Statistical Process Control in Semiconductor Manufacturing

Why It Exists, and the Statistics Behind It

ECE 8803 · AI for Semiconductor Manufacturing

Asif Khan

Associate Professor, School of Electrical and Computer Engineering

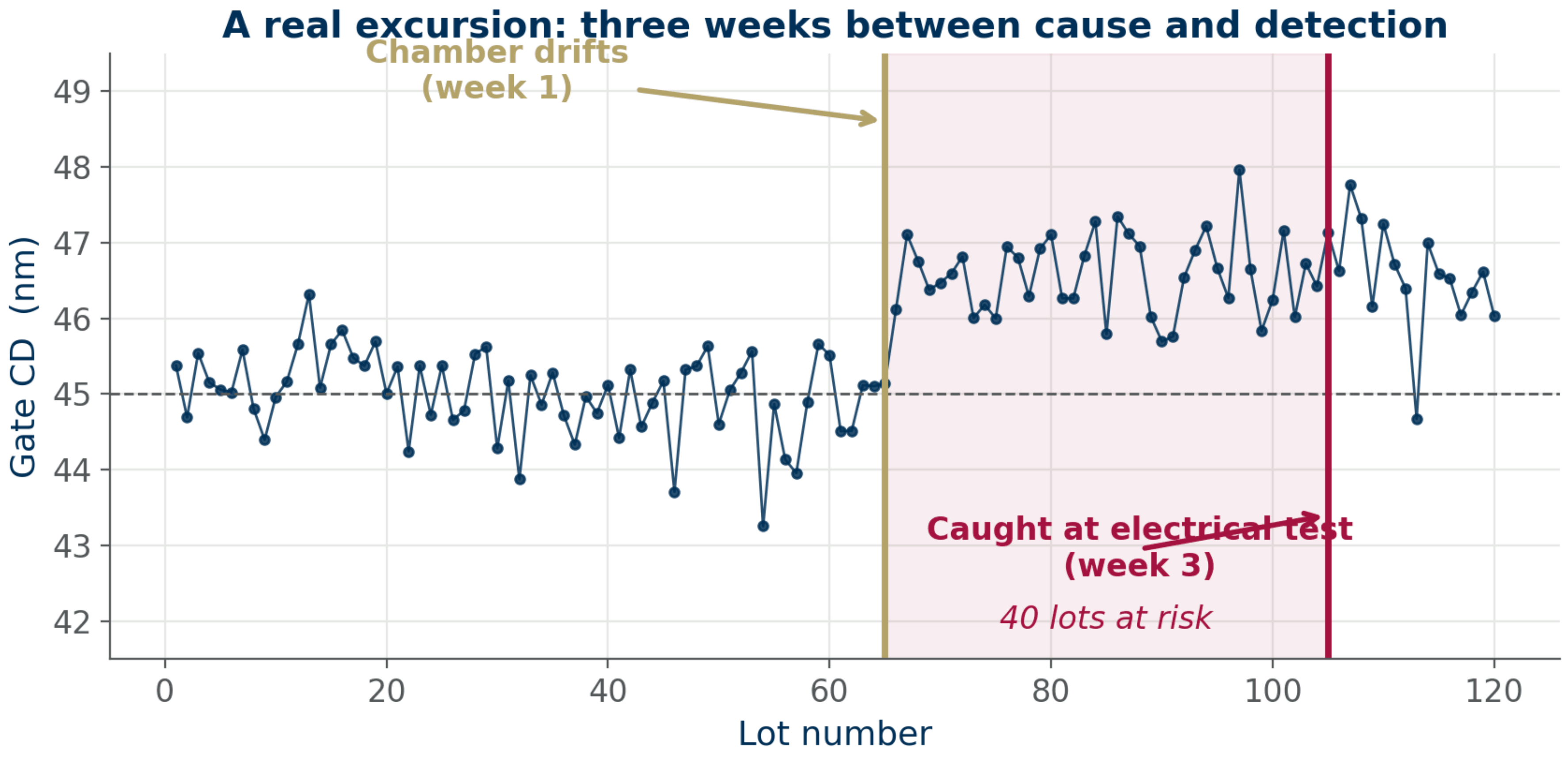
School of Materials Science and Engineering · Georgia Institute of Technology

akhan40@gatech.edu

What you should be able to do by the end of today

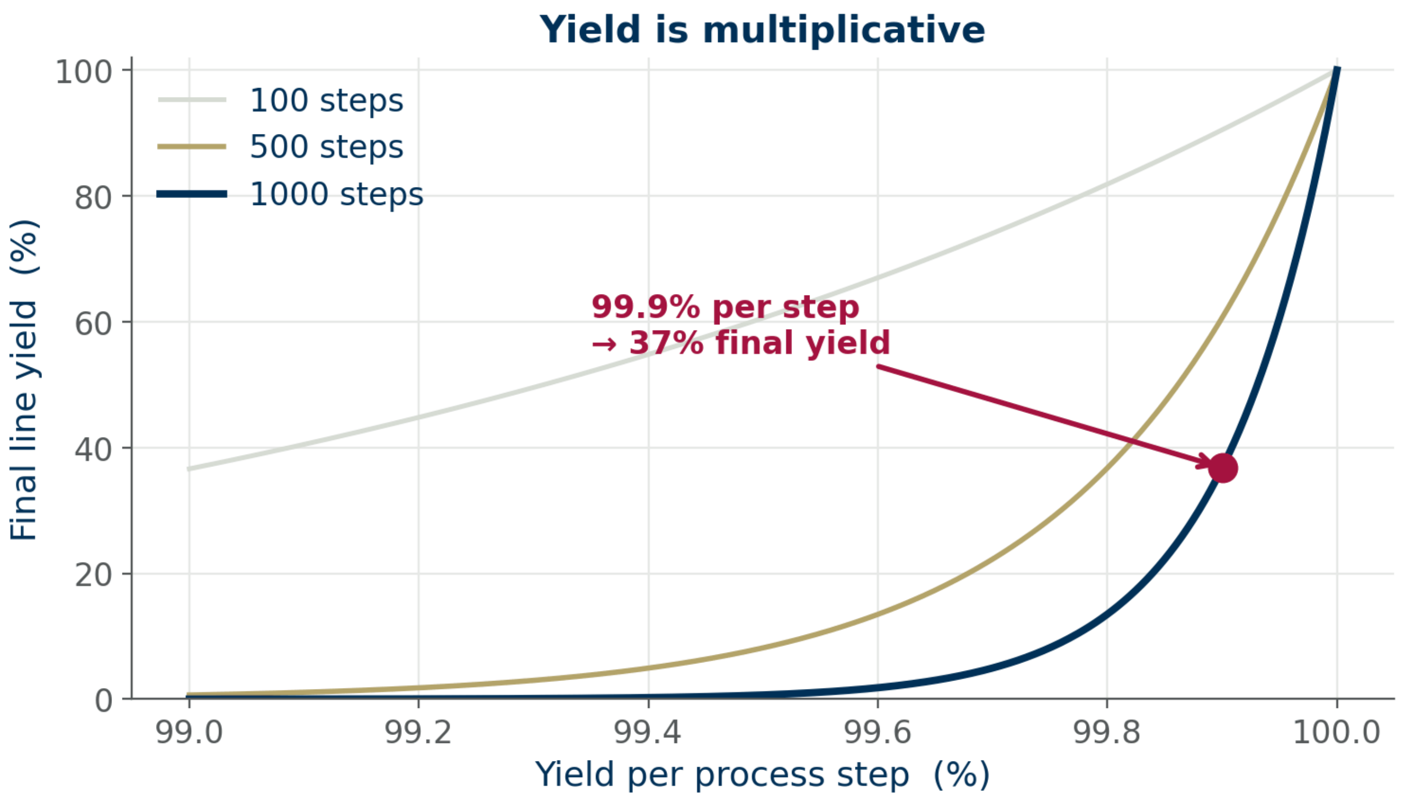
1. Explain why variation — not the average — is what kills yield in a fab
2. Tell common cause from special cause, and name the right response to each
3. Compute and interpret mean, standard deviation, and z-scores on process data
4. Explain why sample averages vary less than individual measurements
5. Frame “is this process out of control?” as a decision with two kinds of error

An excursion nobody saw for three weeks



Nothing was broken. The measurements existed. Nobody was looking at them the right way.

Yield is multiplicative



- A step that is 99.9% good sounds excellent.
- **Run 1000 of them and 63% of your wafers are gone.**
- Each step's yield multiplies, so small per-step losses compound into catastrophic line yield.
- This is why fabs chase the fourth and fifth nines on individual steps that look already-good.
- **Corollary:** there is no such thing as an unimportant step.

Control at every step is not a quality nicety. It is the business model.

Georgia Tech

What poor control actually costs

Direct and immediate

Scrapped material

A 300 mm wafer at a late layer carries months of accumulated value. A 25-wafer lot scrapped after metal-1 is a six-figure write-off.

Tool downtime

An excursion means the tool comes down for diagnosis, cleaning, and qualification. Capacity you paid for and did not use.

Requalification

After any fix you must run monitor wafers and prove the tool is back. Hours to days before production resumes.

These are the costs everyone counts — and they are the smaller half.

THE ENTIRE PROBLEM

Detect a real change quickly, without reacting to noise.

Every method in the next three lectures is a variation on that one sentence.

An experiment

Everyone will produce defects. Nobody will be at fault.

The setup

A container holds 800 white beads and 200 red beads, thoroughly mixed.

Red beads are defects.

Each of you is a production worker. Your job is to produce white beads.

You draw 50 beads with a paddle. Whatever you get is your output for the day.

Management will post the results.

The only rules

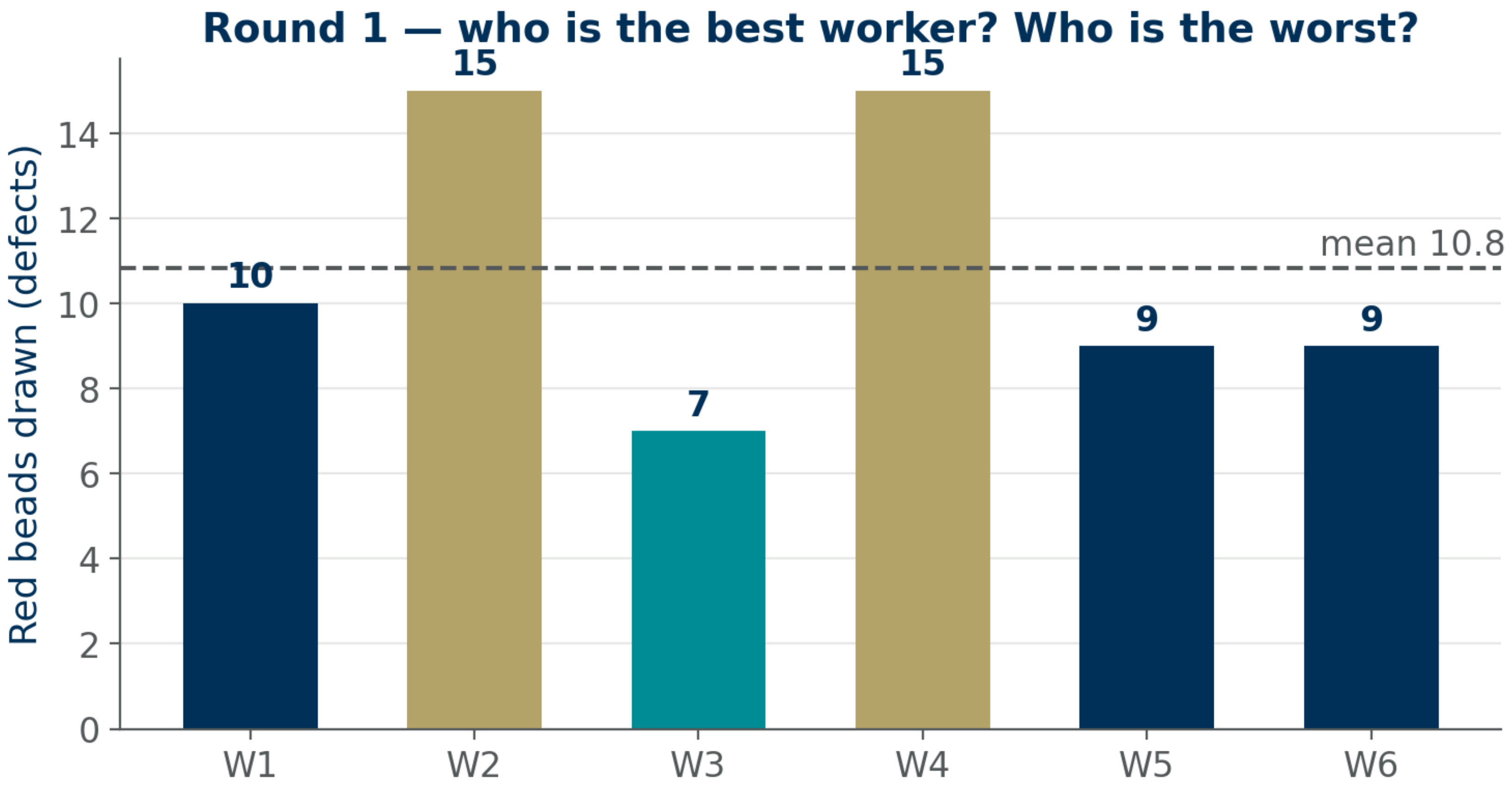
Follow the procedure exactly. Do not tilt the paddle. Do not sort the beads. Work at the same speed as everyone else.

What management believes

Good workers draw fewer red beads. Poor workers draw more. Praise the good ones, coach the poor ones, and quality will improve.

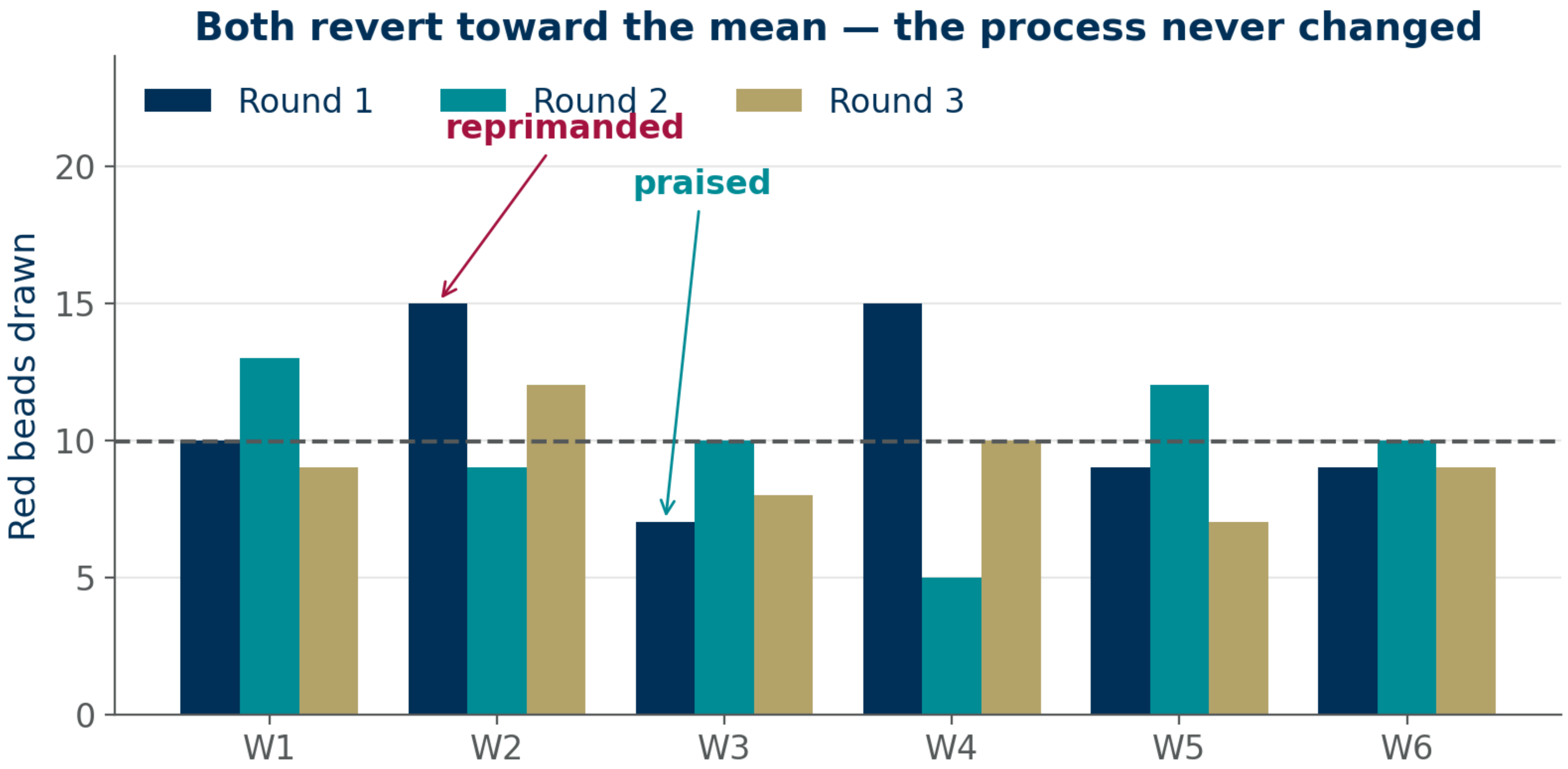
Hold on to your prediction of who will win.

Round 1



Same paddle. Same container. Same procedure. Different results.

Rounds 2 and 3



The praised worker got worse. The reprimanded worker improved. Nobody changed anything.

What actually happened

Every worker drew from the same container, with the same procedure, at the same rate. There was exactly one process.

Common cause variation

Built into the system. Always present. Produces a stable, predictable band of outcomes.

In the fab: chamber-to-chamber differences, gas purity, ambient temperature, metrology noise, incoming wafer variation.

To reduce it you must change the system — a better tool, a tighter recipe, a different measurement method.

Special cause variation

Something specific happened. Not part of the normal system. Shows up as a break in the pattern.

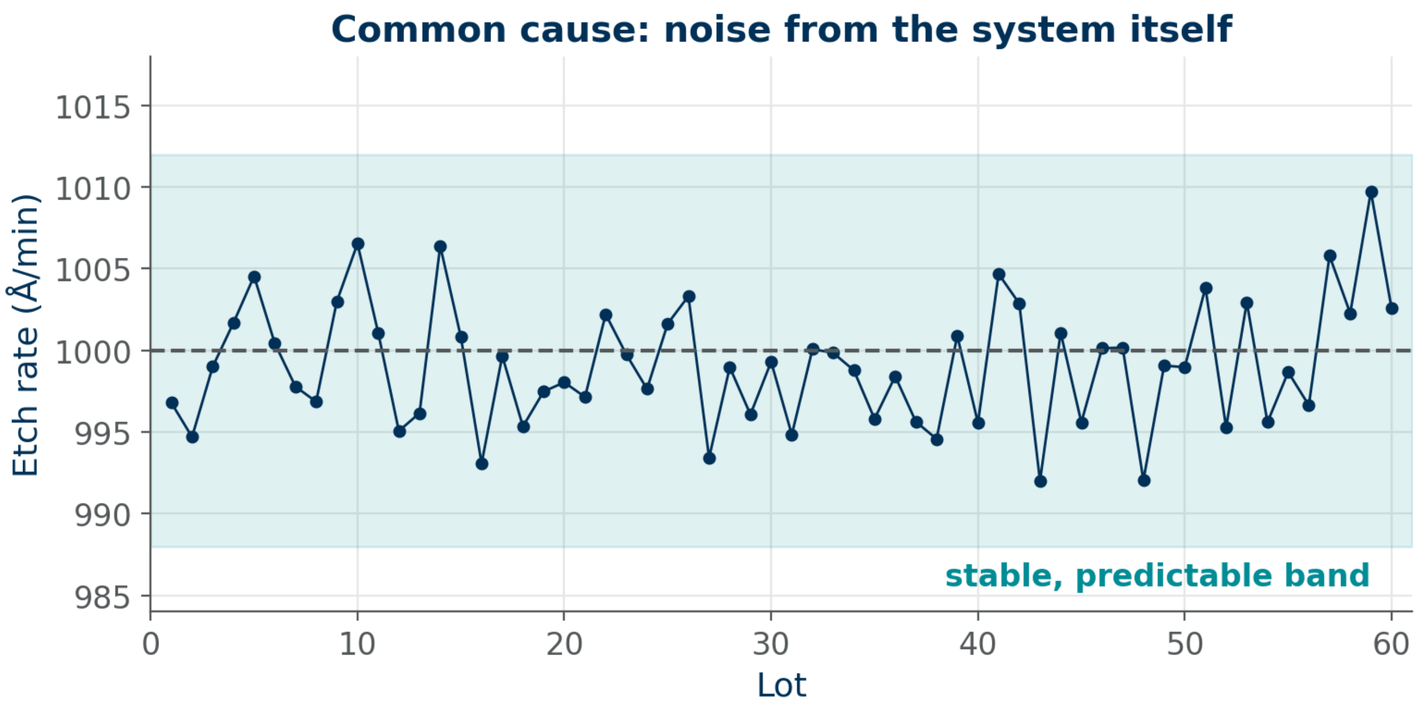
In the fab: a wrong recipe loaded, a gas line leak, a failed RF match, an operator error, a new resist batch.

To fix it you find that specific thing — and the chart tells you roughly when to look.

Different causes demand different responses. Confusing them is expensive in both directions.

Georgia Tech

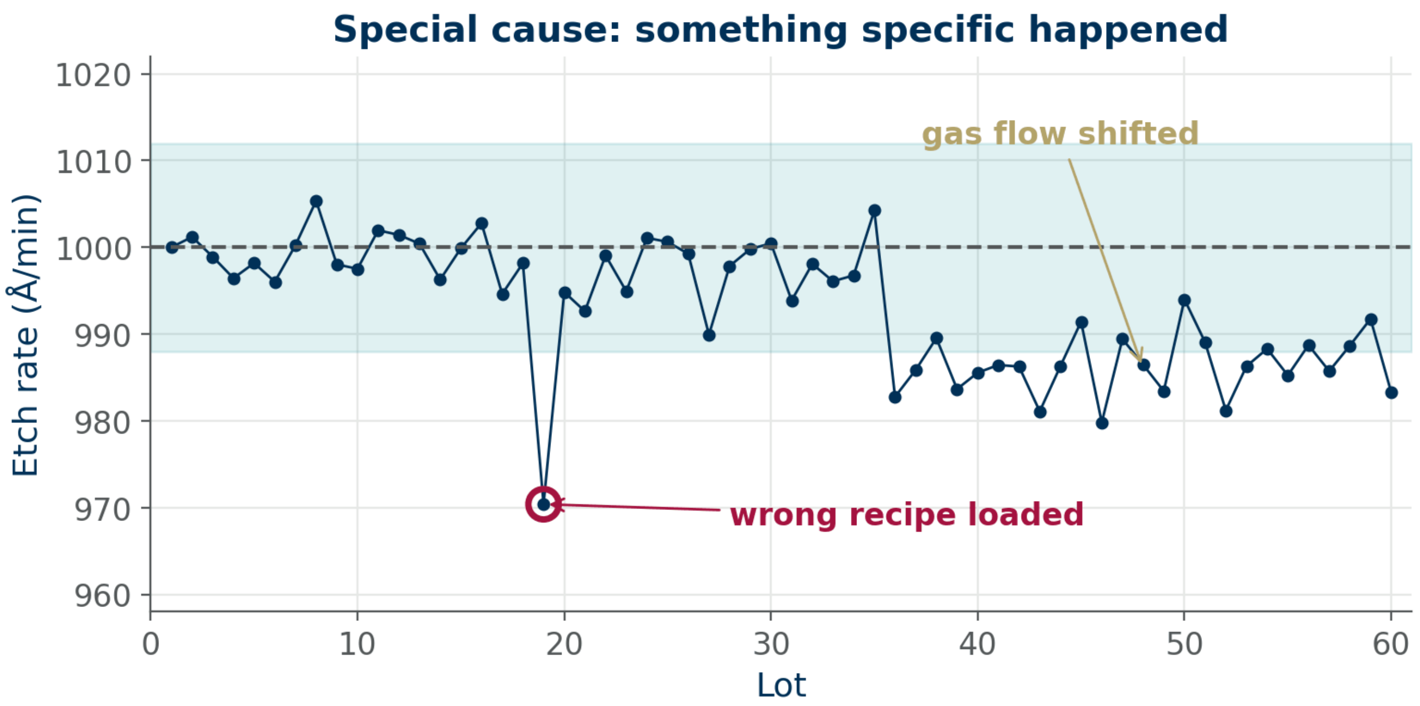
Common cause looks like this



- No point is surprising given the others.
- The band is stable — you could predict tomorrow's range today.
- **Correct response:** leave it alone.
- If the band is too wide for your specs, that is a system problem. Change the tool, the recipe, or the measurement — not this lot.

A stable process is not necessarily a good process. It is a predictable one.

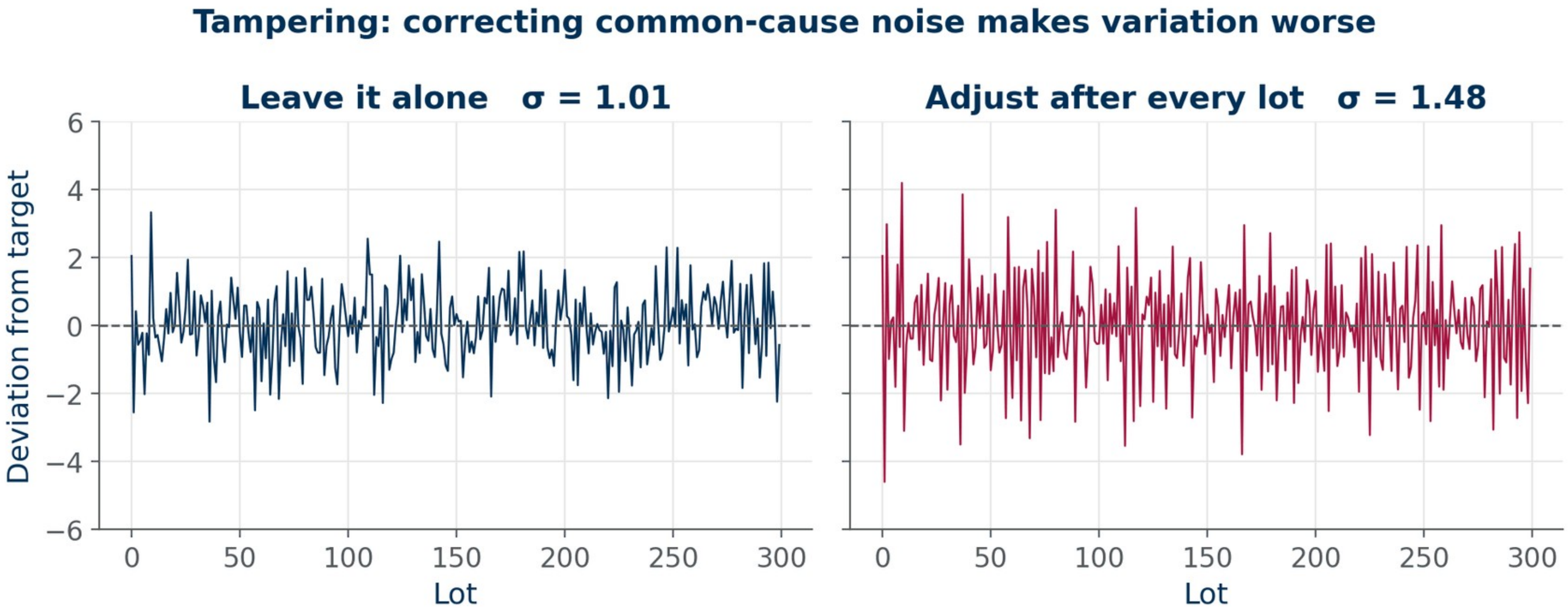
Special cause looks like this



- Two things break the pattern here.
- **A single wild point** — one lot, far outside. Someone loaded the wrong recipe.
- **A sustained shift** — the level moves and stays moved. Gas flow drifted.
- **Correct response:** go find that specific thing.
- The chart does not tell you the cause. It tells you where to look and when it started.

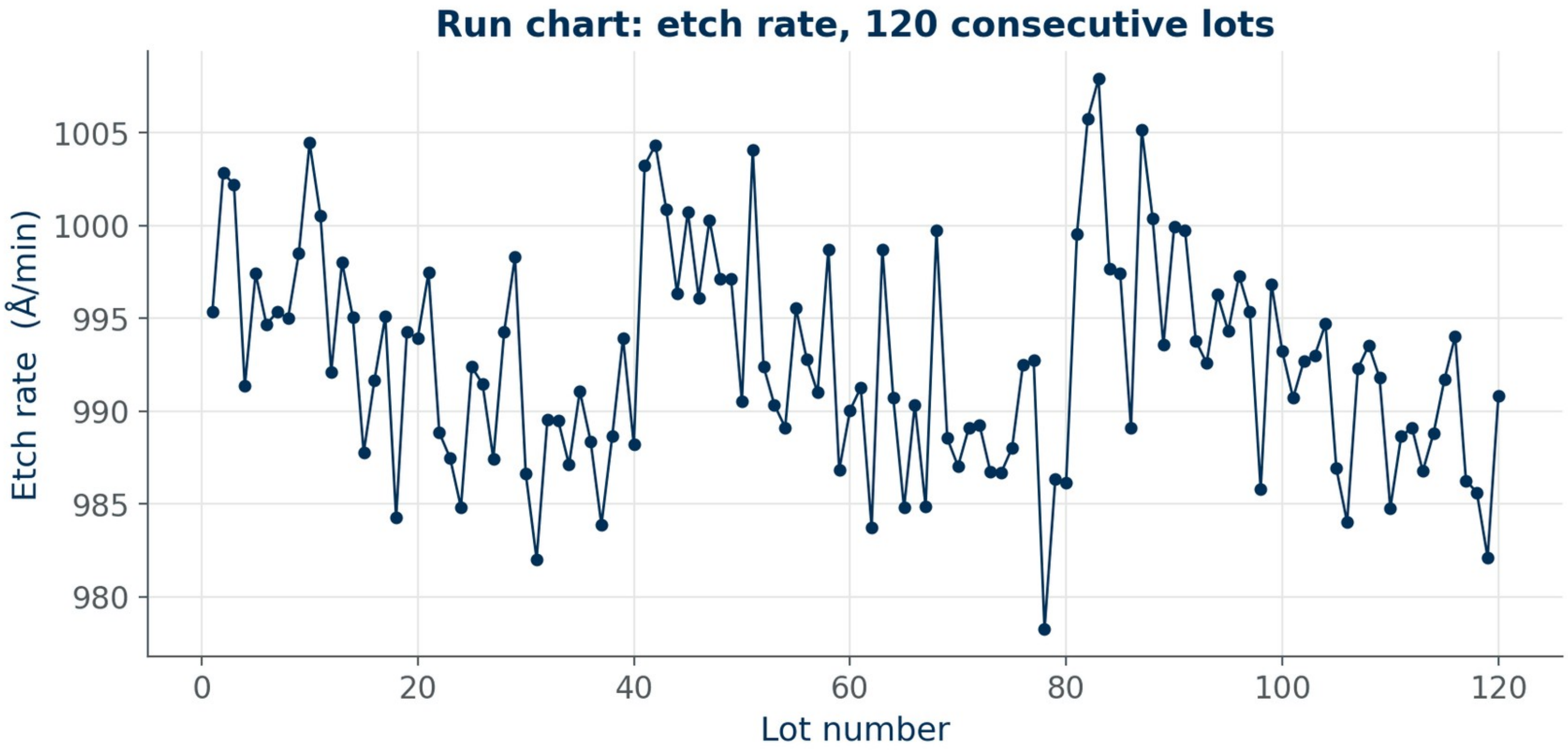
The value of a timestamp: 'it started at lot 36' narrows the search enormously.

Tampering: the cost of over-reacting



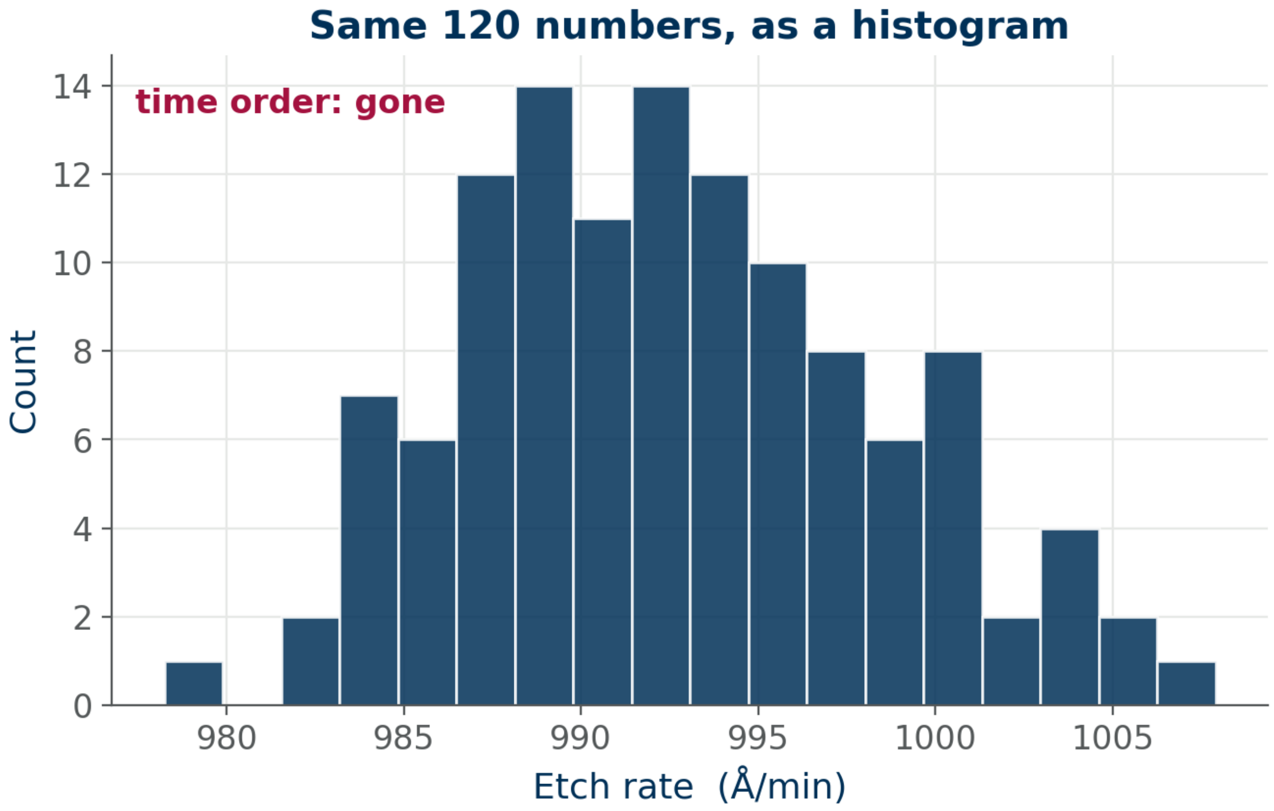
Correcting every deviation more than doubled the variation. Same process, worse output.

Start by plotting the data in time order



Before any statistic, before any limit: what does the process actually do?

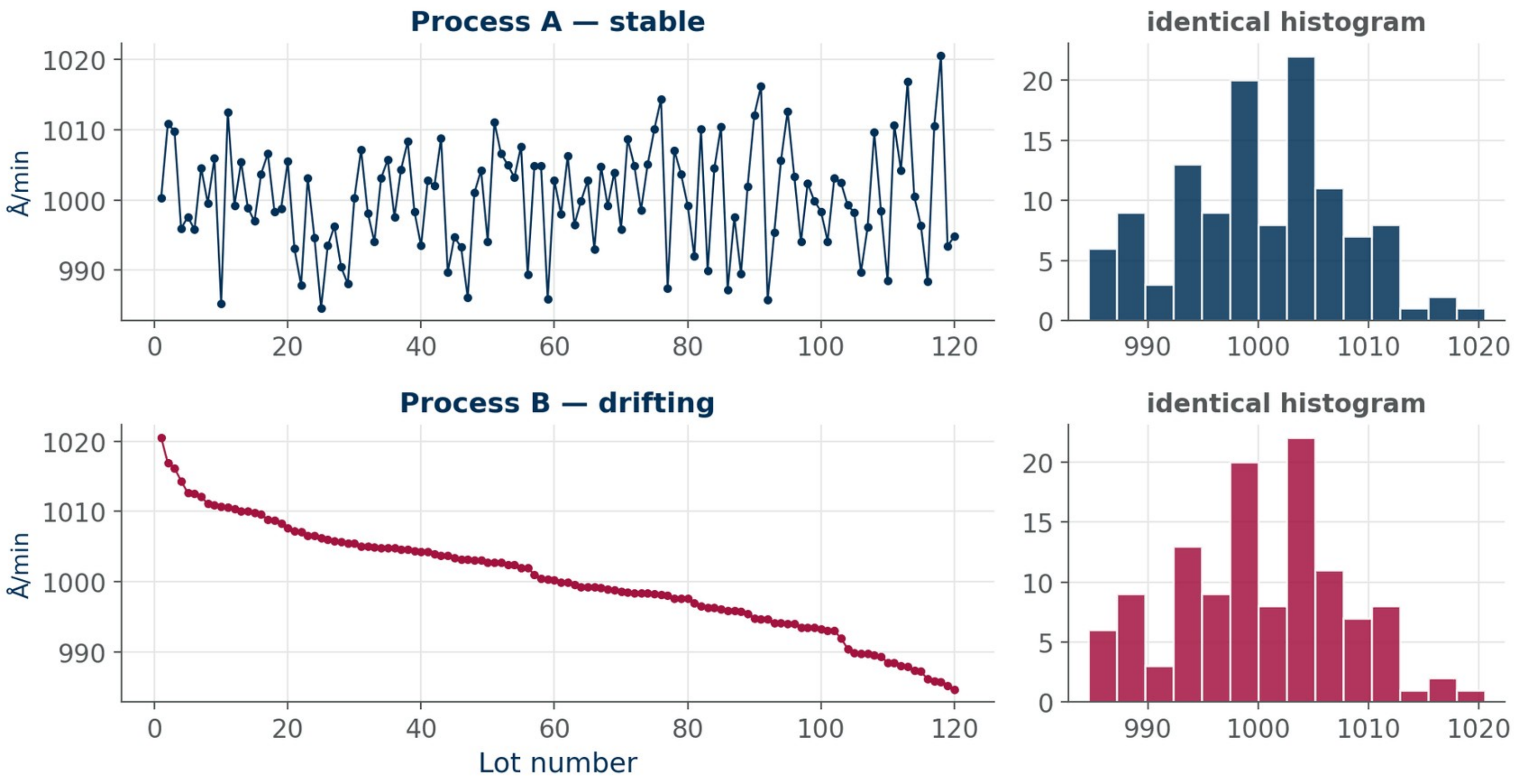
Now collapse it into a histogram



- Same 120 numbers. Nothing has been discarded — except the order they arrived in.
- The histogram tells you the center and the spread.
- **It cannot tell you whether the process is stable.**
- The sawtooth you just saw is completely invisible here.

A histogram is a summary. A run chart is a story.

Two very different processes



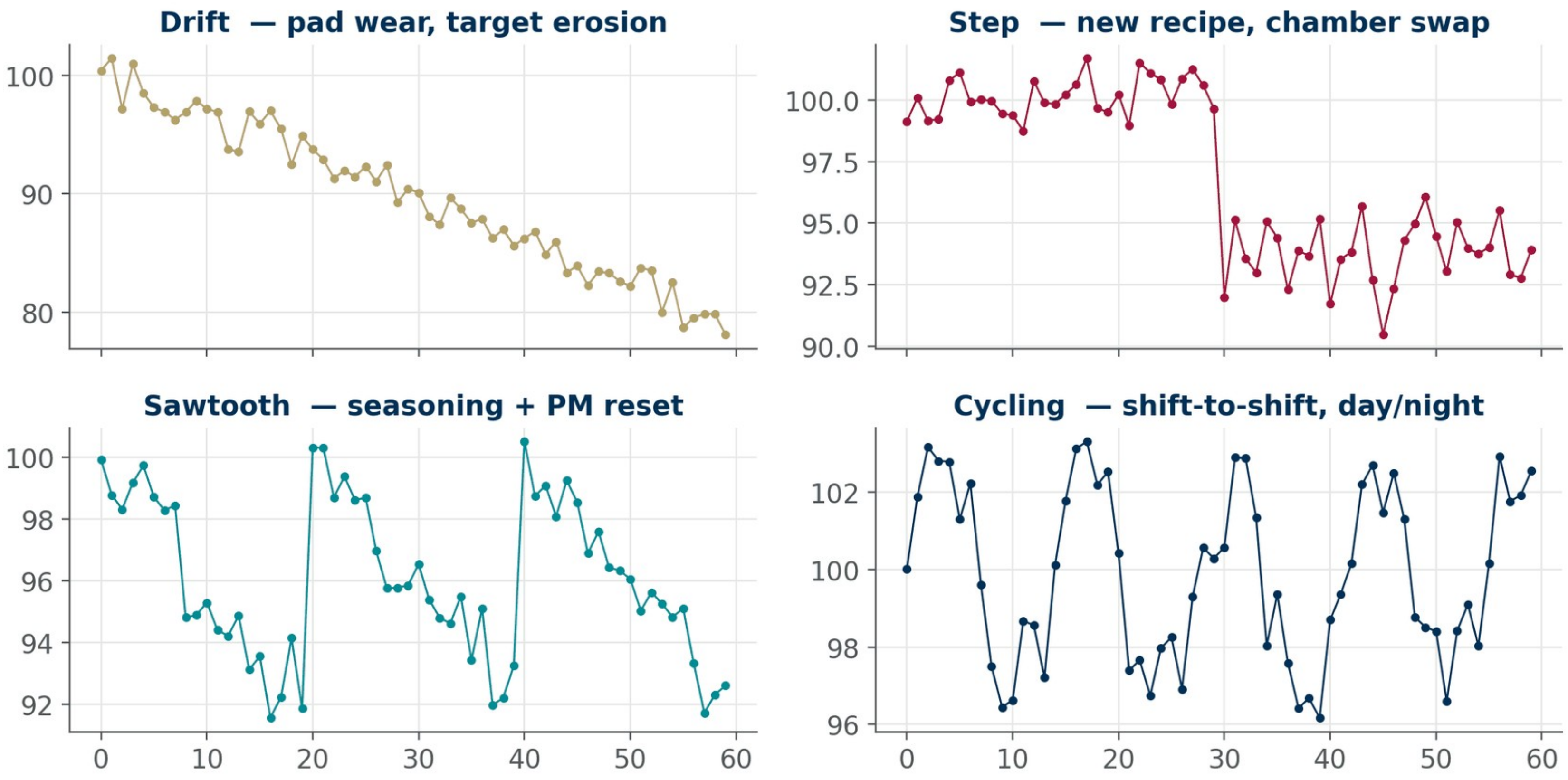
Identical mean. Identical standard deviation. Identical histogram. One of them is failing.

THE CENTRAL IDEA

SPC is about distributions over time.

Not distributions. Not time series. The behaviour of a distribution as time passes.

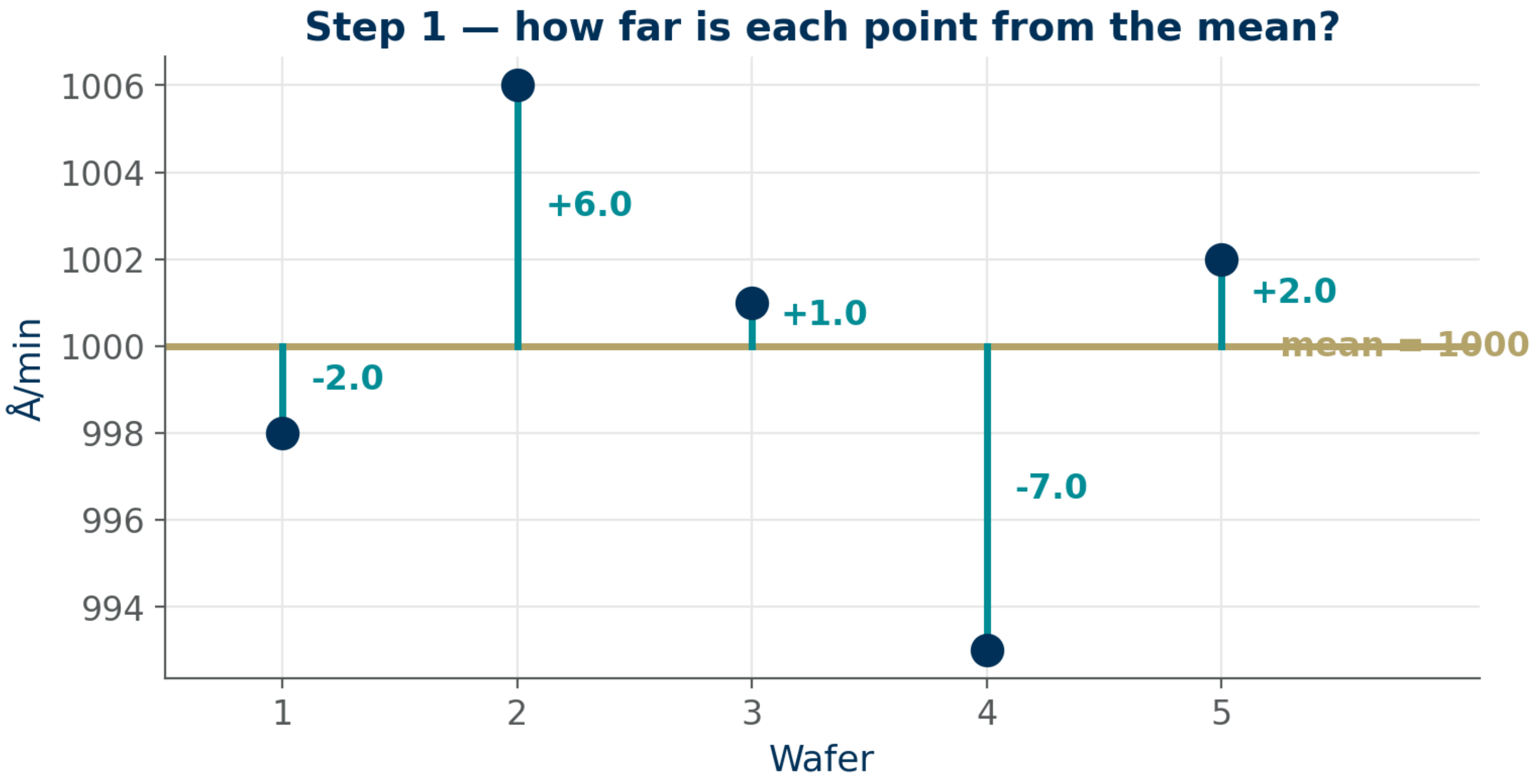
Patterns you will actually see in a fab



Each shape points at a different physical mechanism — the chart narrows the search.

Building the standard deviation — step 1

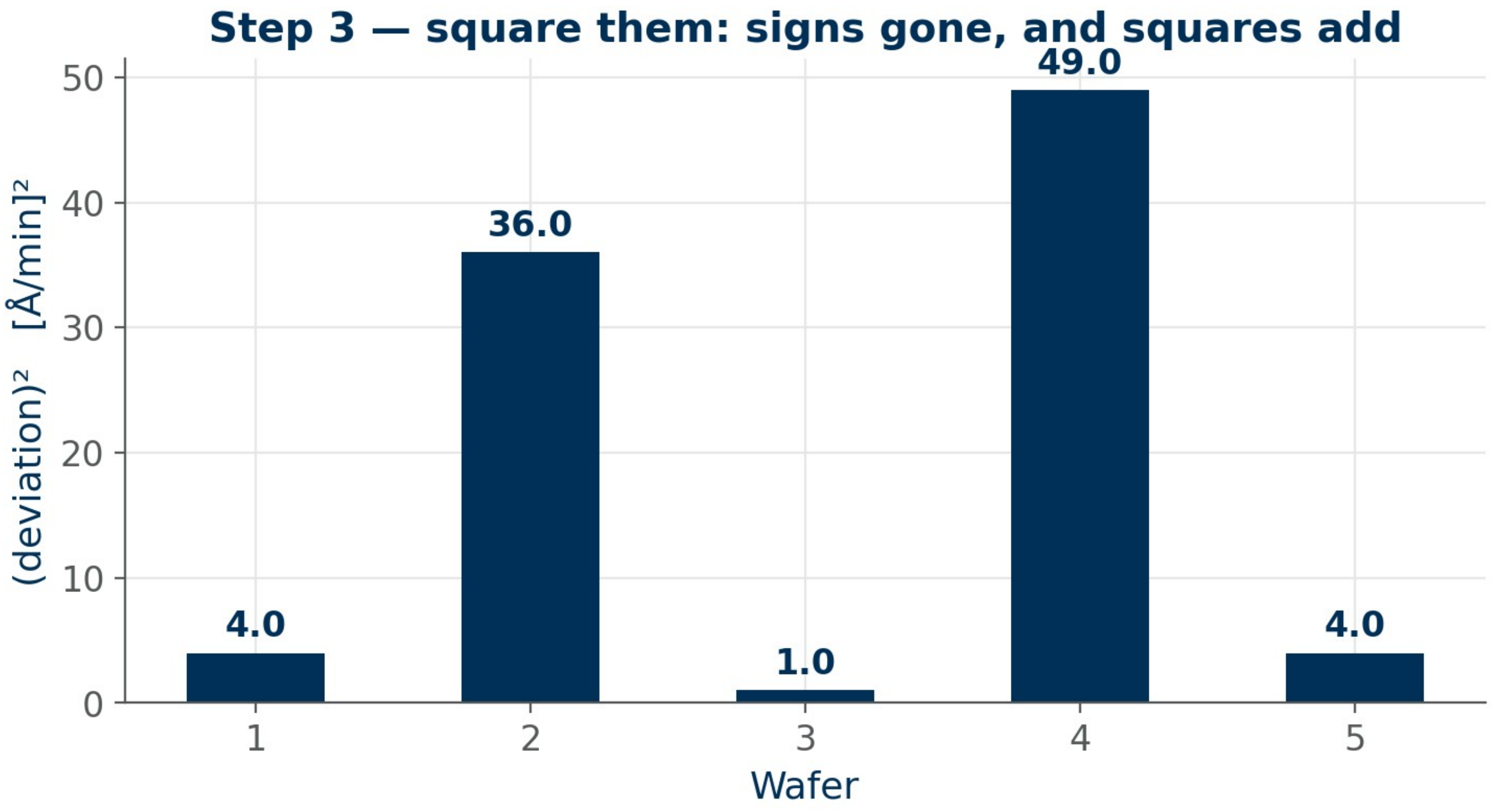
Five wafers from one lot: 998, 1006, 1001, 993, 1002 Å/min



How far is each measurement from the centre? Start with the obvious thing.

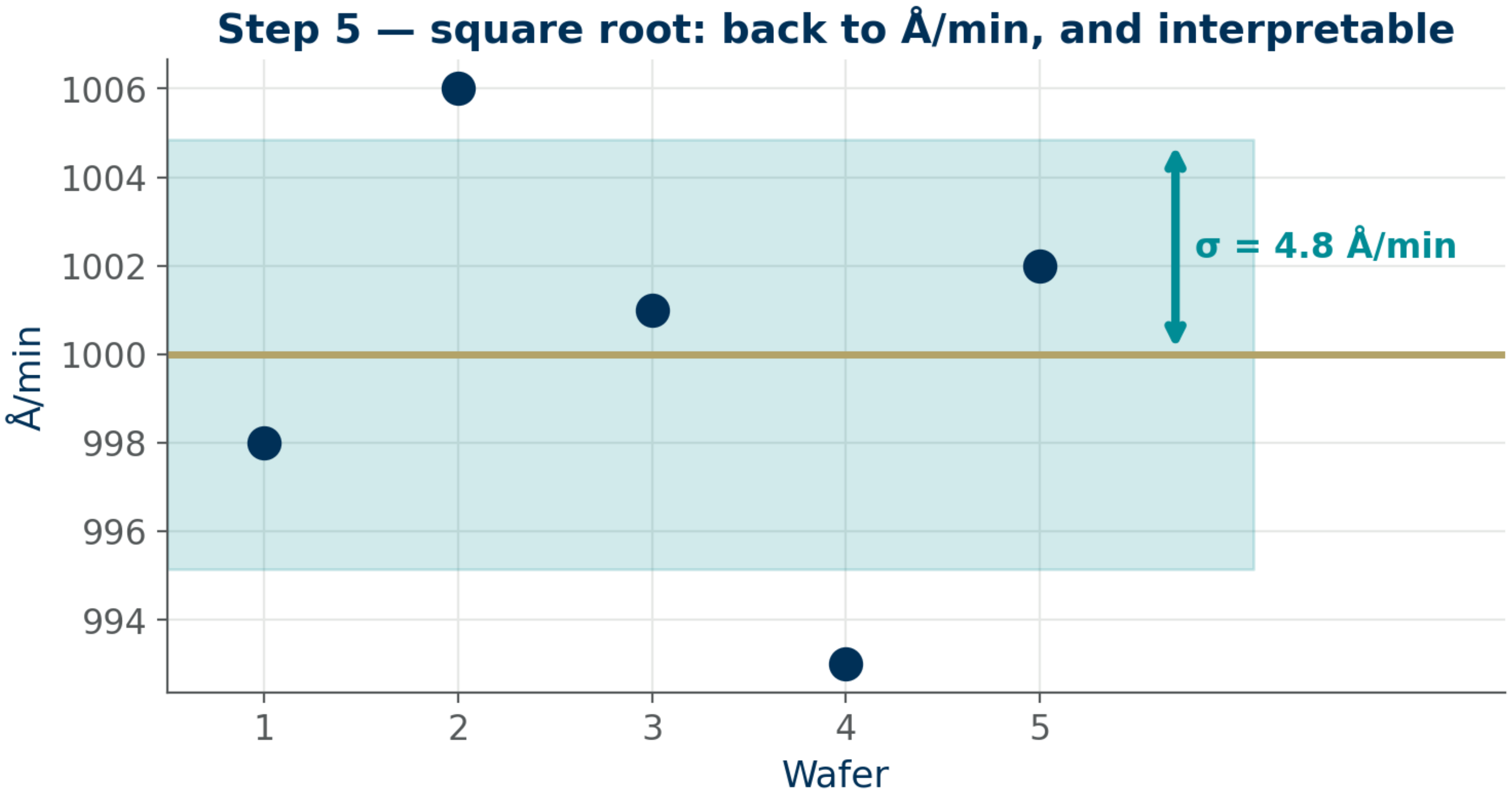
Step 3 — square them

Squares are always positive. And unlike absolute values, squares add.



Why not absolute values? They work — they are just harder to do algebra with, and they do not decompose into variance components.

Step 5 — take the square root



Back to Å/min. Now it means something: the typical distance of a measurement from the centre.

What you now have

Five steps, one number

$$\sigma = \sqrt{\left[\Sigma (x_i - \bar{x})^2 / (n - 1) \right]}$$

What it is

The typical distance from a measurement to the mean, in the original units.

Why $n - 1$

You spent one degree of freedom computing the mean. Four independent deviations remain. With large n it makes almost no difference.

Where it goes next

Control limits are $\pm 3\sigma$. Capability is a tolerance measured in σ. Gauge R&R is a budget of σ^2. It is all this number.

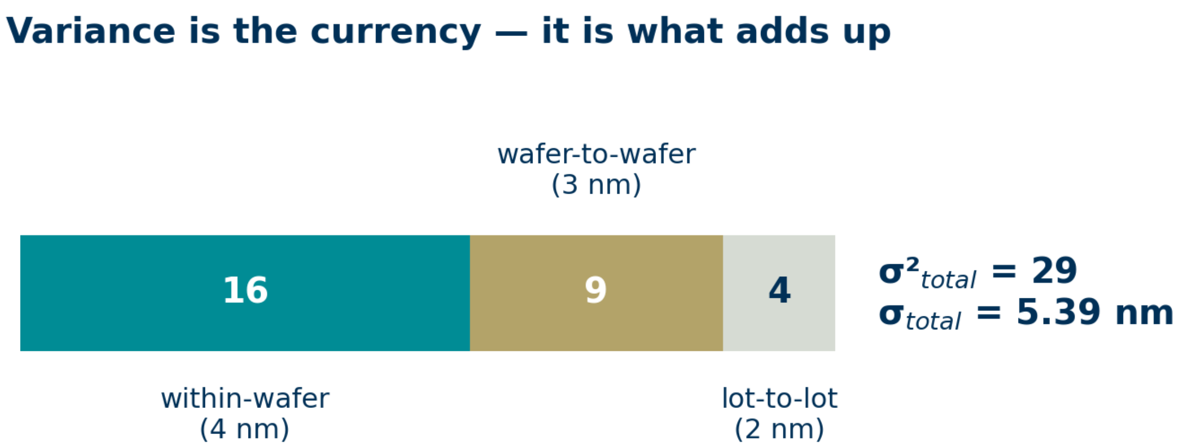
You will not compute this by hand again. You will interpret it constantly.

Georgia Tech

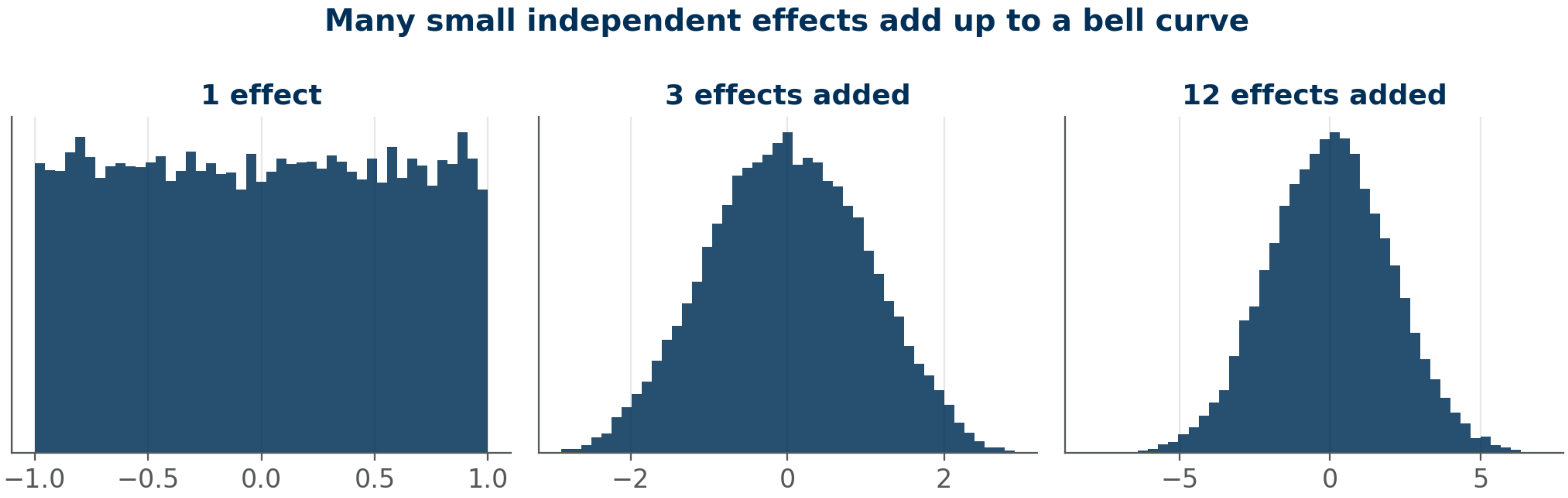
The most useful equation in this course

$$\sigma^2_{\text{total}} = \sigma^2_{\text{within-wafer}} + \sigma^2_{\text{wafer-to-wafer}} + \sigma^2_{\text{lot-to-lot}}$$

Variances add. Standard deviations do not.

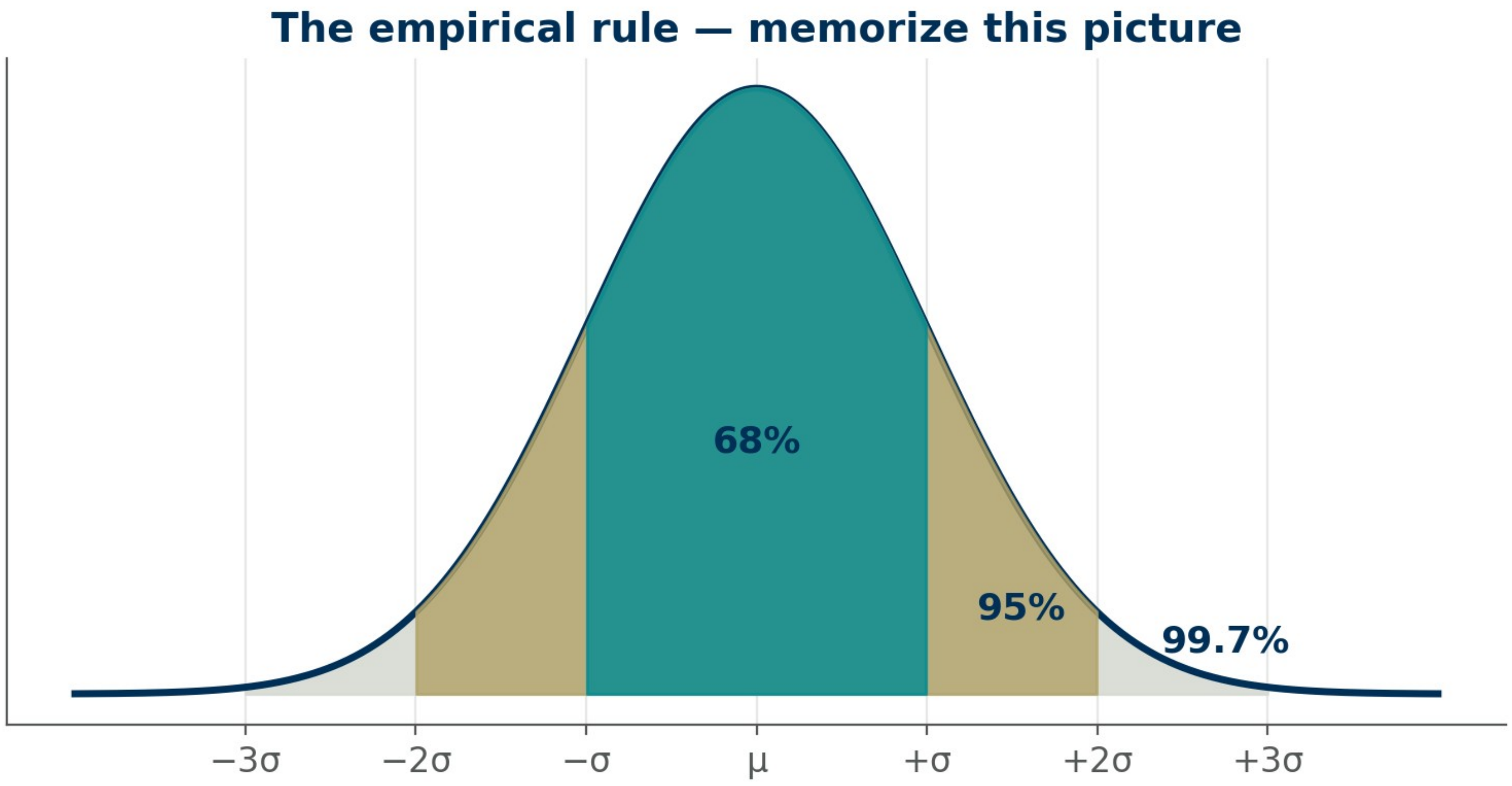


Watch it happen



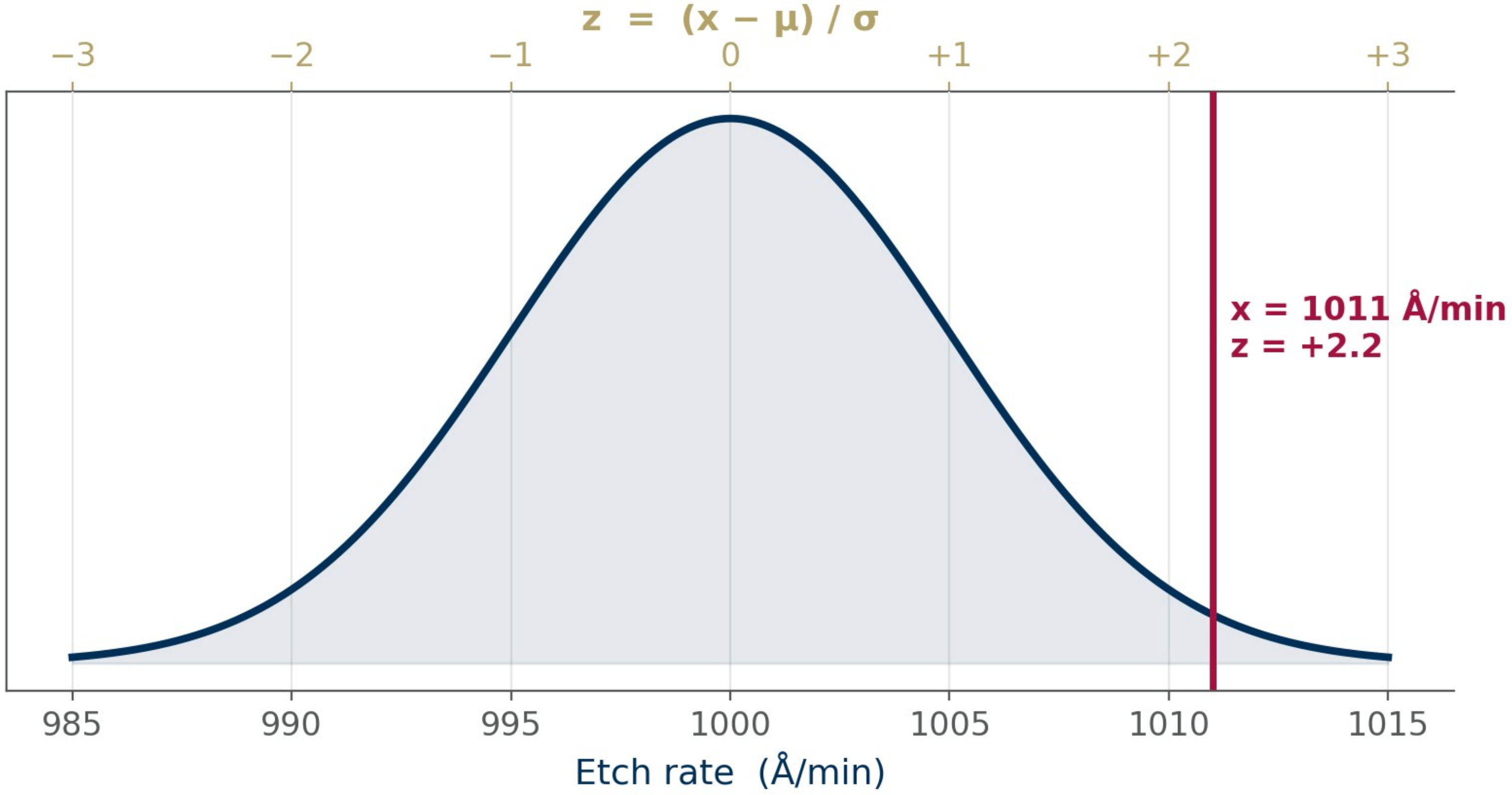
Uniform random effects. Add three and it is already bell-shaped. Add twelve and it is indistinguishable from normal.

The empirical rule



Roughly two thirds within ±1σ, 95% within ±2σ, 99.7% within ±3σ.

The z-score: one ruler for everything



$z = (x - \mu) / \sigma$. "How many standard deviations from centre?"

If you remember one thing from today

One idea. Three names.

Control limits

UCL / LCL sit at

$z = \pm 3$

A point beyond them is a signal.

Lecture 2

Capability

Cpk = (nearest spec limit in σ) / 3

It is a z-score, divided by three.

Lecture 2

Defect rate

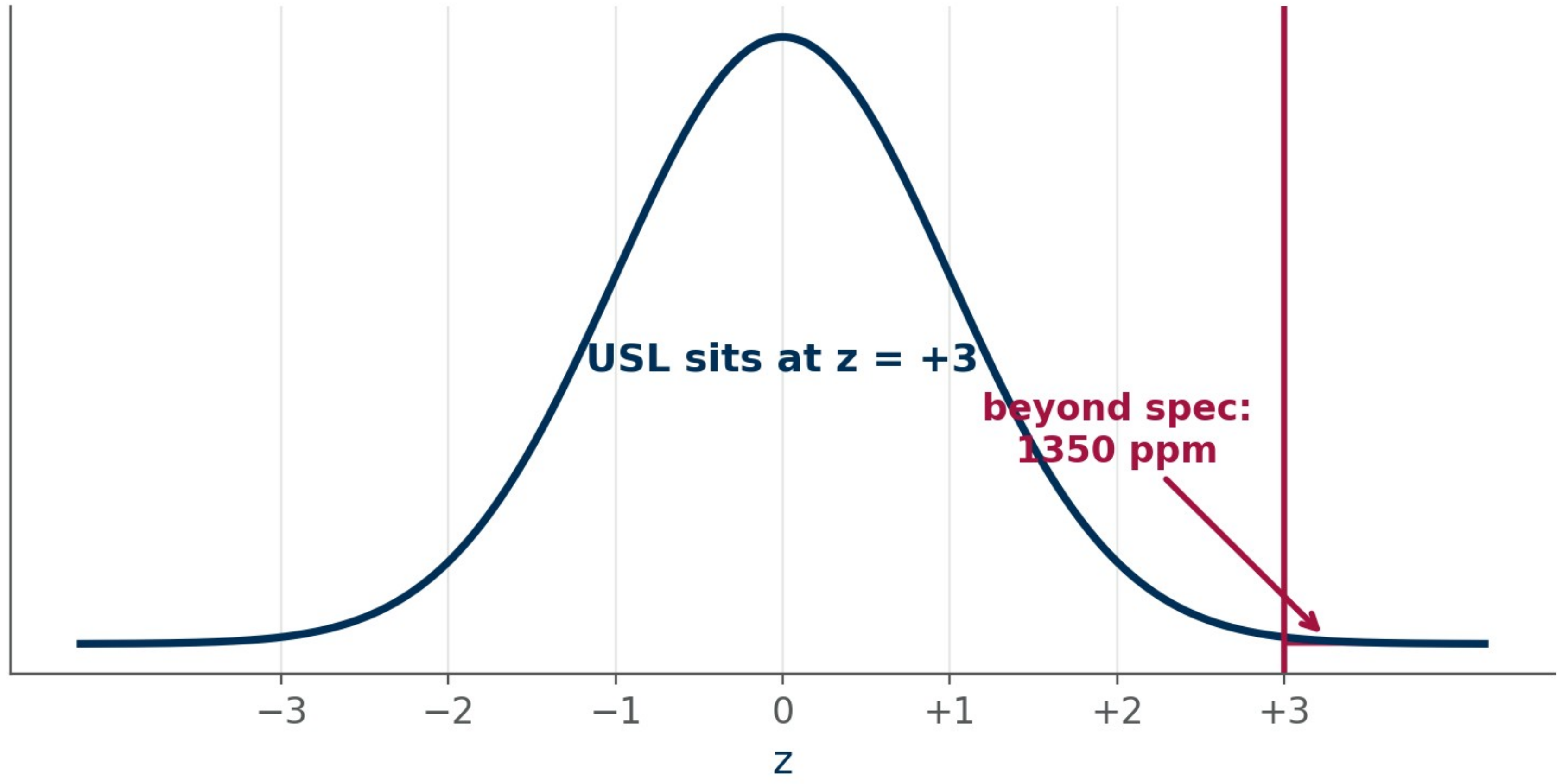
ppm out of spec = the tail area beyond that z

Read straight off a table.

Today

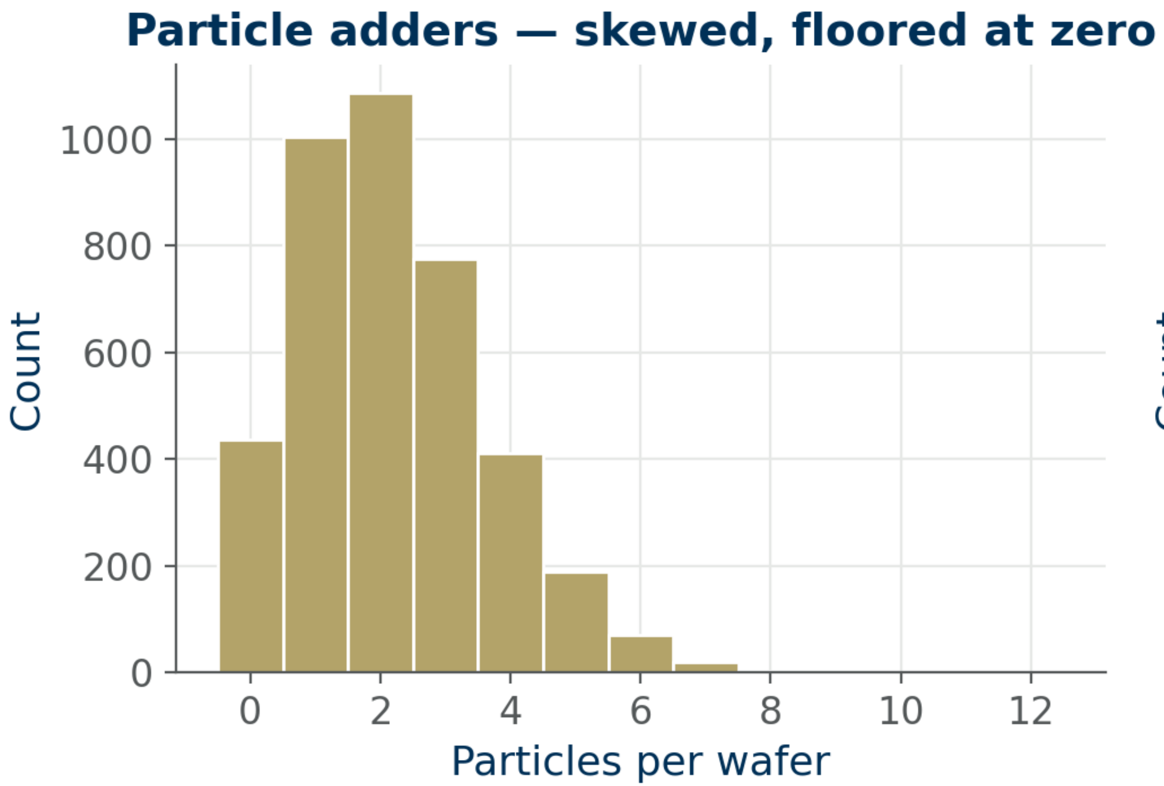
Three different vocabularies, three different textbook chapters, one distance measured in σ.

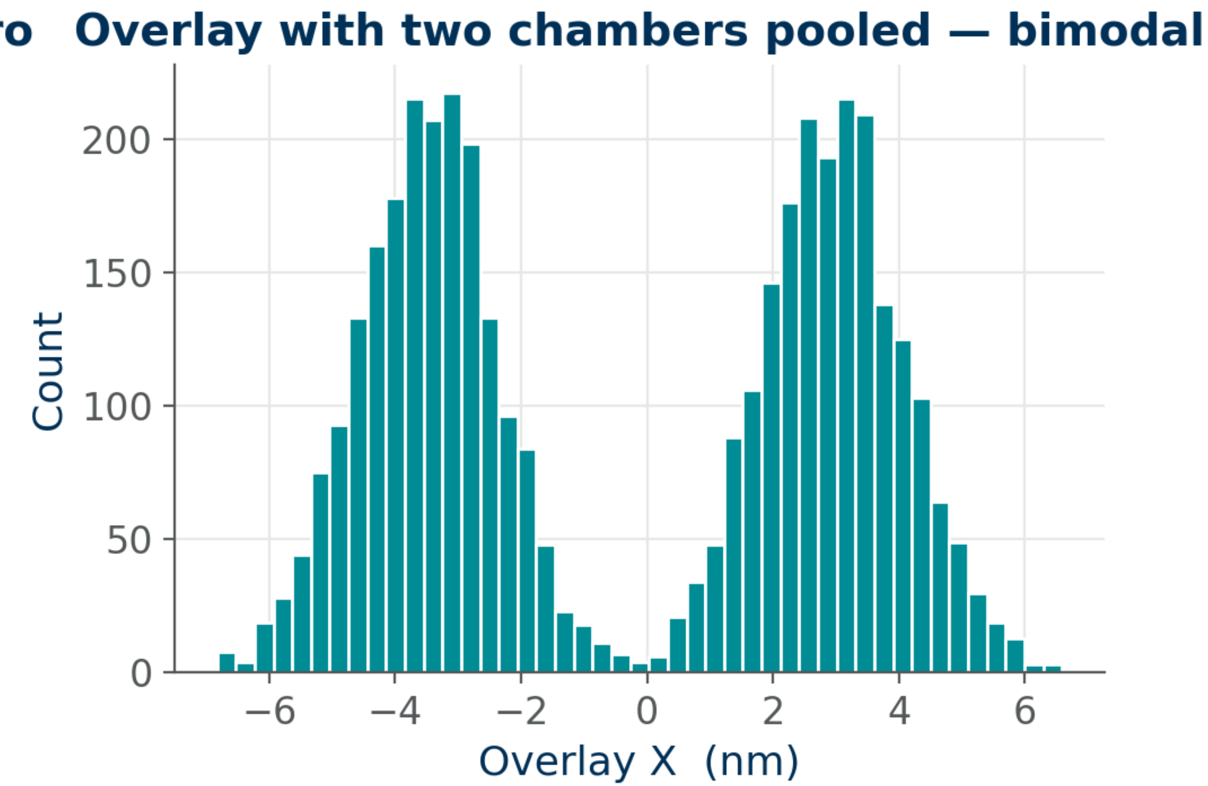
Worked example — one-sided spec



Gate CD target 45 nm, σ = 0.8 nm, upper spec 47.4 nm. z = (47.4 − 45)/0.8 = +3.0 → 1350 ppm above spec.

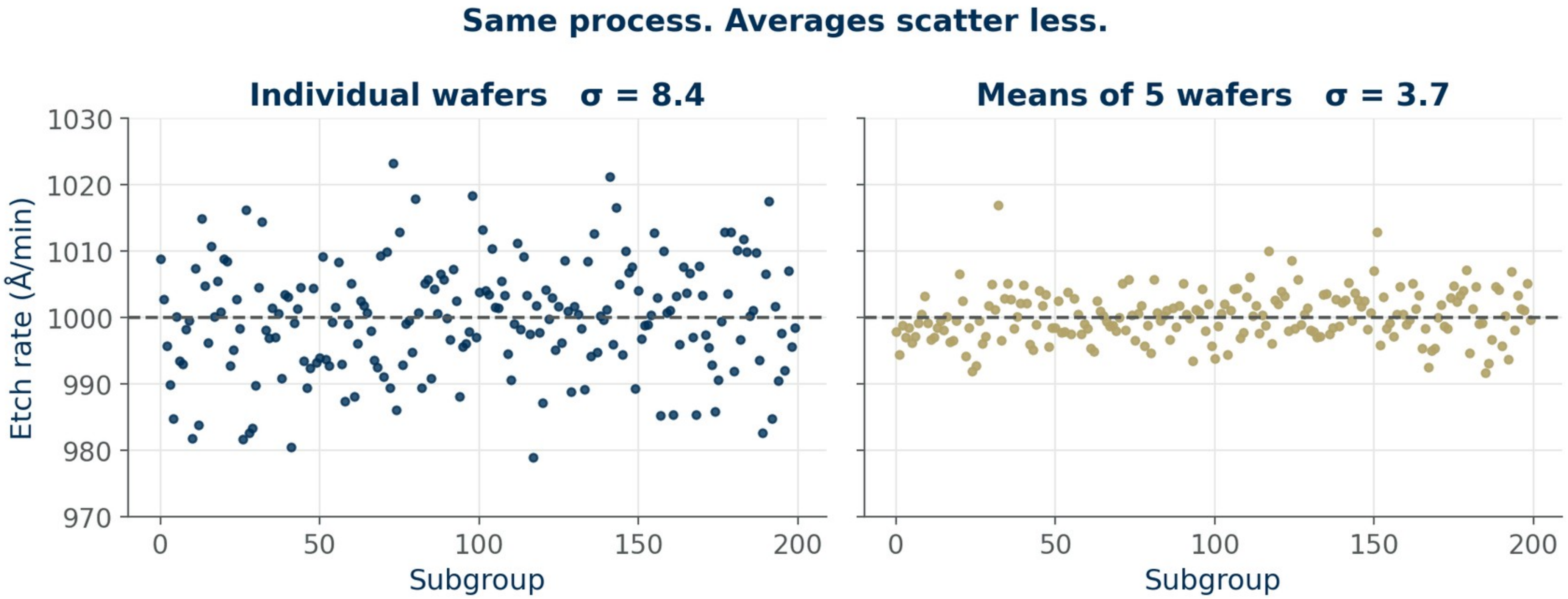
When the normal model does not hold





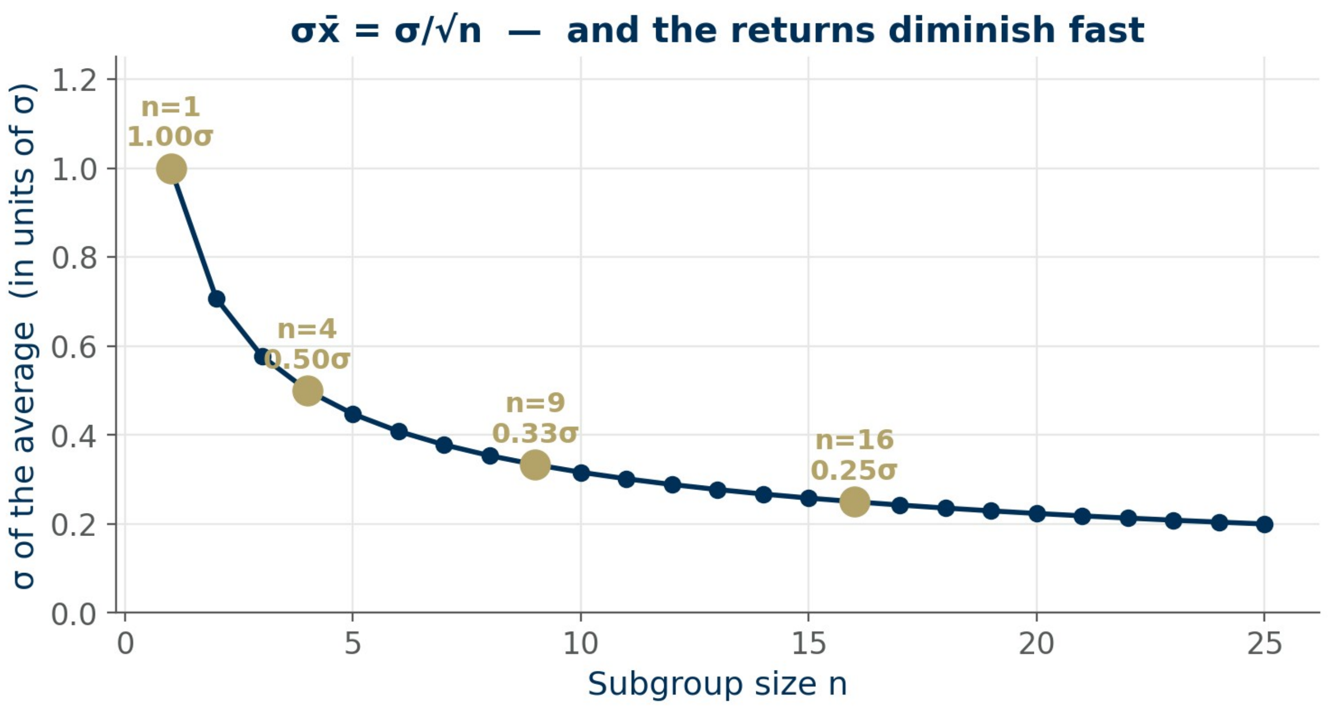
Two failure modes you will meet constantly in fab data.

Same process, two ways of looking at it



Nothing about the process changed. Only what we chose to plot.

Why: σ of the average is σ/√n



- Averaging four wafers halves the noise. Averaging nine cuts it to a third.
- **The returns diminish fast.**
- Going from n = 1 to n = 4 buys you a factor of two. Going from n = 16 to n = 25 buys almost nothing.
- In a fab this is a cost argument as much as a statistical one — every extra measurement is metrology tool time you cannot spend elsewhere.
- Which is why subgroups of 4 or 5 are the near-universal choice.

Subgroup size is chosen where the sensitivity curve flattens against the metrology budget.

The central limit theorem

In words, not symbols

Averages tend toward a normal distribution even when the individual measurements do not.

Why you should care

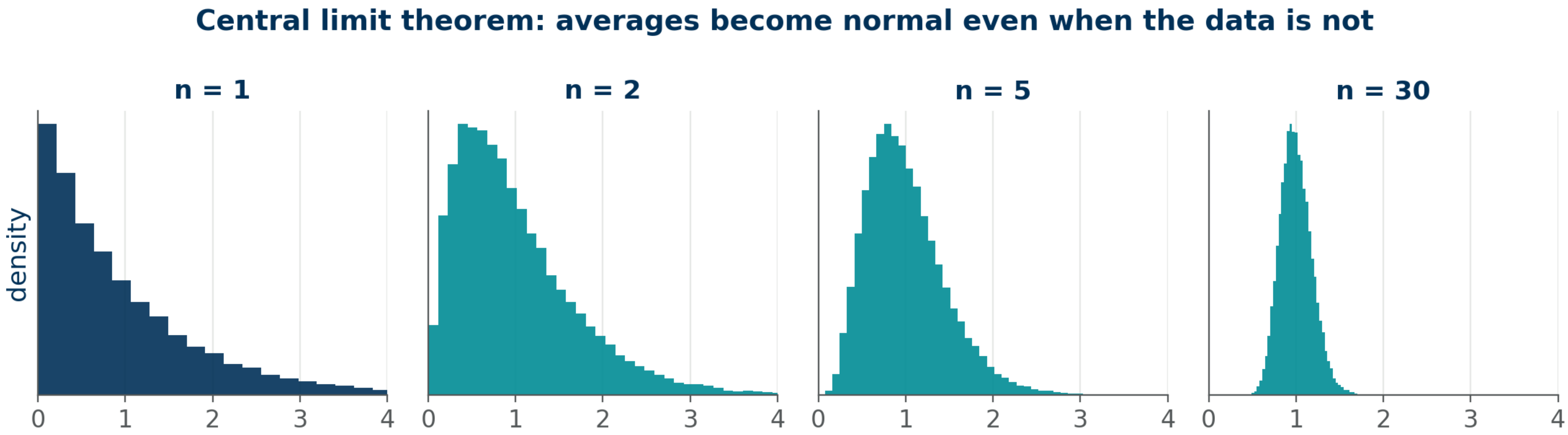
It means the ±3σ machinery we just built applies to averages of almost any measurement, whether or not that measurement is normally distributed.

The catch

It says nothing about individual values. Charts that plot individuals inherit whatever shape the raw data has — including skew.

Georgia Tech

Watch it work on badly non-normal data



The parent distribution is exponential — strongly skewed. By n = 30 the average is nearly symmetric.

The consequence you should write down

$\bar{X}$ charts are robust

They plot averages, so the central limit theorem is working for you. Moderate non-normality in the raw measurements barely affects the false-alarm rate.

You can usually apply them without checking the distribution.

I–MR charts are not

They plot individual measurements, so the chart inherits the raw distribution's shape. Skew produces asymmetric false-alarm rates.

You must check normality — or transform the data first.

And the individuals chart is the one you will use most in a fab,

because metrology is expensive and you often measure one site on one wafer per lot.

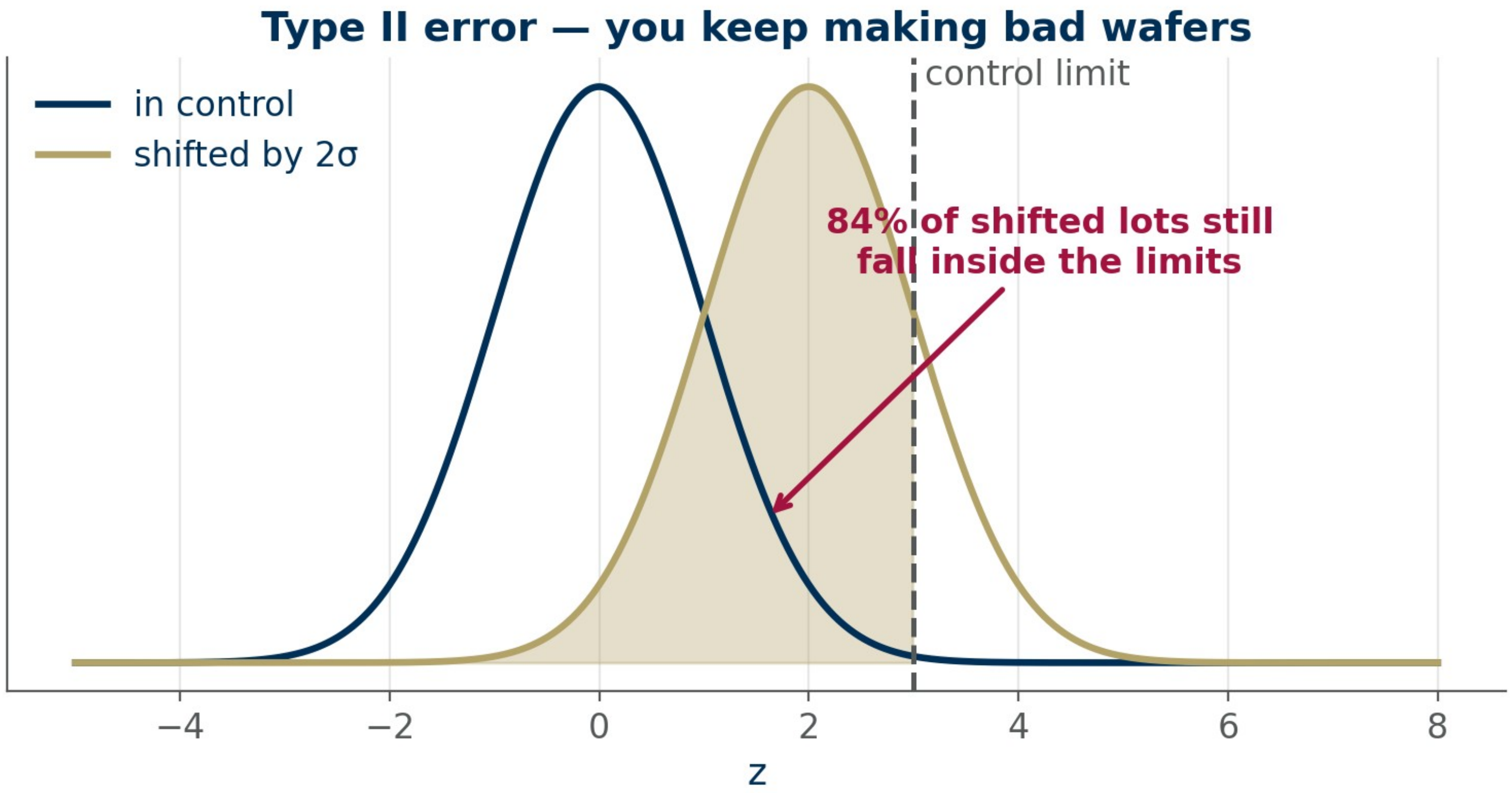
The chart with the weakest assumptions is the one fabs rely on hardest. Check your data.

THE ONLY QUESTION

Has the process changed?

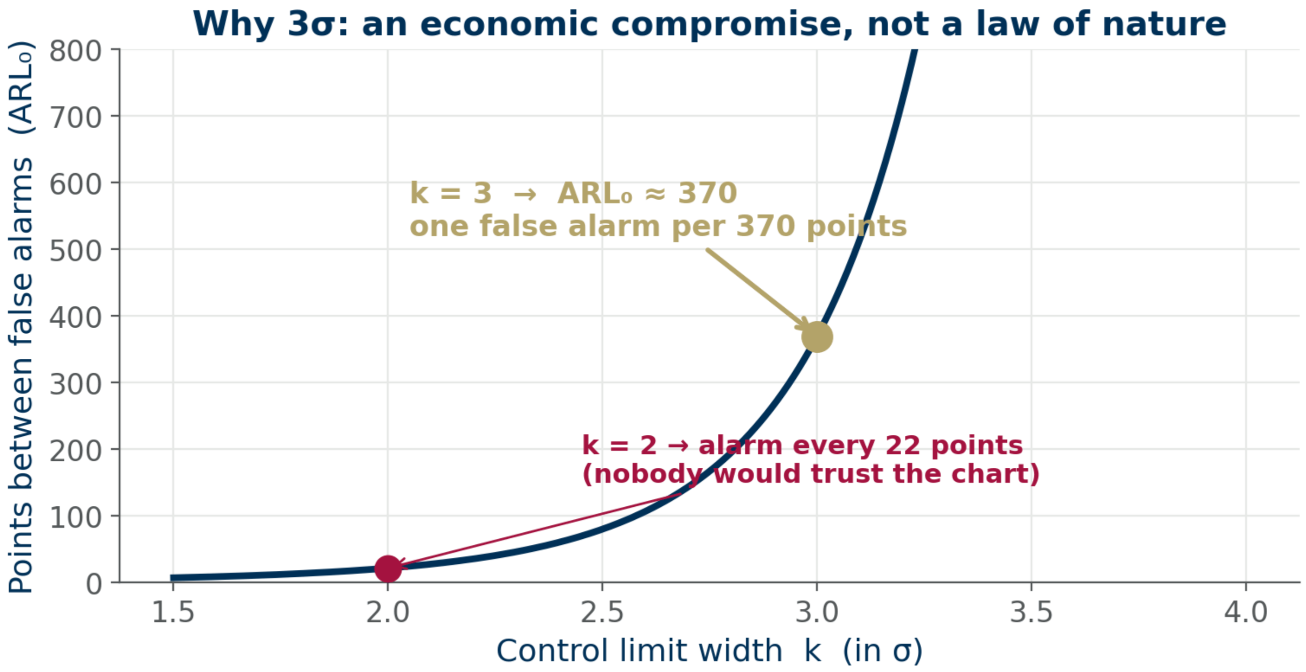
Every control chart, every run rule, every anomaly detector in this course is answering this one question.

Type II error — the missed shift



The process really has shifted by 2σ — and 84% of lots still land inside the limits.

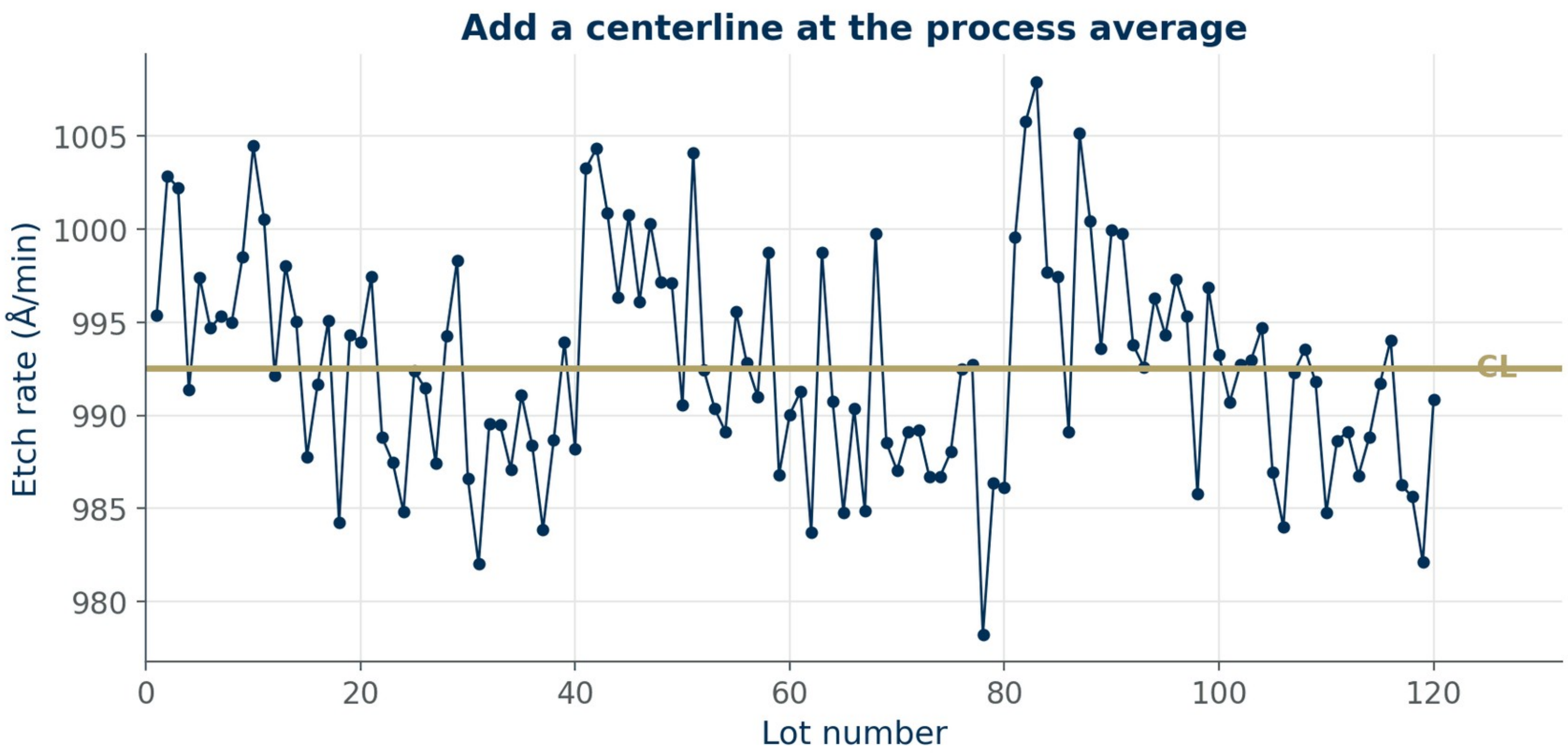
So why 3σ?



- A point beyond $\pm 3\sigma$ happens 0.27% of the time by pure chance.
- **That is one false alarm per 370 points.**
- Roughly one a month if you chart daily — annoying, but tolerable.
- Tighten to 2σ and you alarm every 22 points. The chart gets ignored within a week.
- **Shewhart chose 3 as an economic compromise in 1924.** It is a convention, not a law of nature — and you are allowed to change it if your cost structure differs.

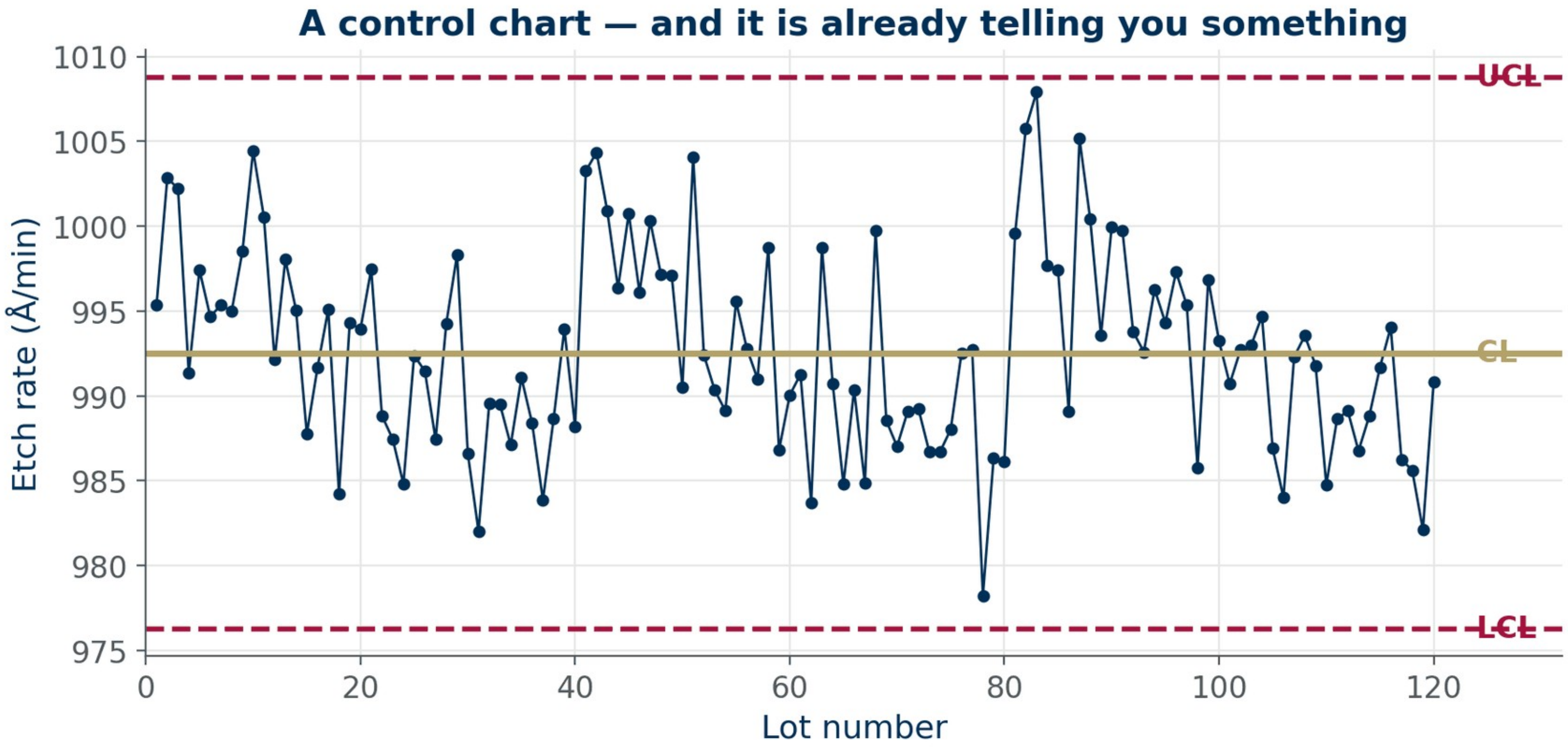
Every threshold in this course is a business decision wearing a statistical costume.

Back to where we started



The same etch rate data. Add one line at the process average.

You have just built a control chart



Next lecture: X̄–R, I–MR and p-charts, what the patterns inside the limits mean, and process capability.

Homework

Due before Lecture 2 · dataset on the course site

1. Compute mean, σ, and z-scores for the supplied etch rate dataset. Estimate ppm beyond each spec limit and report the implied parametric yield.
2. Given σ within-wafer = 2.4 nm, σ wafer-to-wafer = 1.8 nm and σ lot-to-lot = 1.1 nm, find total σ. Then invert it: to reach a total of 2.5 nm, which single term must you attack, and how far?
3. Sampling simulation. From the same dataset, plot the distribution of individual measurements and of means of n = 5. Report the ratio of their standard deviations and compare it to √5.
4. Short answer. Give one fab example of a common cause and one of a special cause. State the correct engineering response to each, and what it would cost to get them backwards.

References and pre-reading

Core texts

May & Spanos, Fundamentals of Semiconductor Manufacturing and Process Control — Ch. 1–4 for the fab context.

Montgomery, Introduction to Statistical Quality Control — Ch. 3 for the statistics, Ch. 5–6 for next lecture.

Deming, Out of the Crisis — the red bead experiment and the funnel.

Before Lecture 2

Read Montgomery Ch. 5.1–5.4.

Work the supplied notebook through the sampling-distribution section — the $\bar{X}$–R chart derivation assumes you have seen $\sigma/\sqrt{n}$ in action.

Bring a question about your own research data if it has a monitoring problem.

Next: Lecture 2 — Control Charts, Out-of-Control Signals, and Process Capability

$\bar{X}$–R · I–MR · p-charts · Western Electric rules · Cp and Cpk

Georgia Tech