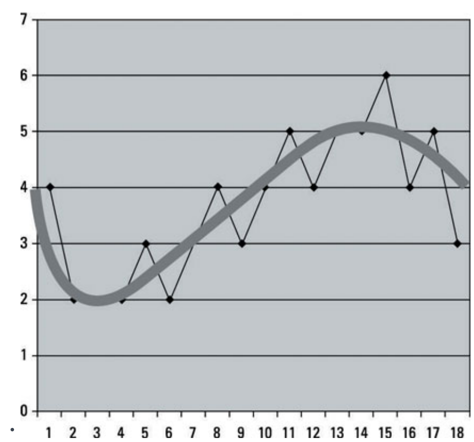
Data Smoothing

To remove irregularity from the data

- Noise removal from normal data
- Data filtering
- Outlier removal etc.

In time series data (Continuous signal)

- Smoothing is usually done to help us better see patterns (trends in time series data)
- smooth out the irregular roughness to see a clearer signal
- Like other DM procedures, smoothing doesn't result into a model,, but it can be a good first step in describing various components of the series which will help to prepare the model
- The implication behind data smoothing is that the data consists of two parts:
    - part (consisting of the core data points) that signifies overall trends or real trends, and
    - part consisting mostly of deviations (noise) — some fluctuating points that result from some volatility in the data.
    - **Data smoothing seeks to eliminate that second part.**



- Data smoothing operates on several assumptions:
    - That fluctuation in data is likeliest to be noise.
    - That the noisy part of the data is of short duration.
    - That the data's fluctuation, regardless of how varied it may be, won't affect the underlying trends represented by the core data points.
- Noise in data tends to be random and its fluctuations should not affect the overall trends drawn from examining the rest of the data.
- Reducing or eliminating noisy data points can clarify real trends and patterns in the data and improv the data's "signal-to-noise ratio."
    - Smoothing will not accurately predict the exact value e.g. price of the next trade for a given stock, but predict the trend

- predicting a general trend can yield more powerful insights than knowing the actual price or its fluctuations.

Smoothing using binning (Already discussed in last class)

Unsorted data for price in dollars
Before sorting: 8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34
First of all, sort the data
After Sorting: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

1. Smoothing the data by equal frequency bins

Bin 1: 8, 9, 15, 16
Bin 2: 21, 21, 24, 26,
Bin 3: 27, 30, 30, 34

- **Smoothing by bin means**

  For Bin 1:
  (8+ 9 + 15 +16 / 4) = 12
  Bin 1 = 12, 12, 12, 12

  For Bin 2
  (21 + 21 + 24 + 26 / 4) = 23
  Bin 2 = 23, 23, 23, 23

  For Bin 3:
  (27 + 30 + 30 + 34 / 4) = 30
  Bin 3 = 30, 30, 30, 30

- **Smoothing by bin boundaries**
  - pick the minimum and maximum value.
  - Put the minimum on the left side and maximum on the right side.
  - Middle values in bin boundaries move to its closest neighbor value with less distance.

Before bin Boundary: Bin 1: 8, 9, 15, 16
Here, 8 is the minimum value and 16 is the maximum value. 9 is near to 8, so 9 will be treated as 8. 15 is more near to 16 and farther away from 8. So, 15 will be treated as 16.

After bin Boundary: Bin 1: 8, 8, 16, 16

Before bin Boundary: Bin 2: 21, 21, 24, 26,
After bin Boundary: Bin 2: 21, 21, 26, 26,

Before bin Boundary: Bin 3: 27, 30, 30, 34
After bin Boundary: Bin 3: 27, 27, 27, 34

The Mean value and Standard Error

Case : A manager of a warehouse wants to know how much a typical supplier delivers in 1000 dollar units. He/she takes a sample of 12 suppliers, at random, obtaining the following results:

Supplier	Amount
1	9
2	8
3	9
4	12
5	9
6	12
7	11
8	7
9	13
10	9
11	11
12	10

- The computed mean or average of the Amount = 10.
- The manager decides to use this as the estimate for *expenditure of a typical supplier.*

Is this a good or bad estimate?

To estimate the judgement we shall calculate the error in the judgement

- The "error" = true amount spent minus the estimated amount.
- The "error squared" is the error above, squared.
- The "SSE" is the sum of the squared errors.
- The "MSE" is the mean of the squared errors (Lower the value, better the estimate).

The estimate = 10

Supplier	$	Error	Error Squared
1	9	-1	1
2	8	-2	4
3	9	-1	1
4	12	2	4
5	9	-1	1
6	12	2	4
7	11	1	1
8	7	-3	9
9	13	3	9
10	9	-1	1
11	11	1	1
12	10	0	0

SSE = 36

MSE = 36/12 = 3

Same way We will calculate the MSE assuming the estimated value to be 7,9 and 12 instead of

Estimator	**7**	**9**	**10**	**12**
SSE	144	48	36	84
MSE	12	4	3	7

Ideally the estimation of **mean** will give the best result

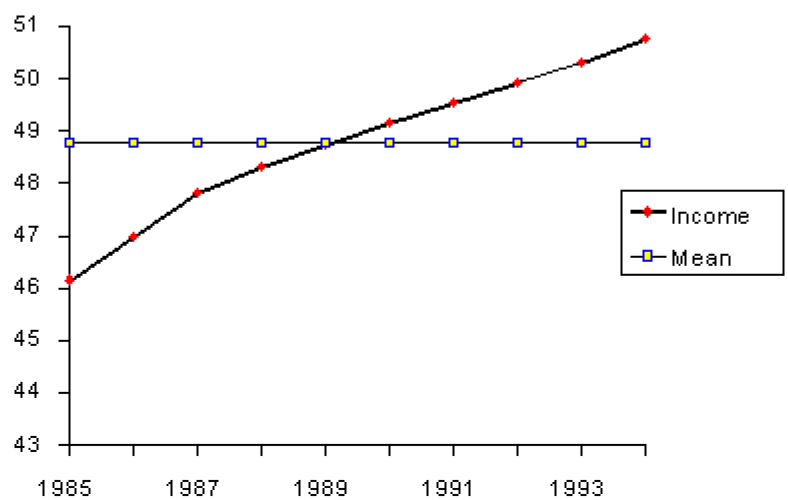
Lets take another example

Data : Income before taxes of a businessman between 1985 and 1994

Year	$ (millions)	Mean	Error	Squared Error
1985	46.163	48.676	-2.513	6.313
1986	46.998	48.676	-1.678	2.814
1987	47.816	48.676	-0.860	0.739
1988	48.311	48.676	-0.365	0.133
1989	48.758	48.676	0.082	0.007
1990	49.164	48.676	0.488	0.239
1991	49.548	48.676	0.872	0.761
1992	48.915	48.676	0.239	0.057
1993	50.315	48.676	1.639	2.688
1994	50.768	48.676	2.092	4.378

The MSE = 1.8129.

Graph showing the Mean and Income line for the data



- ***Can we use the mean to forecast income (Can we smooth the Income line with a mean value?) if the goal is to predict the trend?***

- **The answer is Big NO**

-

Rather than equal, Different weights must be assigned to the recent past and long past data

Example: Average of the values 3, 4, 5 is 4
=3/3 + 4/3 + 5/3
=1 + 1.3333 + 1.6667 = 4.
The multiplier 1/3 is called the *weight*.
General Formula (sum of weights = 1)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \left(\frac{1}{n}\right)x_1 + \left(\frac{1}{n}\right)x_2 + \ldots + \left(\frac{1}{n}\right)x_n$$

Single Moving Average

- Averaging Successive Subsets
-
- Data : (Discussed above)
  - 9, 8, 9, 12, 9, 12, 11, 7, 13, 9, 11, 10
  - Let M= size of Subset = 3
  - For first 3 numbers avg = (9 + 8 + 9) / 3 = 8.667.
- This is "smoothing" is continued by advancing one period and calculating the next average of three numbers, dropping the first number known as Moving average

Results of Moving Average

Supplier	**$**	**MA**	**Error**	**Error squared**
1	9	9	0	0
2	8	8	0	0
3	9	8.667	0.333	0.111
4	12	9.667	2.333	5.444
5	9	10.000	-1.000	1.000
6	12	11.000	1.000	1.000
7	11	10.667	0.333	0.111
8	7	10.000	-3.000	9.000
9	13	10.333	2.667	7.111
10	9	9.667	-0.667	0.444
11	11	11.000	0	0
12	10	10.000	0	0

The MSE = 2.42 as compared to 3 in the previous case

Central Moving Average :

Major change: : place the average in the middle of the time interval of three periods
Note: This works well with odd time periods, but not so good for even time periods.
Smooth with MA's using $M = 4$

		Interim Steps			
Period	**Value**	**MA**	**Centered**	**Final**	**Error**
1	9			9	0
1.5					
2	8			8	0
2.5		9.5			
3	9		9.5	9.5	0.5
3.5		9.5			
4	12		10.0	10	2
4.5		10.5			
5	9		10.750	10.750	1.75
5.5		11.0			
6	12			12	0
6.5					
7	11			11	0

MSE = 1.04

Double Moving Averages for a Linear Trend Process

Moving averages are still not able to handle significant trends when forecasting

- Unfortunately, neither the mean of all data nor the moving average of the most recent M values, when used as forecasts for the next period, are able to cope with a significant trend.

- Double Moving Averages for a Linear Trend Process :
- calculates a second moving average from the original moving average, using the same value for M.

- As soon as both single and double moving averages are available, a computer routine uses these averages to compute a slope and intercept, and then forecasts one or more periods ahead.

Exponential Smoothing

This is a very popular scheme to produce a smoothed Time Series.

Whereas in Single Moving Averages the past observations are weighted equally,

Exponential Smoothing assigns *exponentially decreasing weights* as the observation get older.

In other words, *recent observations are given relatively more weight in forecasting than the older observations*.

S2 = y1

$$S_3 = \alpha y_2 + (1-\alpha) S_2;$$

$$S_t = \alpha y_{t-1} + (1-\alpha) S_{t-1} \quad 0 < \alpha \le 1 \quad t \ge 3\,.$$

- This smoothing scheme begins by setting S2 to y1,
- Si stands for smoothed observation or EWMA (Exponentially weighted Moving Average),
- y stands for the original observation.
- The subscripts refer to the time periods, 1,2,…,n. For the third period,

Setting the first EWMA

The initial EWMA plays an important role in computing all the subsequent EWMAs.
Setting S_2 to y_1 is one method of initialization. Another way is to set it to the target of the process.

Still another possibility would be to average the first four or five observations.

It can also be shown that the smaller the value of α, the more important is the selection of the initial EWMA. The user would be wise to try a few methods, (assuming that the software has them available) before finalizing the settings.

Why is it called "Exponential"?

Why is it called "Exponential"?

Expand basic equation

Let us expand the basic equation by first substituting for S_{t-1} in the basic equation to obtain

$$\begin{aligned} S_t &= \alpha y_{t-1} + (1-\alpha)\left[\alpha y_{t-2} + (1-\alpha) S_{t-2}\right] \\ &= \alpha y_{t-1} + \alpha(1-\alpha) y_{t-2} + (1-\alpha)^2 S_{t-2} \,. \end{aligned}$$

Summation formula for basic equation

By substituting for S_{t-2}, then for S_{t-3}, and so forth, until we reach S_2 (which is just y_1), it can be shown that the expanding equation can be written as:

$$S_t = \alpha \sum_{i=1}^{t-2} (1-\alpha)^{i-1} y_{t-i} + (1-\alpha)^{t-2} S_2 \,, \quad t \geq 2 \,.$$

Expanded equation for S_5

For example, the expanded equation for the smoothed value S_5 is:

$$S_5 = \alpha\left[(1-\alpha)^0 y_{5-1} + (1-\alpha)^1 y_{5-2} + (1-\alpha)^2 y_{5-3}\right] + (1-\alpha)^3 S_2 \,.$$

Illustrates exponential behavior

This illustrates the exponential behavior. The weights, $\alpha(1-\alpha)^t$ decrease geometrically, and their sum is unity as shown below, using a property of geometric series:

$$\alpha \sum_{i=0}^{t-1} (1-\alpha)^i = \alpha\left[\frac{1-(1-\alpha)^t}{1-(1-\alpha)}\right] = 1-(1-\alpha)^t \,.$$

From the last formula we can see that the summation term shows that the contribution to the smoothed value S_t becomes <u>less</u> at each consecutive time period.

Example for $\alpha = 0.3$

Let $\alpha = 0.3$. Observe that the weights $\alpha(1-\alpha)^t$ decrease exponentially (geometrically) with time.

	Value	weight
last	y_1	0.2100
	y_2	0.1470
	y_3	0.1029
	y_4	0.0720

What is the "best" value for α?

How do you choose the weight parameter?

The speed at which the older responses are dampened (smoothed) is a function of the value of α. When α is close to 1, dampening is quick and when α is close to 0, dampening is slow. This is illustrated in the table below.

---------------> towards past observations

α	$(1-\alpha)$	$(1-\alpha)^2$	$(1-\alpha)^3$	$(1-\alpha)^4$
0.9	0.1	0.01	0.001	0.0001
0.5	0.5	0.25	0.125	0.0625
0.1	0.9	0.81	0.729	0.6561

We choose the best value for α so the value which results in the smallest MSE.

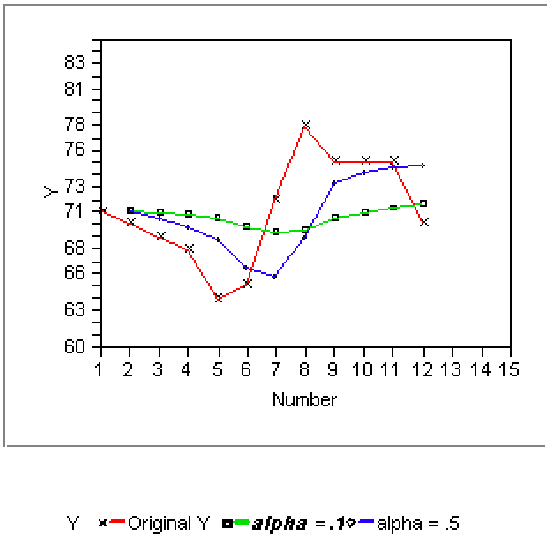
Example : Consider the following data set consisting of 12 observations taken over time:

Time	yt	S(α=0.1)	**Error**	**Error squared**
1	71			
2	70	71	-1.00	1.00
3	69	70.9	-1.90	3.61
4	68	70.71	-2.71	7.34
5	64	70.44	-6.44	41.47
6	65	69.80	-4.80	23.04
7	72	69.32	2.68	7.18
8	78	69.58	8.42	70.90
9	75	70.43	4.57	20.88
10	75	70.88	4.12	16.97
11	75	71.29	3.71	13.76
12	70	71.67	-1.67	2.79

The sum of the squared errors (SSE) = 208.94.
The mean of the squared errors (MSE) is the SSE /11 = 19.0.

The MSE was again calculated for $\alpha=0.5$ and turned out to be 16.29, so in this case we would prefer an α of 0.5. Can we do better? We could apply the proven trial-and-error method. This is an iterative procedure beginning with a range of α between 0.1 and 0.9. We determine the best initial choice for α and then search between $\alpha-\Delta$ and $\alpha+\Delta$. We could repeat this perhaps one more time to find the best α to 3 decimal places.

Exponential Smoothing: Original and Smoothed Values



Pros of Smoothing ☹discussed in last class)

- Data smoothing clears the understandability of different important hidden patterns in the data set.
- Data smoothing can be used to help predict trends. Prediction is very helpful for getting the right decisions at the right time.
- Data smoothing helps in getting accurate results from the data.

Cons of data smoothing

- Data smoothing doesn't always provide a clear explanation of the patterns among the data.
- It is possible that certain data points being ignored by focusing the other data points.

https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm