Ch 6: control charts, continuted.

Stepping back a bit...the idea behind control charts is broader than it appears.

- Use existing data summaries to make decisions about particular data points.
- Using time to help understand a process and how it changes.

Another situation that arises in the real-world is the desire to identify the time at which a process changes.

Statistical control charts can sometimes be used for this, depending on the process. Other times, subgroup averages might not be available or might not make sense.

A time series is a sequence measurements of a variable points ordered in time:

$$y_1, y_2, \ldots, y_N$$

Many time series of interest have measurements taken at regular intervals, such as hourly, daily, monthly, or annually.

Examples:

- daily temperature or precipitation
- daily stock market closing prices
- number of customers who visit a store each day
- number of visits to an urgent care facility each day
- the average snow pack per year
- average speeds of top 100 marathoners each year

A change point of the time series is a time $k \in \{1, \ldots, N-1\}$ at which a change in $y_1, y_2, \ldots, y_N$ occurs. That is, $y_1, \ldots, y_k$ and $y_{k+1}, \ldots, y_N$ differ based on some summary characteristic.

Why do we care about change points? The ability to detect changes is useful for understanding patterns and making decisions.

Examples:

- Is there a change in weather patterns?
- Is there a change the performance of a company you invest in?
- is there a change in the specifications of a product?
- Is there a change in the number of people with some illness?

Techniques that seek to identify such changes are classified as "change point detection" methods. In statistics, these methods seek to detect a change in the underlying distribution that governs a process.

For example, this can be step-like change in the mean, median, slope, standard deviation, or even the correlation between quantitative variables.

One such statistical technique is based on E.S. Page's work in 1954 (Biometrika).

Given $y = (y_1, \ldots, y_N)$.

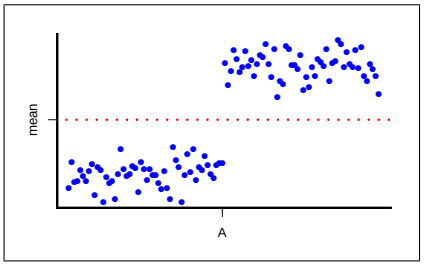
- Find $\bar{y} = \frac{1}{N} \sum_{i=1}^{n} y_i$.
- Create $S = (S_1, \ldots, S_N)$: set $S_1 = y_1 - \bar{y}$, and for $i = 2, \ldots, N$, set $S_i = S_{i-1} + (y_i - \bar{y})$.

Question: the procedure is for finding a change in what parameter?

Let's focus on some intuition building of S with y.

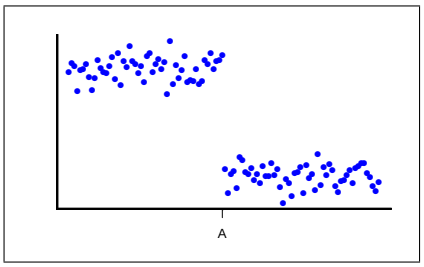
Task 1. Describe, in words, what S_{i-1} is telling us.

Consider a simple case: the time series undergoes a step-like change. For example, the time series could represent the number of daily visits to an emergency department, which holds steady at around 100 people (plus some noise) per day for x days, but then it jumps up at day A to 150 people per day for another x days. Then the average number of visits to the emergency department is right in the middle: 125 visits. Below the time series simulating the described scenario appears....

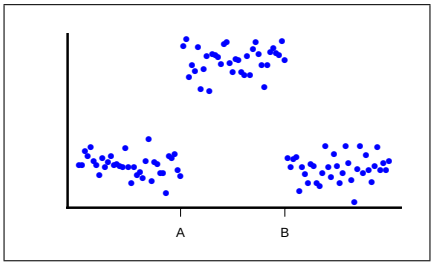


Sketch what you think the new time series S will look like.

Task 2. Let's repeat this thought experiment for a time series with a step-like change in the opposite direction.



Task 3. What will S look like for the time series below?



Task 4.1. We'll check our work. Rather than doing this by hand (each data set has 108 observations!), how do we tell a computer to do it?

- How to write a function in R.

```
myMean = function(x) {
    return (sum(x)/length(x))
}
```

  How to use the myMean function in R:

```
> testVector = c(1,2,3,4,5)
> myMean(testVector)     ## The mean ought to be 15/5=3.
```

- Recall, S is a new time series, based on the original time series. We can represent it as a vector in R. Below we will create a function for producing this vector.

```
my_S_Vector = function(y) {
     ## The argument, or input ``y" is the original time series.
     ## N is the length of the time series
     N = length(y)
     ## Here we need to initialize a vector of zeroes for S.
     ## What does rep(0, times=5) do?
     S =
     ## How do we populate the entries
     ## of the S vector?
     ## Hint:  Set S[1] = (what it says in the notes)
     ##      then use a "for" loop to
     ##      populate S[2], ...  S[N]









     ## What should be returned, aka the output of the function?
     return(      )
}
```

In R, type up the code you wrote down here into the file CPD_activity.R.

Task 4.2. Write out by hand what you should get for the entries of your S vector if the original time series is: `y=c(1,1,1,4,4)`.

$$\bar{y} =$$

i	y_i	$y_i - \bar{y}$	S_i
1			
2			
3			
4			
5			

Task 4.3. Test your `my_S_Vector` function on the vector `c(1,1,1,4,4)`. Compare the result with your by-hand calculation. Do the results agree?

Task 5. Load in the data sets `Toy1.csv, Toy2.csv and Toy3.csv` from your `CPD_activity.R` file and run the `mySvector` function on them. Do the resulting graphs agree with your predictions in Tasks 1-3?

Integral Review

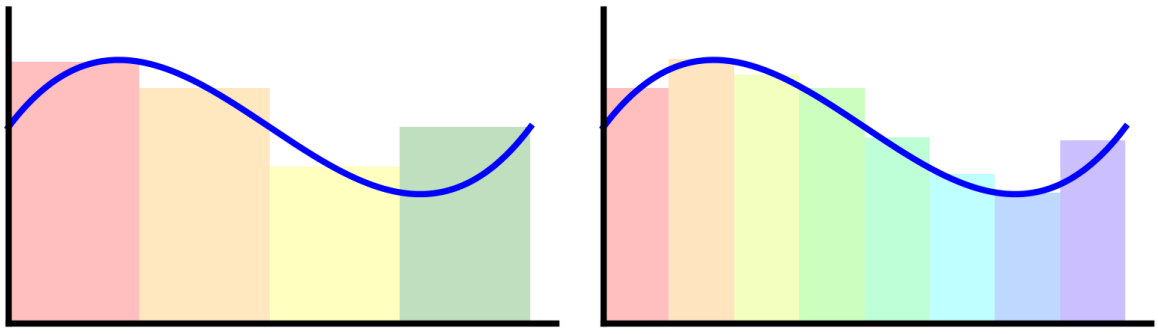
Let's jump ship for a moment and review some Calculus II.

Question: What was Calculus II about?

Question: How are the objects from Calculus II used?

Question: How are the objects from Calculus II defined?

A picture is worth a thousand words.



You may recall that the Riemann Sum is difficult to calculate analytically.

You often need a "trick," even for basic functions.

For example, you may recall the following tricks for integrating the elementary function x^2:

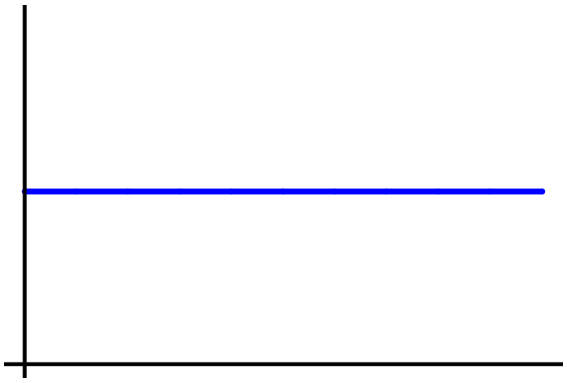
$$\begin{aligned}
\int_0^3 x^2\,dx &= \lim_{n\to\infty} \sum_{i=1}^{n} \left(\frac{3i}{n}\right)^2 \cdot \frac{3}{n} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \frac{9i^2}{n^2} \cdot \frac{3}{n} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \frac{27i^2}{n^3} \\
&= \lim_{n\to\infty} \frac{27}{n^3} \cdot \sum_{i=1}^{n} i^2 \\
&= \lim_{n\to\infty} \frac{27}{n^3} \cdot \frac{n(n+1)(2n+1)}{6} \\
&= \lim_{n\to\infty} \frac{9}{n^2} \cdot \frac{(n+1)(2n+1)}{2} \\
&= \lim_{n\to\infty} \frac{9}{2} \cdot \frac{(n+1)}{n} \cdot \frac{(2n+1)}{n} \\
&= \lim_{n\to\infty} \frac{9}{2} \cdot \left(1+\frac{1}{n}\right)\left(2+\frac{1}{n}\right) \\
&= \frac{9}{2} \cdot (1+0) \cdot (2+0) = 9.
\end{aligned}$$

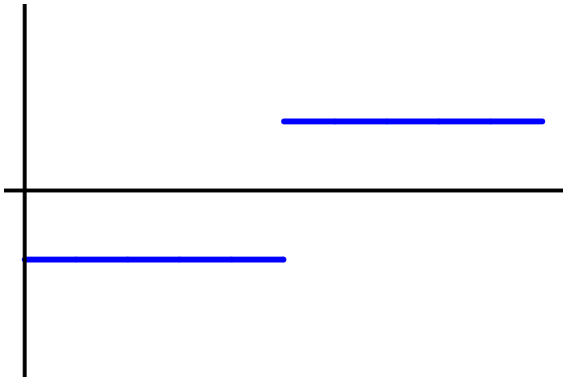
Question: What was the key piece that made calculating definite integrals easier?

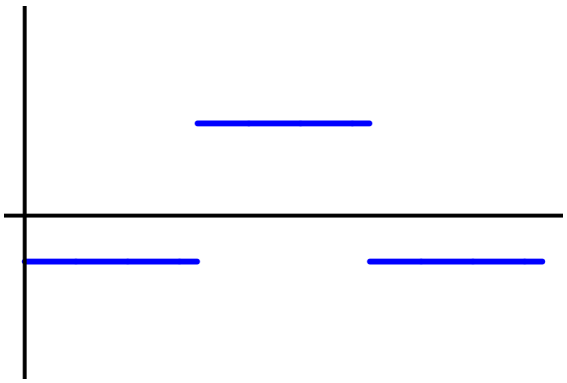
Let's put it to the test:

$$\int_0^3 x^2 \, dx \quad =$$

Task 6. Let's practice sketching some antiderivatives of functions.







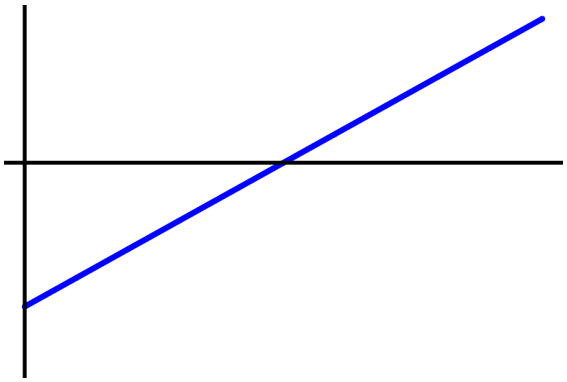
Question: Do any of these shapes look familiar?
Compare with what you did on the computer.

Question: What is the connection?

The connection

$f(x)$ y

Final Task. Let's test the connection with one more example. Sketch an indefinite integral of the function below.



Now, let's load in a time series, graph the time series, and then run your S vector code and graph the resulting time series.

In R, load in the file `Simulated_data.csv`, plot it, and plot its "S" vector.