CS2034B / DH2144B

Data Analytics: Principles and Tools

Western
UNIVERSITY · CANADA

Week 3

Statistics,
Data Preparation & Transformation

Statistics: Basics

Assigned Readings / Tasks

- zyBooks Chapter 5 Statistics Basics

Optional:

- Pearson Product-Moment Correlation
- CORREL function

Measures of Center in Excel

Name	Description	Equation	Excel Function
Mean / Average	The central value of a discrete set of numbers: specifically, the sum of the values divided by the number of values.	$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$	AVERAGE(number1, [number2], ...)
Median	The value separating the higher half from the lower half of a data sample. That is the number in the middle of a set of numbers.	$\{(n + 1) \div 2\}^{th}$ value in an ordered set of odd length. $\frac{\{(n\div 2)\}^{th} \text{ value} + \{(n\div 2+1)\}^{th} \text{ value}}{2}$ in an ordered set of even length.	MEDIAN(number1, [number2], ...)
Mode	The value that appears most often in a set of numbers.	N/A	MODE(number1,[number2],...)

Measures of Center in Excel

Example 1:

	A	B	C	D	E	F	G	H
1	London Ontario Weather					Measures of Center		
2	Date	High (C)	Low (C)				High	Low
3	22-Jan-19	-4	-9			Average:		
4	22-Jan-18	8	1			Median:		
5	22-Jan-17	6	2			Mode:		
	-16	-5	-10					
	-15	-2	-9					
	-14	-14	-24					
	-13	-13	-17					
	-12	2	-11					
	-11	-10	-15					
	-10	2	-4					
	-09	-3	-7					
	-08	-1	-10					
15	22-Jan-07	-4	-8					
16	22-Jan-06	2	-5					
17	22-Jan-05	-12	-19					
18	22-Jan-04	2	-19					
19	22-Jan-03	-12	-16					
20	22-Jan-02	22	8					

Find the Mean (Average), Median and Mode of the high and low temperatures using Excel Functions

Measures of Center in Excel

Example 1:

	A	B	C
1	London Ontario Weather		
2	Date	High (C)	Low (C)
3	22-Jan-19	-4	-9
4	22-Jan-18	8	1
5	22-Jan-17	6	2
6	22-Jan-16	-5	-10
7	22-Jan-15	-2	-9
8	22-Jan-14	-14	-24
9	22-Jan-13	-13	-17
10	22-Jan	[illegible]	[illegible]
11	22-Jan	[illegible]	[illegible]
12	22-Jan	[illegible]	[illegible]
13	22-Jan	[illegible]	[illegible]
14	22-Jan	[illegible]	[illegible]
15	22-Jan	[illegible]	[illegible]
16	22-Jan	[illegible]	[illegible]
17	22-Jan-05	-12	-19
18	22-Jan-04	2	-19
19	22-Jan-03	-12	-16
20	22-Jan-02	22	8

F	G	H
Measures of Center		
	High	Low
Average:	-2.0	-9.6
Median:	-2.5	-9.5
Mode:	2	-9

F	G	H
Measures of Center		
	High	Low
Average:	=AVERAGE(B3:B20)	=AVERAGE(C3:C20)
Median:	=MEDIAN(B3:B20)	=MEDIAN(C3:C20)
Mode:	=MODE(B3:B20)	=MODE(C3:C20)

Measures of Spread

Name	Description	Equation	Excel Function
Maximum	The largest value in the data set.	N/A	MAX(num1, [num2], ...)
Minimum	The smallest value in the data set.	N/A	MIN(num1, [num2], ...)
Range	The difference between the largest and smallest values in the data set.	=Max(X) - Min(X) where X is the dataset	N/A
Mean Absolute Deviation	The average distance from the average.	$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - m(X)\rvert.$	AVEDEV(num1, [num2], ...)
Median Absolute Deviation	The median distance from the median.	$\mathrm{MAD} = \mathrm{median}(\lvert X_i - \tilde{X}\rvert)$ $\tilde{X} = \mathrm{median}(X)$	N/A
Variance	How far a set of numbers is spread out from their average.	$\frac{\sum_{i=1}^{n}(d_i - mean)^2}{n-1}$	VAR.S(num1,[num2],...) VAR.P(num1,[num2],...)
Standard Deviation	The the square root of the variance. Measure used to quantify the amount of variation in a data set.	$\sqrt{variance(X)}$	STDEV.S(num1,[num2],...) STDEV.P(num1,[num2],...)

Measures of Spread

Example 2 Part 1:

Find the Maximum, Minimum and Range of the Highs.

	A	B	C	D	E	F	G	H	I
1	London Ontario Weather								
2	Date	High (C)		Abs Difference from Mean	Abs Difference from Median	Difference from Mean Squared		Measures of Center	
3	22-Jan-19	-4						Average:	-2.0
4	22-Jan-18	8						Median:	-2.5
5	22-Jan-17	6						Mode:	2
6	22-Jan-16	-5							
7	22-Jan-15	-2							
8	22-Jan-14	-14						Measures of Spread	
9	22-Jan-13	-13						Maximum:	
10	22-Jan-12	2						Minimum:	
11	22-Jan-11	-10						Range	
12	22-Jan-10	2							
13	22-Jan-09	-3						Mean Absolute Deviation	
14	22-Jan-08	-1						Median Absolute Deviation	
15	22-Jan-07	-4							
16	22-Jan-06	2						Variance:	
17	22-Jan-05	-12						Standard Deviation:	
18	22-Jan-04	2							
19	22-Jan-03	-12							
20	22-Jan-02	22							

Measures of Spread

Example 2 Part 1:

Find the Maximum, Minimum and Range of the Highs.

London Ontario Weather	
Date	High (C)
22-Jan-19	-4
22-Jan-18	8
22-Jan-17	6
22-Jan-16	-5
22-Jan-15	-2
22-Jan-14	-14
22-Jan-13	-13
22-Jan-12	2
22-Jan-11	-10
22-Jan-10	2

Abs Difference from Mean	Abs Difference from Median	Difference from Mean Squared

Measures of Center	
Average:	-2.0
Median:	-2.5
Mode:	2

Measures of Spread	
Maximum:	22.0
Minimum:	-14.0
Range	36.0

Mean Absolute Deviation	
Median Absolute Deviation	
Variance:	
Standard Deviation:	

Measures of Spread	
Maximum:	=MAX(B3:B20)
Minimum:	=MIN(B3:B20)
Range	=I9-I10

Measures of Spread

Example 2 Part 2:

Find the absolute differences and use them to find the Mean and Median Absolute Deviations. Don't use AVEDEV.

	A	B	C	D	E	F	G	H	I
1	London Ontario Weather								
2	Date	High (C)		Abs Difference from Mean	Abs Difference from Median	Difference from Mean Squared		Measures of Center	
3	22-Jan-19	-4						Average:	-2.0
4	22-Jan-18	8						Median:	-2.5
5	22-Jan-17	6						Mode:	2
6	22-Jan-16	-5							
7	22-Jan-15	-2							
8	22-Jan-14	-14						Measures of Spread	
9	22-Jan-13	-13						Maximum:	22.0
10	22-Jan-12	2						Minimum:	-14.0
11	22-Jan-11	-10						Range	36.0
12	22-Jan-10	2							
13	22-Jan-09	-3						Mean Absolute Deviation	
14	22-Jan-08	-1						Median Absolute Deviation	
15	22-Jan-07	-4							
16	22-Jan-06	2						Variance:	
17	22-Jan-05	-12						Standard Deviation:	
18	22-Jan-04	2							
19	22-Jan-03	-12							
20	22-Jan-02	22							

Measures of Spread

Example 2 Part 2:

Find the absolute differences and use them to find the Mean and Median Absolute Deviations. Don't use AVEDEV.

	A	B	C	D	E	F	G	H	I
1	London Ontario Weather								
2	Date	High (C)		Abs Difference from Mean	Abs Difference from Median	Difference from Mean Squared		Measures of Center	
3	22-Jan-19	-4		=ABS(B3-I3)	=ABS(B3-I4)			Average:	-2.0
4	22-Jan-18	8		=ABS(B4-I3)	=ABS(B4-I4)			Median:	-2.5
5	22-Jan-17	6		=ABS(B5-I3)	=ABS(B5-I4)			Mode:	2
6	22-Jan-16	-5		=ABS(B6-I3)	=ABS(B6-I4)				
7	22-Jan-15	-2		=ABS(B7-I3)	=ABS(B7-I4)				
8	22-Jan-14	-14		=ABS(B8-I3)	=ABS(B8-I4)			Measures of Spread	
9	22-Jan-13	-13		=ABS(B9-I3)	=ABS(B9-I4)			Maximum:	22.0
10	22-Jan-12	2		=ABS(B10-I3)	=ABS(B10-I4)			Minimum:	-14.0
11	22-Jan-11	-10		=ABS(B11-I3)	=ABS(B11-I4)			Range	36.0
12	22-Jan-10	2		=ABS(B12-I3)	=ABS(B12-I4)				
13	22-Jan-09	-3		=ABS(B13-I3)	=ABS(B13-I4)			Mean Absolute Deviation	=AVERAGE(D3:D20)
14	22-Jan-08	-1		=ABS(B14-I3)	=ABS(B14-I4)			Median Absolute Deviation	=MEDIAN(E3:E20)
15	22-Jan-07	-4		=ABS(B15-I3)	=ABS(B15-I4)				
16	22-Jan-06	2		=ABS(B16-I3)	=ABS(B16-I4)			Variance:	
17	22-Jan-05	-12		=ABS(B17-I3)	=ABS(B17-I4)			Standard Deviation:	
18	22-Jan-04	2		=ABS(B18-I3)	=ABS(B18-I4)				
19	22-Jan-03	-12		=ABS(B19-I3)	=ABS(B19-I4)				
20	22-Jan-02	22		=ABS(B20-I3)	=ABS(B20-I4)				

Measures of Spread

Example 2 Part 3:

Find the variance and standard deviation without using VAR.S or STDEV.S.

	A	B	C	D	E	F	G	H	I
1	London Ontario Weather								
2	Date	High (C)		Abs Difference from Mean	Abs Difference from Median	Difference from Mean Squared		Measures of Center	
3	22-Jan-19	-4		2	2			Average:	-2.0
4	22-Jan-18	8		10	11			Median:	-2.5
5	22-Jan-17	6		8	9			Mode:	2
6	22-Jan-16	-5		3	3				
7	22-Jan-15	-2		0	1				
8	22-Jan-14	-14		12	12			Measures of Spread	
9	22-Jan-13	-13		11	11			Maximum:	22.0
10	22-Jan-12	2		4	5			Minimum:	-14.0
11	22-Jan-11	-10		8	8			Range	36.0
12	22-Jan-10	2		4	5				
13	22-Jan-09	-3		1	1			Mean Absolute Deviation	6.6
14	22-Jan-08	-1		1	2			Median Absolute Deviation	4.5
15	22-Jan-07	-4		2	2				
16	22-Jan-06	2		4	5			Variance:	
17	22-Jan-05	-12		10	10			Standard Deviation:	
18	22-Jan-04	2		4	5				
19	22-Jan-03	-12		10	10				
20	22-Jan-02	22		24	25				

Measures of Spread

Example 2 Part 3:

Find the variance and standard deviation without using VAR.S or STDEV.S.

	A	B	C	D	E	F	G	H	I
1	London Ontario Weather								
2	Date	High (C)		Abs Difference from Mean	Abs Difference from Median	Difference from Mean Squared		Measures of Center	
3	22-Jan-19	-4		2	2	=D3^2		Average:	-2.0
4	22-Jan-18	8		10	11	=D4^2		Median:	-2.5
5	22-Jan-17	6		8	9	=D5^2		Mode:	2
6	22-Jan-16	-5		3	3	=D6^2			
7	22-Jan-15	-2		0	1	=D7^2			
8	22-Jan-14	-14		12	12	=D8^2		Measures of Spread	
9	22-Jan-13	-13		11	11	=D9^2		Maximum:	22.0
10	22-Jan-12	2		4	5	=D10^2		Minimum:	-14.0
11	22-Jan-11	-10		8	8	=D11^2		Range	36.0
12	22-Jan-10	2		4	5	=D12^2			
13	22-Jan-09	-3		1	1	=D13^2		Mean Absolute Deviation	6.6
14	22-Jan-08	-1		1	2	=D14^2		Median Absolute Deviation	4.5
15	22-Jan-07	-4		2	2	=D15^2			
16	22-Jan-06	2		4					
17	22-Jan-05	-12		10					
18	22-Jan-04	2		4	5	=D18^2			
19	22-Jan-03	-12		10	10	=D19^2			
20	22-Jan-02	22		24	25	=D20^2			

Variance:	=SUM(F3:F20)/(COUNT(F3:F20)-1)
Standard Deviation:	=SQRT(I16)

Measures of Spread

Example 2 Part 3:

Find the variance and standard deviation without using VAR.S or STDEV.S.

	A	B	C	D	E	F
1	London Ontario Weather					
2	Date	High (C)		Abs Difference from Mean	Abs Difference from Median	Difference from Mean Squared
3	22-Jan-19	-4		2	2	4
4	22-Jan-18	8		10	11	100
5	22-Jan-17	6		8	9	64
6	22-Jan-16	-5		3	3	9
7	22-Jan-15	-2		0	1	0
8	22-Jan-14	-14		12	12	144
9	22-Jan-13	-13		11	11	121
10	22-Jan-12	2		4	5	16
11	22-Jan-11	-10		8	8	64
12	22-Jan-10	2		4	5	16
13	22-Jan-09	-3		1	1	1
14	22-Jan-08	-1		1	2	1
15	22-Jan-07	-4		2	2	4
16	22-Jan-06	2		4	5	16
17	22-Jan-05	-12		10	10	100
18	22-Jan-04	2		4	5	16
19	22-Jan-03	-12		10	10	100
20	22-Jan-02	22		24	25	576

	H	I
2	Measures of Center	
3	Average:	-2.0
4	Median:	-2.5
5	Mode:	2
8	Measures of Spread	
9	Maximum:	22.0
10	Minimum:	-14.0
11	Range	36.0
13	Mean Absolute Deviation	6.6
14	Median Absolute Deviation	4.5
16	Variance:	79.5
17	Standard Deviation:	8.9

Probability Mass Functions

- A **probability mass function (pmf)** assigns the probability that a discrete random variable is exactly equal to some value.

Example:
Alice, a teaching assistant for CS2034, has been keeping track of how many students attend her office hours. She has found the following **pmf** for the number of students attending her hours each week:

x = # Students	0	2	5	10
p(x)	0.25	0.45	0.13	0.17

What is the likely hood of Alice having 5 students attend in a given week?

Probability Mass Functions

- A **probability mass function (pmf)** assigns the probability that a discrete random variable is exactly equal to some value.

Example:

Alice, a teaching assistant for CS2034, has been keeping track of how many students attend her office hours. She has found the following **pmf** for the number of students attending her hours each week:

13%

x = # Students	**0**	**2**	**5**	**10**
p(x)	0.25	0.45	$\boxed{0.13}$	0.17

What is the likely hood of Alice having 5 students attend in a given week?

Probability Mass Functions

- The **cumulative distribution function (cdf)** is the probability that for any number y, the observed value of the random variable will be at most y or **p**($x \le y$).

Example:

The following table now also contains the **cdf** for the values Alice calculated.

x = # Students	**0**	**2**	**5**	**10**
p(x)	0.25	0.45	0.13	0.17
F(x)	0.25	0.70	0.83	1.00

What is the likely hood of Alice having 0 to 2 (inclusive) students attend her office hours in a given week?

Probability Mass Functions

- The **cumulative distribution function (cdf)** is the probability that for any number y, the observed value of the random variable will be at most y or **p**(x ≤ y).

Example:

The following table now also contains the **cdf** for the values Alice calculated.

x = # Students	0	2	5	10
p(x)	0.25	0.45	0.13	0.17
F(x)	0.25	$\boxed{0.70}$	0.83	1.00

70%

What is the likely hood of Alice having 0 to 2 (inclusive) students attend her office hours in a given week?

Expected Value

- The **expected value (μ)** is the sum of the possible values of X multiplied by the probability of the value.

Example:

How many students can Alice expect to attend her office hours each week?

x = # Students	0	2	5	10
p(x)	0.25	0.45	0.13	0.17

Expected value (μ) =

Expected Value

- The **expected value (μ)** is the sum of the possible values of X multiplied by the probability of the value.

Example:
How many students can Alice expect to attend her office hours each week?

x = # Students	**0**	**2**	**5**	**10**
p(x)	0.25	0.45	0.13	0.17

Expected value (μ) = 0.25*0 + 0.45*2 + 0.13*5 + 0.17*10
= 3.25

Variance and Standard Deviation

- The **variance** of a discrete random variable, X, is a measure of the spread of a distribution. The variance is calculated using the following equation:

$$\sigma^2 = V(X) = \sum (x - \mu)^2 * p(x).$$

- The **standard deviation** is a measure of the spread in the units of the original random variable. The standard deviation is the square root of the variance.

$$\sqrt{\sigma^2}$$

Variance and Standard Deviation

Example:

Find the **variance** and **standard deviation** of the number of students in the last example.

Expected value (μ) = 3.25

x = # Students	**0**	**2**	**5**	**10**
p(x)	0.25	0.45	0.13	0.17
$x-\mu$				
$(x-\mu)^2$				
$p(x)*(x-\mu)^2$				

Variance =

Standard Deviation =

Variance and Standard Deviation

Example:

Find the **variance** and **standard deviation** of the number of students in the last example.

Expected value (μ) = 3.25

x = # Students	0	2	5	10
p(x)	0.25	0.45	0.13	0.17
x-μ	-3.25	-1.25	1.75	6.75
$(x-\mu)^2$				
$p(x)*(x-\mu)^2$				

Variance =

Standard Deviation =

Variance and Standard Deviation

Example:

Find the **variance** and **standard deviation** of the number of students in the last example.

Expected value (μ) = 3.25

x = # Students	**0**	**2**	**5**	**10**
p(x)	0.25	0.45	0.13	0.17
$x-\mu$	-3.25	-1.25	1.75	6.75
$(x-\mu)^2$	10.56	1.56	3.06	45.56
$p(x)*(x-\mu)^2$				

Variance =

Standard Deviation =

Variance and Standard Deviation

Example:

Find the **variance** and **standard deviation** of the number of students in the last example.

Expected value (μ) = 3.25

x = # Students	**0**	**2**	**5**	**10**
p(x)	0.25	0.45	0.13	0.17
x-μ	-3.25	-1.25	1.75	6.75
$(x-\mu)^2$	10.56	1.56	3.06	45.56
$p(x)*(x-\mu)^2$	2.64	0.70	0.40	7.75

Variance =

Standard Deviation =

Variance and Standard Deviation

Example:

Find the **variance** and **standard deviation** of the number of students in the last example.

Expected value (μ) = 3.25

x = # Students	**0**	**2**	**5**	**10**
p(x)	0.25	0.45	0.13	0.17
x-μ	-3.25	-1.25	1.75	6.75
$(x-\mu)^2$	10.56	1.56	3.06	45.56
$p(x)*(x-\mu)^2$	2.64	0.70	0.40	7.75

Variance = 2.64 + 0.70 + 0.40 + 7.75

= 11.49

Standard Deviation =

Variance and Standard Deviation

Example:

Find the **variance** and **standard deviation** of the number of students in the last example.

Expected value (μ) = 3.25

x = # Students	**0**	**2**	**5**	**10**
p(x)	0.25	0.45	0.13	0.17
$x-\mu$	-3.25	-1.25	1.75	6.75
$(x-\mu)^2$	10.56	1.56	3.06	45.56
$p(x)*(x-\mu)^2$	2.64	0.70	0.40	7.75

Variance = 2.64 + 0.70 + 0.40 + 7.75

= 11.49

Standard Deviation = $\sqrt{11.49}$ = 3.39

Excel Example

Example 3:

Redo the calculations we just did in Excel including finding the **cdf** values. Don't hardcode any values, the results should change if the original data changes.

	A	B	C	D	E	F	G	H
1	**x = # Students**	**0**	**2**	**5**	**10**		**Expected Value (u):**	=B2*B1+C2*C1+D2*D1+E2*E1
2	**p(x)**	0.25	0.45	0.13	=1-0.83		**Variance:**	=SUM(B7:E7)
3	**F(x)**	=B2	=B3+C2	=C3+D2	=D3+E2		**Standard Deviation:**	=SQRT(H2)
4								
5	**x - u:**	=B1-H1	=C1-H1	=D1-H1	=E1-H1			
6	**(x - u)^2:**	=B5^2	=C5^2	=D5^2	=E5^2			
7	**p(x)*(x-u)^2:**	=B2*B6	=C2*C6	=D2*D6	=E2*E6			

	A	B	C	D	E	F	G	H
1	**x = # Students**	**0**	**2**	**5**	**10**		**Expected Value (u):**	3.25
2	**p(x)**	0.25	0.45	0.13	0.17		**Variance:**	11.49
3	**F(x)**	0.25	0.7	0.83	1		**Standard Deviation:**	3.39
4								
5	**x - u:**	-3.25	-1.25	1.75	6.75			
6	**(x - u)^2:**	10.56	1.56	3.06	45.56			
7	**p(x)*(x-u)^2:**	2.64	0.70	0.40	7.75			

Binomial Distribution

- The **binomial distribution** models how many times an event occurs in a certain number of trials, with the assumption that the probability of the event is the same for each trial.
- The binomial probability can be found using the equation:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

where **n** is the number of trials, **k** is the number of successes, and **p** is the probability of success.

- In Excel we can use the BINOM.DIST function:

```
BINOM.DIST(k, n, p, cumulative)
```

where **cumulative** is TRUE or FALSE and controls if BINOM.DIST returns the **pmf** (FALSE) or **cdf** (TRUE).

Binomial Distribution

Example 4:

If we flip a fair coin 6 times, what are the odds of getting heads 0 times, 1 time, 2 times, etc.?

	A	B	C	D	E	F	G	H
1	**Attempts:**	6						
2	**Probability of Heads:**	0.5						
3								
4	**x = #of heads**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
5	**p(x)**							
6								

Use the Excel BINOM.DIST function to solve this.

Binomial Distribution

Example 4:

If we flip a fair coin 6 times, what are the odds of getting heads 0 times, 1 time, 2 times, etc.?

	A	B	C	
1	**Attempts:**	=COUNT(B4:H4)-1		
2	**Probability of Heads:**	0.5		
3				
4	**x = #of heads**	**0**	**1**	**2**
5	**p(x)**	=BINOM.DIST(B4,B1,B2,FALSE)	=BINOM.DIST(C4,B1,B2,FALSE)	=BINOM.DIST(
6				

	A	B	C	D	E	F	G	H
1	**Attempts:**	6						
2	**Probability of Heads:**	0.5						
3								
4	**x = #of heads**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
5	**p(x)**	0.02	0.09	0.23	0.31	0.23	0.09	0.02

Negative Binomial Distribution

- What if we want to know how many times we have to attempt a trial before success?
- **Negative Binomial Distribution** models the number of failures in a sequence of trials before a success occurs.
- In Excel we can use the NEGBINOM.DIST function:

```
NEGBINOM.DIST(f, s, p, cumulative)
```

where **f** is the number of failures, **s** is the number of successes, **p** is the probability of success and cumulative is the same TRUE/FALSE value as used in BINOM.DIST.

Negative Binomial Distribution

Example 5:

Bob is a "Hardcore Gamer" and wants to know how many monsters he has to defeat to get an rare item in his game. Each monster has a 5% chance to drop the item when slain.

How many monsters must Bob defeat to have a 90% chance of obtaining the item?

	A	B	C	D	E
1	Probability of Epic Loot:	0.05		90% Chance:	
2					
3	x = # Monsters Slain	F(x)			
4	1				
5	2				
6	3				
7	4				
8	5				
9	6				
10	7				
11	8				
12	9				

Negative Binomial Distribution

Example 5:

Bob is a "Hardcore Gamer" and wants to know how many monsters he has to defeat to get an rare item in his game. Each monster has a 5% chance to drop the item when slain.

How many monsters must Bob defeat to have a 90% chance of obtaining the item?

	A	B	C	D	E
1	**Probability of Epic Loot:**	0.05		**90% Chance:**	=INDEX(A4:A153,MATCH(0.9,B4:B153,1))
2					
3	**x = # Monsters Slain**	**F(x)**			
4	1	=NEGBINOM.DIST(A4-1,1,B1,TRUE)			
5	2	=NEGBINOM.DIST(A5-1,1,B1,TRUE)			
6	3	=NEGBINOM.DIST(A6-1,1,B1,TRUE)			
7	4	=NEGBINOM.DIST(A7-1,1,B1,TRUE)			
8	5	=NEGBINOM.DIST(A8-1,1,B1,TRUE)			
9	6	=NEGBINOM.DIST(A9-1,1,B1,TRUE)			
10	7	=NEGBINOM.DIST(A10-1,1,B1,TRUE)			
11	8	=NEGBINOM.DIST(A11-1,1,B1,TRUE)			
12	9	[illegible]			

Negative Binomial Distribution

Example 5:

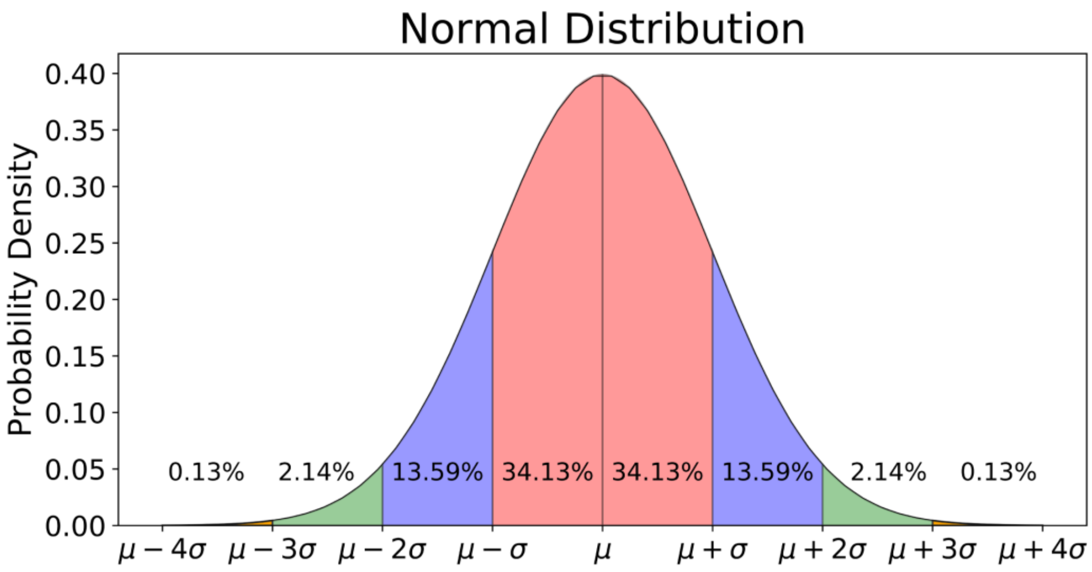
Bob is a "Hardcore Gamer" and wants to know how many monsters he has to defeat to get an rare item in his game. Each monster has a 5% chance to drop the item when slain.

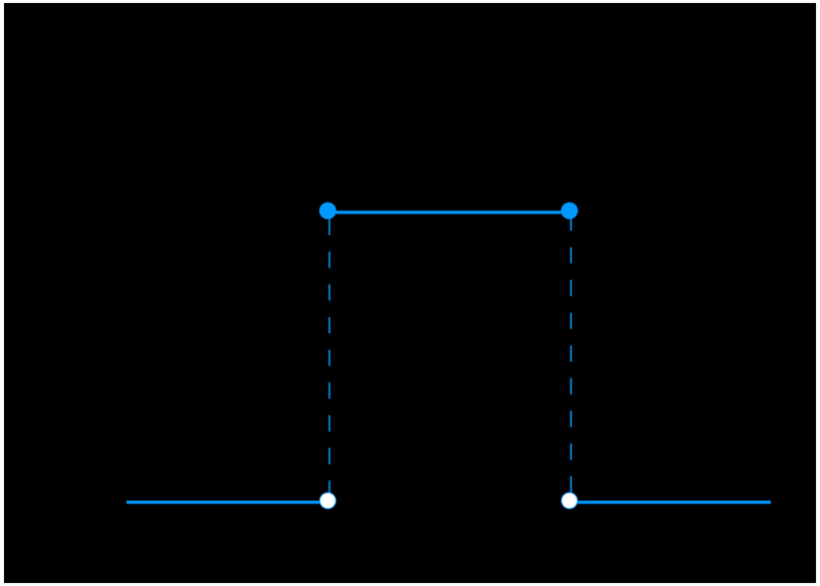
How many monsters must Bob defeat to have a 90% chance of obtaining the item?

	A	B	C	D	E
1	**Probability of Epic Loot:**	0.05		**90% Chance:**	44
2					
3	**x = # Monsters Slain**	**F(x)**			
4	1	0.05			
5	2	0.10			
6	3	0.14			
7	4	0.19			
8	5	0.23			
9	6	0.26			
10	7	0.30			
11	8	0.34			
12	9	0.37			

Many Other Distributions

- Don't need to know them all but should know about **Binomial**, **Negative Binomial**, **Normal** and **Uniform**.





Pearson Product-Moment Correlation

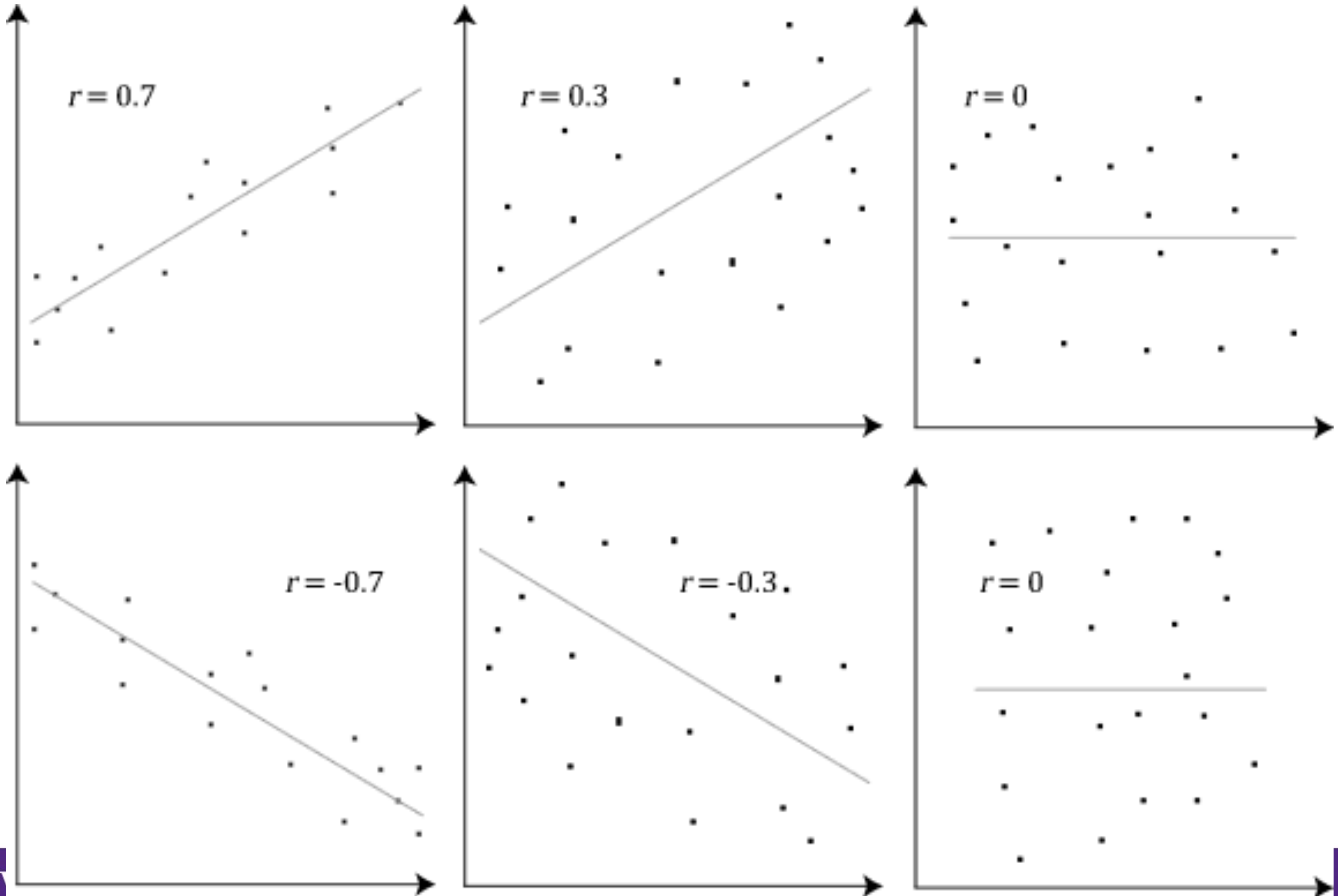
- How can we determine how correlated (the strength of their connection) two variables are?
- Could just graph it and "eyeball" it.
- Better to use something less subjective like the Pearson Product-Moment Correlation.
- This is the CORREL function in Excel:

```
CORREL(array1, array2)
```

Pearson Product-Moment Correlation

- Value between -1 and 1.
- -1 being a perfect negative correlation, 0 being no correlation at all, 1 being a perfect correlation.
- Basically attempts to draw a line of best fit through the data.
- Only works for linear correlations.

Pearson Product-Moment Correlation



Pearson Product-Moment Correlation

Example 6:

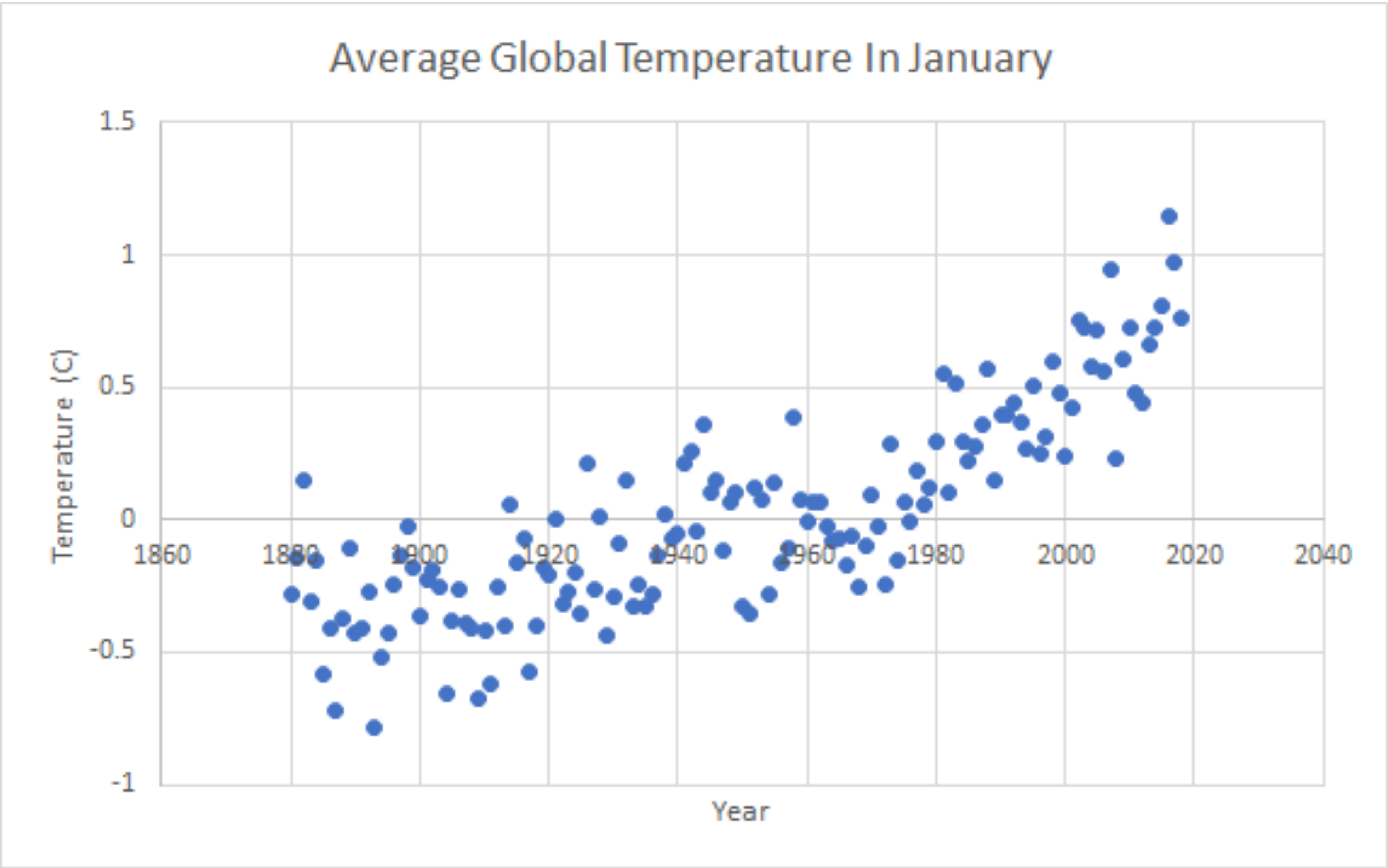
Is there a correlation between average global temperatures in January and the time (year)? Is it getting hotter or colder in January each year?

	A	B
1	**Year**	**Average Global January Temperature (C)**
2	1880	-0.28
3	1881	-0.14
4	1882	0.15
5	1883	-0.31
6	1884	-0.15
7	1885	-0.58
8	1886	-0.41
9	1887	-0.72
10	1888	-0.37
11	1889	-0.11
12	1890	-0.43
13	1891	-0.41
14	1892	-0.27
15	1893	-0.78
16	1894	-0.52
17	1895	-0.43
18	1896	-0.24
19	1897	-0.13
20	1898	-0.02
21	1899	-0.18
22	1900	-0.36
23	1901	-0.23

Pearson Product-Moment Correlation

Example 6:

We could try "eyeballing" it with a graph:



Pearson Product-Moment Correlation

Example 6:

Better to try CORREL.

Correlation:	=CORREL(A2:A140,B2:B140)

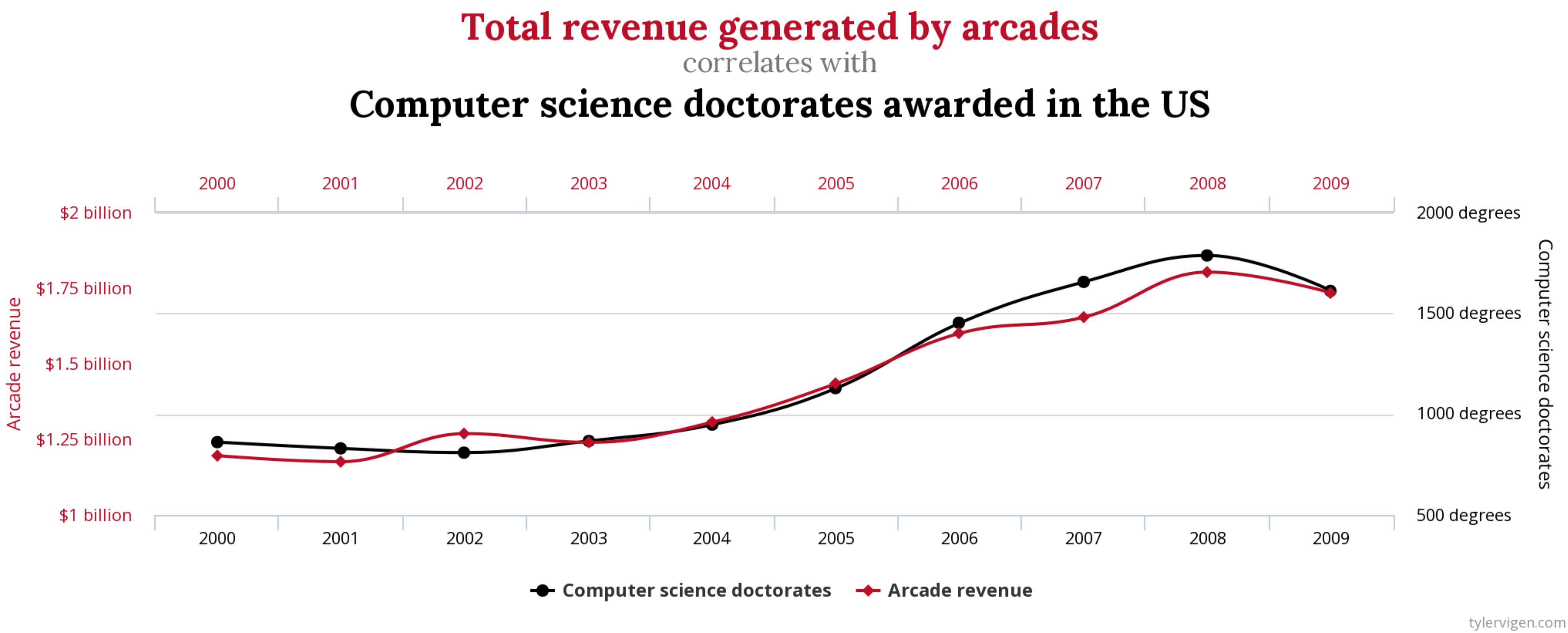
Correlation:	0.826240325

Pearson Product-Moment Correlation

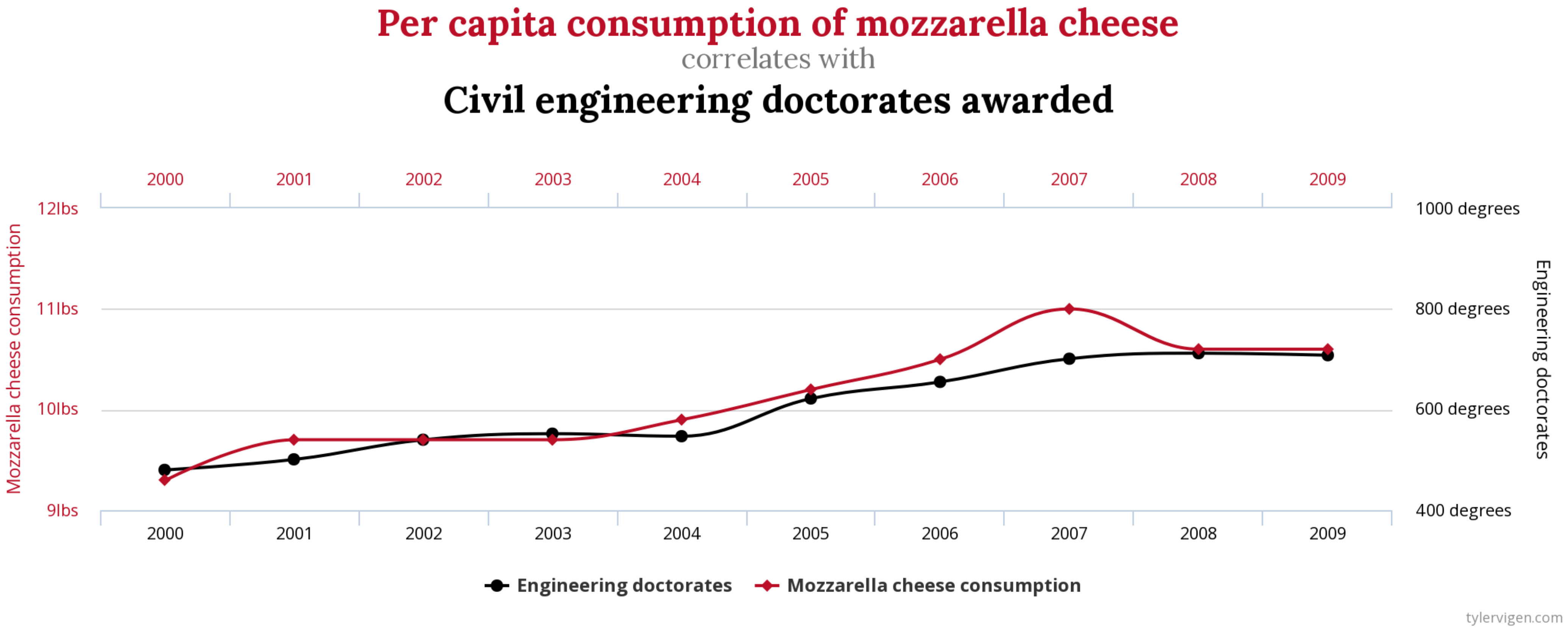
Example 6:

- Does this mean it is getting hotter each year?
- Not necessarily. Finding a correlation is just a starting point. We need to create and test our hypothesis.
- More about testing hypotheses in zyBook Chapter 5 Statistics Basics.

Do Correlations Imply Causation?



Do Correlations Imply Causation?



Data Preparation & Transformation

Data Preparation

- Data scientists call it data wrangling, data munging, data janitor work.
- According to interviews and expert estimates, data scientists spend from 50% to 80% of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.
- Article: *For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights*

Data Preparation

Situation

- Start with a text file, or
- Cut and paste from a web page or document

Problems

- Data incomplete, or irregular format
- Paste tries to put data all into one cell
- The file would take too long to edit by hand

Data Preparation

Example From Lab 2:

Want to copy data from article on Cambridge University wine spending to Excel

Assigned Readings / Tasks

- Introduction to Regular expressions using Atom
- RegexOne: https://regexone.com
- Tab-separated values (Wikipedia)