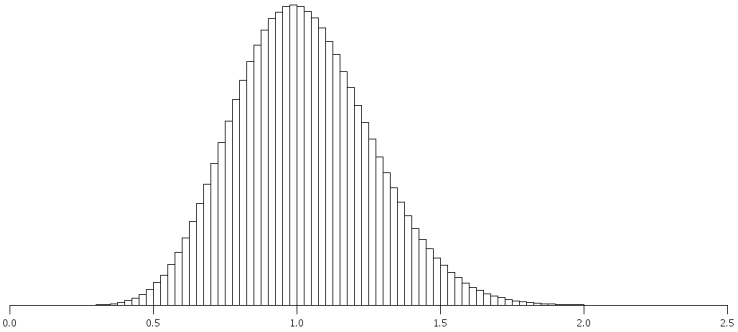


TRIANGULAR

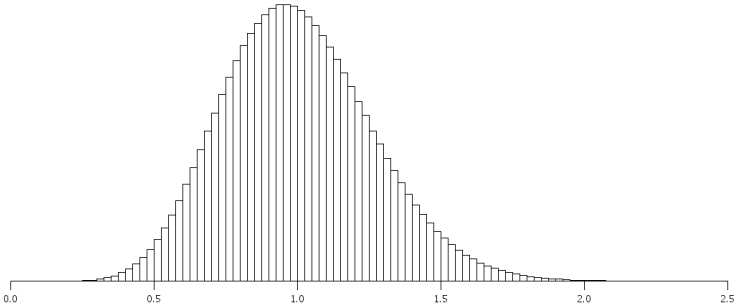
EXPONENTIAL



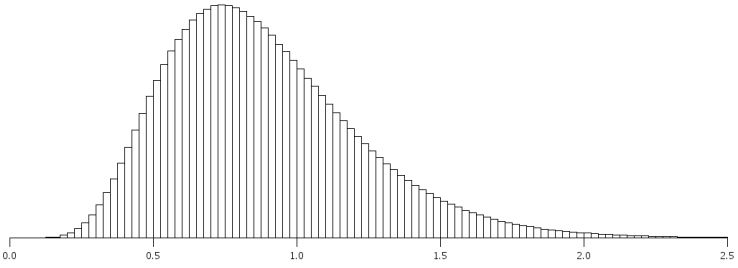
NORMAL



UNIFORM



TRIANGULAR



EXPONENTIAL

4. "If your data are normally distributed then the probability of a false signal is 0.0027"

There was a fairly full introduction to this simulation study on pages 63–64 and so I can now move straight into the details of the study.

The main question for me to consider was what size of subgroups should I try and how many subgroups should I use in calculating the control limits? Since you have now seen something of the *Japanese Control Chart* in Part B, I decided to be initially guided by what the Japanese workers did. Investigating their chart which, as you will recall, used subgroups of size $n = 4$ throughout, it appeared that they generally used between 10 and 15 subgroups to compute the limits. Therefore I chose to use 12 subgroups of size 4 in the simulation. Next, although $n = 4$ is quite a common choice of subgroup size, I decided to repeat the simulation twice: once with $n = 2$ and then with $n = 6$. (It is rare to find people using subgroups any larger than 6.) After a little trial and error I found that to use one million replications each time was sufficient to produce very clear pictures. The resulting histograms are shown on the next page, and I have clearly marked the 0.0027 probability on the horizontal axes.

So what was all this about normality of the data implying that the probability of a false signal ***is*** 0.0027?! The histograms are *very* spread out. As you can see, I chose a horizontal axis which stretches from zero probability right up to 0.025 (nearly ten times 0.0027) and even that was insufficient to cover all the results!

After seeing these histograms, I could imagine some statisticians complaining that to use only 12 subgroups for computing the limits was insufficient. I therefore repeated the whole simulation using 40 subgroups instead. 40 subgroups is a far larger number than most people use in practice. Those histograms are shown on page 70.

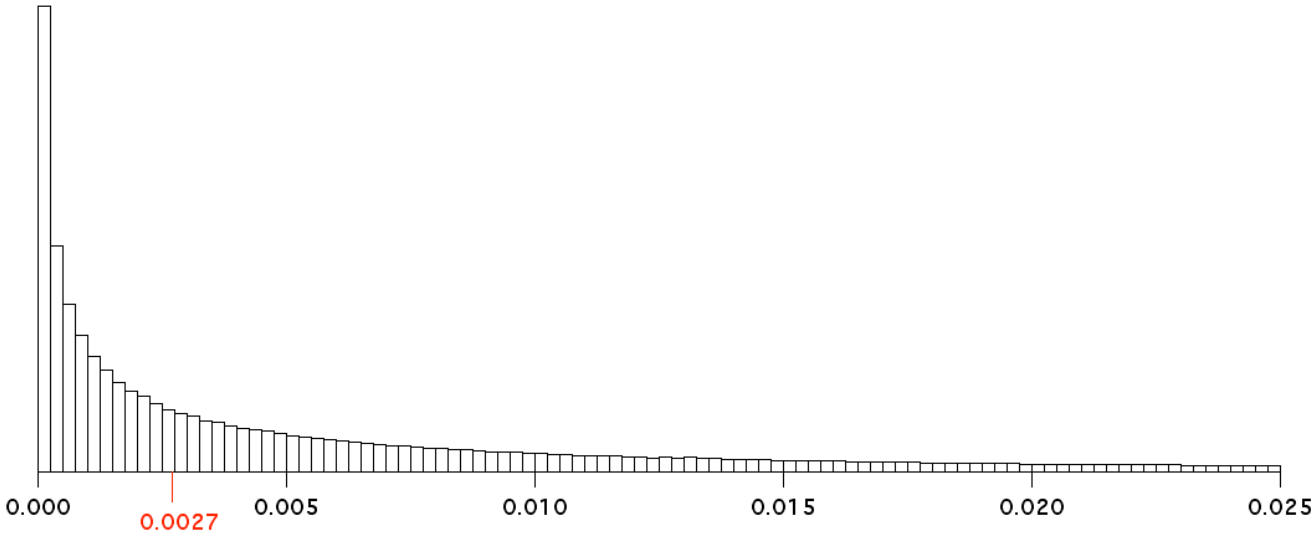
At least, with these latter histograms, one can see some evidence of the one thing we know about the values of the probability of a false signal. We can actually see that it is indeed feasible for the probability of a false signal to have a (very) long-term average of around 0.0027—admittedly not so with subgroups of size 2 but looking not unlikely with subgroups of size 4 or 6. However, this is already using far more data than is at all usual in practice. The disadvantages of using long baselines with subgrouped data are very much the same as with one-at-a-time data: recall that the latter are the subject of the first part of the Technical Section which follows on page 71.

Quite simply, if considering the probability of a false signal when using an $\bar{X}$-chart, even with 40 subgroups (way beyond what is usual), really all that can be said about that probability is that it is likely to be less than somewhere between 0.01 and 0.02 depending on the size of the subgroups.

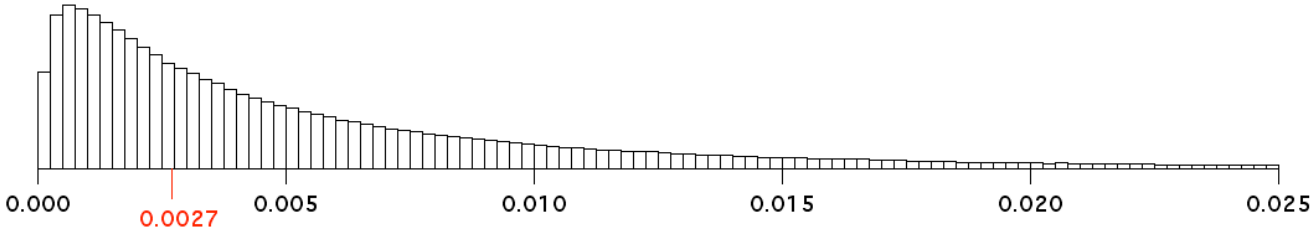
The nonsense of claiming that that probability ***is*** 0.0027 is surely plain for all to see.

The control chart as Shewhart created and developed it does *not* have "a fine ancestry of highbrow statistical theorems". But it *does* work.

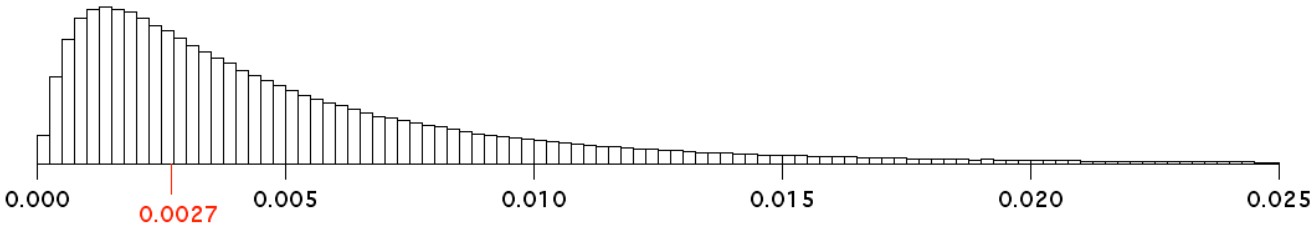
If you have read Balaji Reddie's "Contributions", particularly his pages 32–33 in "Some Lessons from History", you will know that some of the content here has been based on an article that I wrote around 20 years ago: it was titled *Two Superstations*. A subsequent article was titled *More Superstitions*. However, unlike what we have covered here, that further article involved not control charts but the topic known as "six-sigma" quality. Despite this, Balaji was very keen that I also make this article available to *12 Days to Deming* students since he had found it to be of particular interest to his own students and other contacts. If you also might be interested, you will find it beginning on Appendix page 43.



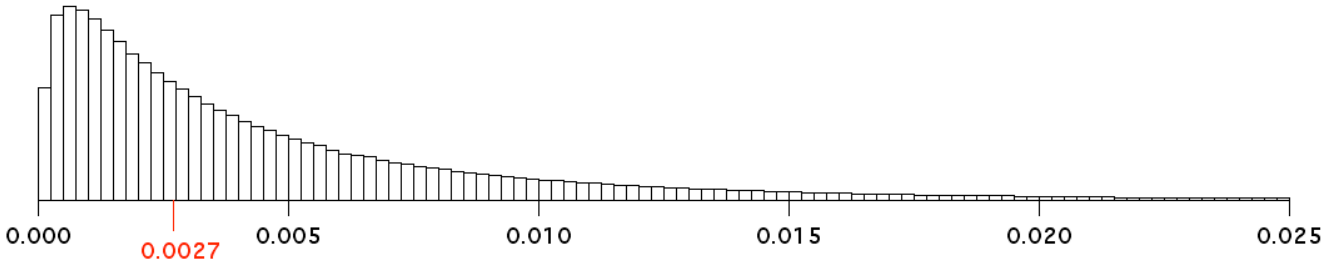
12 subgroups of size 2



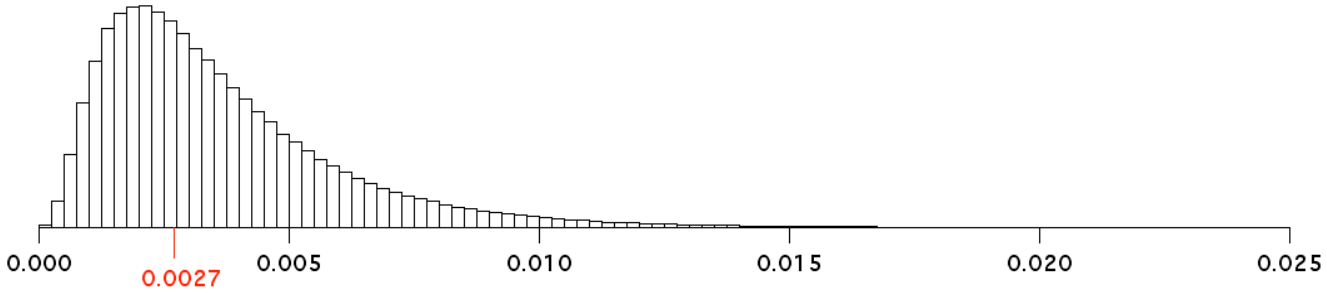
12 subgroups of size 4



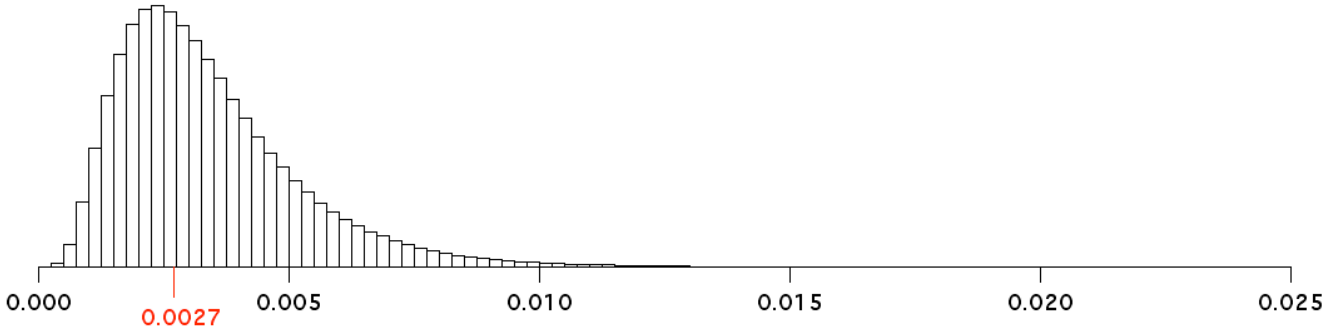
12 subgroups of size 6



40 subgroups of size 2



40 subgroups of size 4



40 subgroups of size 6

PART F: TECHNICAL SECTION

1. Length of the baseline

When introducing control charts (for one-at-a-time data) to delegates at my seminars, reactions were usually very positive, even from those who started out by saying such things as 'I can't do Statistics' or, worse still, telling me in advance that they hated the subject! A little while later there were instead expressions of relief, even surprise, when they discovered how straightforward the technique is, how relatively simple are the calculations involved and, before long, how they were able to interpret what the charts were telling them. As you might imagine, discussions on a set of processes such as those on Day 3 page 19 were exceedingly helpful for the latter. Further, the delegates could usually quickly understand the wisdom of basing the measurement of variation in an ongoing process on moving ranges, even those who were familiar with the standard deviation through some basic course on Statistics.

The one thing they often remained understandably uneasy about was the matter of choosing the baseline, i.e. the number of data to use for computing the control limits. The kind of guidance that I gave them might still not satisfy them—they might want to know the *reasons* for my guidance. It may well be that the same is true of you. If so then I hope this discussion will provide you with some thoughts and information that will be helpful to you when *you* are faced with deciding what length of baseline to try.

There was some brief discussion on Day 3 about the length of the baseline in Technical Aids 8 (page 17) and 9 (page 27). In practice, the choice of baseline length has to partly depend on how quickly the data are coming in and, if slowly, how soon you want to make at least a tentative start on the control chart. There are no "rules" on this matter. But for broad guidance I'd usually suggest, say, maybe 12 to 15 if the data are coming in fairly quickly (e.g. as in the Funnel Experiment), or if rather slower then perhaps around 10. Monthly data are of course rather a pain—perhaps initially just 5 or 6 there.

> In the case-study that I retold in *ST*, Don Wheeler did once use a baseline of length 4—but that was when there *were* only four data-points before the process was deliberately changed—what else could he do?!

If you have a conventional Statistics background, you may well be a little startled by how short my suggested baseline lengths are. The general sense in conventional Mathematical Statistics concerning sample sizes used when involved with methods of statistical inference (such as hypothesis testing and forming confidence intervals which were briefly introduced in the "crash-course" on pages 54–55) is that "the more data, the better". The conventional statistician may have carried out calculations on how many data are needed to estimate a parameter (such as the mean) to within a certain precision with some high degree of confidence—and come up with sample sizes in the hundreds if not the thousands! The same can happen in computations concerned with the "power" of hypothesis tests. As usual, other readers who are not familiar with these matters do *not* need to know about them! For here we are dealing with an entirely different kind of problem. In those traditional types of calculations it is effectively assumed that there is a virtually limitless pot of data available, totally unlike our basic situation here of data being generated (often rather slowly) over time. In those mathematical exercises, the object is to get a correct number. But heed well what Don Wheeler has repeated dozens of times in his own seminars: "We are not so much concerned with getting right numbers: we are concerned with taking right actions." And that is a very different—and much more important—matter.

If and when you have the time, you could gain some useful experience quite quickly about the pros and cons of using different lengths of baseline by carrying out some experimentation on the Funnel Experiment data that you generated in Major Activity 3–h. Try using some different baselines from whatever you chose in Part A of these Optional Extras, and see what happens. That is, examine the ways in which different baseline lengths may affect your judgment of what is happening with the four Rules, and when.

As you know by now, I have always been quite keen on using computer simulation studies when faced with problems that are difficult or impossible to solve just by Mathematics: simulation studies do not "solve" problems but they can throw light on them. On pages 82–84 in the second edition of my book of *Statistics Tables* (which I've been abbreviating by *ST*), I included details of a simulation study that I carried out to help me get some "feel" for the pros and cons of different baseline lengths used in computing control limits for the usual type of control chart using one-at-a-time data. I'll discuss below some of the results which came from that simulation study.

However, I should first point out that simulation studies do have some similar drawbacks to mathematical solutions, in particular that it is necessary to make some choices and assumptions in the details to be used in the design of these studies—just as that is necessary in order to carry out mathematical derivations. (So yes, e.g. I confess that my simulation study did involve generating data from normal distributions—which were introduced in the "crash-course"). Of course, as with mathematical derivations, there is no expectation that the results obtained will *exactly* reflect what will happen in practice when such choices and assumptions do not hold. However, with choices and assumptions that are made with an eye on the kind of things which might be expected to *approximately* reflect practical situations, it is reasonable to hope that the main results from both Mathematics and computer simulations will at least roughly indicate the general lines of what will happen in practice—otherwise, of course, they're not much use! The assumption that I needed to make in this study was that the process remained in statistical control throughout the baseline period. However, I shall also briefly discuss the situation where this assumption does not hold.

Firstly I'll look at the possibility of "false alarms", i.e. signals that a special cause exists when in fact the process has remained stable. False alarms can be costly. A signal, i.e. a point outside the control limits, is a signal that guides you start looking for a special cause. And that can be time-consuming and expensive—and even more so if there is no special cause to find, and then perhaps even kid ourselves that we've found one! So here's the first of two rather similar questions for the simulation study to tackle:

(A) How often will false alarms occur when the control chart is being used "live"?

With the assumptions made in the simulation study, and with the variety of baseline lengths as shown, here are the percentages of false alarms (i.e. signals that occur *if the process remains stable*) when the control chart is "up and running" (i.e. after the baseline period ends):

Baseline length	4	6	10	15	20	30	50	100
Prob of any false alarm(s)	7.65%	4.28%	2.17%	1.34%	0.99%	0.68%	0.48%	0.35%

As is immediately obvious, that percentage is high for short baselines but improves as the baseline length increases. This, of course, coincides with the conventional statistician's almost automatic expectation—and for similar reasons. The longer the baseline, i.e. the greater the number of data from which the control limits are computed, the closer they will be to those which would have been computed directly from the normal distribution that is being assumed, and so the better they will "fit" the data that are being generated. So that's no surprise, and I'll therefore move straight on to the next question.

(B) What is the likelihood of there being any signals (i.e. false alarms) *during* the baseline period?

Let's look straightaway at results from the simulation study:

Baseline length	4	6	10	15	20	30	50	100
Prob of any false alarm(s)	0	1.3%	2.1%	3.5%	4.8%	7.4%	12.2%	23.3%

So yes, it is possible to have signals *within* the baseline period. Indeed, if you read Part A of these Optional Extras then you will have seen some on page 13 when considering Rule 4 of the Funnel. But, of course, those were *justified* signals: the process was already out of control. And fortunately that is almost always

the case if you get a signal within the baseline. But not *always*, as the short table of probabilities shows. So, if you are doubtful about whether a signal is or is not a false alarm, you might be sensible to wait and recompute the control limits after you have recorded, say, another two or three data, and then check again. That's especially the case if you are using a particularly short baseline. My reason for that advice is that a false alarm *within the baseline* is liable to be even more costly than a false alarm *after* the baseline period. For, of course, the control limits are supposed to guide us about *when* the process goes out of control—but now we have an indication that the process is out of control already! So not only would time and money be fruitlessly spent on searching for the reason for that signal—it would be illogical to even continue using this control chart.

As you can see from the short table of probabilities, the probability of one or more false alarms occurring during the baseline steadily *rises* as the baseline length increases. Because of the problems that such false alarms cause, this immediately indicates that it is unwise to use a very long baseline. Seeing that, as just pointed out, it would be illogical to extend the control limits beyond the baseline and continue to use the chart if there *are* already any points outside the limits during the baseline, my computer program was written to reflect common practice by *only* including those cases where the chart was clear of signals during the baseline for the purposes of answering Questions **(A)** and also **(C)** below.

But why *does* the probability of false alarms occurring within the baseline behave in the way shown? In particular, why is the probability actually *zero* with that really short baseline of 4? Remembering that the control limits are calculated directly from the data that are within the baseline (using that familiar method using the 2.66), it turns out to be arithmetically *impossible* for any of the four values to lie outside those control limits. If you like, try it for yourself with *any* set of four numbers that you care to choose: never mind how weird a selection of numbers you'd like to dream up, you'll find that that remains true! But it does become *just* possible with a baseline length of 5 and then, as you've seen, ever more possible with longer baselines. Why is that? One obvious reason is that, the longer the baseline, the more opportunities there are for a false alarm to occur within it.

So, in summary, we had evidence with Question **(A)** that to use very *short* baselines is dangerous, and now we have evidence with Question **(B)** that to use very *long* baselines is also dangerous! Such conflicting evidence is not unexpected: there are conflicting interests in play here and so the conclusion is that we shall have to finish up with some kind of compromise between them. But what will guide our choice of such a compromise?

Our third question may help. Here, instead of false alarms, we focus on *justified* signals, i.e. points which fall outside the control limits *after* the process has *changed*. Then, obviously, we would like to get a signal without much delay so that the search for a special cause is (correctly) triggered sooner rather than later. To examine this we'll use a traditional method for describing the sensitivity of a control chart which is to compute the average number of data occurring after a process change up to and including when the chart gives its first signal; this is known as the Average Run Length (ARL). So our third question is:

(C) How do the Average Run Lengths behave for different lengths of baseline?

Remembering that we are generating data from a normal distribution, I considered three cases of a sudden process change: a shift in the process average by an amount of σ, 2σ or 3σ, σ being the standard deviation of that normal distribution. Whereas the answers to the first two questions might not have surprised you, these results may do so:

Baseline length	4	6	10	15	20	30	50	100
ARL following shift σ	7.1	10.1	14.9	19.6	23.1	28.0	33.8	39.7
ARL following shift 2σ	3.21	3.73	4.35	4.83	5.14	5.53	5.91	6.25
ARL following shift 3σ	1.888	1.949	1.992	2.016	2.028	2.040	2.049	2.055

So the ARLs steadily *increase (worsen)* as the baseline gets longer. This might indeed initially appear rather surprising since the conventional statistician's "obvious" argument is that, the more data we use to derive the control limits, the more "accurate" and therefore "better" the chart will be. But that argument reflects the Mathematical Statistician's almost automatic way of thinking rather than considering what actually happens in practice. Remember that, in line with common and sensible practice, one does not usually continue with some newly-computed control limits if there are any points outside those limits *during that baseline*. As already argued, what would be the logic of extending the control limits into the future if we have been sent a signal that the process is likely to be out of control already? So, reflecting that "common and sensible practice", recall that in the computer simulation I did not continue with any such cases as regards answering Questions **(A)** and **(C)**.

Let's consider in more detail what happens if we are using a long baseline. Clearly, the longer the baseline, the greater is the chance that a signal will occur during it—as has already been pointed out, there are then simply more opportunities for it to do so (whether or not the process stays in control). But if the process *does* actually stay in control throughout the baseline then that signal is, of course, a false alarm. We've seen that the results for Question **(B)** confirm the increasing chance of such a false alarm with longer baselines: in fact, except for the very shortest baselines, those percentages increase by about 0.25% for each extra data-point included in the baseline. So, as examples, if the baseline-length is 40 then there is about a 1-in-10 chance of there being such a false alarm, while if the baseline-length is 100 then the chance rises to almost 1 in 4. Now, naturally the control limits will vary when computed from different sets of baseline data: with some data-sets the gap between the limits will be "fairly typical", but with others the gap will be either relatively narrow or relatively wide. So in which cases are those false alarms within the baseline (i.e. the cases which will *not* be considered in Question **(C)**) most likely to occur? Surely it will be those where the gap between the control limits is relatively *narrow.* But those discontinued cases are the very ones that would have been the most likely to produce signals when the process goes *out* of control! *That's* the combination of theory and practice which explains why the ARL increases as the baseline gets longer.

Thus, whereas Questions **(A)** and **(B)** produced evidence in favour of longer or shorter baselines respectively, I suggest that the evidence in Question **(C)** provides a valid casting vote! *Very* short baselines need care because of the evidence in **(A)**, but otherwise the weight of evidence points to the wisdom of using reasonably short baselines rather than the longer ones that may appeal to a conventional statistician.

Finally, recall where we saw our first control chart: it was in the Red Beads Experiment. The control limits there were computed from all 24 of the available data. So why don't we forget all this stuff about baseline lengths and use "all of the available data" in "live" control charting? Of course, that would mean recomputing the control limits every time a new data-value arrives. That would be tedious manually, but a computer could easily do it for us. The reason we don't is, believe it or not, that this is liable to actually *reduce* the ability of the chart to produce useful information. As evidence of that, I'll finish here with the following sad memory:

> This recalls the occasion when a delegate came up to me with tears of gratitude after I had emphasised this point in the seminar. Her company had installed some quality-control software on their computers. This software had been written to use "all the available data" in the manner I have just described. I repeat that It is, of course, very easy and not at all tedious for a computer to immediately recompute the control limits after each new reading arrives. The trouble was that the lady's process was slowly trending upward (which was a very undesirable state of affairs with her process) but the consequence of using that software was that the control limits *also* kept trending upward—and so she never obtained a value above the (current up-to-date) Upper Control Limit because it also kept moving further up and away! Her boss refused to consider any action unless the chart produced an officially out-of-control signal. So it was my poor delegate who was increasingly being blamed for the worsening behaviour of this process.

On Day 7, Deming cited "We installed quality control" as an "Obstacle to the Transformation". There's always more to learn!

2. "Expected" values—the great misnomer!

NB Some of the mathematical material on this topic and others in this Technical Section may be seen as overly basic by those who are experienced in Mathematical Statistics. But remember that this material is all written for a quite general audience—so feel free to skim over any of it that is already familiar to you.

Some more notation (mathematical shorthand)

Let's start by revisiting the development of the two illustrations on pages 37–42 in Part D of (a) tossing two coins and counting the number of Heads and of (b) throwing three dice and counting the number of sixes. I said there that these were simple examples of what are known as *binomial* distributions. In the final section of these Optional Extras I shall develop some main ideas about the whole family of binomial distributions.

We also saw that, with the usual symmetry assumptions (i.e. a 50-50 chance of Head or Tail at each toss of a coin and a 1 in 6 chance of getting a six when a die in thrown), we finished up with the following probability distributions:

$$\begin{aligned} &\text{Probability of no Heads} &&= \tfrac{1}{4} \\ &\text{Probability of 1 Head and 1 Tail} &&= \tfrac{1}{2} \\ &\text{Probability of 2 Heads} &&= \tfrac{1}{4} \end{aligned}$$

and

$$\begin{aligned} &\text{Probability of 0 sixes} = \tfrac{125}{216} && \text{Probability of 1 six} = \tfrac{75}{216} \\ &\text{Probability of 2 sixes} = \tfrac{15}{216} && \text{Probability of 3 sixes} = \tfrac{1}{216}. \end{aligned}$$

We then moved on to expressing the mean μ and the variance σ^2 of the first of these distributions as

$$\mu = 0 \times \tfrac{1}{4} + 1 \times \tfrac{1}{2} + 2 \times \tfrac{1}{4} = 1$$

and

$$\sigma^2 = (0-1)^2 \times \tfrac{1}{4} + (1-1)^2 \times \tfrac{1}{2} + (2-1)^2 \times \tfrac{1}{4} = \tfrac{1}{2}.$$

If you then carried out the voluntary exercise on page 42 of computing the mean and variance of the second distribution, this is what you should have obtained:

$$\mu = 0 \times \tfrac{125}{216} + 1 \times \tfrac{75}{216} + 2 \times \tfrac{15}{216} + 3 \times \tfrac{1}{216} = \tfrac{1}{2}$$

and

$$\sigma^2 = (0-\tfrac{1}{2})^2 \times \tfrac{125}{216} + (1-\tfrac{1}{2})^2 \times \tfrac{75}{216} + (2-\tfrac{1}{2})^2 \times \tfrac{15}{216} + (3-\tfrac{1}{2})^2 \times \tfrac{1}{216}$$

$$= (-\tfrac{1}{2})^2 \times \tfrac{125}{216} + (\tfrac{1}{2})^2 \times \tfrac{75}{216} + (\tfrac{3}{2})^2 \times \tfrac{15}{216} + (\tfrac{5}{2})^2 \times \tfrac{1}{216}$$

which, after some careful arithmetic, comes out as $\sigma^2 = \tfrac{5}{12}$.

If we denote the random variable concerned in each case by X then we have been calculating μ and σ^2 by

$$\mu = \text{the sum of all values of } \{x \text{ multiplied by the probability that } X = x\}$$

and

$$\sigma^2 = \text{the sum of all values of } \{(x-\mu)^2 \text{ multiplied by the probability that } X = x\}.$$

As you might suspect, there exists some mathematical shorthand for such expressions. Typically it's

$$\mu = \sum x.\text{Prob}(X = x) \quad \text{and} \quad \sigma^2 = \sum \left(x - \mu\right)^2.\text{Prob}(X = x).$$

Obviously enough, there I have abbreviated "the probability of" by "Prob( )". And, as you can see, "Σ" is shorthand for "the sum of all values of". (Confusingly, Σ is actually the *capital* Greek letter "sigma"!) Further, because "x" is the mathematician's favourite letter and so is likely to appear in many such expressions, in order to avoid yet further confusion I am now abbreviating "multiplied by" by just a simple dot rather than the usual multiplication sign which, of course, looks very much like the letter x! Indeed, often the dot is omitted as long as that doesn't cause any ambiguity. All of this is very common notation in the books, and I have introduced it here since we shall need it in some of what follows.

"Expected" values

As you know, a concept that we have used quite a lot in Parts C and D is that of what happens "in the long term" when we consider random samples getting larger and larger, in particular "long-term average values" and "long-term proportions". We developed the idea of the *probability* of an event as being the long-term proportion of times that the specified event occurs when considering random samples of, say, trials of some operation or procedure. Or if we measure or count something—let's denote it by X (again, the mathematician's favourite letter!)—then the long-term average $\bar{X}$ of the recorded values of X gives us the "true mean" μ of X's probability distribution which, as we have now seen, can be expressed as

$$\mu = \sum x.\text{Prob}(X = x).$$

And then we also had the variance σ^2 expressed as $\sum \left(x - \mu\right)^2.\text{Prob}(X = x)$.

We shall make further use of this concept of long-term average values in this Technical Section. But "long-term average value" is also rather a mouthful. And so, not surprisingly, there's also a standard notation (shorthand) for that! The notation is "E[ ]". So, for example, μ can now be expressed as $\text{E}[X]$. Similarly, σ^2 can be expressed as $\text{E}[(X-\mu)^2]$. This concept can be generalised to any so-called *function* of X, g(X) say, such as $g(X) = X^2$ or $g(X) = (X+1)(X+2)$—remember the latter means $(X+1)$ *multiplied by* $(X+2)$. A "function" of X simply means anything that can be evaluated once the value of X is known. So then

$$\text{E}[g(X)] = \sum g(x).\text{Prob}(X = x).$$

Unfortunately, that E[ ] notation is an abbreviation for surely one of the biggest misnomers in the whole subject of Statistics! It stands for "Expectation" or "Expected Value". What's unfortunate about that is that $\text{E}[X]$ is very often either a very rare value or sometimes an *impossible* value of X, so it seems rather odd to then call it the *expected* value of X! The same unfortunate fact is also true for the Expected Value of most functions g(X) of X.

With a discrete distribution, sometimes $\text{E}[X]$ *is* a possible value of X. For example, when we considered tossing two coins and counting the number of Heads, then μ, i.e. $\text{E}[X]$ (the long-term average value of X), turned out to be equal to 1. However, when we then moved on to considering the number of sixes when three dice were thrown, we had $\text{E}[X]$ as being equal to $\frac{1}{2}$—hardly an "expected" value of a number which can actually only be 0, 1, 2 or 3! And if a random variable X has a continuous probability distribution then we reached the conclusion in Part D that the probability of X being equal to *any* specified value is zero—so, again, whatever μ turns out to be, it will hardly be an "expected" value of X!

However, the language and the notation of expected values is so common—pretty much universal—that we are stuck with it. So let's get on and use it. As you will soon see, expected values (long-term average

values) are very helpful in explaining some of the mysteries that have emerged both in the main text and earlier in these Optional Extras.

Combining expected values

A straightforward yet vital property of expected values is that they are *additive*. If we express this property in terms of "long-term averages" you will soon see how straightforward it is. Suppose we are considering the *sum* of two random variables: $X + Y$. Then surely the long-term average value of $X + Y$ must be equal to the long-term average value of X plus the long-term average value of Y. Think about it—how could it be otherwise? So simple, yet so incredibly useful: $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$. And it's not restricted to adding just two random variables together: the same thought-process works for the sum of 3 or 4 or *any* number of random variables—or, indeed, *functions* of random variables! It also includes subtraction rather than just addition: e.g. $\mathrm{E}[X-Y] = \mathrm{E}[X] - \mathrm{E}[Y]$.

So, if expected values can be *added* together, perhaps a natural follow-on question is whether they can be *multiplied* together. That's not quite so straightforward: the answer is sometimes Yes and sometimes No. But, as an interim step, what happens if we want to find the expected value of some *multiple* of a random variable, i.e. what is $\mathrm{E}[cX]$ where c is some constant value, e.g. 2? In this case there is no difficulty. Again thinking of long-term average values, it is surely obvious that if we *double* all of our sampled values of X then we also *double* their average. Actually, we could have used the additivity property to verify this particular case, i.e. $\mathrm{E}[2X] = \mathrm{E}[X+X] = \mathrm{E}[X] + \mathrm{E}[X] = 2\mathrm{E}[X]$. But the general result is true as well: $\mathrm{E}[cX]$ *is* equal to $c\,\mathrm{E}[X]$ for *any* value of c.

But what about the expected value of the *product* of two random variables? ("Product" is the jargon for *multiplying* two or more things together.) So is $\mathrm{E}[XY]$ equal to $\mathrm{E}[X].\mathrm{E}[Y]$? In order to answer that question we need to introduce yet more statistical terminology, but at least in this case the statistical interpretations of the words used are pretty consistent with their ordinary English meanings. It is the concept of the *independence* or otherwise of two (or more) events. "Independence" of two events simply implies that the occurrence or non-occurrence of either event has absolutely no influence on the occurrence or nonoccurrence of the other one. The statistical definition of independence of two events A and B is simply that the probability of both events occurring is equal to the product of their individual probabilities, i.e. in shorthand, $\mathrm{Prob}(\mathrm{A}\ \textit{and}\ \mathrm{B}) = \mathrm{Prob}(\mathrm{A}).\mathrm{Prob}(\mathrm{B})$; this is often called the "simple multiplication rule".

Now, the fact that the statistical definition of independence of events is true of events that are truly *independent* of each other in the ordinary English sense of the word is so obvious that we have already used it several times in this material without comment! An early instance was in Part D near the bottom of page 41 where, having assumed that there was a one-in-six chance of a six showing when a die was thrown, we immediately deduced the probability distribution of the number of sixes showing when *three* dice were thrown. The first probability calculated there was of all three dice showing a six:

$$\begin{aligned}\mathrm{Prob}(3\text{ sixes}) &= \mathrm{Prob}(\text{six on the first die } \textit{and} \text{ six on the second die } \textit{and} \text{ six on the third die})\\ &= \mathrm{Prob}(\text{six on the first die}).\mathrm{Prob}(\text{six on the second die}).\mathrm{Prob}(\text{six on the third die})\\ &= \tfrac{1}{6}\times\tfrac{1}{6}\times\tfrac{1}{6} = \tfrac{1}{216}.\end{aligned}$$

In other words, we were (quite reasonably!) *assuming* that the three events "six on the first die" and "six on the second die" and "six on the third die" were all independent of each other and thus implicitly using the multiplication rule for independent events.

But back to the original question: is the *expected value* of the product of two (or more) random variables equal to the product of their individual expected values? I.e., in the case of two random variables, X and Y,

is $E[XY]$ equal to $E[X].E[Y]$? As you might expect from the build-up, (a) there is an entirely analogous concept of the independence of *random variables* to the above concept of independence of *events*, and (b) the answer to the question about the expected value of a product of random variables is Yes *if* the random variables are independent (but otherwise No, except by an occasional fluke). If that seems obvious to you then you can move straight on to this page's final short paragraph. Otherwise, I'll sketch a proof for you.

In order to prove this important multiplication rule for expected values, i.e. that if X and Y are *independent* random variables then $E[XY] = E[X].E[Y]$, it's easiest to first verify it for a couple of very simple discrete random variables. This will show a pattern that is then easily extended to more general cases (although rather lengthy to write down).

So let's suppose X can take on just two possible values x_1 and x_2 with probabilities p_1 and p_2 respectively, and similarly Y can take on just two possible values y_1 and y_2 with probabilities q_1 and q_2 respectively. Then, summing over all possible outcomes, we have

$$\begin{aligned} E[XY] &= \sum xy.\,\text{Prob}\,(X = x \text{ and } Y = y) \\ &= x_1 y_1.\,\text{Prob}\,(X = x_1 \text{ and } Y = y_1) + x_1 y_2.\,\text{Prob}\,(X = x_1 \text{ and } Y = y_2) \\ &+ x_2 y_1.\,\text{Prob}\,(X = x_2 \text{ and } Y = y_1) + x_2 y_2.\,\text{Prob}\,(X = x_2 \text{ and } Y = y_2) \end{aligned}$$

which, if X and Y are *independent* random variables, can be rewritten

$$\begin{aligned} & x_1 y_1.\,\text{Prob}\,(X = x_1).\,\text{Prob}\,(Y = y_1) + x_1 y_2.\,\text{Prob}\,(X = x_1).\,\text{Prob}\,(Y = y_2) \\ & + x_2 y_1.\,\text{Prob}\,(X = x_2).\,\text{Prob}\,(Y = y_1) + x_2 y_2.\,\text{Prob}\,(X = x_2).\,\text{Prob}\,(Y = y_2) \end{aligned}$$

which can then be cleverly rewritten as

$$\left\{ (x_1.\,\text{Prob}\,(X = x_1) + x_2.\,\text{Prob}\,(X = x_2) \right\}.\left\{ (y_1.\,\text{Prob}\,(Y = y_1) + y_2.\,\text{Prob}\,(Y = y_2) \right\}.$$

Multiply out all the terms in that expression if you don't believe me! And this is indeed equal to $E[X].E[Y]$. As I said, that pattern of proof can be extended to *any* discrete probability distributions.

The same result is true of *continuous* random variables. But, of course, then we cannot use that same method of verification. A similar approach to the proof *can* be developed but only in terms of the branch of Mathematics known as Calculus. So, as I do not intend to also attempt to provide you with a crash-course in Calculus, I'm afraid you'll just have to trust me!

3. Proofs and uses of expected values

There are quite a few little results and a couple of big ones in this section. The proof of later results often depends on one or more of the results proved earlier. So, to keep track, it will be wise for me to identify the results in a way that will enable you to easily trace back. I shall do that in red print on the right-hand side of the page.

To begin with, let's restate the results about expected values derived in the previous section. First we had the additivity property. We started with the simplest result that, for any random variables X and Y, we had $E[X+Y] = E[X]+E[Y]$. I then pointed out that this also works with subtraction and can be extended to more than two random variables and even to functions of them. Here is a selection of simple additivity results:

$$E[X+Y] = E[X] + E[Y]; \quad E[X-Y] = E[X] - E[Y];$$
$$E[X^2+Y^2] = E[X^2] + E[Y^2]; \quad E[X+Y+Z+\dots] = E[X] + E[Y] + E[Z] + \dots \qquad \{E1\}$$

It is worth emphasising that these results are entirely general: they apply whether *or not* the random variables are independent.

Then we had the simple and pretty obvious result about a multiple of a random variable:

$$\text{For any (constant) value c, } E[cX] = cE[X]. \qquad \{E2\}$$

Finally, there was the important multiplication rule:

$$\text{If X and Y are } \textit{independent} \text{ random variables then } E[XY] = E[X].E[Y]. \qquad \{E3\}$$

Similarly to the additivity property, this is also extendable to more than two independent random variables.

Next we shall derive some results about *variances*. As you will recall, the variance σ^2 of a random variable X is defined as $E[(X-\mu)^2]$ where $\mu = E[X]$ (and remember that σ itself is the *standard deviation*). Firstly, we'll find a useful alternative expression for σ^2.

We have $\sigma^2 = E[(X-\mu)^2]$ which, multiplying out $(X-\mu)^2$, gives

$$\sigma^2 = E[X^2 + \mu^2 - 2\mu X] .$$

Using both {E1} and {E2} and the obvious fact that the expected value of a constant is simply that constant, this gives

$$\sigma^2 = E[X^2] + E[\mu^2] - 2E[\mu X] = E[X^2] + \mu^2 - 2\mu.E[X] ,$$
$$\text{i.e. } \sigma^2 = E[X^2] + \mu^2 - 2\mu^2, \text{ which gives } \sigma^2 = E[X^2] - \mu^2. \qquad \{V1\}$$

Secondly, let's consider the variance of a multiple cX of X. From {E2} we know that $E[cX] = cE[X]$. So, using {V1},

$$\text{the variance of } cX = E[(cX])]^2 - (E[cX])^2 = c^2E[X^2] - c^2(E[X])^2 = c^2(E[X^2] - c^2\mu^2$$
$$= c^2(E[X^2] - \mu^2) = c^2\sigma^2 . \qquad \{V2\}$$

This is, of course, consistent with the standard deviation being a measure of variability since, as you would expect, it gives the standard deviation of cX as $c\sigma$.

Next, variances also have an additivity property. However, it isn't as all-embracing as {E1}: in general, it applies only to *independent* random variables. So if X and Y *are* independent random variables then, using all of {E1}, {E2}, {E3} and {V1} we have

$$\begin{aligned}\text{the variance of } (X+Y) &= \mathrm{E}\left[(X+Y)^2\right] - \left(\mathrm{E}[X+Y]\right)^2 \\ &= \mathrm{E}\left[X^2+Y^2+2XY\right] - \left(\mathrm{E}[X]+\mathrm{E}[Y]\right)^2 \\ &= \mathrm{E}[X^2] + \mathrm{E}[Y^2] + 2\mathrm{E}[XY] \; - \left\{\mathrm{E}[X]^2 + \mathrm{E}[Y]^2 + 2\mathrm{E}[X]\mathrm{E}[Y]\right\} \\ &\quad \text{which, since } X \text{ and } Y \text{ are independent,} \\ &= \mathrm{E}[X^2] + \mathrm{E}[Y^2] + 2\mathrm{E}[X]\mathrm{E}[Y]) \; - \left\{\mathrm{E}[X]^2 + \mathrm{E}[Y]^2 + 2\mathrm{E}[X]\mathrm{E}[Y]\right\} \\ &= \mathrm{E}[X^2] + \mathrm{E}[Y^2] + 2\mathrm{E}[X]\mathrm{E}[Y]) \; - \mathrm{E}[X]^2 - \mathrm{E}[Y]^2 - 2\mathrm{E}[X]\mathrm{E}[Y] \\ &= \left(\mathrm{E}[X^2] - \mathrm{E}[X]^2\right) + \left(\mathrm{E}[Y^2] - \mathrm{E}[Y]^2\right)\end{aligned}$$

i.e. the variance of $(X+Y)$ = the variance of X + the variance of Y. {V3}

Similarly to {E1}, this additivity property can be extended to three or more independent random variables.

The purpose of producing all these relatively small results is to be able to prove some *big* results that have been quoted and used previously. The first of these provides the variance of the mean $\bar{X}$ of a random sample of n values of the random variable X. As usual, we'll denote the mean and variance of X by μ and σ^2.

Previously we have regarded the fact that $\mathrm{E}[\bar{X}] = \mu$ as obvious directly through our considerations of long-term sampling. We could now instead effectively prove it using {E1}, {E2} and {E3}:

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\tfrac{1}{n}\textstyle\sum X\right] = \tfrac{1}{n}\mathrm{E}\left[\textstyle\sum X\right] = \tfrac{1}{n}\textstyle\sum \mathrm{E}[X] = \tfrac{1}{n}\textstyle\sum \mu.$$

Remembering that the sum is over n terms, this is therefore simply $\frac{1}{n} n\mu = \mu$.

However, rather than using the shorthand summation symbol Σ, I suspect that such lines of mathematics might be more easily understood by reverting to longhand! So let's denote our sample by $X_1, X_2, \ldots, X_n$ and rewrite the above as

$$\begin{aligned}\mathrm{E}[\bar{X}] = \mathrm{E}\left[\tfrac{1}{n}(X_1 + X_2 + \ldots + X_n)\right] = \tfrac{1}{n}\mathrm{E}\left[X_1 + X_2 + \ldots + X_n\right] &= \tfrac{1}{n}\left(\mathrm{E}[X_1] + \mathrm{E}[X_2] + \ldots + \mathrm{E}[X_n]\right) \\ &= \tfrac{1}{n}\left(\mu + \mu + \ldots + \mu\right) = \tfrac{1}{n} n\mu = \mu. \qquad \{E4\}\end{aligned}$$

Now let's move on to the variance of $\bar{X}$. Having obtained these recent results, this now turns out to be very easy to find by using {V2} and {V3}. Abbreviating "the variance of" by "var", and remembering that X_1, X_2, ..., X_n are all independent random variables having variance σ^2, we have

$$\begin{aligned}\mathrm{var}(\bar{X}) = \mathrm{var}\left(\tfrac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) &= \tfrac{1}{n^2}\,\mathrm{var}(X_1 + X_2 + \ldots + X_n) \\ &= \tfrac{1}{n^2}\left\{\mathrm{var}(X_1) + \mathrm{var}(X_2) + \ldots + \mathrm{var}(X_n)\right\} \\ &= \tfrac{1}{n^2}\left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right) = \tfrac{1}{n^2} n\sigma^2.\end{aligned}$$

Thus we have proved the very important result (as used, in particular, in the statement of the Central Limit Theorem in terms of Z on page 51) that

the variance of $\bar{X}$ is $\frac{1}{n}\sigma^2$. {V4}

Finally as far as variances are concerned, let's return to that mystery of why, given a random sample of size n, the sample variance s^2 was defined with the divisor $n-1$ rather than n as might have been expected, i.e.

$$s^2 = \frac{1}{n-1}\sum (X-\bar{X})^2.$$

Let's now see why it is defined that way.

Clearly, if we knew the value of μ, it would have made sense to use it in the expression for the sample variance rather than $\bar{X}$. $\bar{X}$ is there because, in general, we *wouldn't* know the value of μ. Now, the role that we want s^2 to play is as an *estimator* of the unknown value of σ^2, in the same way that we use $\bar{X}$ as an *estimator* of the unknown value of μ. The difference between them is that, whereas with the "obvious" estimator of μ, i.e. $\bar{X}$, we know it is true that $\mathrm{E}[\bar{X}] = \mu$, the expected value of the "obvious" estimator of σ^2 turns out to be slightly smaller than σ^2. Very annoying! What *would* have been true is that, in the unlikely event that we *knew* the value of μ and therefore were able to use it when estimating σ^2, all would have been well: yes, the expected value of the "obvious" estimator of σ^2 in *those* unlikely circumstances would indeed have been σ^2. That's very simple to verify, so let's do that first.

$$\begin{aligned}
\mathrm{E}\left[\tfrac{1}{n}\textstyle\sum (X-\mu)^2\right] &= \tfrac{1}{n}\mathrm{E}\left[\textstyle\sum (X-\mu)^2\right] = \tfrac{1}{n}\mathrm{E}\left[(X_1-\mu)^2 + (X_2-\mu)^2 + \ldots + (X_n-\mu)^2\right] \\
&= \tfrac{1}{n}\left(\mathrm{E}\left[(X_1-\mu)^2\right] + \mathrm{E}\left[(X_2-\mu)^2\right] + \ldots + \mathrm{E}\left[(X_n-\mu)^2\right]\right) \\
&= \tfrac{1}{n}\left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right) \\
&= \sigma^2.
\end{aligned}$$

However, let's now investigate $\mathrm{E}\left[\sum (X-\bar{X})^2\right]$ without any divisor for the moment.

A useful trick here is to subtract and then add back μ within the brackets, like this:

$$\begin{aligned}
\sum (X-\bar{X})^2 &= \sum (X-\mu-\bar{X}+\mu)^2 = \sum \left\{(X-\mu)-(\bar{X}-\mu)\right\}^2 \\
&= \sum \left\{(X-\mu)^2+(\bar{X}-\mu)^2-2(X-\mu)(\bar{X}-\mu)\right\}.
\end{aligned}$$

I think it would be wise to return to longhand to sort this out. It's the amateur way, but also the safer way! So

$$\begin{aligned}
&\sum \left\{(X-\mu)^2+(\bar{X}-\mu)^2-2(X-\mu)(\bar{X}-\mu)\right\} \\
= \;& (X_1-\mu)^2 + (\bar{X}-\mu)^2 - 2(\bar{X}-\mu)(X_1-\mu) \\
&+ (X_2-\mu)^2 + (\bar{X}-\mu)^2 - 2(\bar{X}-\mu)(X_2-\mu) \\
&+ \quad\ldots\quad\quad\ldots\quad\quad\ldots \\
&+ (X_n-\mu)^2 + (\bar{X}-\mu)^2 - 2(\bar{X}-\mu)(X_n-\mu).
\end{aligned}$$

Let's add up these terms column by column. The first column is $(X_1-\mu)^2 + (X_2-\mu)^2 + \ldots + (X_n-\mu)^2$. The second column's terms are all the same, and so their sum is simply $n(\bar{X}-\mu)^2$. The final column has $2(\bar{X}-\mu)$ as a factor throughout, and so the sum is $-2(\bar{X}-\mu)(X_1+X_2+\ldots+X_n-n\mu)$.

And now let's find the expected value of $\sum (X-\bar{X})^2$. It's the sum of the expected values of all those terms.

The expected value of the first column is $E[(X_1-\mu)^2] + E[(X_2-\mu)^2] + \ldots + E[(X_n-\mu)^2$. But every one of these n terms is simply σ^2, and so therefore their sum is $n\sigma^2$. Next, the expected value of the second column is $E[n(\bar{X}-\mu)^2] = nE[(\bar{X}-\mu)^2]$. But that's just n times the variance of $\bar{X}$ which we know from {V4} to be σ^2/n. So the expected value of the second column is just σ^2. The sum of the values in the third column includes $X_1+X_2+\ldots+X_n$ which is, of course, equal to $n\bar{X}$. So the total in the third column can now be expressed as $-2(\bar{X}-\mu)(X_1+X_2+\ldots+X_n-n\mu) = -2(\bar{X}-\mu)(n\bar{X}-n\mu)$, i.e. simply $-2n(\bar{X}-\mu)^2$. But the expected value of that is clearly $-2n$ times the variance of $\bar{X}$, so this simply boils down to $-2\sigma^2$. The expected value of the whole expression is therefore $n\sigma^2 + \sigma^2 - 2\sigma^2 = (n-1)\sigma^2$. *That's* why the sample variance is defined as

$$s^2 = \frac{1}{n-1}\sum(X-\bar{X})^2 :$$

it's so that $E[s^2] = \sigma^2$. As we saw on page 65, an estimator whose expected value is equal to the thing that it is trying to estimate is known as an *unbiased* estimator. Mathematical statisticians are understandably keen to have unbiased estimators—as the name they've given it suggests! And, in fact, it is the aim to devise *unbiased* estimators which underlies most of the previously mysterious facts that have been quoted both in the main text and earlier in these Optional Extras, particularly involving those "control-chart constants". In light of that, we'll take another look in the next section at all of the control-chart constants that we've seen.

However, to conclude here, there was yet another example of the use of an unbiased estimator in Part B of these Optional Extras: see page 20. It was adjustment of the MAD to make it comparable with, i.e. on the same scale as, the standard deviation. Yes, as indicated there, in the same way that we now know that s^2 is an unbiased estimator of σ^2, we need to scale up the MAD by a factor of 1.253 in order that its square also becomes an unbiased estimator of σ^2. The only difference is that, whereas the divisor of $n-1$ *always* serves the purpose using s^2, the computation of that 1.253 factor is derived using the normal distribution assumption, as is the case with almost all of the control-chart constants. This is for the usual reason that the mathematics just can't be carried out without some such assumption.

There are two "asides" worth mentioning here.

First, the fact that s^2 is an *unbiased* estimator of σ^2 might sound impressive. But, actually, unbiasedness is not a particularly *strong* property for an estimator (although Mathematical Statisticians are pretty keen on it). All it says is, of course, that the estimator gets closer and closer to the thing being estimated as $n \to \infty$. But our sample sizes n are usually rather smaller than that! The property of unbiasedness says *nothing* about how *close* the estimator is likely to be to the item of interest when using ordinary sample sizes. Nevertheless I guess that, if you were fortunate enough to be dealing with very large samples, you would naturally regard it as beneficial for the unbiasedness criterion to be satisfied rather than for the estimator to get closer and closer to something *else* in the long term!

Secondly, the fact that s^2 is an unbiased estimator of σ^2 does *not* imply that s is an unbiased estimator of σ. Unfortunate as it may be, the property of unbiasedness does *not* carry over when operations such as taking the square root or squaring an estimator are involved. The fact that the sample variance has long been defined in such a way that the sample standard deviation is *not* in general an unbiased estimator for σ is yet a further indication of the Mathematical Statistician's preference for the variance rather than the standard deviation, despite the fact that it is the standard deviation which is the version that directly reflects *variability*, i.e. the characteristic in which we are interested. However, when we now move on to looking at control-chart constants, the criterion used *is* that they are based on unbiased estimators for σ rather than for σ^2—a contrast which I suggest is further evidence of how control charts have been developed according to what makes the better sense in practice rather than simply following mathematical tradition.

4. Why are control-chart constants what they are?

Let's run one-by-one through the various control-chart constants that we have seen.

Our first control charts were those associated with data from the Experiment on Red Beads. The details of why the control limits were computed in the particular way used there will be explained in the final part of this Technical Section.

Then on Day 3 we studied the type of control chart more generally used to analyse one-at-a-time data: the moving-range method using the familiar number 2.66. On page 65 we saw that there is a direct connection between the 2.66 and the value of h for $n = 2$ in the table on page 20. So let's move straight on to considering that quantity.

h was introduced on page 20 as the conversion factor by which a subgroup range R needs to be divided in order to scale it down to a number which is comparable to the standard deviation (when the latter exists). The same is true with the average range $\bar{R}$. As with all of the control-chart constants to be covered in this section, the values of h are derived under the assumption that the data are normally distributed (in which case, of course the standard deviation *does* exist). The reasons for this are the same as usual: (a) the mathematics cannot be carried out without some such assumption, and (b) the values of h derived using that assumption have been found to be pretty useful in practice. Following what has recently been discussed, you can probably recall the criterion we use to derive the values of h. They are the values that result in R/h, and equivalently $\bar{R}/h$, being *unbiased* estimators of σ, i.e. such that $\mathrm{E}[R/h] = \sigma$, where σ is the standard deviation of the assumed normal distribution.

We then moved on to consider situations where we have subgrouped (a-few-at-a-time) data. This is where the $\bar{X}$-R chart is commonly used, comprising both the $\bar{X}$-chart and the R-chart. These two charts, rather obviously, have their Central Lines at $\bar{\bar{X}}$ and $\bar{R}$ respectively, but how far away from them are the control limits? In both cases the answer is in the form of a multiple of $\bar{R}$. In the case of the $\bar{X}$-chart the multiplier is H, tabulated on page 21: the control limits are placed at a distance of $H\bar{R}$ either side of the Central Line.

As a matter of fact, with what you know now, you could compute the values of H yourself! Using Shewhart's 3σ-guidance along with the knowledge that the standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$ and that $\bar{R}/h$ is an unbiased estimator of σ, it turns out that

$$H = \frac{3}{h\sqrt{n}}.$$

But it would be tedious to have to work that out every time you wanted to construct an $\bar{X}$-chart! This is why the table of values of H was included.

As mentioned above, the Central Line of the R-chart is at $\bar{R}$ and the Upper Control Limit is another multiple of $\bar{R}$. The multiple this time is h_2 which is tabulated on page 22. The derivation of h_2 follows similar lines as previously. First, the standard deviation of R is computed in terms of σ, then an unbiased estimator of σ based on $\bar{R}$ is derived, and then that is multiplied by 3 following Shewhart's guidance. I trust you are glad that, long ago, other people did all that work for you!

There is one final type of control chart that I would like to mention to you. This takes us back to one-at-a-time data and is an interesting variant on the familiar chart based on control limits that are placed a distance of $2.66\overline{MR}$ either side of $\bar{X}$. An annoying problem that can sometimes occur, especially if you are using a relatively short baseline, is that there might be just one item of data in the baseline which is substantially higher or lower than all the rest of the baseline data. For ease of description, let's suppose it's

higher. Unless this is either the very first or the very last piece of data in the baseline, it will result in *two* of the moving ranges (one to its left and the other to its right) becoming considerably greater than all the rest. This will have two consequences: one is that the Central Line $\bar{\bar{X}}$ will be quite a lot higher than it would have been otherwise, and the other is that the control limits will be considerably further apart. In particular, this "double whammy" will put the UCL *much* higher than otherwise. Of course, if that troublesome item is *very* much higher than everything else then it may still finish up above the UCL despite the amount by which the latter has been raised, in which case one would have no hesitation in regarding it as a special-cause signal. But, quite often, the troublesome item will have had such a strong influence on the UCL that it finishes up below it. And then it becomes difficult to interpret.

In this sort of case, the sample mean cannot really do a very good job of genuinely reflecting the typical kind of values being recorded: it will still be a lot lower than the awkward value, but it will now be noticeably higher than all the rest of the data. In such a case (both with control-chart work and in other analyses) an alternative measure of "average" is sometimes used—one which isn't so prone to those effects. This is the *median* of the data. Imagine that your sample of data is rearranged from lowest value to highest value. Then, if the sample size *n* is an odd number, the median is defined as the central number in that rearranged list; whereas if *n* is even then the median is defined as halfway between the two middle numbers. One can rearrange the moving ranges in just the same way and thus produce the *median moving range*. It's easy to see that both the sample's median and its median moving range will be largely unaffected by the nuisance value: that high value will be at the top end of the ordered list of data, a long way away from affecting the median. Similarly, the *two* unusually high moving ranges will be at the top end of the ordered list of moving ranges, thus again leaving the median moving range essentially unaffected by those exceptionally high moving ranges. These facts are the motivation for sometimes using the alternative type of control chart which has the sample median as its Central Line, and with control limits computed using the median moving range.

The underlying theory about medians is, as you would expect, different from that about means. The consequence is that we will need to use a different multiplier from the 2.66 when computing the control limits for this alternative type of control chart. Again the theory assumes a normal distribution and again the resulting method is consistent both with (a) Shewhart's 3σ-guidance and (b) using an unbiased estimator of σ. With this estimator being based on the *median* moving range rather than the *mean* moving range, that different multiplier is 3.145. So the control limits are set at a distance of 3.145 times the median moving range either side of the Central Line (which, recall, is now placed at the sample median).

As usual, there are "pros and cons" regarding this alternative method. I have already discussed the important "pro": its relative resistance to the effect of a "nuisance value" in the data. The main "con" that I have found when using this method is that (except when there *are* such nuisance values) the control limits tend to be further apart than in the usual method, thus reducing the chart's ability to signal *real* special causes. This effect is not huge, but I found it to be more than a little annoying. I therefore finished up using this alternative method only when I was analysing a process that had some tendency to produce nuisance values, particularly if I was using a relatively short baseline. So my suggestion is for you to at least keep this method at the back of your mind or, perhaps preferably, try it out on some of your own data in order to get a "feel" for whether it might be useful to you or not. It's just as mathematically "valid" as the standard method, so *you* are free to judge whether it is the pro or the con that appears to be the more important as regards analysing your own data.

> Just in case the similarity had occurred to you, there is *no* connection between that constant 3.145 and the one usually represented by the Greek letter π ("pi") which is used e.g. to compute the circumference of a circle. The fact that they are virtually equal is just a fluke. However, if you *are* familiar with π then you could, of course, use it as a handy reminder of the sort of constant to use when constructing a control chart for medians—the difference between them will hardly be noticed!

5. More on the binomial and normal distributions

Both the binomial and normal distributions were introduced in Part D. The normal distribution was covered quite fully, so there is not much to add here: I shall simply show you how to use commonly-available tables to find probabilities.

However, only two very simple binomial distributions were introduced in Part D: those involving the count of Heads when two coins are tossed and then the number of sixes when three dice are thrown. Here we shall firstly tackle binomial distributions more generally.

The binomial distribution

If it is a while since you read that material on the simple binomial distributions at the beginning of Part D then I suggest, in order to put yourself back in the picture, you skim through those first few pages (from page 37) now before continuing with this more general treatment.

As in those introductory cases, binomial distributions are concerned with a number of repeated independent trials of some procedure or operation etc in each of which we will classify the outcome in just one of two ways (we had either Head or Tail in the first illustration and either a six or not a six in the second). I'll follow fairly common practice in the books by referring to these two possibilities as Success and Failure respectively (although, e.g. in some inspection procedure, "Success" might refer to "defective" and "Failure" to "non-defective"!). "Successes" are simply what we decide to count.

The big general question to tackle now is as follows. If we denote the probabilities of Success or Failure at each trial by p and q respectively (where obviously $q = 1 - p$), and we carry out n independent trials of the operation, etc, what is the probability that the total number X of Successes obtained is equal to 0 or 1 or 2 or any specified number up to the maximum of n?

So, in shorthand, how can we compute Prob $(X = x)$ for $x = 0, 1, 2, \ldots, n$? Referring back to those early pages of Part D, there are two steps in this computation: one is easy but the other one can be quite difficult. The easy step is to compute the probability of the number of Successes as being equal to x *when those Successes occur in specific positions* in the sequence of Successes and Failures. So suppose the following sequence contains a total of n letters comprising x S's and $n-x$ F's:

$$\text{SFFSSFSFFF ... FSFFFSSF}.$$

That is, the first trial produced a Success, the second and third trials produced Failures, and so on. Seeing that these are *independent* trials with fixed probabilities of Success and Failure, we can use the simple multiplication rule (page 77) many times over to obtain the probability:

$$\text{Prob}\,(\text{SFFSSFSFFF ... FSFFFSSF}) = p^x q^{n-x}.$$

The same will, of course, be true for *any* particular sequence x S's and $n-x$ F's. Therefore the next question is: how many such sequences are there? Ah: that's the tricky bit! It was easy enough with those simple introductory illustrations on the early pages of Part D, but it's not so easy in general. Now, *I* know the answer to that question, but how can I verify that answer for you?

The best way to do this that I can think of is to employ a very useful technique called "mathematical induction". Mathematical induction works in a case such as this where we are trying to prove that a result is true for *all* values of n when (a) it's easy to prove it for some small starting-value of n, usually 0 or 1, and then

(b) if we *assume* it to be true for a particular value of n then we can subsequently *prove* it to be true for $n+1$. I believe you'll soon see the logic!

Let's say (a) it's easy to see that the result is true for $n=1$. Then we no longer have to *assume* it's true for $n=1$: we *know* it's true! But that means we can use (b) to *prove* it true for $n=2$. But then we no longer have to *assume* it's true for $n=2$: we *know* it's true! But that means we can use (b) to *prove* it true for $n=3$. And so on, and so on, and ... .

So what *is* this result I want to prove to you? For the time being I'll use the letter r rather than the letter x since otherwise there's a big danger of getting confused between the letter x and multiplication signs. The result to be proved is that the number of sequences containing r S's and $n-r$ F's is

$$\frac{n\times(n-1)\times\ \ldots\ \times(n-r+2)\times(n-r+1)}{r\times(r-1)\times\ \ldots\ \times 2\times 1}\quad=\quad\binom{n}{r}.$$

The "shorthand" for this big fraction that I've shown on the right-hand side is known as a *binomial coefficient*. Notice that there are exactly r terms in both the top and bottom of that big fraction. Let me show that to you by spreading out the denominator like this:

$$\frac{n\times(n-1)\times\ \ldots\ \times(n-r+2)\times(n-r+1)}{r\times(r-1)\times\ \ldots\times\quad 2\quad\times\quad 1}.$$

Some particular values of this expression are easy to check. For example, if $r = 1$ then we get just the single term n at the top and the single term 1 at the bottom, with the answer n. That's obviously correct since the one S can be anywhere in the available n places, thus corresponding to n sequences altogether. A further easy check is with $r = n$, in which case the top and the bottom of this fraction are identical, thus giving the answer 1. That's also obviously true, since there's only one sequence consisting entirely of S's. There is just one exceptional case where the big fraction entirely disappears! That's when $r = 0$. But that corresponds to when there are no S's at all in the sequence, i.e. we have all F's. There is obviously only one such sequence, and so the binomial coefficient is *defined* as being equal to 1 when $r = 0$.

So now, moving on to the induction process, let's *assume* that the above binomial coefficient *is* the correct number of sequences of length n which contain r S's (for all possible values of r) and see what happens with sequences of length $n+1$. The extra place at the right-hand end of the sequence can, of course, be filled with either an S or an F. Let's suppose first that it's an S. Then, in order for there to be r S's in the sequence of length $n+1$, there must be $r-1$ S's in the first n places. The number of such sequences is

$$\binom{n}{r-1}\quad=\quad\frac{n\times(n-1)\times(n-2)\times\ \ldots\ \times(n-r+2)}{(r-1)\times(r-2)\times\ \ldots\ \times 2\times 1}\quad.$$

Now suppose the extra place at the right-hand end is an F. Then, for there to be r S's in the sequence of length $n+1$, there must be r S's in the first n places. The number of such sequences is, of course,

$$\binom{n}{r}\quad=\quad\frac{n\times(n-1)\times\ \ldots\ \times(n-r+2)\times(n-r+1)}{r\times(r-1)\times\ \ldots\ \times 2\times 1}\quad.$$

So now we hope that adding together these two binomial coefficients will give us a total which is equal to the binomial coefficient corresponding to the number of sequences of length $n+1$ which contain r S's. Let's see. We have

$$\frac{n\times(n-1)\times(n-2)\times\ \ldots\ \times(n-r+2)}{(r-1)\times(r-2)\times\ \ldots\ \times 2\times 1}\quad+\quad\frac{n\times(n-1)\times\ \ldots\ \times(n-r+2)\times(n-r+1)}{r\times(r-1)\times\ \ldots\ \times 2\times 1}$$

$$= \frac{n\times(n-1)\times(n-2)\times\ \ldots\ \times(n-r+2)}{r\times(r-1)\times(r-2)\times\ \ldots\ \times 2\times 1}\left\{r+(n-r+1)\right\}$$

$$= \frac{n\times(n-1)\times(n-2)\times\ \ldots\ \times(n-r+2)}{r\times(r-1)\times(r-2)\times\ \ldots\ \times 2\times 1}(n+1) \ = \ \binom{n+1}{r}.$$

It works! All the induction proof needs now is a starting value for n. $n=1$ is a good choice! For then there are just two possible sequences: the one comprising a single S and the one comprising a single F. The latter is the exceptional case in the middle of the previous page, and the former is the case where both r and n are equal to 1, also covered on the previous page. So the proof that the number of sequences containing r S's and $n-r$ F's is given by the binomial coefficient as defined opposite is now complete.

Therefore, combining both the initial easy part of the above work and now the second rather more demanding part, we have the complete statement of the probability distribution for the number X of Successes in n independent trials as:

$$\text{Prob}\,(X=x) = \binom{n}{x} p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

Hooray! Now, that's all very well, but that will still involve a pretty unpleasant amount of arithmetic to actually get numerical values for all those probabilities! Also, the prospect of computing the mean and variance of the distribution using the formulae

$$\mu \ = \ \sum x.\,\text{Prob}\,(X=x) \ \text{ and } \sigma^2 = \sum\left(x-\mu\right)^2.\,\text{Prob}\,(X=x)$$

as shown at the top of page 76 is also not particularly appealing! Let's deal with this latter issue first.

Fortunately, we can bypass those formulae by taking advantage of the additivity properties for both means and variances proved in Section 3 (pages 79 and 80). X, the number of successes, can of course be considered as

$$X = \ X_1 + X_2 + \ldots + X_n$$

where X_1 = 1 or 0 according as the first trial yields S or F respectively, X_2 = 1 or 0 according as the second trial yields S or F respectively, and so on through all the n trials. So, in fact, X_1, X_2, ..., X_n are all independent simple binomial random variables with $n=1$.

There's no difficulty in calculating the mean and variance of each of those! We have $\text{Prob}\,(X_1=0) = q$ and $\text{Prob}\,(X_1=1) = p$, and so clearly we have $\text{E}[X_1] = p$, as is also true of $\text{E}[X_2]$, $\text{E}[X_3]$, ..., $\text{E}[X_n]$. Thus $\mu = \text{E}[X] = np$. The variance is almost as easy. Recall the useful result {V1} on page 79: $\sigma^2 = \text{E}\,[X^2]-\mu^2$. Also, of course, with both possible values 0 and 1, X_1^2 is very fortunately *equal to* X_1! That being the case, $\text{E}\,[X_1^2] = \text{E}[X_1] = np$. So this gives us the variance of X_1 as $\text{E}\,[X_1^2] - \text{E}[X_1]^2 = p - p^2 = p\,(1-p) = pq$. The same will be true of the variances of each of X_2, ..., X_n. And so finally we can use the additivity property for variances (see {V3} on page 80) to obtain the variance of X, i.e. σ^2, as npq. Those important results are worth displaying:

For the binomial distribution as defined above, $\mu = np$ and $\sigma^2 = npq = np(1-p)$.

Before returning to the matter of how to easily obtain numerical values for all those binomial probabilities (binomial coefficients and all!) there is one loose end to tidy up. So far in these Optional Extras we have revisited all of the various control-chart constants and types of control limits that we had encountered during the course—except for just one case. That exception is the control limits used on the control charts constructed to examine data from the Red Beads Experiment. On Day 2 page 20 we saw:

Technical Aid 1

One of the earliest applications of Shewhart's invention of the control chart was for batch inspection of mass production processes. In such inspection, samples (batches) of n items from the process's output are regularly drawn and inspected, and the number X of defective items recorded. After several samples have been inspected, the control limits are computed as follows.

Using the statistician's traditional shorthand for averages, $\bar{X}$ represents the average number of defectives found in the samples so far, while $\bar{p} = \bar{X} \div n$ is the average *proportion* of defectives in the samples. Shewhart's guidance about control limits then leads to the upper and lower limits being placed at

$$\text{UCL} = \bar{X} + 3\sqrt{\bar{X}(1-\bar{p})} \quad \text{and} \quad \text{LCL} = \bar{X} - 3\sqrt{\bar{X}(1-\bar{p})}\,.$$

We have just shown that, for a binomial distribution, $\sigma^2 = np(1–p)$ which is, of course, saying that the standard deviation $\sigma = \sqrt{np(1-p)}$. So the distance from the Central Line shown in that Technical Aid takes the form of Shewhart's "3σ" except that it uses an *estimate* of σ since p is not known. This is a similar kind of argument to that used for the other various types of control limits we have seen: indeed, an exact value of "σ" often doesn't even exist in practice as the conventional statistician understands it. But it does here. And when it does exist then what we use is a sensible unbiased estimator of it. As I have said before, this is not an exact science!

In fact, in this case, a further approximation has been made. If you think about it you will realise that X = the number of red beads in the paddle *cannot* be exactly binomially distributed, whatever assumptions you might care to make. One of the assumptions is that the sample (of size 50 in the case of the Red Beads Experiment) can be expressed as the sum of simple binomial random variables $X = X_1 + X_2 + ... + X_{50}$ where each one of these 50 variables has probability p of being a red bead *irrespective* of what happens elsewhere in the sample. For this argument let's imagine that the 50 holes in the paddle are numbered 1, 2, ..., 50. But if, say, there is a red bead in Hole Number 1 then there are now just 3,999 beads left—799 of them red and 3,200 of them white—compared with the 4,000 beads that we started with of which 800 were red. So, with a red bead in Hole Number 1, the proportion of red beads available for Hole Number 2 is very slightly *less than* the 0.2 that we started with. And so on. However, when relatively small samples are drawn from relatively large populations, it is usual to ignore such little matters! In any case, as already discussed in Part E, it is highly unlikely that the sample of 50 beads in the paddle can truthfully even be considered as a *random* sample, which is another good reason for not worrying too much about such a minor complication!

However, Mathematical Statisticians might be interested in how to compute the probabilities of the number of red beads in the paddle *if* we make all the assumptions that they might like to make but now, in addition, take that complication into account. Actually, with what you have now learned, *you* could tell them! The probability of x red beads in the paddle will surely be the number of possible selections of x red beads from the container multiplied by the number of possible selections of $50-x$ white beads and then divided by the total number of different selections of 50 beads out of the 4,000 available. That probability can thus be expressed in terms of binomial coefficients as

$$\frac{\binom{800}{x}\binom{3200}{50-x}}{\binom{4000}{50}}\,.$$

If you'd like to impress the Mathematical Statistician even further, you can tell him that the probability distribution comprising those delightful probabilities is called the *hypergeometric* distribution! And if you'd like

to go the whole distance then you can also provide him with details of how good the binomial distribution is as an approximation to that hypergeometric distribution by showing him the following table of probabilities:

x =	0	1	2	3	4	5	6	7
Hypergeometric	0.00%	0.02%	0.10%	0.42%	1.25%	2.91%	5.49%	8.68%
Binomial	0.00%	0.02%	0.11%	0.44%	1.28%	2.95%	5.54%	8.70%

x =	8	9	10	11	12	13	14	15
Hypergeometric	11.72%	13.71%	14.07%	12.79%	10.37%	7.55%	4.96%	2.96%
Binomial	11.69%	13.64%	13.98%	12.71%	10.33%	7.55%	4.99%	2.99%

x =	16	17	18	19	20	21	22	23 etc
Hypergeometric	1.60%	0.79%	0.36%	0.15%	0.06%	0.02%	0.01%	0.00%
Binomial	1.64%	0.82%	0.37%	0.16%	0.06%	0.02%	0.01%	0.00%

Well—this part of these Optional Extras *is* titled the "Technical Section"!

That brings us to the final issue to be discussed regarding the binomial distribution: how to obtain those binomial probabilities without getting involved with too much unwieldy arithmetic. The answer is to have a suitable set of Statistics Tables by your side. And there, as you might suspect, I must declare an interest!

The better of my two little books of Statistics Tables for this purpose is *Elementary Statistics Tables* (*EST*). At the very beginning of *EST* there are four pages of tables of probabilities $\text{Prob}(X = x)$ in binomial distributions. They cover all values of n up to 20 and an extensive range of values of p: 0.01 to 0.10 and 0.90 to 0.99 in steps of 0.01, 0.15 to 0.85 in steps of 0.05, and the fractions $\frac{1}{6}$, $\frac{1}{3}$, $\frac{2}{3}$ and $\frac{5}{6}$.

However, sometimes one wants the probability that X lies in some *interval* of values rather than the probability of just a single value. That could, of course, involve adding up quite a number of individual probabilities. To avoid the need for that there are also four pages (covering the same range of values of n and p) of the cumulative distribution function (cdf) of X. Tables of the cdf are even more important for the normal distribution, as we shall soon see. The cdf (let's denote it by $F(x)$) gives the probability that X is *less than or equal to* x: $F(x) = \text{Prob}(X \leqslant x)$. The advantage of the cdf is that you only need to look up just two probabilities in order to find the probability that X lies in an interval, never mind how wide the interval is.

For example, suppose you want the probability that X takes some value between 4 and 8 inclusive. Then the only table entries that you need look up are $F(8)$ and $F(3)$ since, clearly,

$$\text{Prob}(4 \leqslant X \leqslant 8) = \text{Prob}(X \leqslant 8) - \text{Prob}(X \leqslant 3).$$

Careful: don't subtract $\text{Prob}(X \leqslant 4)$! On page 48 I also mentioned that sometimes we need the probability that X lies in a "one-sided" interval, i.e. the probability that X is *at most* some number or X is *at least* some number. The first of those two options is simply the cdf value at that number, e.g.

$$\text{Prob}(X \leqslant 8) = F(8).$$

Or if you wanted the probability that X is *at least* 8 then that would be

$$\text{Prob}(X \geqslant 8) = 1 - F(7).$$

Finally, since the tables only go up to $n = 20$, how can we find probabilities for larger values of n without needing to do a lot of arithmetic? For this purpose, and recalling the Central Limit Theorem, it is often possible to use tables of the normal distribution to obtain good approximations to binomial probabilities. So I'll return to that matter in the bottom half of page 91 after now describing how to use widely-available tables of the normal distribution.

The normal distribution

Part D includes a fairly extensive introduction and discussion on the normal distribution. So effectively all that is left is (as promised on page 50) to introduce you to the tables of the normal distribution that you will find in all introductory books on Statistics and plenty of other sources as well. In *EST* the main table is on pages 18–19 and in *ST* it's on pages 34–35.

As you are now well aware from Part D, since the normal distribution is a *continuous* distribution we cannot now sensibly consider probabilities of individual values. So we have already pointed out that it is the probability of the normal random variable lying in an *interval* (including the "one-sided" type of interval just mentioned on the previous page) which does make sense. Indeed the pictures on page 49 have already shown you examples of this. So what are the details?

The published tables invariably apply directly just to the *standard normal* distribution, i.e. N(0,1), the normal distribution having mean 0 and variance 1. Fortunately, following on from some of the development in Part D, that is sufficient for us to be able to find probabilities in *any* normal distribution.

But one thing at a time. Let's first familiarise ourselves with just finding probabilities in N(0,1). The standard normal distribution is such an important distribution that it is often given a special notation which usually applies *only* to N(0,1) and to no other distribution. And that is, of course, the notation you will find in both *EST* and *ST.* A N(0,1) random variable is almost always denoted by Z rather than X, and its cdf is usually denoted by Φ (capital "phi"). The tables mostly found in the books are tables of $\Phi(z)$. Here is an abbreviated version of such a table:

z	0	1	2	3	4	5	6	7	8	9
−3.	0.0013	0010	0007	0005	0003	0002	0002	0001	0001	0000
−2.	0.0228	0179	0139	0107	0082	0062	0047	0035	0026	0019
−1.	0.1587	1357	1151	0968	0808	0668	0548	0446	0359	0287
−0.	0.5000	4602	4207	3821	3446	3085	2743	2420	2119	1841
0.	0.5000	5398	5793	6179	6554	6915	7257	7580	7881	8159
1.	0.8413	8643	8849	9032	9192	9332	9452	9554	9641	9713
2.	0.9772	9821	9861	9893	9918	9938	9953	9965	9974	9981
3.	0.9987	9990	9993	9995	9997	9998	9998	9999	9999	1.00

This short table provides values of $\Phi(z)$ for z ranging from –3.9 to +3.9 in steps of 0.1. The usual full published tables provide values of z in steps of 0.01 with some additional proportional parts allowing close approximations to $\Phi(z)$ for z expressed to three decimal places. However, this brief table is sufficient to get you started on finding probabilities in normal distributions.

Let's read off a few values. For a start, $\text{Prob}(Z \leqslant 1) = 0.8413$ and $\text{Prob}(Z \leqslant 1.5) = 0.9332$. Remembering that probabilities are represented by areas under the normal curve, you might like to check the value of $\text{Prob}(Z \leqslant 1)$ approximately by adding up the relevant areas in the pictures on page 49. Let's also read off $\text{Prob}(Z \leqslant -1) = 0.1587$. Notice that this is equal to $1 - \text{Prob}(Z \leqslant +1)$. That is bound to be true because of the *symmetry* of the normal curve. But $1 - \text{Prob}(Z \leqslant 1) = \text{Prob}(Z \geqslant 1)$, so this is simply confirming that the area under the standard normal curve to the right of $z = +1$ is equal to the area to the left of $z = -1$.

It is also worth noting that, when considering probabilities in a normal distribution, as is the case with any *continuous* distribution, we do not have to worry as to whether we should write, for example, $\text{Prob}(Z \geqslant 1)$ or $\text{Prob}(Z > 1)$—these probabilities will be the *same* as each other because of the feature that, in continuous distributions, the probability of any single value is zero! That is, of course, wholly different from what

happens with discrete distributions. Recall the sentence from page 89 when we were discussing binomial distributions: "if you wanted the probability that X is *at least* 8 then that would be $\text{Prob}(X \geqslant 8) = 1 - F(7)$", i.e. the probability that $X \geqslant 8 = 1 - \text{Prob}(X \leqslant 7)$, *not* $1 - \text{Prob}(X \leqslant 8)$!

Let's give just one further illustration. Suppose that, for some reason, you wanted to find the probability that Z lies in the interval between –1.3 and 1.8: $\text{Prob}(-1.3 \leqslant Z \leqslant 1.8)$. All you have to do is look up both $\Phi(-1.3)$ and $\Phi(1.8)$ in the table, giving respectively 0.0968 and 0.9641, and subtract one from the other: 0.9641 – 0.0968 = 0.8673. This is because $\Phi(1.8)$ gives the *total* area to the left of 1.8 under the standard normal curve and $\Phi(-1.3)$ gives the area to the left of –1.3 which is the part of $\Phi(1.8)$ that we don't want to be included in our interval.

Now let's see how to obtain probabilities for *any* normal distribution $N(\mu,\sigma^2)$, not just N(0,1). Perhaps it's worthwhile to take yet another look at page 49 to remind yourself of the extremely fortunate fact that those areas representing the probabilities under *any* normal curve stay the same, irrespective of *which* normal distribution we have. In particular, comparing the random variable, X say, which has a $N(\mu,\sigma^2)$ distribution, with our N(0,1) random variable Z for which we can now use that table of values of $\Phi(z)$ to find probabilities, we have the exceedingly useful fact that

$$\text{Prob}(X \leqslant x) \;=\; \text{Prob}\left(\frac{X-\mu}{\sigma} \leqslant \frac{x-\mu}{\sigma}\right) \;=\; \text{Prob}\left(Z \leqslant \frac{x-\mu}{\sigma}\right) \;=\; \Phi\left(\frac{x-\mu}{\sigma}\right).$$

That's to say the cdf of X, $F(x) = \text{Prob}(X \leqslant x)$, can be found by "standardising" the value x by forming

$$z \;=\; \frac{x-\mu}{\sigma}$$

and looking up the value of $\Phi(z)$ at *that* value of z in the table.

As an example, if X has the $N(5, 2^2)$ distribution, i.e. is normally distributed with mean $\mu = 5$ and standard deviation $\sigma = 2$, then $F(8) = \text{Prob}(X \leqslant 8) = \Phi(1.5) = 0.9332$ since 1.5 is the *standardised* version of 8, obtained by subtracting μ and dividing by σ, i.e. subtracting 5 and then dividing by 2.

And then, following the various illustrations you've already seen, you can now find probabilities of anything you need, involving *any* normal distributions, just using a table of values of $\Phi(z)$.

Finally, as promised on page 89, let's see how probabilities for binomial distributions with larger n ($n > 20$ in the case of both *EST* and *ST*) than contained in your book of Statistics Tables can be found, again without involving a lot of arithmetic. As stated there, the Central Limit Theorem can often be used. In words rather than symbols, the Central Limit Theorem says that the distribution of a sample mean becomes more and more like normal as the sample size increases. Referring back to where we introduced the Central Limit Theorem on page 51, we then made use of what we now know to be the mean and standard deviation of the sample mean to carry out the "standardising" operation (subtracting the mean and dividing by the standard deviation) in order to produce an approximate N(0,1) random variable whose probabilities can be found from the tables. Here it's more convenient to consider the distribution of X rather than $\bar{X}$ since it is, of course, X which has the binomial distribution, not $\bar{X}$. That's not a problem: X is just a multiple of $\bar{X}$ and so has the same-shaped distribution. We just have to be sure to use the mean and standard deviation of X rather than $\bar{X}$ in the standardising operation. So our variable that is now well-approximated by N(0,1) is

$$Z \;=\; \frac{X - np}{\sqrt{npq}}\,.$$

The really useful aspect of the Central Limit Theorem for practical purposes is what the computer simulations then showed on pages 52–54: i.e. that for reasonably symmetric distributions the tendency toward

normality becomes evident for quite small, sometimes *very* small, values of n. The binomial distribution is *exactly* symmetric if $p = q = \frac{1}{2}$ but become increasingly unsymmetric as one of p or q gets close to 0 and the other gets close to 1. The general guidance to be found in the books is that the normal distribution provides good approximations to binomial probabilities as long as $np \geqslant 5$ if $p \leqslant q$ (or $nq \geqslant 5$ if $q < p$).

But the binomial distribution is *discrete* while the normal distribution is *continuous*. So how exactly do we find those good approximations to binomial probabilities from tables of the (standard) normal distribution? There's a good clue in the way that histograms were drawn e.g. on pages 38–39. Those histograms have boxes which are *centred* on the actual integer values of the random variable, e.g. the box representing $X = 2$ stretches from $1\frac{1}{2}$ to $2\frac{1}{2}$. So we now do the equivalent here. To find the probability (to within a good approximation) that $X = x$ we find the probability under the relevant normal distribution between $x - \frac{1}{2}$ and $x + \frac{1}{2}$. By "relevant normal distribution" I mean the normal distribution which has the same mean and variance as the binomial distribution.

Let's look at a couple of examples. Consider the binomial distribution having $n = 100$ and $p = \frac{1}{5}$. What is the probability that $X = 22$? The mean and variance of X are $\mu = np$ and $\sigma^2 = np(1-p)$ respectively which give us $\mu = 100 \times \frac{1}{5} = 20$ and $\sigma^2 = 100 \times \frac{1}{5} \times \frac{4}{5} = 16$, i.e. $\sigma = 4$. In the corresponding normal distribution we want the area between $21\frac{1}{2}$ and $22\frac{1}{2}$. Standardising these two values (i.e. subtracting 20 and dividing by 4) gives us $\frac{3}{8}$ and $\frac{5}{8}$ respectively, i.e. 0.375 and 0.625. Since these numbers involve three decimal places, we can't read these directly from the brief table on page 90, so I need to use more detailed standard normal tables to tell you that the probabilities are 0.6462 and 0.7340 (although you could get these roughly from the table on page 90 by interpolating between the entries for 0.3 and 0.4 and between 0.6 and 0.7). Subtracting 0.6462 from 0.7340 then gives us the approximate probability of $X = 22$ as 0.0878.

Finding the probability of X lying in some interval is no more difficult. For example, suppose we want to find a good approximation to $\text{Prob}(18 \leqslant X \leqslant 25)$. This corresponds to the area between $17\frac{1}{2}$ and $25\frac{1}{2}$ or, standardising, between $-\frac{5}{8} = -0.625$ and $\frac{11}{8} = 1.375$ whose entries in the standard normal table are 0.2660 and 0.9155. And so we finish up with $\text{Prob}(18 \leqslant X \leqslant 25)$ being approximately 0.9155 – 0.2660 = 0.6495.

There's just one piece of the jigsaw left to fit in. You'll recall the guidance that it's reasonable to use the normal distribution as an approximation if $np \geqslant 5$ (with $p \leqslant q$). So how about when $n > 20$ but $np < 5$ (again with $p \leqslant q$)? Let's take $n = 100$ again but now with $p = 0.02$. There is a well-known discrete probability distribution that works well in these cases. It's the *Poisson* distribution which is tabulated on *EST* pages 14–16. Here you can simply enter the table of the Poisson distribution at the appropriate value of μ which is $np = 2$. First, let's consider the probability of a particular value, say $X = 3$. The table of the Poisson distribution gives $\text{Prob}(X = 3)$ as 0.1804. Secondly, let's consider the probability of X lying between 3 and 5 inclusive. Here you have two options. Firstly, you could just look up the probabilities of $X = 3$, 4 and 5 and add them up, giving 0.1804 + 0.0902 + 0.0361 = 0.3067. But that could get quite tedious for wider ranges of values. So alternatively, on *EST* page 17, there is a chart from which we can read off the values of $\text{Prob}(X \geqslant x)$ with excellent accuracy for probabilities near 0 or 1 and reasonably confidently to two decimal places for probabilities near $\frac{1}{2}$. **NB:** That's not a misprint! The chart does indeed provide the probabilities $\text{Prob}(X \geqslant x)$ which is the opposite way round from the cdf's $\text{Prob}(X \leqslant x)$. But I won't bore you by trying to explain why this is the way that such Poisson charts have traditionally been constructed! To obtain an approximation to $\text{Prob}(3 \leqslant X \leqslant 5)$, the chart gives $\text{Prob}(X \geqslant 3)$ and $\text{Prob}(X \geqslant 6)$ as about 0.32 and 0.015 respectively so that $\text{Prob}(3 \leqslant X \leqslant 5)$, i.e. Prob(3, 4 or 5), is around 0.32 – 0.015 = 0.305.

To give you some idea of the accuracy of these various approximations, I have also computed these probabilities directly from the exact expression for binomial probabilities a third of the way down page 87. For the case of $n = 100$ and $p = 0.2$ I obtained $\text{Prob}(X = 22) = 0.0849$ (compared with 0.0878) and for an interval I obtained $\text{Prob}(18 \leqslant X \leqslant 25) = 0.6413$ (rather than 0.6495). Then for the case of $n = 100$ and $p = 0.02$ I obtained $\text{Prob}(X = 3) = 0.1823$ (rather than 0.1804) and $\text{Prob}(3 \leqslant X \leqslant 5) = 0.3078$ (rather than the 0.3067 or 0.305). I'd say that both methods using the Poisson approximation to the Binomial work pretty well!