

Visualizing data

Why it's important
How to do it well

Comparison is the primary occupation of scientists

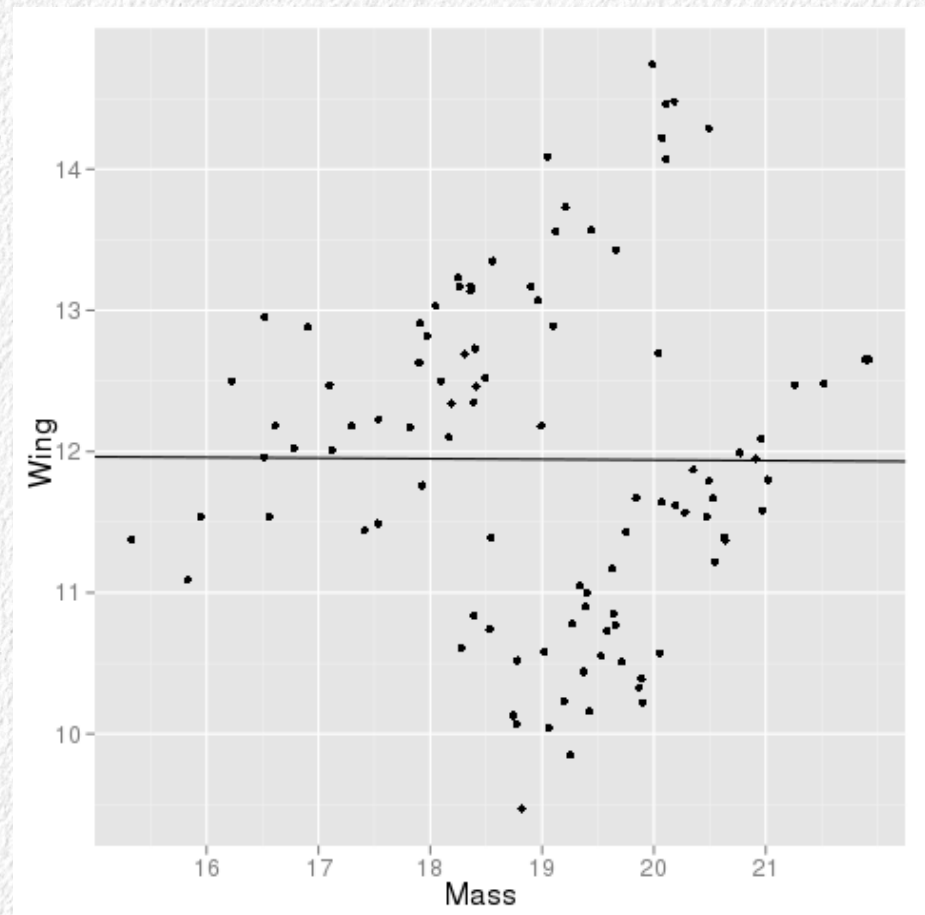
- Compare treatment groups to control
- Compare treatments to one another
- Correct conclusions only possible when the correct comparison is made
- Visualizing experimental data makes comparison easy

Not just pretty pictures

- Visualization is an important part of correct data analysis
 - Deriving knowledge from information
 - Deriving meaning from data – especially large data sets
- Visualization is an important part of communicating results to others
 - We're visual creatures
 - A picture is worth a thousand words

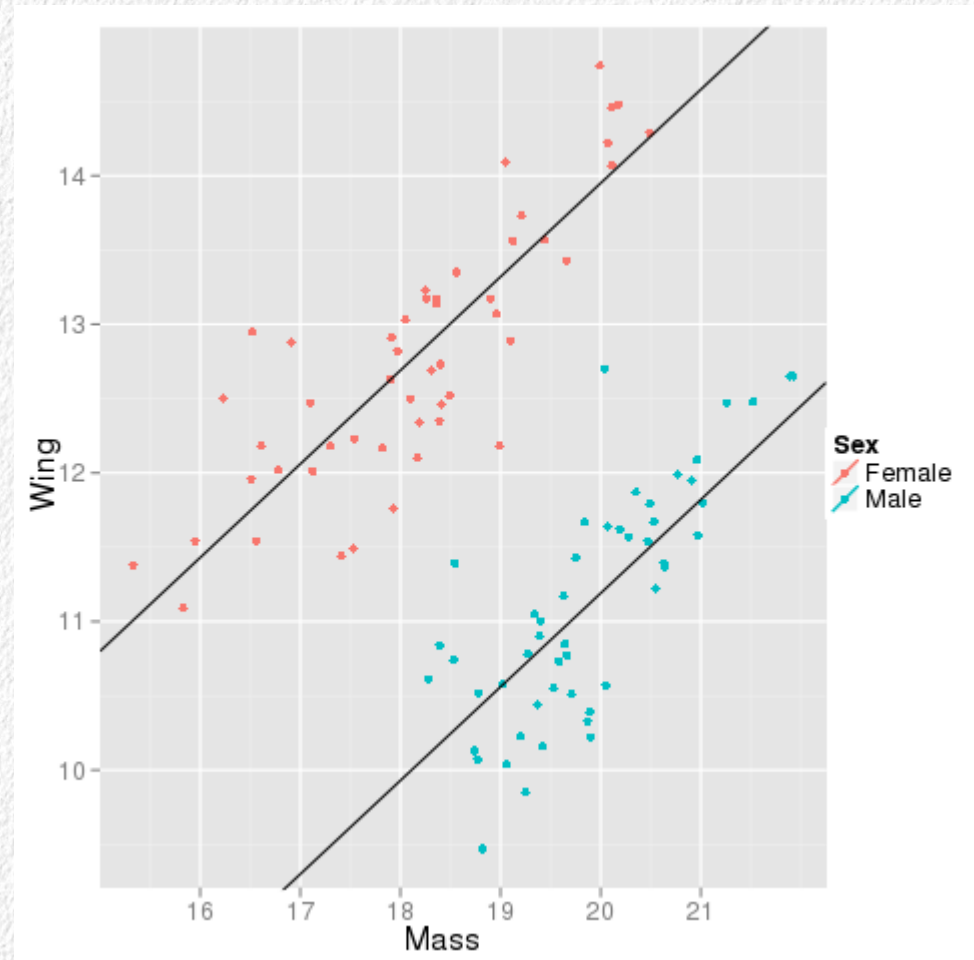
Example: what is the relationship between mass and wing length?

- Expect bigger birds to have bigger wing spans
- Flat line – mass is not related to wing length in this bird
- What's wrong with this picture?



Grouped by sex

- Group by sex, and fit a line to each sex's data swarm
- Now the pattern is apparent

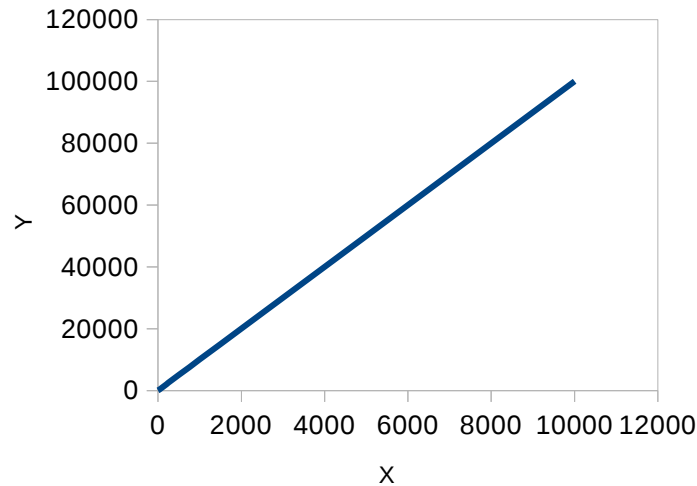


Using graphs to understand nature of relationship between variables

- You can use graphical methods to get an idea of the functional relationship between variables
- If we want to predict how a response variable changes when we make a change in a predictor, we need to know the correct functional form
- Different functions are straight lines on plots with logarithmic axes
 - Log-log plots – both x and y are on log scales, power functions will be linear
 - Semi-log plots – one linear axis, one log-scale axis
 - Exponential relationships are linear when y-axis is log-scale
 - Logarithmic relationships are linear when x-axis is log-scale

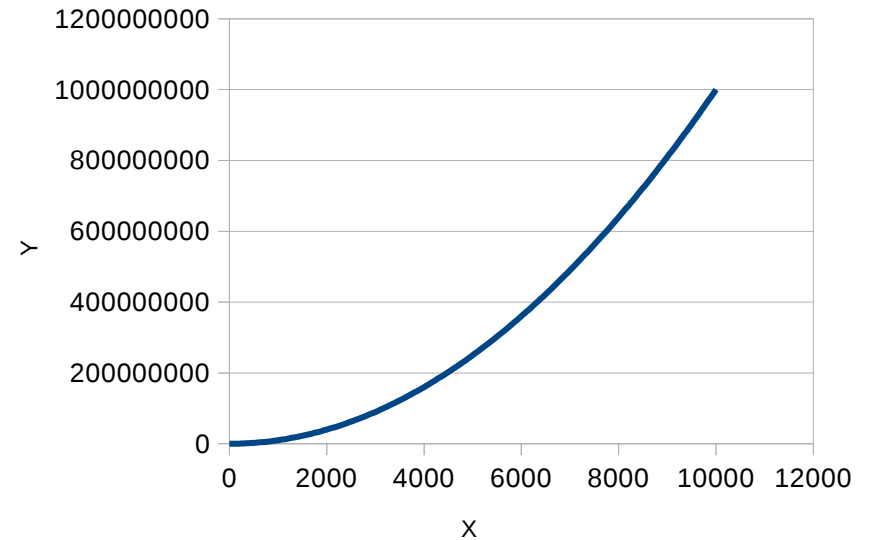
Three power function relationships between x and y

$$Y = aX^1$$

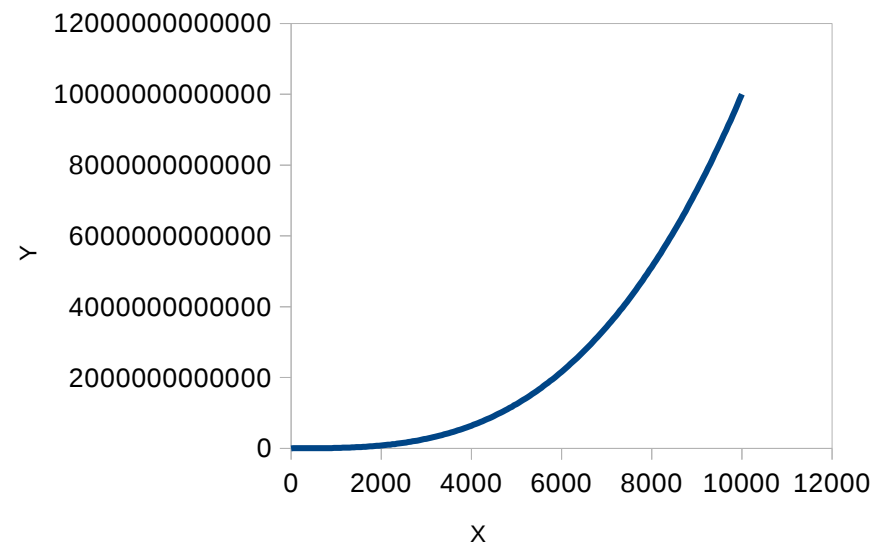


$$Y = aX^b$$

$$Y = aX^2$$



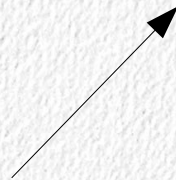
$$Y = aX^3$$



Log Y scales linearly with log X

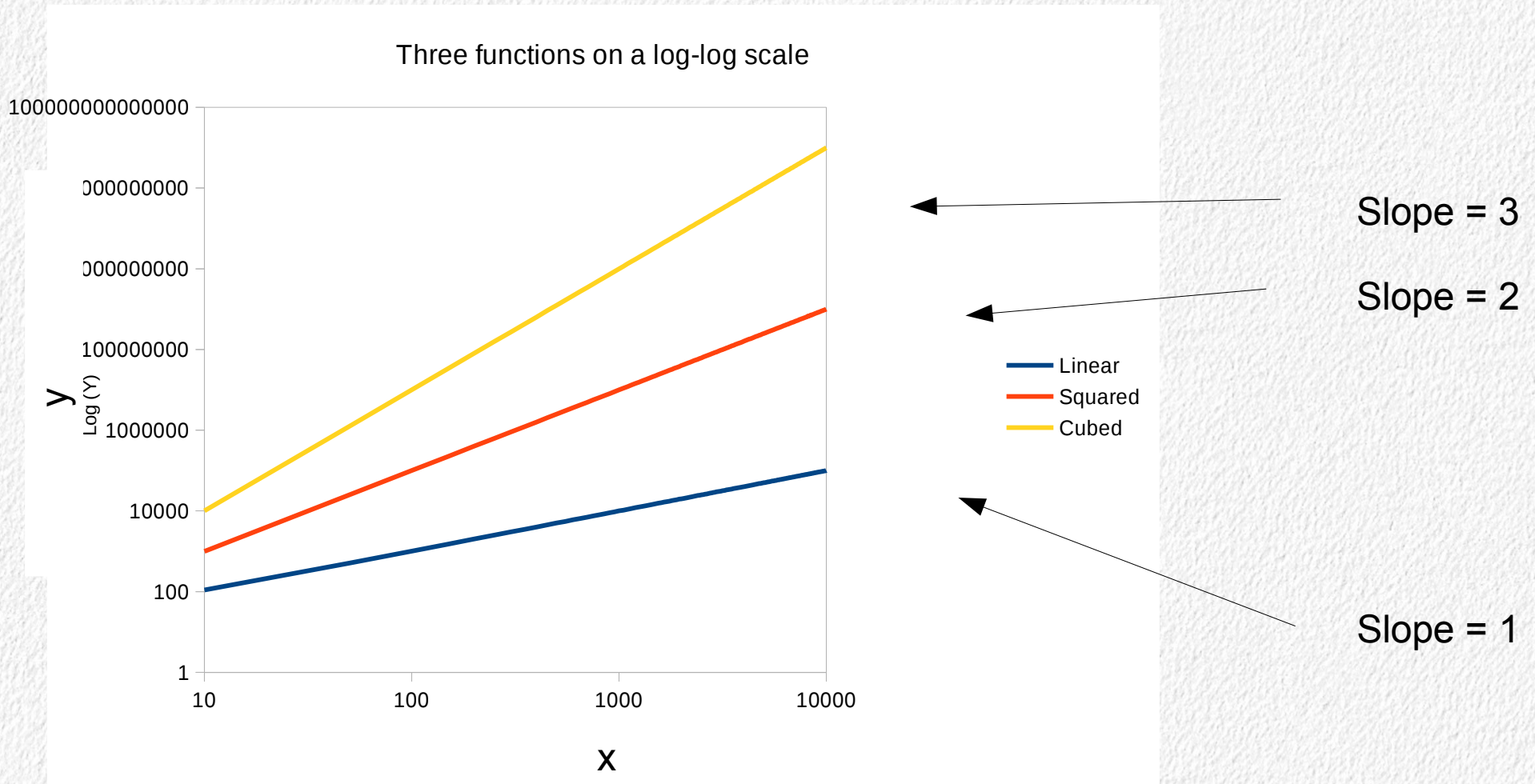
$$Y = a X^b$$

$$\log(Y) = \log(a) + b \log(X)$$



$$y = b + mx$$

Log-log plots

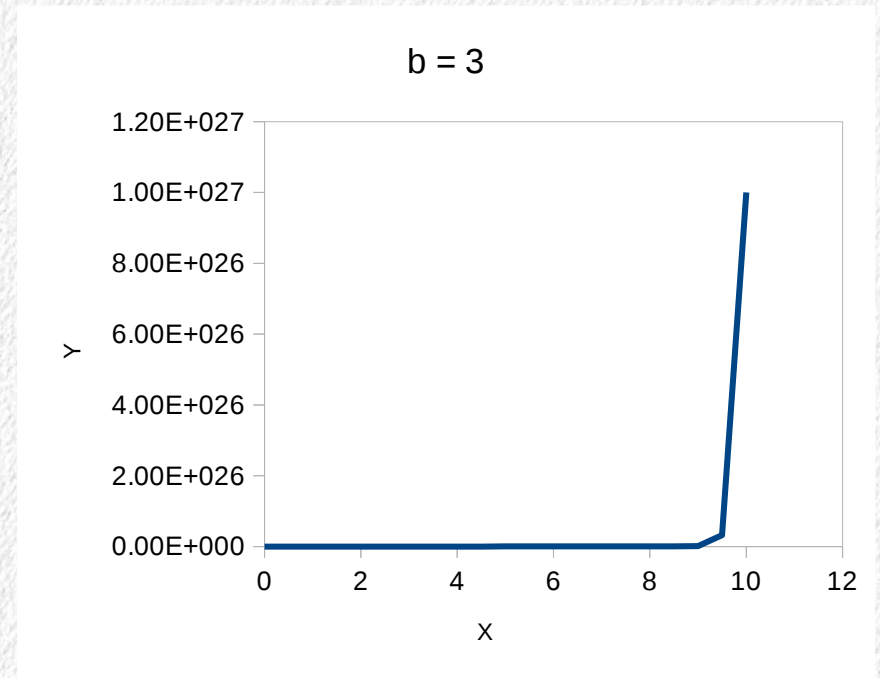
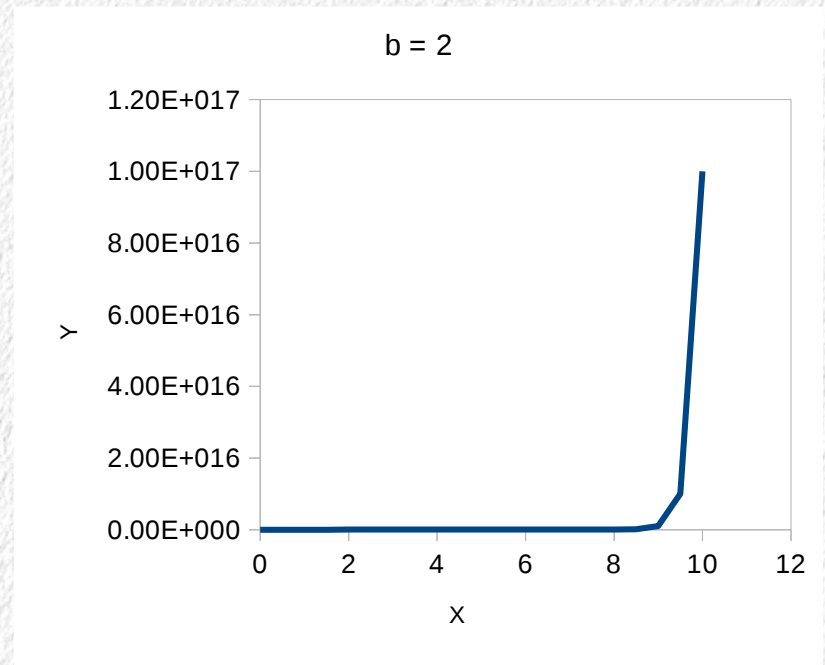


Both axes on a log scale

If data are a straight line on a log-log plot, then the relationship is a power function

Exponential relationships

$$Y = a 10^{bX}$$

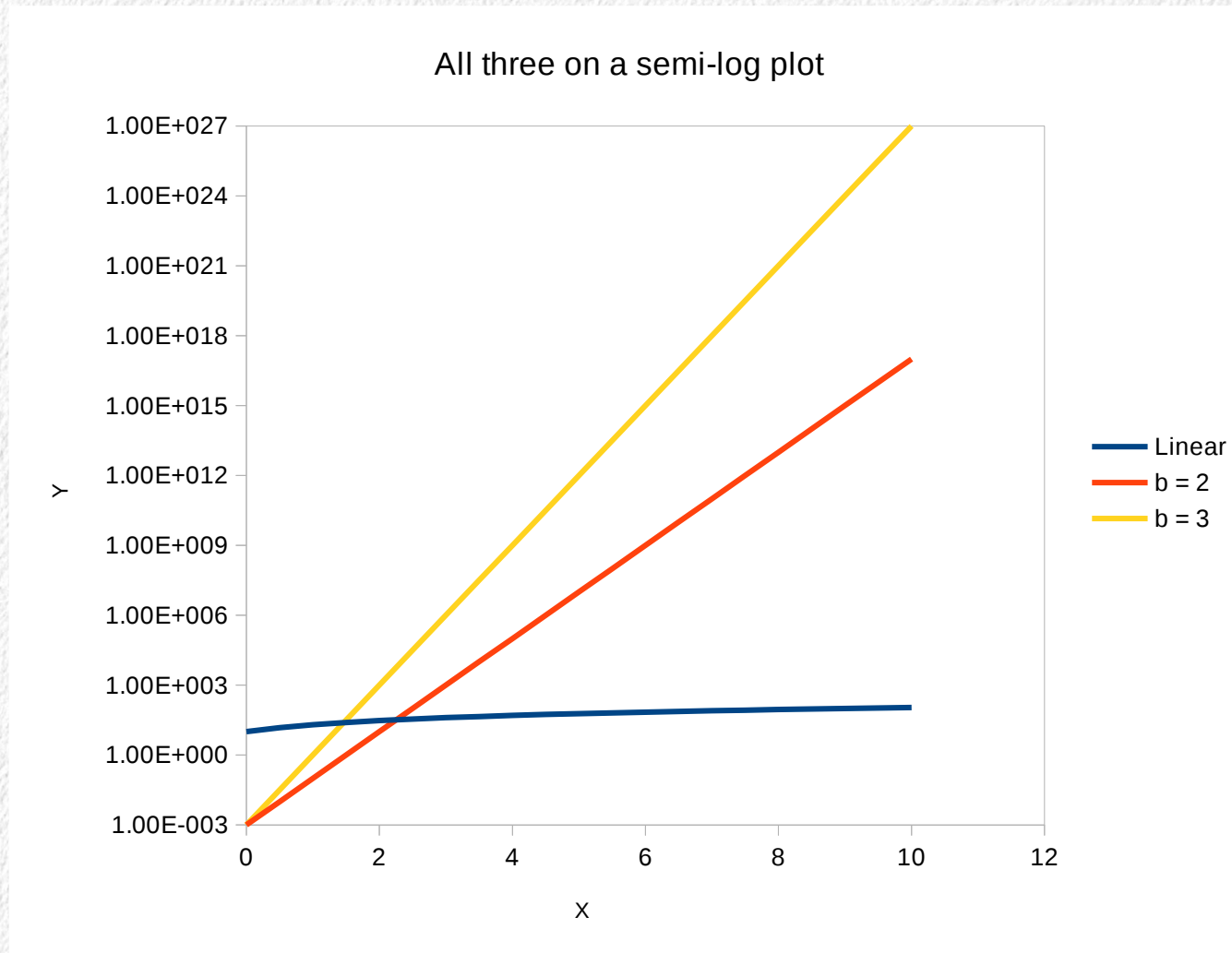


Log of Y is linear with X

$$Y = a 10^{bX}$$

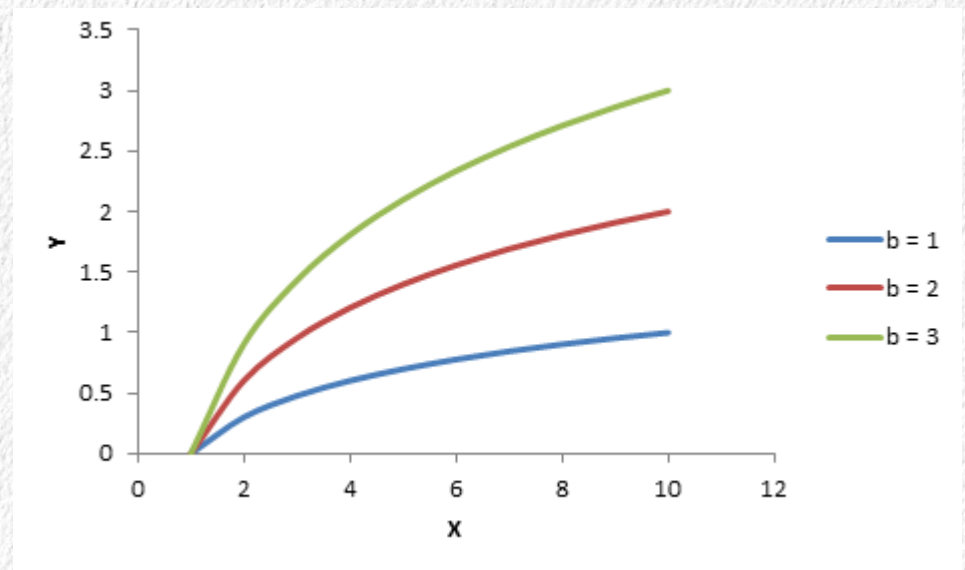
$$\log(Y) = \log(a) + bX$$

On a semi-log plot – y-axis on log scale



X is on a linear scale

Logarithmic relationships



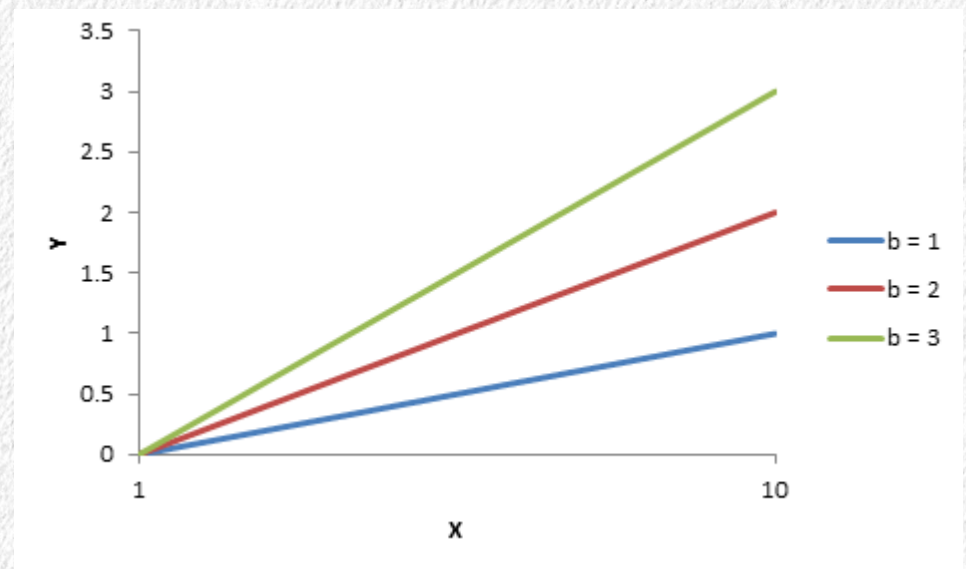
$$10^Y = aX^b$$

Y is related to the log of X

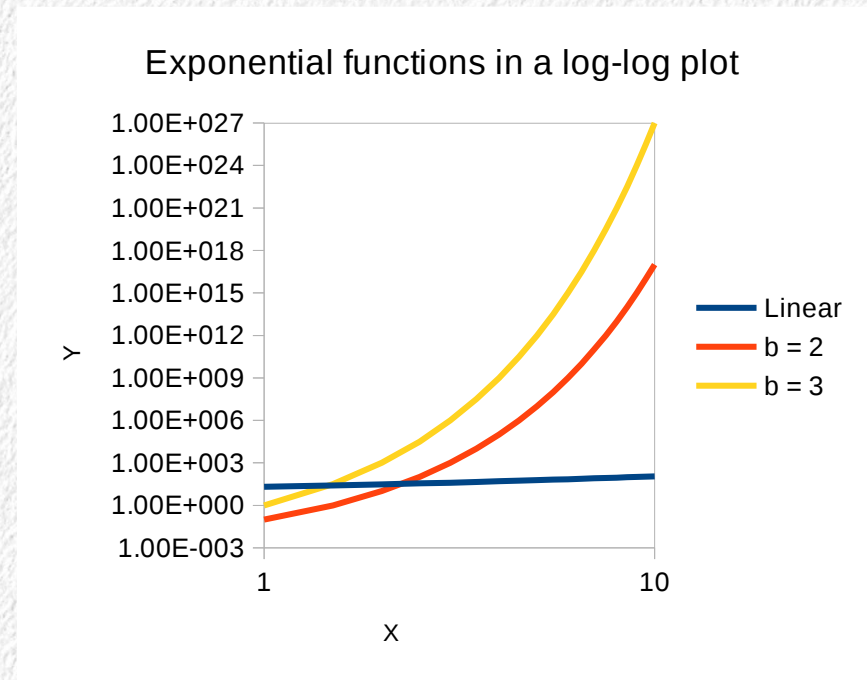
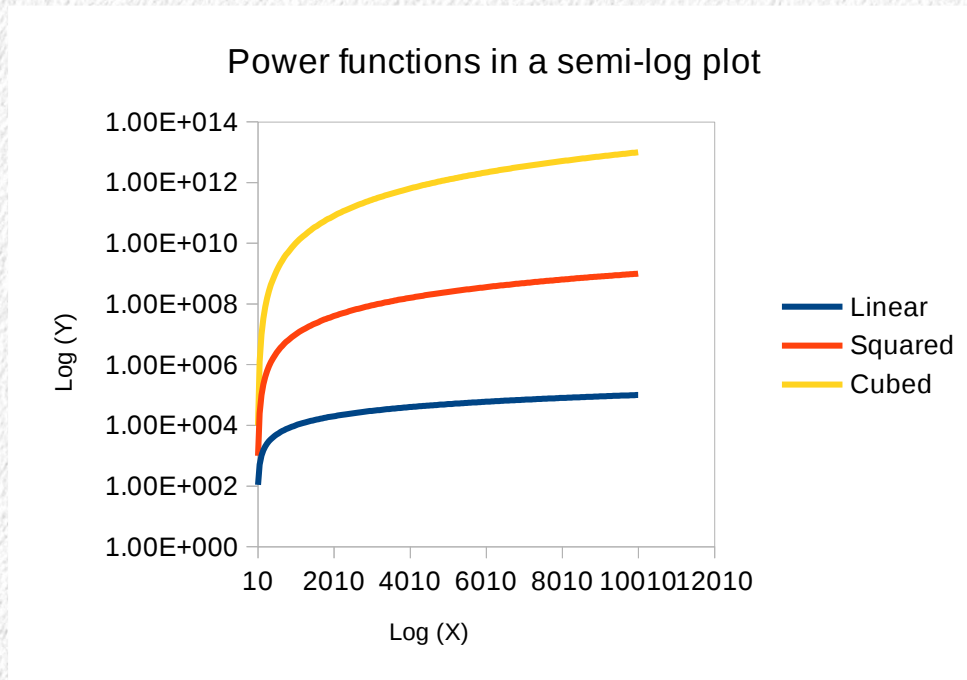
$$10^Y = aX^b$$

$$Y = \log(a) + b \log(X)$$

Y is on a linear scale, x is logarithmic



Wrong axis scales for the data lead to curved lines












Thus, changing the scale from linear to logarithmic can help you diagnose the functional relationship between variables

Once we know the relationship, we can have Excel give us the equation for the line

Common graph types

- The graph you should use depends on the type of data you will display
- Common graph types (i.e. those supported by Excel) cover most of the basic data display tasks
- Less common, task-specific graph types (not supported by Excel) are used in various fields of Biology
 - Statistical packages
 - Dedicated graphing software

Excel's graph types

Graph type		Use
Column		A numeric variable plotted at levels of a categorical variable
Bar		A horizontal column chart
Histogram		Displaying distribution of data (not a distinct graph type in Excel)
Line		Values of a numeric variable displayed at the same levels of a categorical variable.
Pie		Proportions, percentages
Area		Line graph with the area below the lines shaded
Scatter		Relationship between two numeric variables
Surface		A three dimensional surface, with categorical x,y and numeric z
Bubble		A scatter plot with symbol size set to display a third variable
Radar		Each numeric variable is a ray, each observation is plotted on each ray, with points connected

Column chart

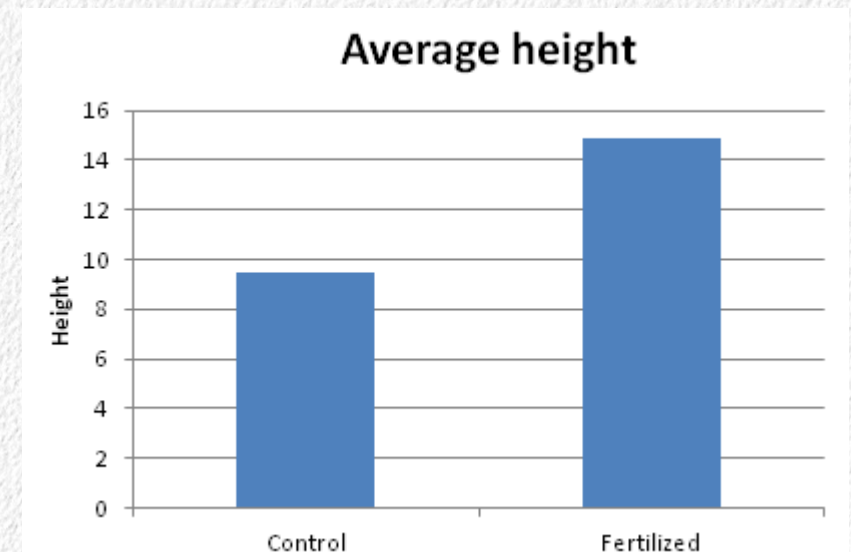
Data in Excel

	A	B
1	Treatment	Height
2	Control	11.5
3	Control	7.0
4	Control	10.9
5	Control	13.0
6	Control	7.2
7	Control	5.6
8	Control	9.0
9	Control	10.8
10	Control	9.7
11	Control	9.8
12	Fertilized	14.5
13	Fertilized	12.8
14	Fertilized	15.9
15	Fertilized	14.7
16	Fertilized	16.5
17	Fertilized	14.8
18	Fertilized	14.4
19	Fertilized	14.6
20	Fertilized	16.7
21	Fertilized	13.8
22		

Pivot table for graphing

	D	E
	Average of Height	
	Treatment	Total
	Control	9.442980293
	Fertilized	14.88775842
	Grand Total	12.16536935

Graph of pivot table data



One categorical variable (grouping)
One numeric variable (response)

Grouped column chart

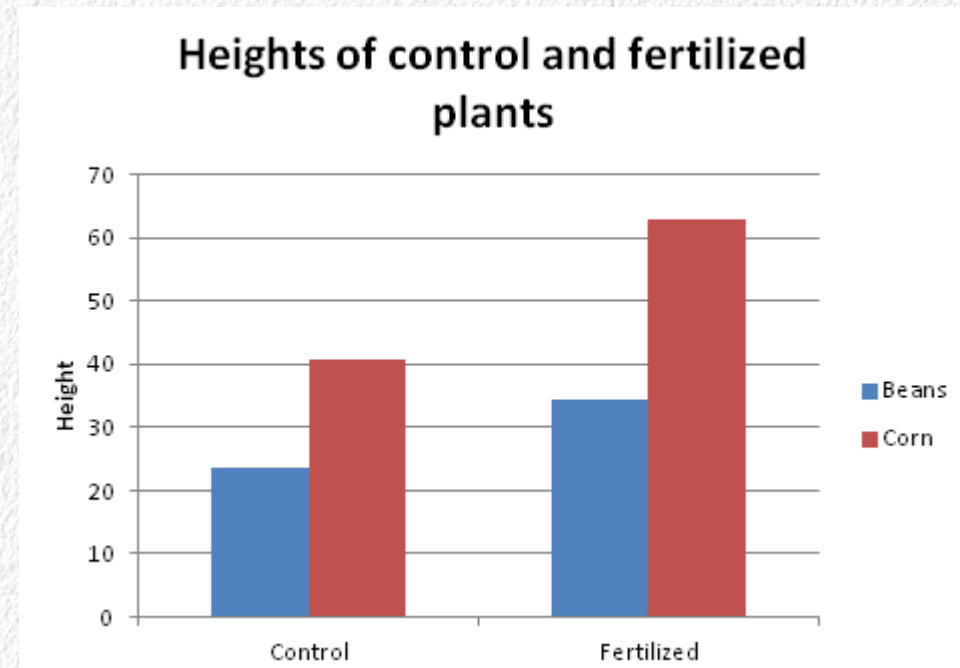
Data in Excel

	A	B	C
1	Treatment	Plant	Height
2	Control	Corn	41.7
3	Control	Corn	37.4
4	Control	Corn	41.8
5	Control	Corn	44.6
6	Control	Corn	37.2
7	Control	Beans	24.2
8	Control	Beans	18.9
9	Control	Beans	25.5
10	Control	Beans	22.8
11	Control	Beans	26.8
12	Fertilized	Corn	53.1
13	Fertilized	Corn	60.5
14	Fertilized	Corn	71.9
15	Fertilized	Corn	64.9
16	Fertilized	Corn	63.3
17	Fertilized	Beans	32.7
18	Fertilized	Beans	32.3
19	Fertilized	Beans	39.3
20	Fertilized	Beans	34.1
21	Fertilized	Beans	32.9
22			

Pivot table for graphing

	E	F	G	H
	Average of Height	Plant		
	Treatment	Beans	Corn	Grand Total
	Control	23.64382879	40.56847979	32.10615429
	Fertilized	34.2673524	62.7506732	48.5090128
	Grand Total	28.9555906	51.6595765	40.30758355

Graph of pivot table data



Two categorical variables (treatment group, plant type)

One numeric variable (response)

Histogram

Excel is **not a good choice** for histograms – use MINITAB, or equivalent

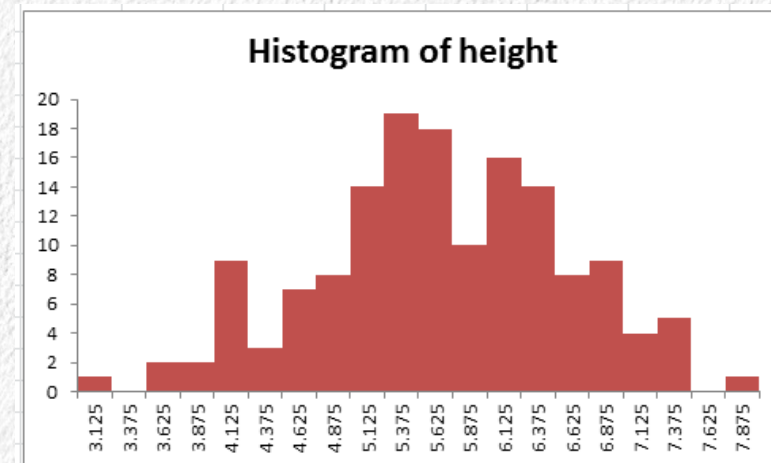
Data in Excel

	A	B
1	Height	
2	5.32	
3	4.53	
4	4.41	
5	5.76	
6	5.42	
7	4.68	
8	5.72	
9	5.23	
10	7.67	
11	4.89	
12	3.81	
13	5.81	
14	5.43	
15	5.54	
16	6.72	
17	5.49	
18	6.7	
19	4.95	
20	3.76	
21	6.84	
22	6.71	
23	5.17	

Bins, midpoints, and frequencies

Bins	Midpoints	Frequency
3	3.125	1
3.25	3.375	0
3.5	3.625	2
3.75	3.875	2
4	4.125	9
4.25	4.375	3
4.5	4.625	7
4.75	4.875	8
5	5.125	14
5.25	5.375	19
5.5	5.625	18
5.75	5.875	10
6	6.125	16
6.25	6.375	14
6.5	6.625	8
6.75	6.875	9
7	7.125	4
7.25	7.375	5
7.5	7.625	0
7.75	7.875	1

Bar chart of freq's, no gap between bars



Line

Data in Excel

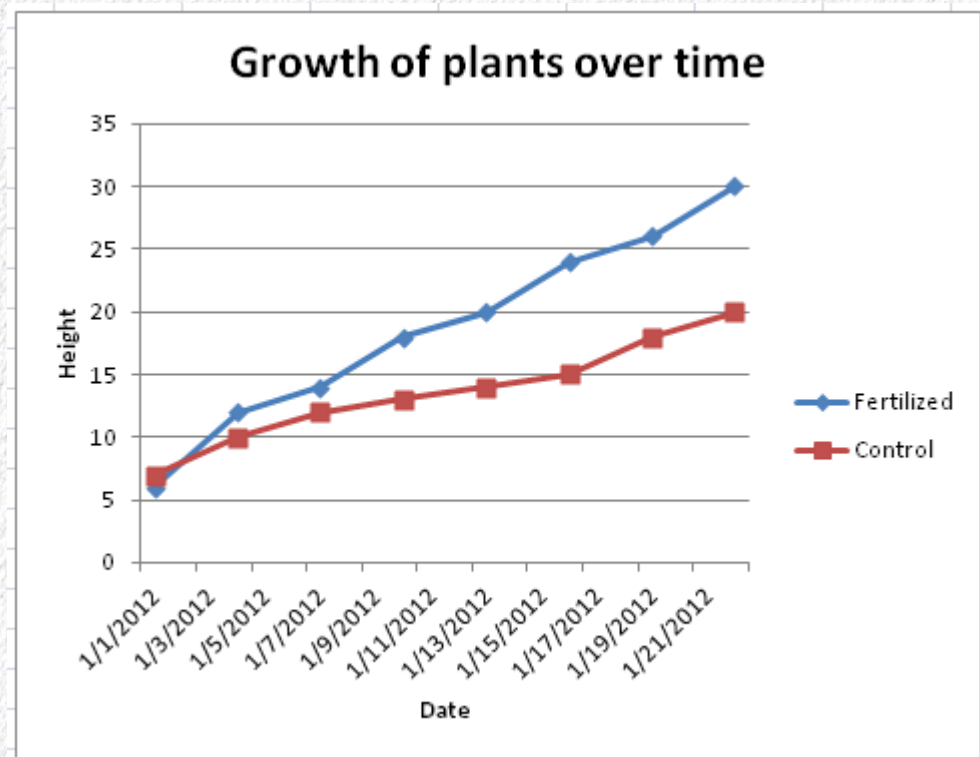
	A	B	C
1	Date	Fertilized	Control
2	1/1/2012	6	7
3	1/4/2012	12	10
4	1/7/2012	14	12
5	1/10/2012	18	13
6	1/13/2012	20	14
7	1/16/2012	24	15
8	1/19/2012	26	18
9	1/22/2012	30	20
10			

X axis can be numbers, dates, ordinal categories

*BUT, regardless, axis is treated as a **categorical** axis*

Order on the graph is the same as the order in the sheet

Graph

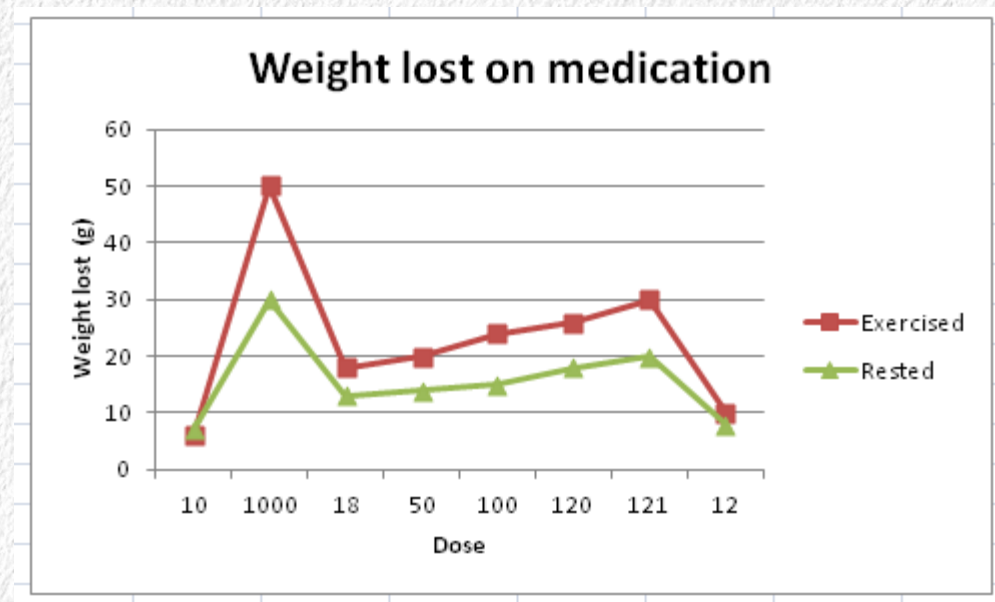


Line graphs are easy to misuse

Data in Excel

	A	B	C
1	Dose	Exercised	Rested
2	10	6	7
3	1000	50	30
4	18	18	13
5	50	20	14
6	100	24	15
7	120	26	18
8	121	30	20
9	12	10	8
10			

Graph

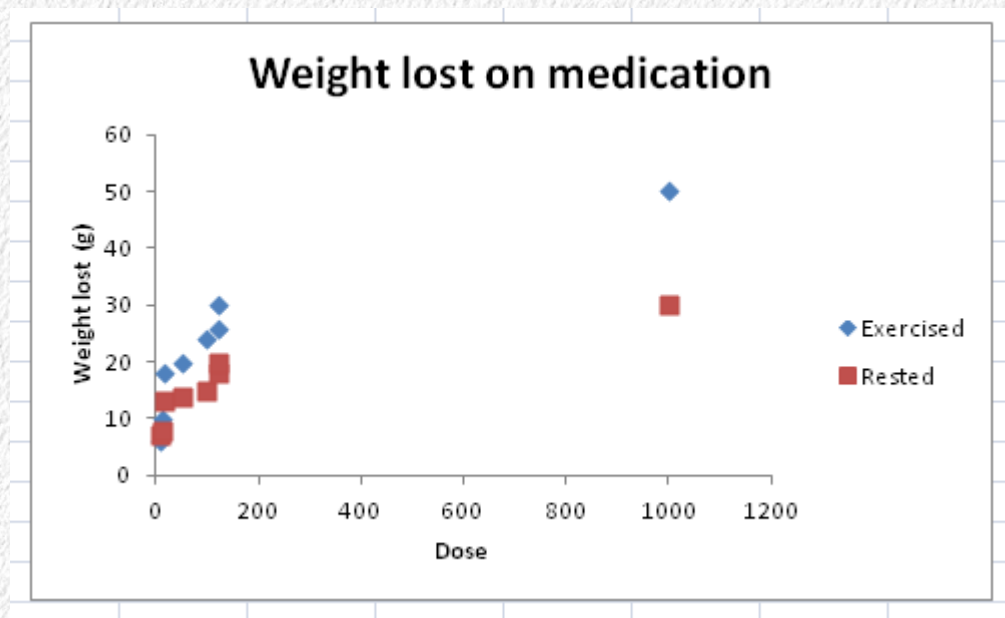


X categories are not in order

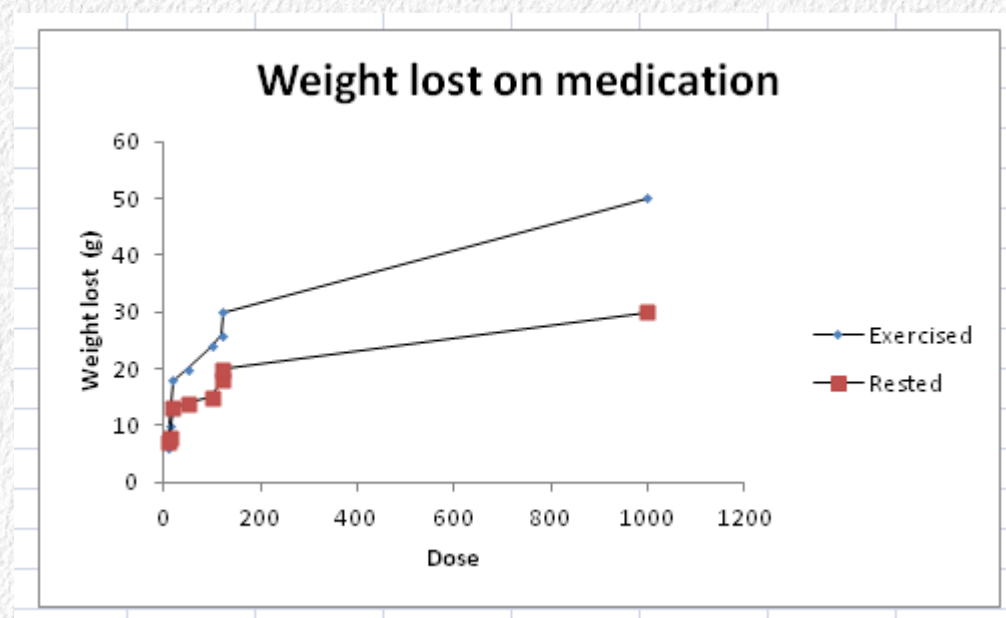
Amount of spacing between them is not even, but they're plotted as though the spacing is the same

Scatter

Without connecting lines

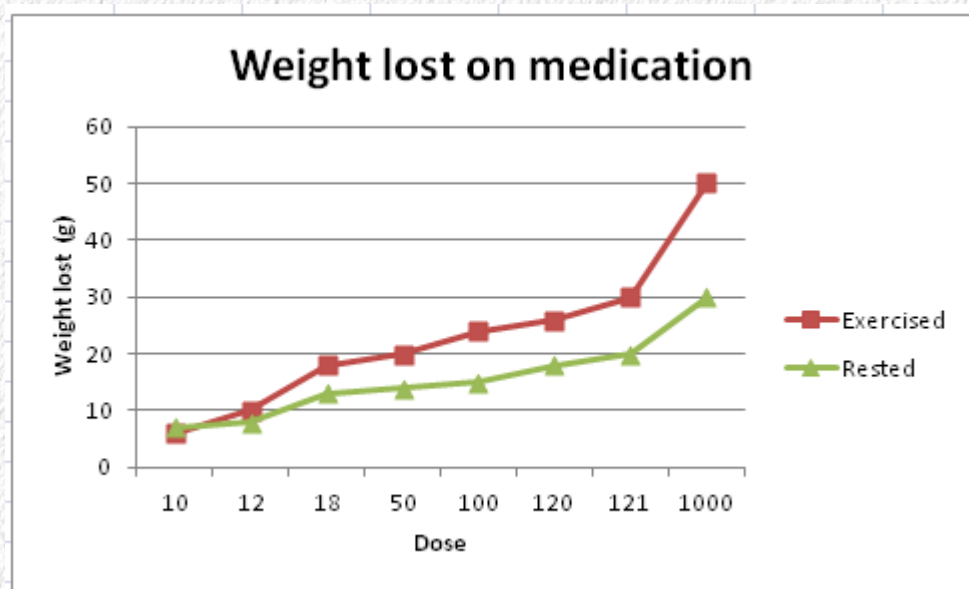


With connecting lines

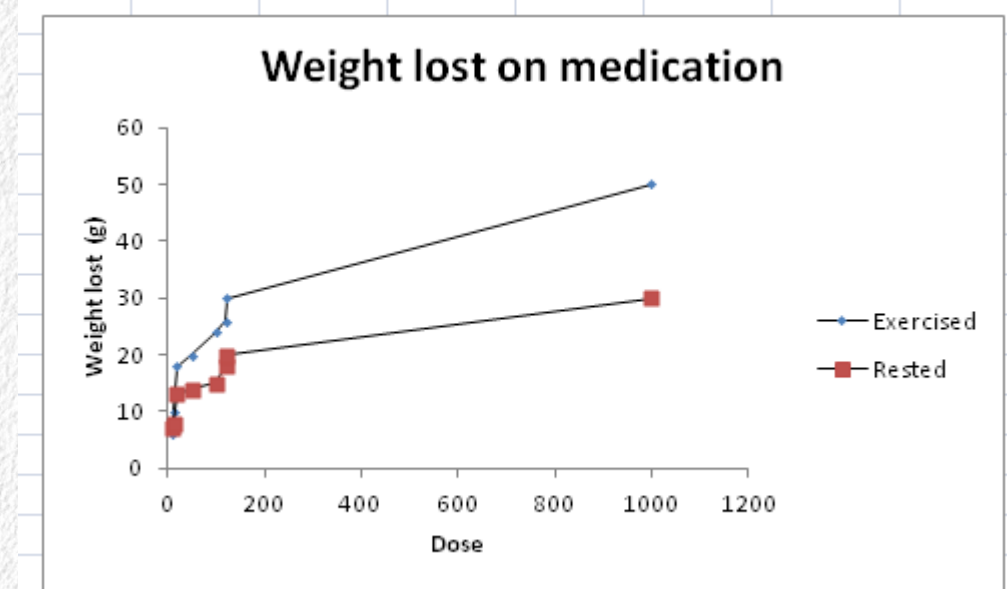


Line graphs vs. scatter plots with connected points?

Line graph



Scatter plot

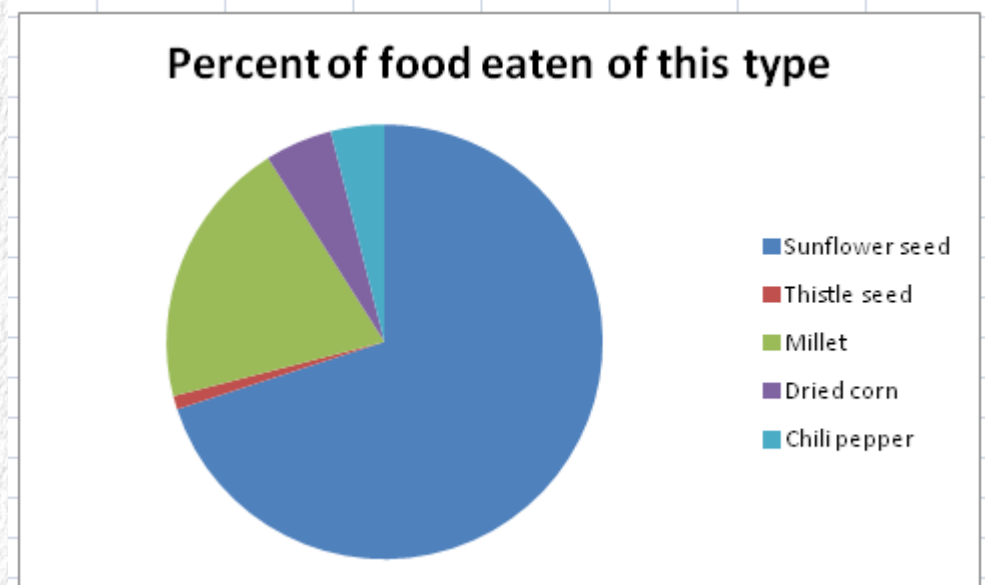


Pie chart

Data in Excel

Food	Percent of food eaten of this type
Sunflower	70%
Thistle seed	1%
Millet	20%
Dried corn	5%
Chili pepper	4%

Graph of percentages



*Proportions of total
Percentages
Relative frequencies*

*Counts will be converted to
proportions of the total (okay)*

*Any other numbers used will be
converted to proportions (not okay)*

High order data sets

- Some data sets have many different variables
- Flat screens only have 2 dimensions – two variables easy to display
- Adding variables means adding dimensions we don't have
 - 3-d graphs use depth cueing (perspective tricks)
 - Plot “slices” through the data
 - Use symbols/lines

Surface plot – depth cueing

Graph of values in body of matrix

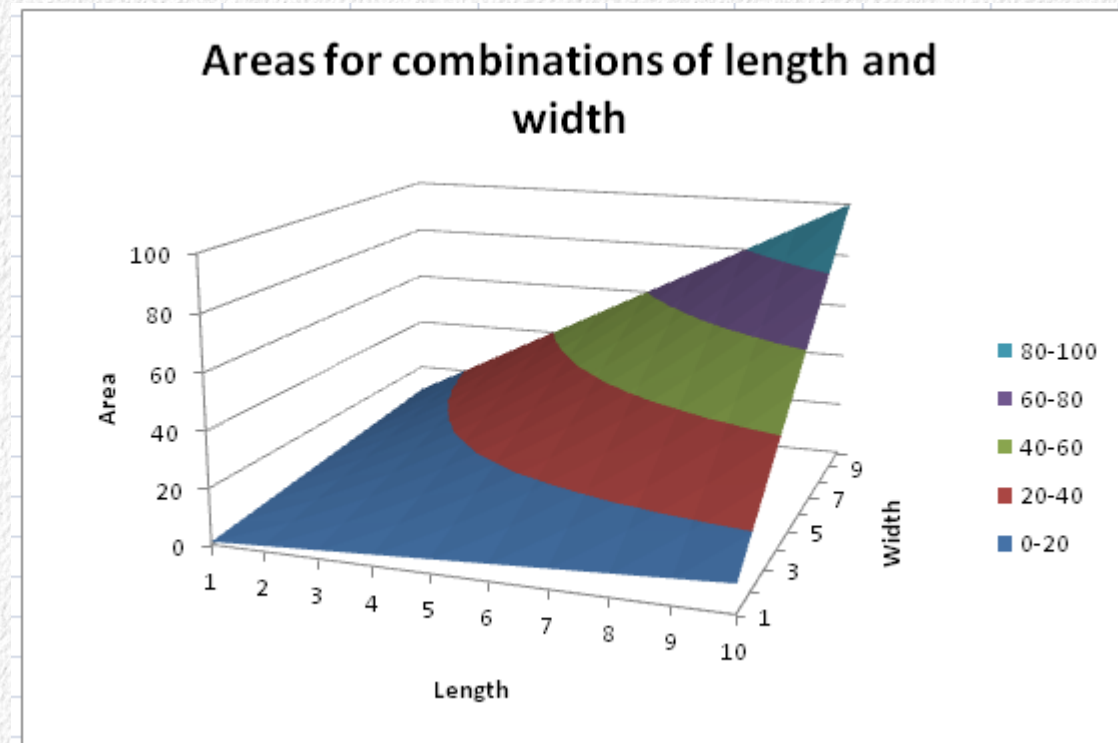
Data in Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1		Length										
2	Width		1	2	3	4	5	6	7	8	9	10
3		1	1	2	3	4	5	6	7	8	9	10
4		2	2	4	6	8	10	12	14	16	18	20
5		3	3	6	9	12	15	18	21	24	27	30
6		4	4	8	12	16	20	24	28	32	36	40
7		5	5	10	15	20	25	30	35	40	45	50
8		6	6	12	18	24	30	36	42	48	54	60
9		7	7	14	21	28	35	42	49	56	63	70
10		8	8	16	24	32	40	48	56	64	72	80
11		9	9	18	27	36	45	54	63	72	81	90
12		10	10	20	30	40	50	60	70	80	90	100
13												

Three dimensions

X and Y are row and column labels,
treated as categorical

Z is numeric, in body of matrix



Plotting slices across a third variable

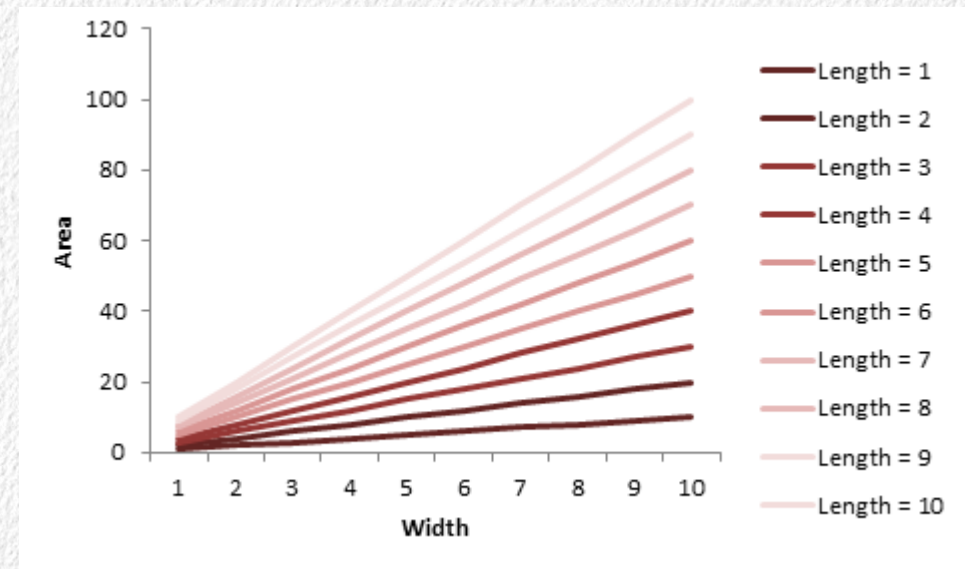
- Can group data based on levels of a third variable
- You can see the effect of grouping by comparing the groups
- Example: plotting the length, width, area data as a set of lines

Slices through the data – lines defined by lengths

Each length is a series, width on x-axis

Data in Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1			Length									
2	Width		1	2	3	4	5	6	7	8	9	10
3		1	1	2	3	4	5	6	7	8	9	10
4		2	2	4	6	8	10	12	14	16	18	20
5		3	3	6	9	12	15	18	21	24	27	30
6		4	4	8	12	16	20	24	28	32	36	40
7		5	5	10	15	20	25	30	35	40	45	50
8		6	6	12	18	24	30	36	42	48	54	60
9		7	7	14	21	28	35	42	49	56	63	70
10		8	8	16	24	32	40	48	56	64	72	80
11		9	9	18	27	36	45	54	63	72	81	90
12		10	10	20	30	40	50	60	70	80	90	100
13												



Line colors selected to indicate lengths

Symbol properties

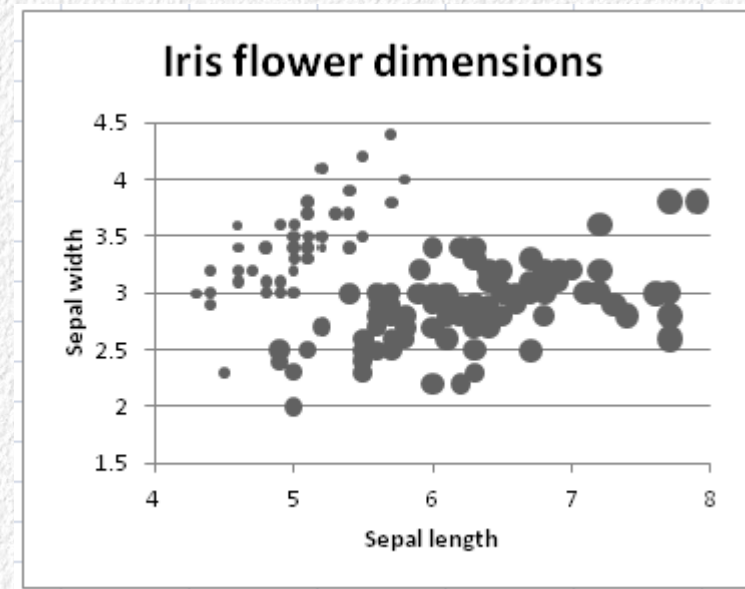
- Can subset the data and display with:
 - Symbol size
 - Symbol color
 - Symbol type

Bubble chart

Data in Excel

	A	B	C
1	Sepal.Length	Sepal.Width	Petal.Length
2	5.1	3.5	1.4
3	4.9	3	1.4
4	4.7	3.2	1.3
5	4.6	3.1	1.5
6	5	3.6	1.4
7	5.4	3.9	1.7
8	4.6	3.4	1.4
9	5	3.4	1.5
10	4.4	2.9	1.4
11	4.9	3.1	1.5
12	5.4	3.7	1.5
13	4.8	3.4	1.6
14	4.8	3	1.4
15	4.3	3	1.1
16	5.8	4	1.2
17	5.7	4.4	1.5
18	5.4	3.9	1.3
19	5.1	3.5	1.4
20	5.7	3.8	1.7
21	5.1	3.8	1.5
22	5.4	3.4	1.7
23	5.1	3.7	1.5
24	4.6	3.6	1

*Chart - symbol
size is proportional to
petal length*



Radar charts – multiple numeric axes

- Each ray is a different variable
- Each data point is plotted on each ray, with lines connecting

Radar plot with four axes

Data in Excel

	A	B	C	D
1	Sepal.Len	Sepal.Wic	Petal.Len	Petal.Width
2	5.1	3.5	1.4	0.2
3	4.9	3	1.4	0.2
4	4.7	3.2	1.3	0.2
5	4.6	3.1	1.5	0.2
6	5	3.6	1.4	0.2
7	5.4	3.9	1.7	0.4
8	4.6	3.4	1.4	0.3
9	5	3.4	1.5	0.2
10	4.4	2.9	1.4	0.2
11	4.9	3.1	1.5	0.1
12	5.4	3.7	1.5	0.2
13	4.8	3.4	1.6	0.2
14	4.8	3	1.4	0.1
15	4.3	3	1.1	0.1
16	5.8	4	1.2	0.2
17	5.7	4.4	1.5	0.4
18	5.4	3.9	1.3	0.4
19	5.1	3.5	1.4	0.3
20	5.7	3.8	1.7	0.3
21	5.1	3.8	1.5	0.3
22	5.4	3.4	1.7	0.2
23	5.1	3.7	1.5	0.4
24	4.6	3.6	1	0.2
25	5.1	3.3	1.7	0.5

