Stats Review

February 10, 2010

Pedro Wolf

Today

- Review Statistics
- Data?
  - Creating measures of constructs
- Descriptive Statistics
- The Correlation Coefficient
- The Student's T-test

What Data to Collect

- FPOT
- On the bottom level you have the undifferentiated real world
- On the other end you have a theory (analytical world)
- The purpose of research is to compare that theory (analytic world) to the undifferentiated real world
- The first step in that comparison is data collection

Things, Events, and Data

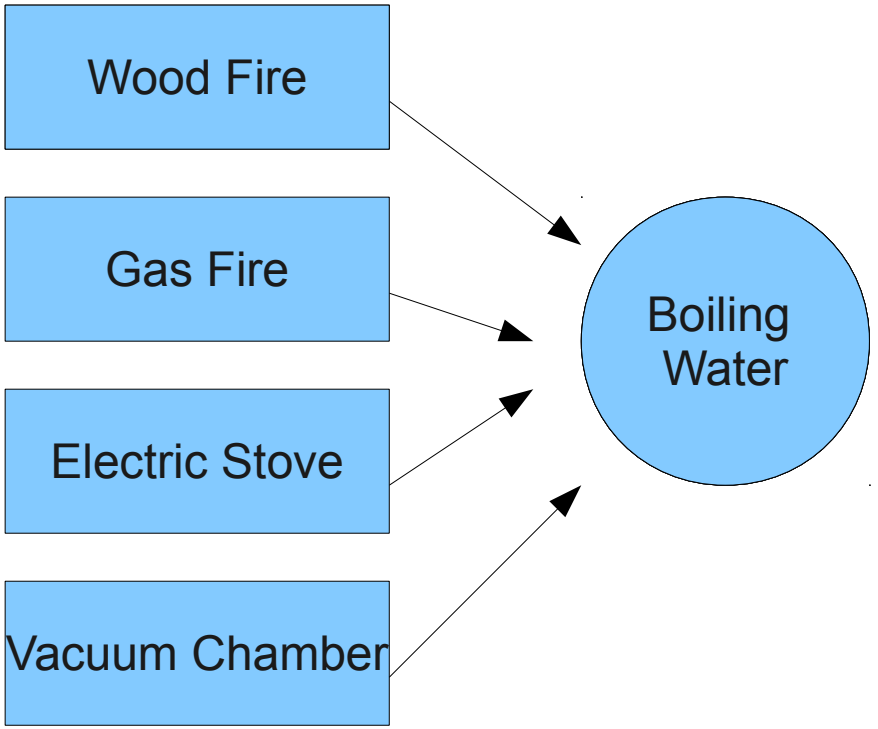
- The undifferentiated real world is extremely complex and fluid
  - e.g. where do I begin and end?
  - you see me as a distinct thing because things like me have been a selective pressure on your ancestors
  - At the same time we can't perceive other potential “things”
- To overcome this we define and extract things from the real world
- In your research these things and events are data

The Latent Variable and Cause

- Necessary and Sufficient Causes
  - Necessary- if *x* is a necessary cause of y then the presence of *y* implies the presence of *x*.
  - Sufficient- if *x* is a sufficient cause of *y,* then the presence of *x* implies the presence of *y*, but another cause *z* could also cause *y.*
  - In science we are looking for necessary causes
- You can never see a “necessary cause” directly.
  - I put a gallon of water in a pot, what can I do to make it boil?
    - I can build a fire under the pot
    - I can decrease the surrounding pressure
    - If I wanted water to not boil at let's say 213 degrees Fahrenheit I could increase the pressure
    - I can put it on a stove and turn a dial
  - What causes water to boil?
    - pressure, stove, fire, pressure and heat...?

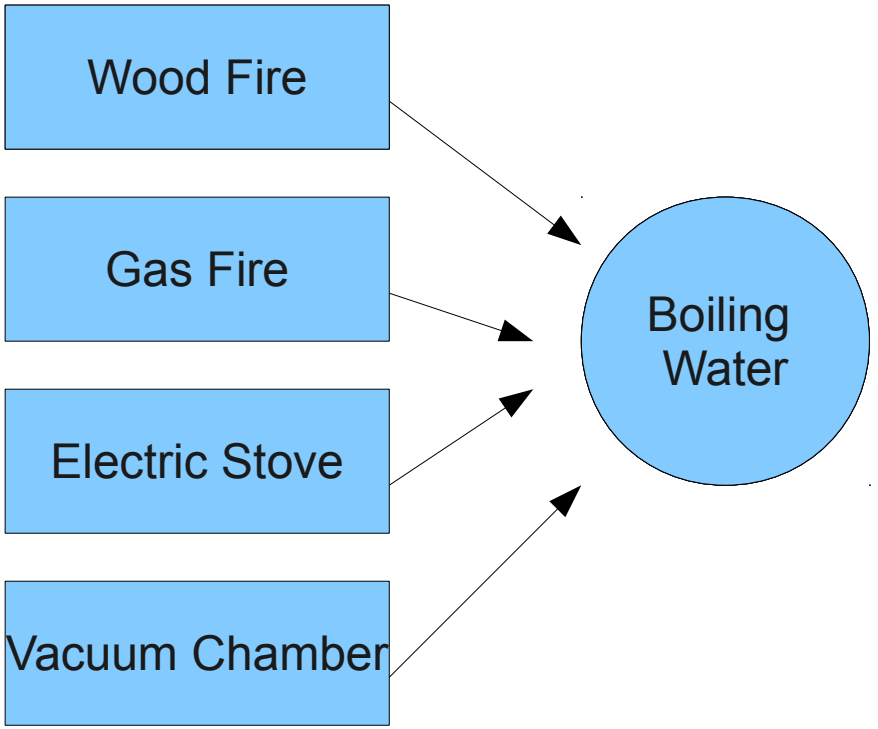
Latent Variable

- Are these the causes of water boiling? What kind?



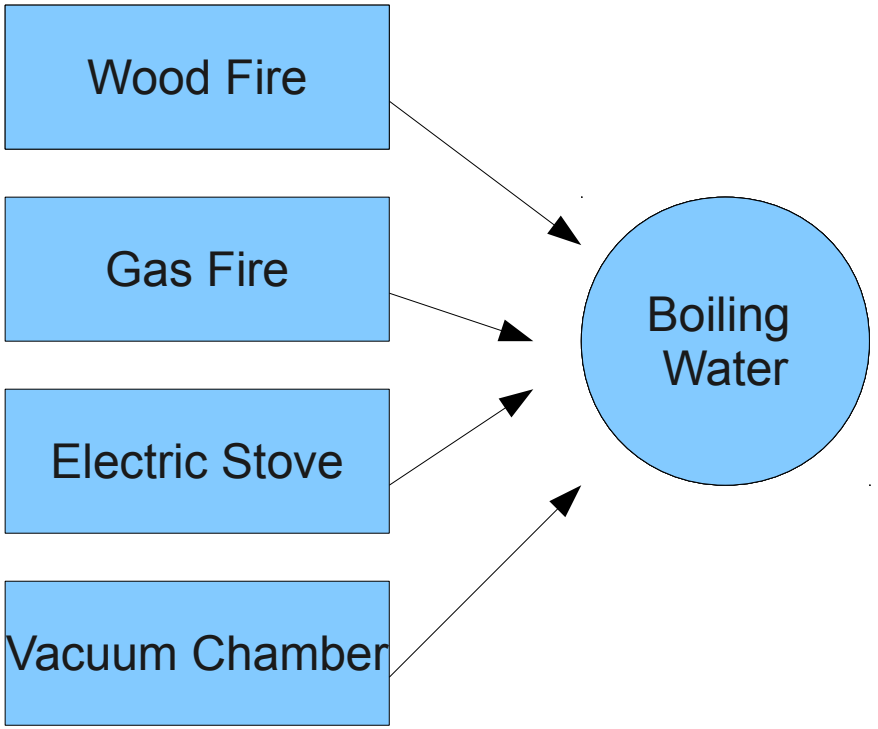
Latent Variable

- They are all sufficient causes



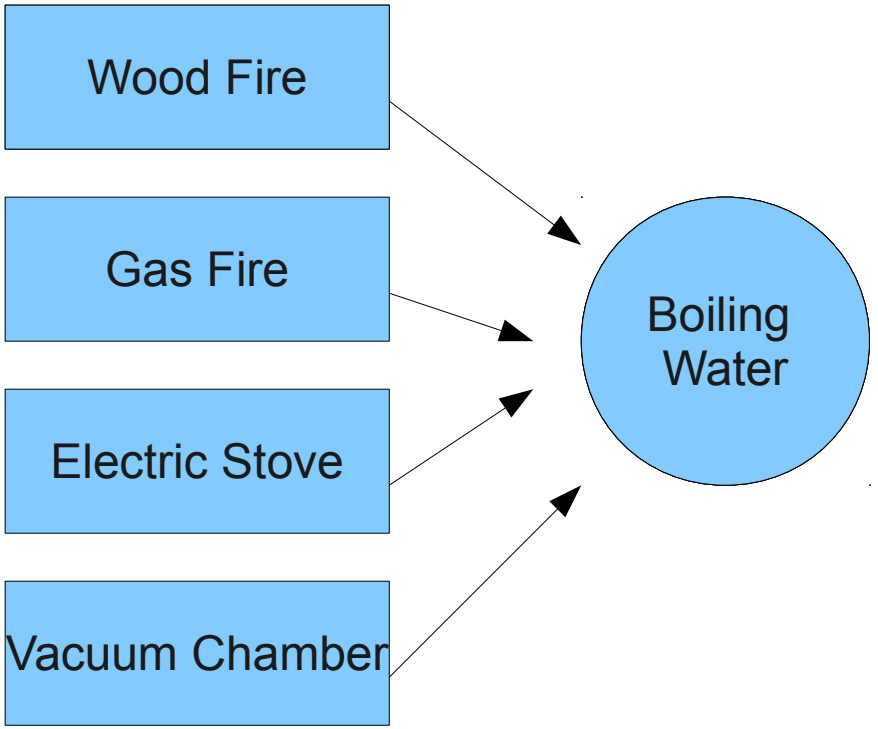
Latent Variable

- What is the necessary cause?



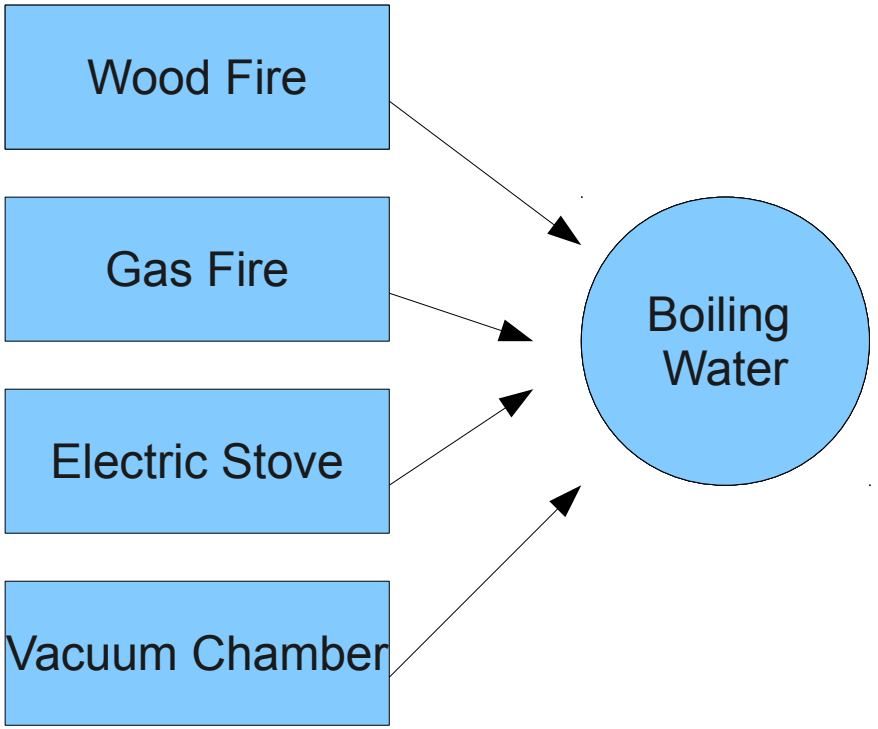
Latent Variable

- They all manipulate heat.



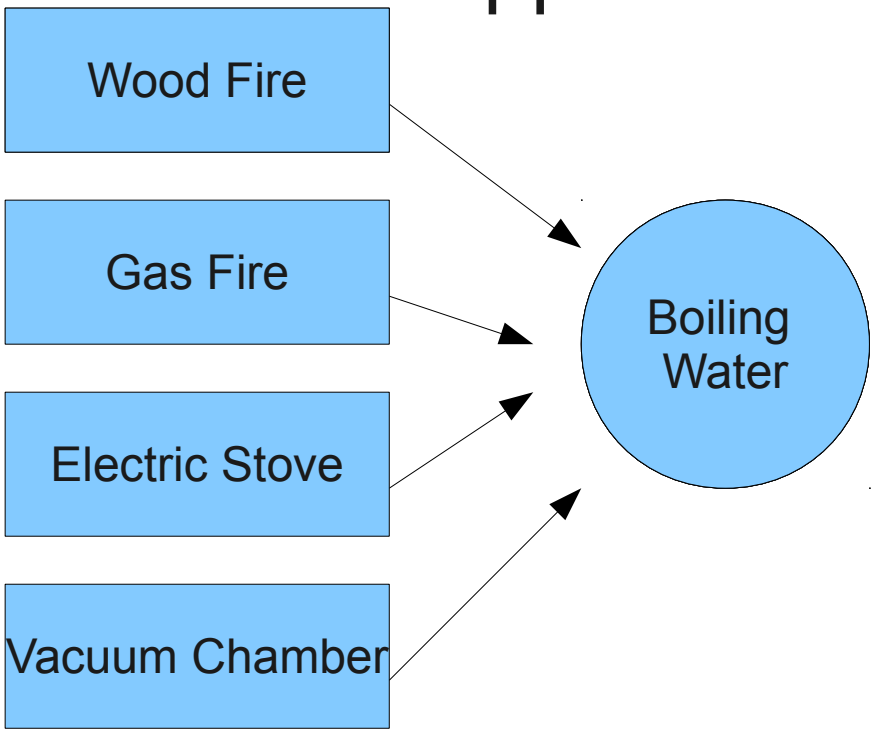
Latent Variable

- Heat causes molecules to move



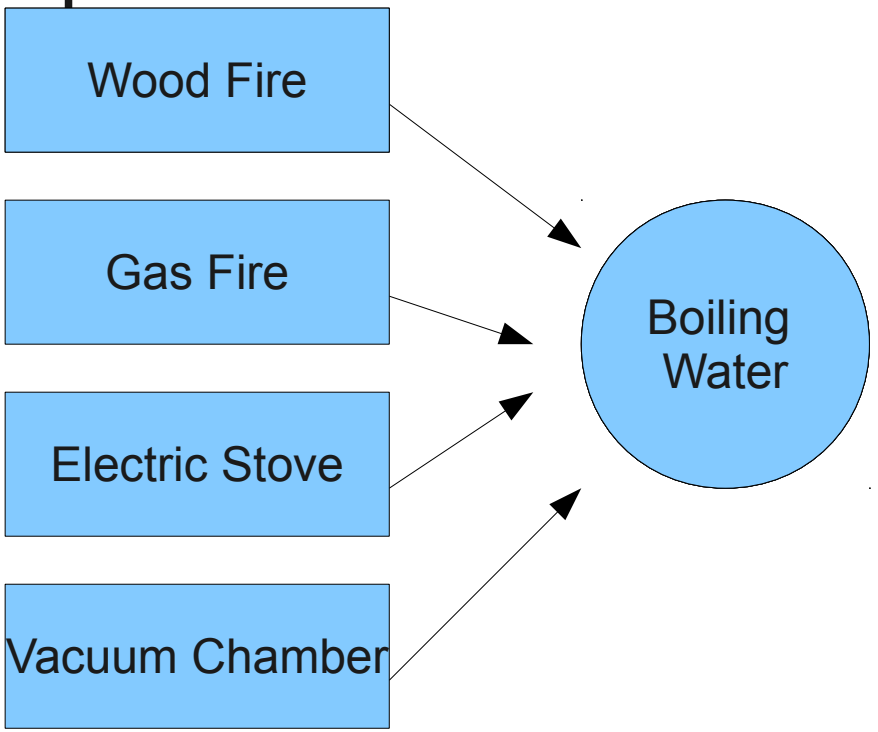
Latent Variable

- When the molecular motion is strong enough a phase transition happens



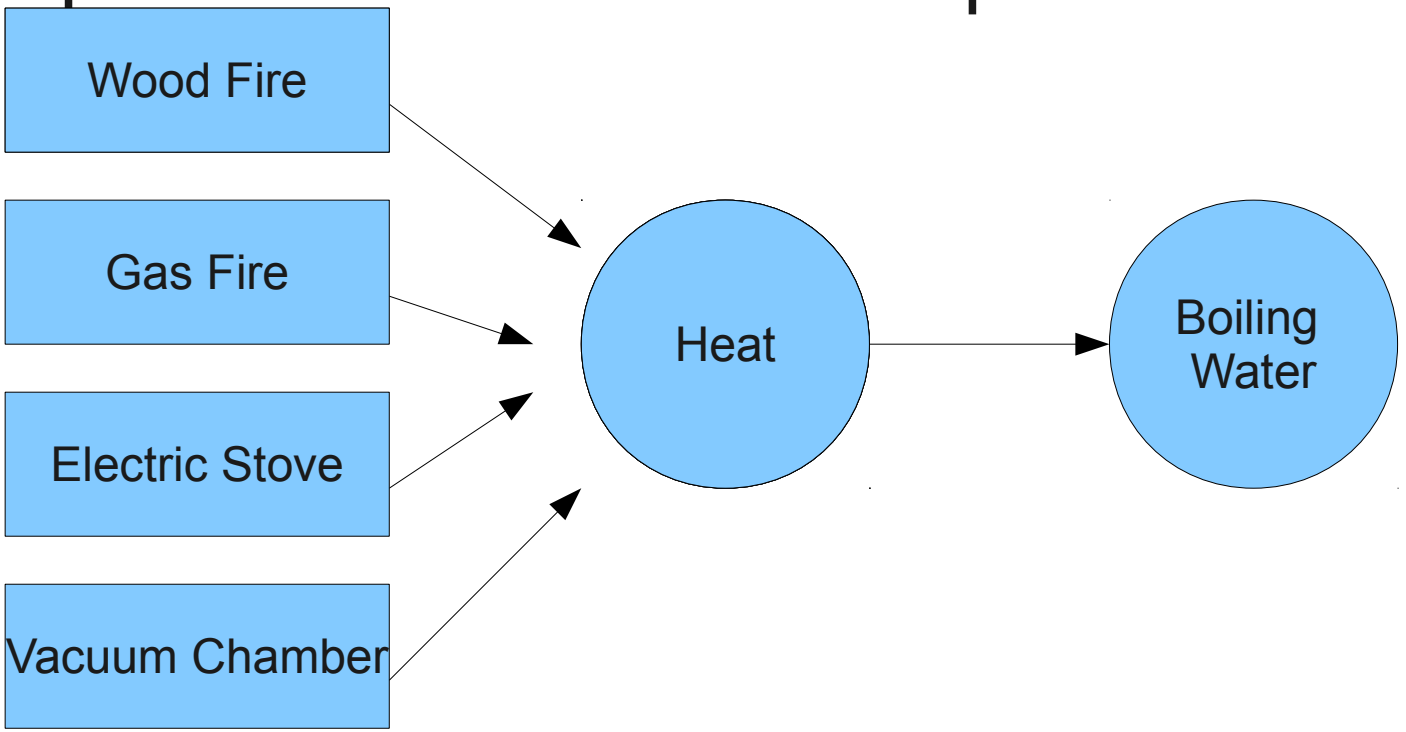
Latent Variable

- This happens when the vapor pressure of the liquid equals the environmental pressure



Latent Variable

- This happens when the vapor pressure of the liquid equals the environmental pressure



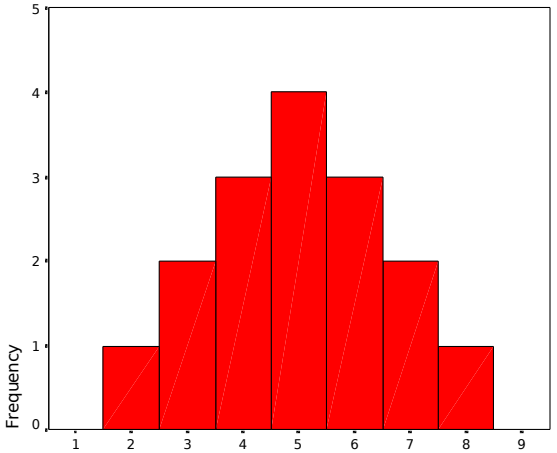
Selecting Things and Events

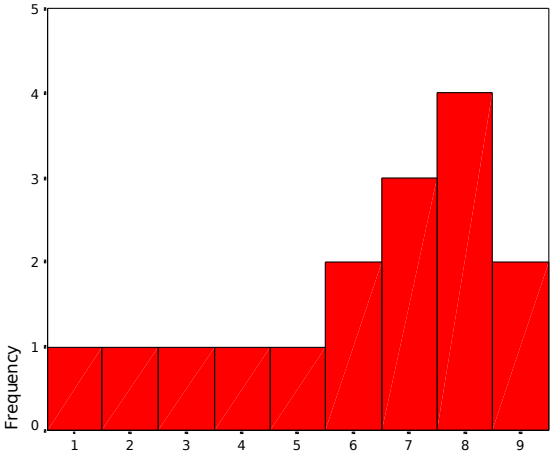
- Selecting things and events 1st step in linking analytical world with real world
- Things and events need to be linked with "empirical constructs" which are part of your hypothesis
  - If your studying heat and boiling water, how are you measuring heat, how are you measuring boiling?
  - You have to select the appropriate things which make up your events, which eventually become your data

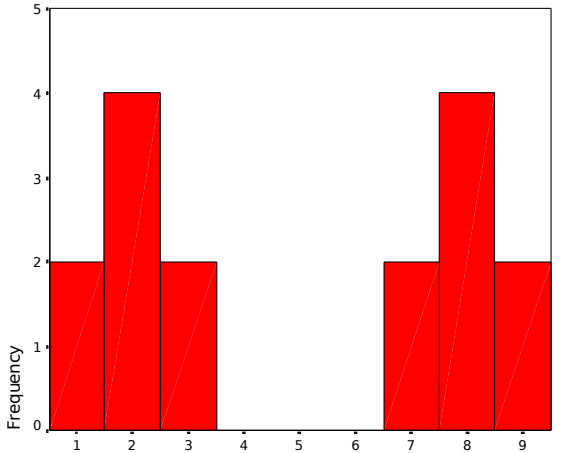
Your data is collected, now what do you do?

- First thing you do is calculate the descriptive statistics.
- Measures of central tendency
  - Mean- the mean is the score located at the exact mathematical center of a distribution
  - median- the median is the middle score of the data; the score that divides the data in half
  - mode-the mode is the score that has the highest frequency in the data
- Measures of dispersion
  - range- the range indicates the distance between the two most extreme scores in a distribution
  - standard deviation- The standard deviation indicates the “average deviation” from the mean, the consistency in the scores, and how far scores are spread out around the mean

Measures of Central Tendency







Mean

$$\overline{X} = \frac{\Sigma X}{N}$$

- The *X* with the bar over it is the mean of *X*
- The sigma is code for add all the values of X up
- The *N* is the number of values for *X you have*
- We are not doing [illegible] and.
  - Go to the course [illegible] and download todays data set

Mean in Excel

- http://www.u.arizona.edu/~wolfp/psych297a.htm
- Or google Pedro Wolf and click on the link under my picture.
-

	A	B	C	D
1	Y	Jan H	Jan A	
2	1996	69.5	53.6	
3	1997	63.6	52.4	
4	1998	67.1	53.2	
5	1999	70	53.6	
6	2000	70.5	55	
7	2001	65.6	49.7	
8	2002	73.6	51.8	
9	2003	73.6	58.1	
10	2004	64.6	52.9	
11	2005	66	54.4	
12	2006	69.5	54.5	
13	2007	60.7	48.6	
14	2008	63.6	51.7	
15	2009	69.1	55.5	
16	2010	66	53.9	
17				

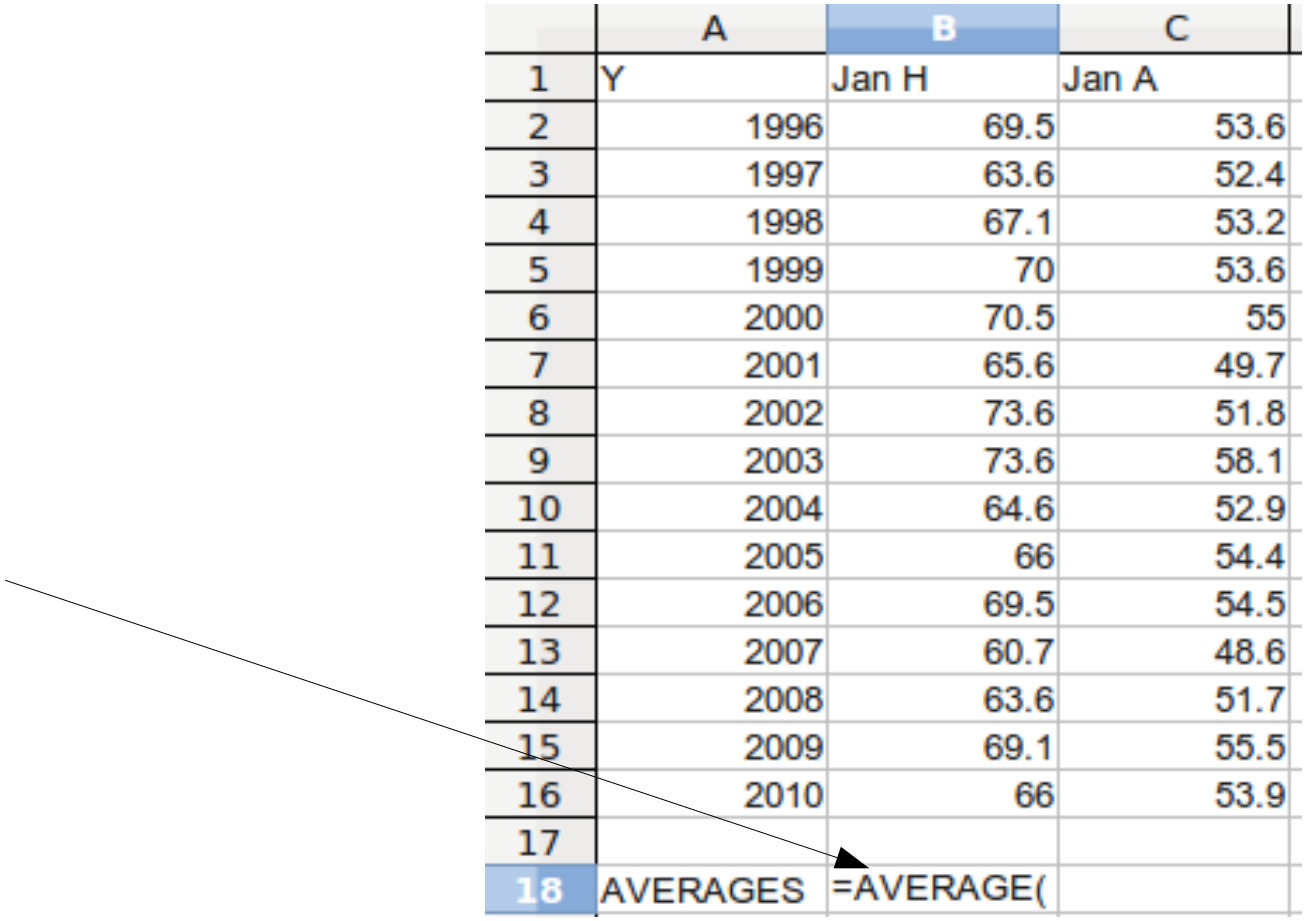
Mean In Excel

- First you may want to label your rows which contain averages
-

	A	B	C
1	Y	Jan H	Jan A
2	1996	69.5	53.6
3	1997	63.6	52.4
4	1998	67.1	53.2
5	1999	70	53.6
6	2000	70.5	55
7	2001	65.6	49.7
8	2002	73.6	51.8
9	2003	73.6	58.1
10	2004	64.6	52.9
11	2005	66	54.4
12	2006	69.5	54.5
13	2007	60.7	48.6
14	2008	63.6	51.7
15	2009	69.1	55.5
16	2010	66	53.9
17			
18	AVERAGES		

Mean In Excel

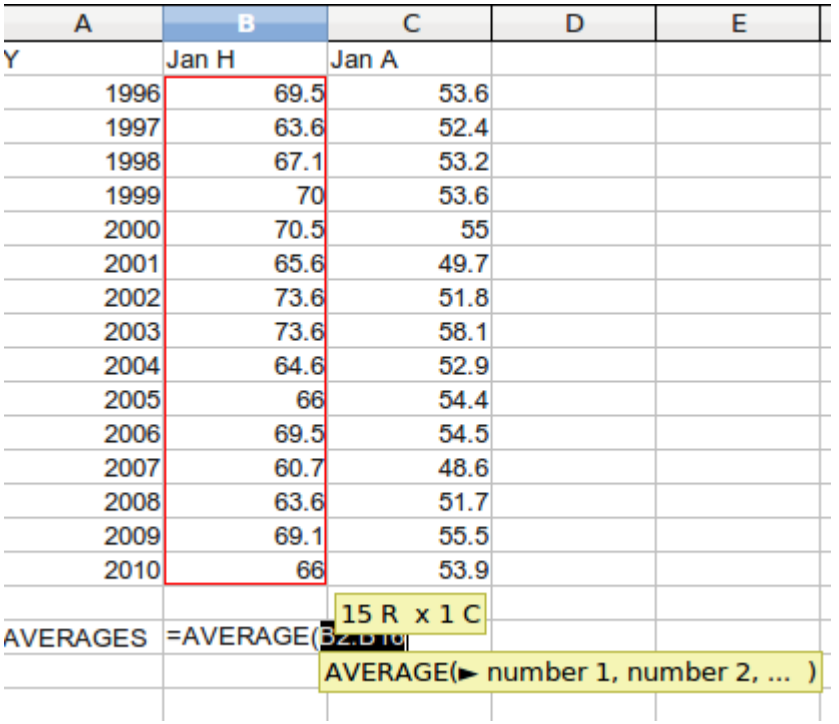
- In the row under the column you want to take the average for type =AVERAGE(
-



	A	B	C
1	Y	Jan H	Jan A
2	1996	69.5	53.6
3	1997	63.6	52.4
4	1998	67.1	53.2
5	1999	70	53.6
6	2000	70.5	55
7	2001	65.6	49.7
8	2002	73.6	51.8
9	2003	73.6	58.1
10	2004	64.6	52.9
11	2005	66	54.4
12	2006	69.5	54.5
13	2007	60.7	48.6
14	2008	63.6	51.7
15	2009	69.1	55.5
16	2010	66	53.9
17			
18	AVERAGES	=AVERAGE(	

Mean In Excel

- Now highlight the numbers you want to average
- To do this just place your cursor over the first cell in the column and drag to the last



Mean In Excel

- Now just finish by either hitting enter or close the parantheses
- What's our answer?

Mean In Excel

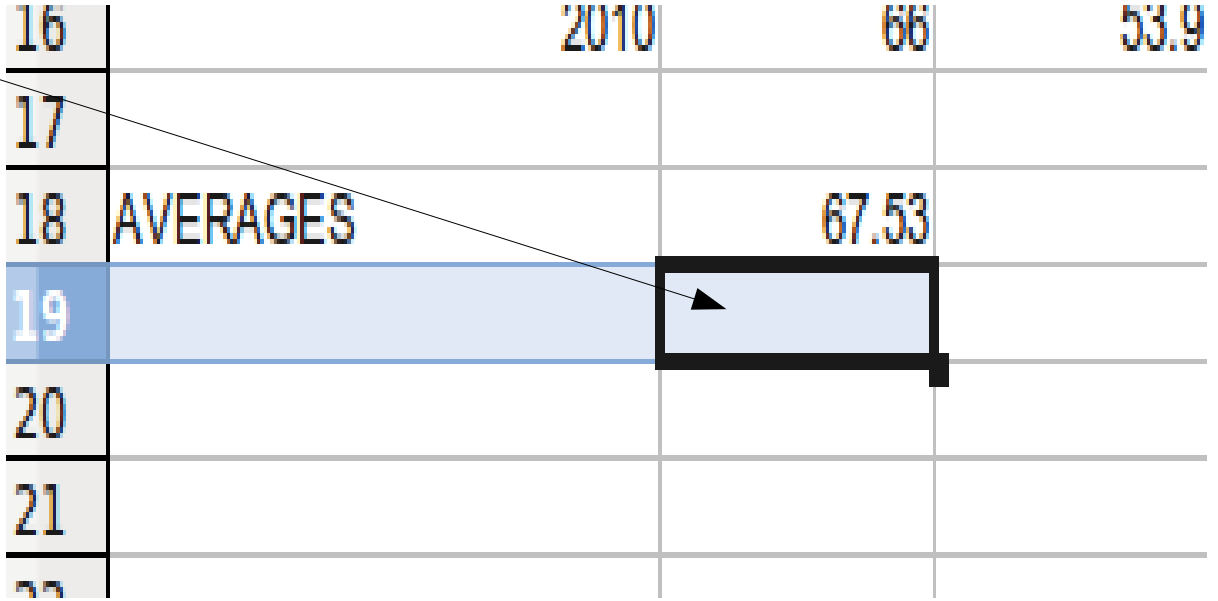
- Now just finish by either hitting enter or close the parantheses
- What's our answer?

Mean In Excel

A	B	C
Y	Jan H	Jan A
1996	69.5	53.6
1997	63.6	52.4
1998	67.1	53.2
1999	70	53.6
2000	70.5	55
2001	65.6	49.7
2002	73.6	51.8
2003	73.6	58.1
2004	64.6	52.9
2005	66	54.4
2006	69.5	54.5
2007	60.7	48.6
2008	63.6	51.7
2009	69.1	55.5
2010	66	53.9
AVERAGES	67.53	

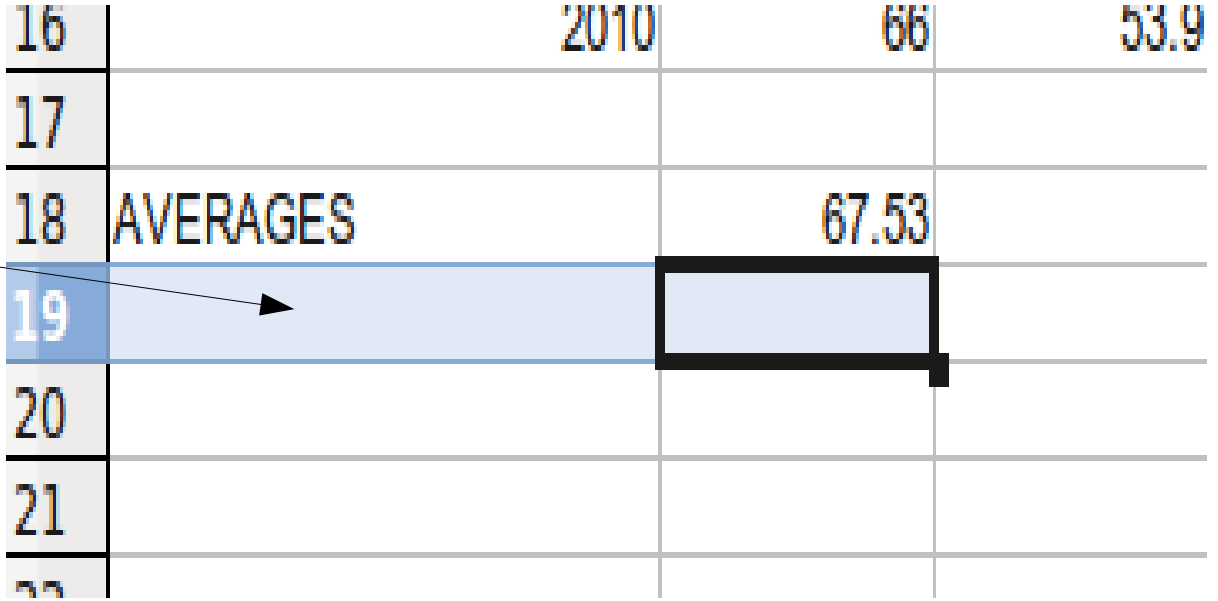
The Standard Deviation

- Select a cell
- Label it
- Type =stdev(
- Highlight the cells
- Hit enter



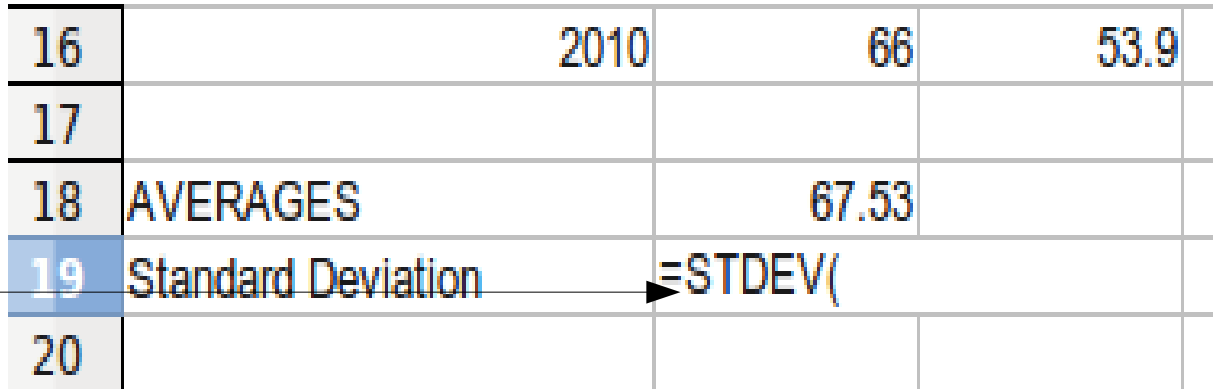
The Standard Deviation

- Select a cell
- Label it
- Type =stdev(
- Highlight the cells
- Hit enter



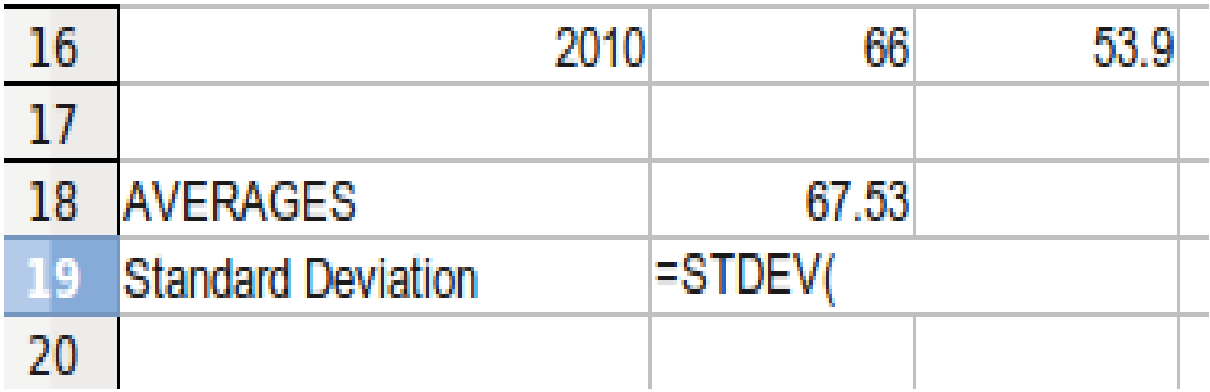
The Standard Deviation

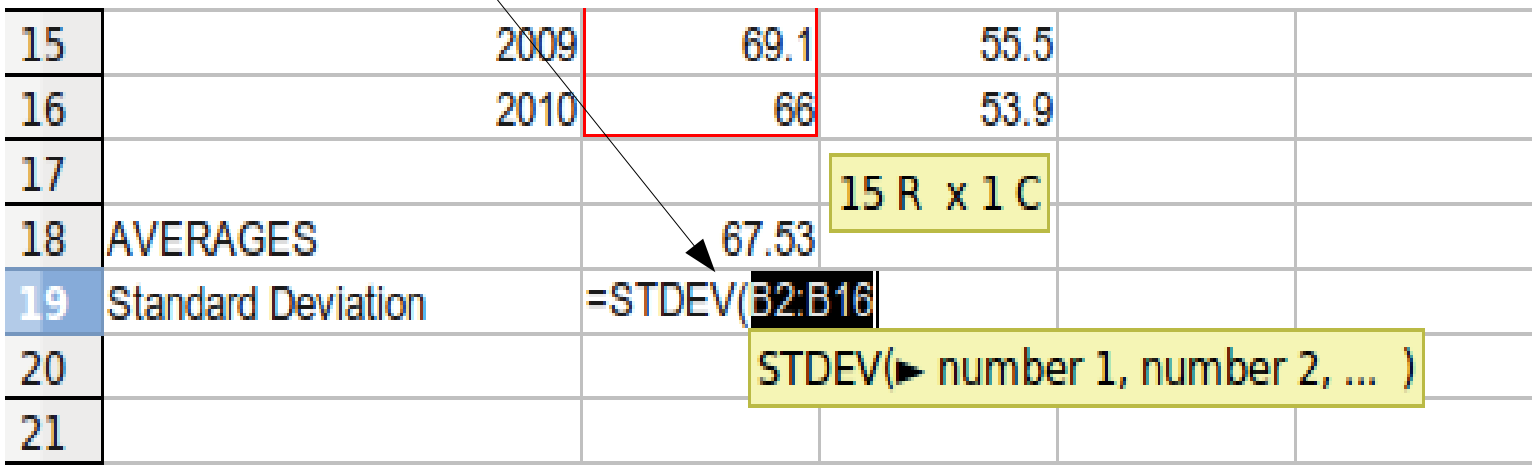
- Select a cell
- Label it
- Type =stdev(
- Highlight the cells
- Hit enter



The Standard Deviation

- Select a cell
- Label it
- Type =stdev(
- Highlight the cells
- Hit enter





The Correlation Coefficient as a Cause Detecting Tool

- The correlation coefficient by itself can not detect a cause
- There are different types of correlation coefficients.
  - These different correlation coefficients are used for different types of data
  - Review:
  - What's the difference between Nominal, Ordinal, Interval, and Ratio type data?

The Correlation Coefficient as a Cause Detecting Tool

- Nominal
  - r_{phi} (phi coefficient) Two dichotomous variables
  - r_b (biserial *r*) One dichotomous variable with continuity assumed
  - r_t (tetrachoric) Two dichotomous variables in which underlying continuity can be assumed
- *Ordinal*
  - r_s (Spearman *r*) Ranked data. Both measures must be at least ordinal.
  - τ (Kendall's tau or rank order correlation
- Interval or ratio
  - *Pearson r* Both scales interval or ratio
- We will primarily use the *Pearson r*

Hypothesis

- We want to test a global warming hypothesis for Tucson.
- This past January seemed particularly cold
- I have data from 1996 to 2010 on both January average high (Jan H) and the January average for the daily averages (Jan A)
- To test this hypothesis we are going to see if there is a correlation between the passage of time and these temperatures using a correlation coefficient

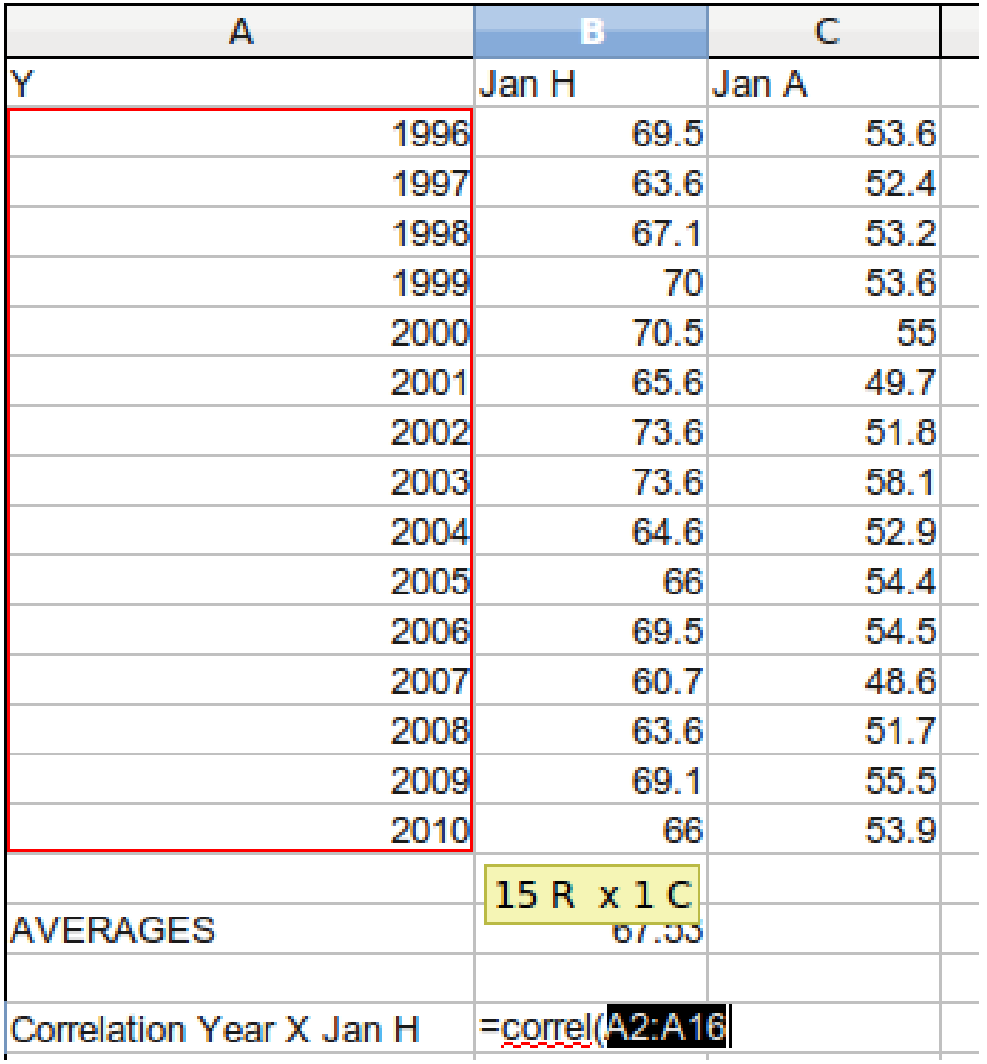
Correlation Coefficient In excel

- choose a cell
- Label it
- Type =correl(

	A	B	C	D
	Y	Jan H	Jan A	
	1996	69.5	53.6	
	1997	63.6	52.4	
	1998	67.1	53.2	
	1999	70	53.6	
	2000	70.5	55	
	2001	65.6	49.7	
	2002	73.6	51.8	
	2003	73.6	58.1	
	2004	64.6	52.9	
	2005	66	54.4	
	2006	69.5	54.5	
	2007	60.7	48.6	
	2008	63.6	51.7	
	2009	69.1	55.5	
	2010	66	53.9	
	AVERAGES	67.53		
	Correlation Year X Jan H	=correl(		
		CORREL(► Data_1, Data_2)		

Correlation Coefficient In excel

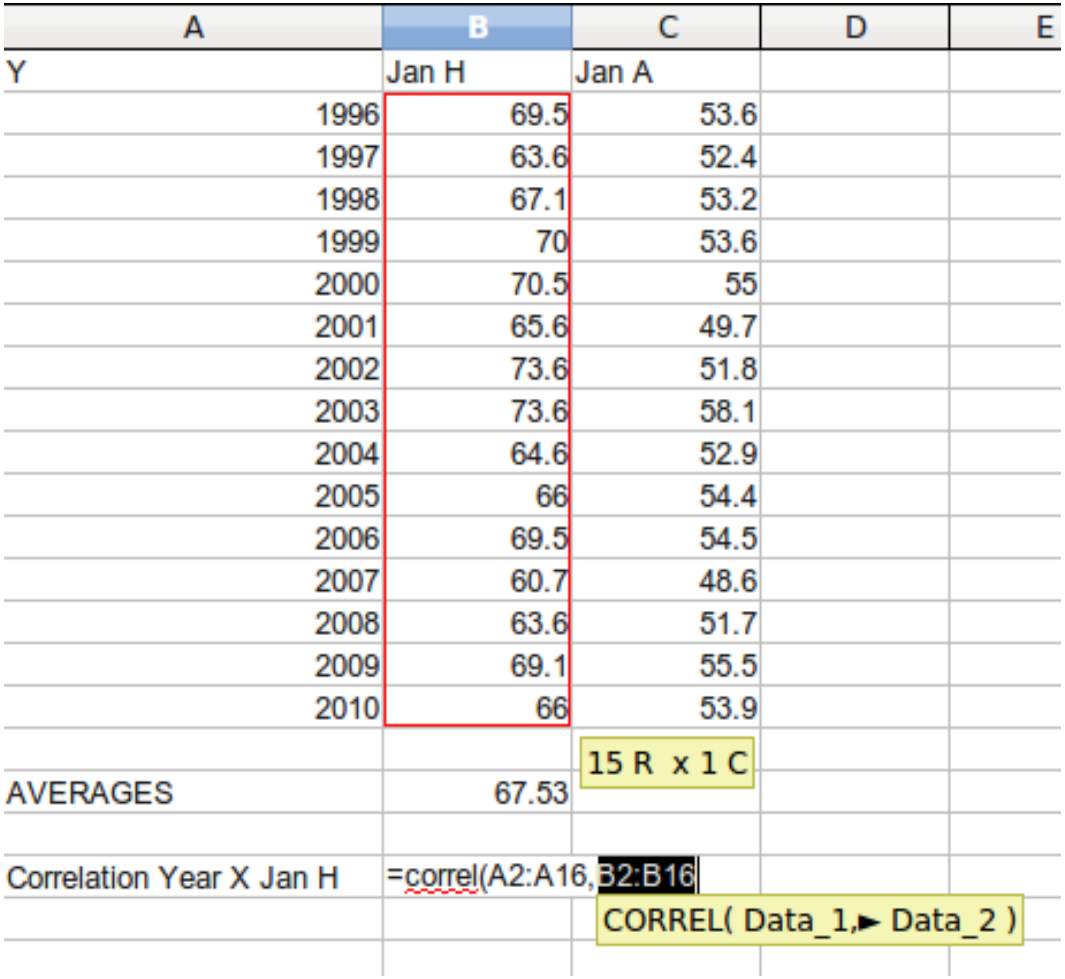
- Highlight all values for Y



A	B	C
Y	Jan H	Jan A
1996	69.5	53.6
1997	63.6	52.4
1998	67.1	53.2
1999	70	53.6
2000	70.5	55
2001	65.6	49.7
2002	73.6	51.8
2003	73.6	58.1
2004	64.6	52.9
2005	66	54.4
2006	69.5	54.5
2007	60.7	48.6
2008	63.6	51.7
2009	69.1	55.5
2010	66	53.9
	15 R x 1 C	
AVERAGES	67.33	
Correlation Year X Jan H	=correl(A2:A16	

Correlation Coefficient In excel

- type a comma and highlight all values for Jan H and hit enter



	A	B	C	D	E
	Y	Jan H	Jan A		
	1996	69.5	53.6		
	1997	63.6	52.4		
	1998	67.1	53.2		
	1999	70	53.6		
	2000	70.5	55		
	2001	65.6	49.7		
	2002	73.6	51.8		
	2003	73.6	58.1		
	2004	64.6	52.9		
	2005	66	54.4		
	2006	69.5	54.5		
	2007	60.7	48.6		
	2008	63.6	51.7		
	2009	69.1	55.5		
	2010	66	53.9		
			15 R x 1 C		
	AVERAGES	67.53			
	Correlation Year X Jan H	=correl(A2:A16,B2:B16			
			CORREL(Data_1,► Data_2)		

Correlation Coefficient In excel

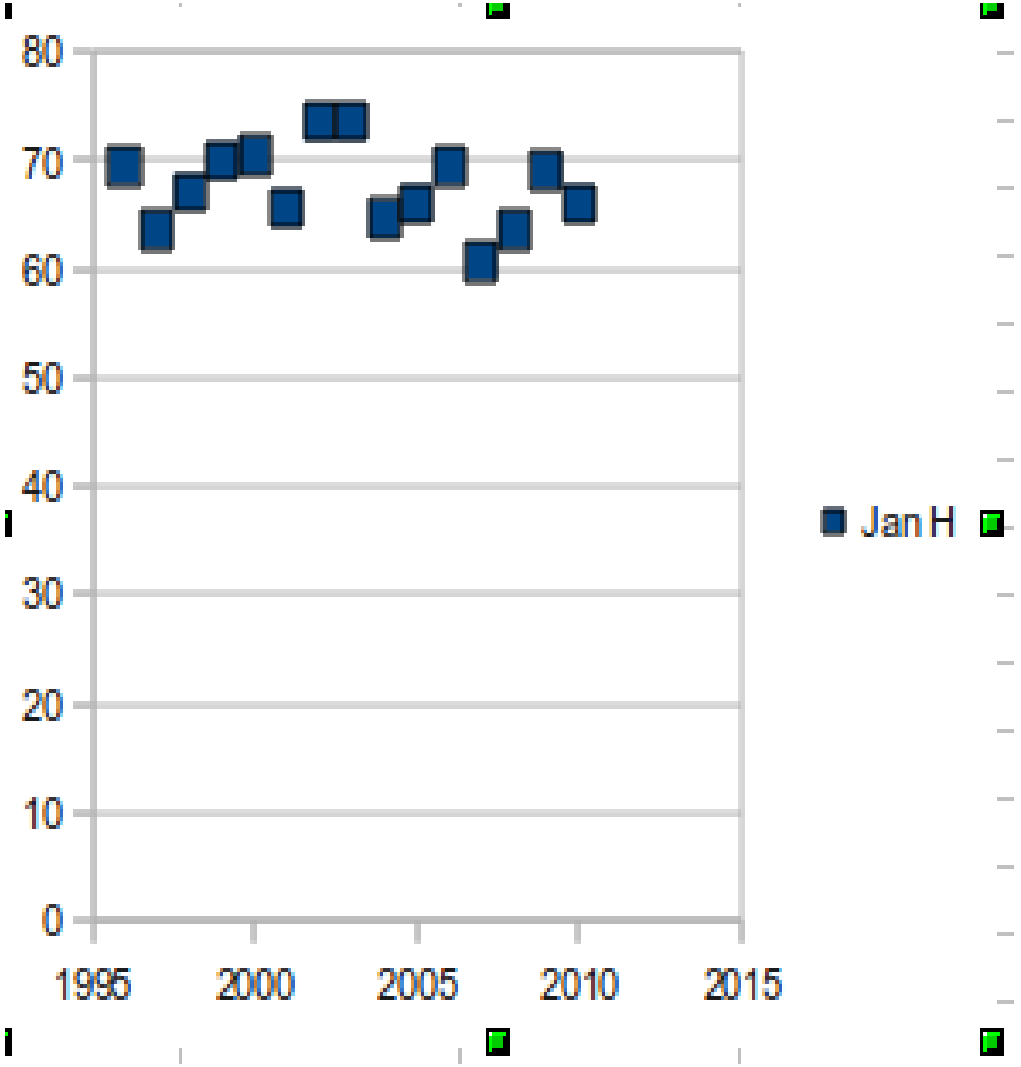
- Does this result support the global warming hypothesis?

A	B	C
Y	Jan H	Jan A
1996	69.5	53.6
1997	63.6	52.4
1998	67.1	53.2
1999	70	53.6
2000	70.5	55
2001	65.6	49.7
2002	73.6	51.8
2003	73.6	58.1
2004	64.6	52.9
2005	66	54.4
2006	69.5	54.5
2007	60.7	48.6
2008	63.6	51.7
2009	69.1	55.5
2010	66	53.9
AVERAGES	67.53	
Correlation Year X Jan H	-0.25	

Let's see what the data look like?

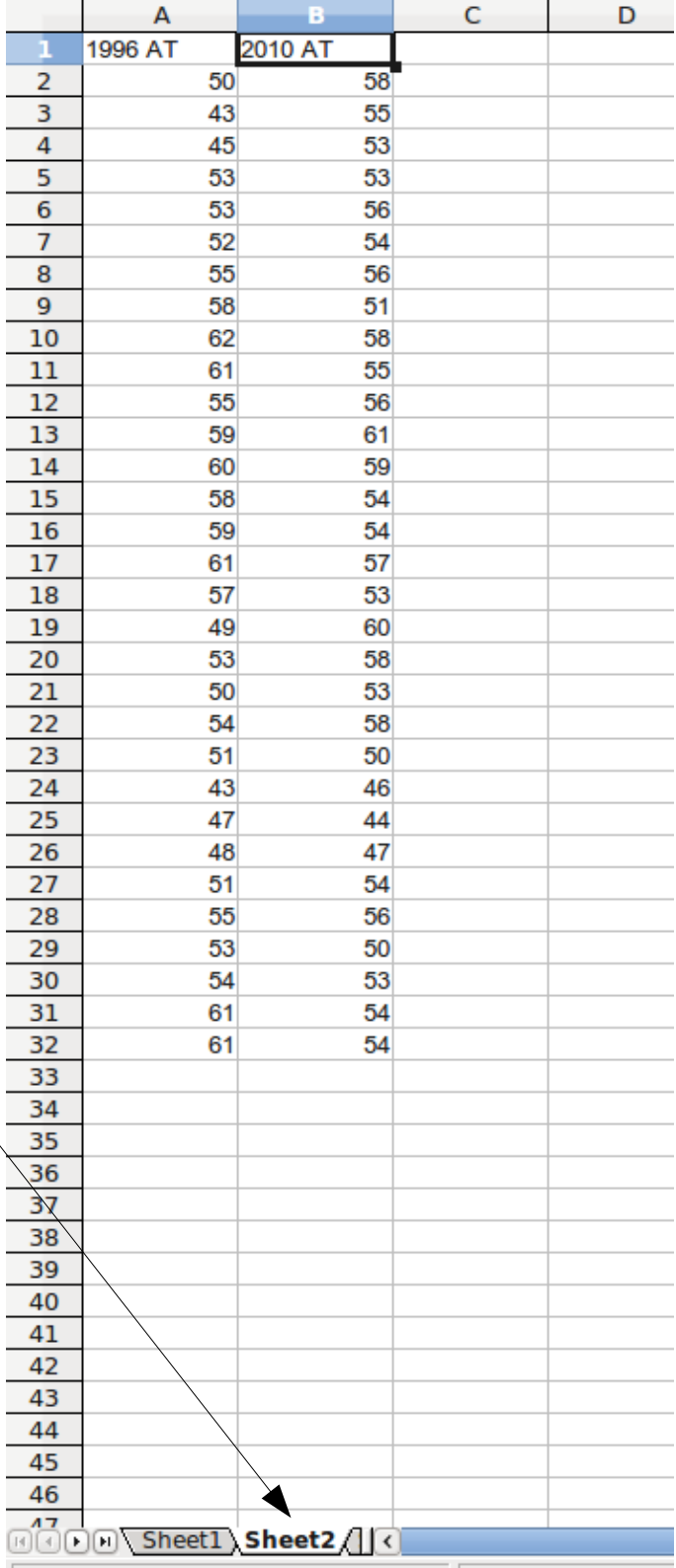
- Highlight the two columns
- Click on the chart
- Select Scatterplot

What do the data look like?



Another test of the hypothesis

- Select sheet two

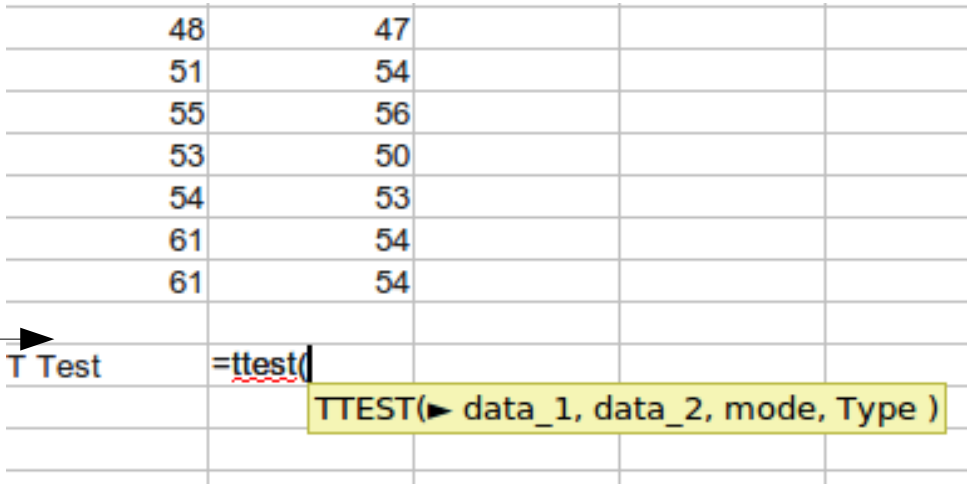


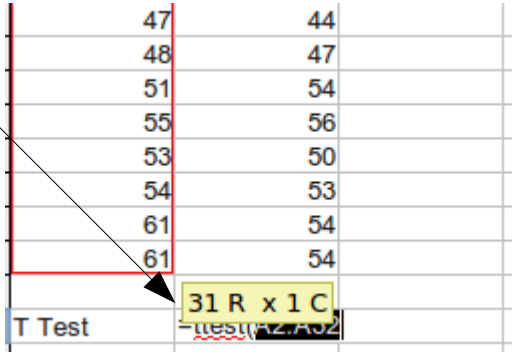
	A	B	C	D
1	1996 AT	2010 AT		
2	50	58		
3	43	55		
4	45	53		
5	53	53		
6	53	56		
7	52	54		
8	55	56		
9	58	51		
10	62	58		
11	61	55		
12	55	56		
13	59	61		
14	60	59		
15	58	54		
16	59	54		
17	61	57		
18	57	53		
19	49	60		
20	53	58		
21	50	53		
22	54	58		
23	51	50		
24	43	46		
25	47	44		
26	48	47		
27	51	54		
28	55	56		
29	53	50		
30	54	53		
31	61	54		
32	61	54		
33				
34				
35				
36				
37				
38				
39				
40				
41				
42				
43				
44				
45				
46				

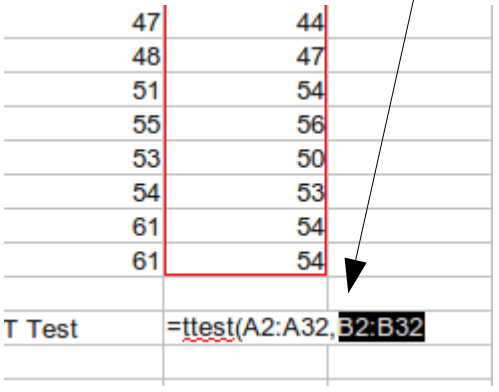
Sheet1 Sheet2

A t-test in excel

- Again select a cell
- Label it
- Type =ttest(
- Highlight your first group
- Type a comma
- Highlight the second

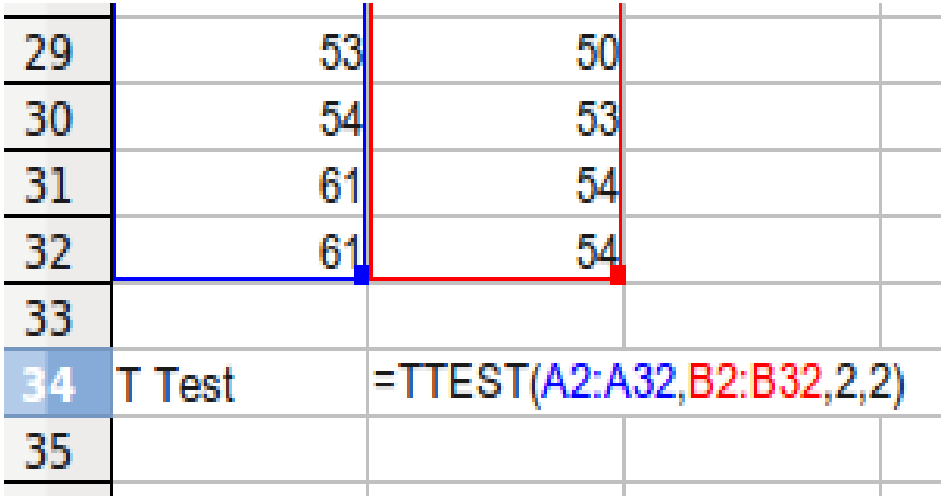




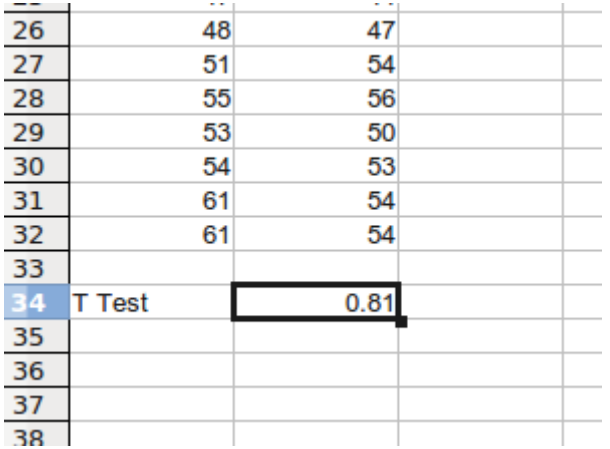


A t-test in excel

- type a comma
- Select one or two tails by entering a one
- type a comma
- Select the type of t-test- we are going to assume equal variances so type a 2



Answer



26	48	47
27	51	54
28	55	56
29	53	50
30	54	53
31	61	54
32	61	54
33		
34	T Test	0.81
35		
36		
37		
38		

Homework

- Write a rough draft for an introduction for next week
  - Rubric will be posted
- Also turn in a complete spreadsheet in an excel file
- All the variables need descriptive statistics
  - Mean and standard deviation
- I want correlations between all the variables in sheet 1
- I want a 2 tailed t test assuming equal variances in sheet 2
- All due next Tuesday by 10 p.m.