A Crash Course to Descriptive Statistics

by Jason Schneider

I give credit, with much gratitude, to Dr. Jeffrey Witzel who taught the content of descriptive statistics, intended for linguistics, in which I use the notes from his lectures while I attended graduate school at UTA.

Measures of Central Tendency

Measures of central tendency

- indicate the “typical” behavior of the data

Measures of Central Tendency

1. Mean
-aka, average
-usually symbolized as M

$$M = \frac{\sum X}{N}$$

M = mean

X = individual scores

N = number of scores

$\sum$ = sum of

What is the mean for the values below?
(77+75+70+55+4+65+61)/7
=AVERAGE(range)

Measures of Central Tendency

2. Mode

-the most frequent score

-note that there could be more than one most frequent score.

bimodal (if there are two peaks)

trimodal (if there are three peaks)….

What is the mode for the values from the data?

2, 4, 6, 2, 5, 2, 67, 2, 8, 2, 4, 9, 10, 22, 45

=MODE(range)

Measures of Central Tendency

3. Median

-the value that half of the values fall below and the other half fall above

e.g. the median of 100, 95, 83, 71, 61, 57, 30 is...

71

e.g. the median of 100, 95, 83, 75, 71, 61,57, 30 is...

73

What is the median for the values from the data?

2, 4, 6, 2, 5, 2, 67, 2, 8, 2, 4, 9, 10, 22, 45

=MEDIAN(range)

Measures of Central Tendency

4. Midpoint

-the value halfway between the highest value and the lowest value

$$\text{Midpoint} = \frac{\text{High} + \text{Low}}{2}$$

e.g. if the highest score was 100 and the lowest was 30, then the midpoint is....65

What is the midpoint for 3, 5, 8, 6, 3?

=(MAX(range)+MIN(range))/2

Measures of Dispersion

Measures of dispersion

- indicate how data varied with the typical distribution of data

Measures of Dispersion

1. Range
-the number of points between the highest value and the lowest value +1

What is the range for 3, 5, 8, 6, 3?
=MAX(range)-MIN(range)+1

Measures of Dispersion

2. High and low values
-the highest value and the lowest value
-shows where the range falls on the continuum of possible values

What are the high and low values 3, 5, 8, 6, 3?
=MAX(range)
=MIN(range)

Measures of Dispersion

3. Variance

-the average of the squared difference between each value and the mean

$$\text{Variance} = \frac{\sum(X-M)^2}{N}$$

The difference is squared so that the values above and below the mean do not cancel each other out.

$(77-69)^2 =$ $(61-69)^2 =$

$(8)^2 = 64$ $(-8)^2 = 64$

What is the variance for values below?

77+75+70+55+4+65+61

=VARP(range)

Measures of Dispersion

	A	B	C	D	E
1		**X**	**(X-M)^2 =(X-E2)^2**		
2		77	355.59183673	**mean =AVERAGE(B2:B8)**	**58.14285714**
3		75	284.16326531	**variance =VARP(B2:B8)**	**539.5510204**
4		70	140.59183673	**manual variance =C9/E5**	**539.5510204**
5		55	9.8775510204	**length =ROWS(B2:B8)**	**7**
6		4	2931.4489796		
7		65	47.020408163		
8		61	8.1632653061		
9	**Sum =SUM(B2:B8) =SUM(C2:C8)**	**407**	**3776.8571429**		

Let’s pretend that we have 30 or more data points to use VARP. However, the calculated variance is wrong here because we should use VAR, not VARP.

Measures of Dispersion

3. Variance

-In addition to

$$\text{Variance} = \frac{\sum(X-M)^2}{N}$$ `=VARP(range)`

you can also use the following formula

$$\text{Variance} = \frac{\sum(X-M)^2}{N-1}$$ `=VAR(range)`

→ The first one (the *N* formula) is for the whole population, while the second one (*N*-1 formula) is for a sample.

→ Population – whole group; sample – portion of the group

→ As a rule of thumb… 30+ values (the *N* formula);

under 30 values (*N*-1 formula)

Measures of Dispersion

4. Standard deviation

-aka *SD*

-generally considered the most reliable estimate of dispersion.

$$SD = \sqrt{\frac{\sum(X-M)^2}{N}}$$

Does this look familiar?

SD = square root of variance

The standard deviation shows how much the data deviates from the mean of the dataset.

What is the SD for the values below?

2, 4, 6, 2, 5, 2, 67, 2, 8, 2, 4, 9, 10, 22, 45

=STDEVP(range)

Measures of Dispersion

Think what a large standard deviation might indicate. If there is considerable distance between the data points from the mean of the dataset, this might indicate that our data significantly varies from each other and might be the result of differing factors, making our data the result of different things and not the one thing we might expect. This typically indicates that we have outliers, too.

Measures of Dispersion

	A	B	C	D	E	F
1		77	77	74	69	69
2		75	75	75	68	68
3		70	70	70	70	68
4		55	55	67	67	67
5		4	60	60	60	63
6		65	65	65	65	65
7		61	61	61	61	61
8	**Standard Deviation**	**25.08936409**	**8.173709305**	**5.912053869**	**3.903600292**	**2.968084199**

Notice that the data points become more similar and closer as the standard deivation becomes smaller because the differences in data from the mean of the dataset lessons as the data points become more alike.

Measures of Dispersion

4. Standard deviation

-In addition to

$$SD = \sqrt{\frac{\sum (X-M)^2}{N}}$$ `=STDEVP(range)`

you can also use the following formula

$$SD = \sqrt{\frac{\sum (X-M)^2}{N-1}}$$ `=STDEV(range)`

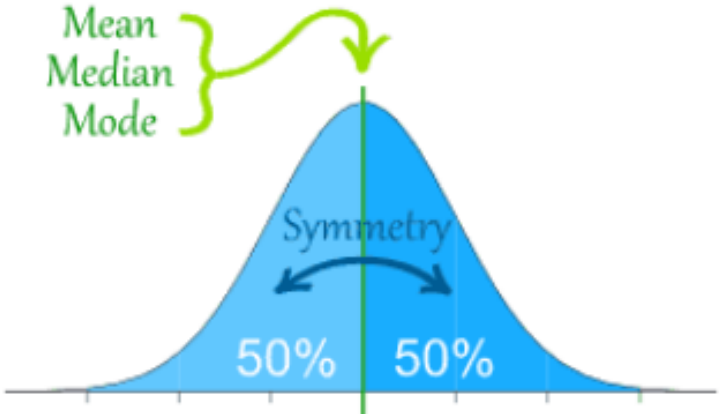
→ The first one (the *N* formula) is for the whole population, while the second one (*N*-1 formula) is for a sample.

→ As a rule of thumb… 30+ values (the *N* formula);

under 30 values (*N*-1 formula)

Normal Distributions

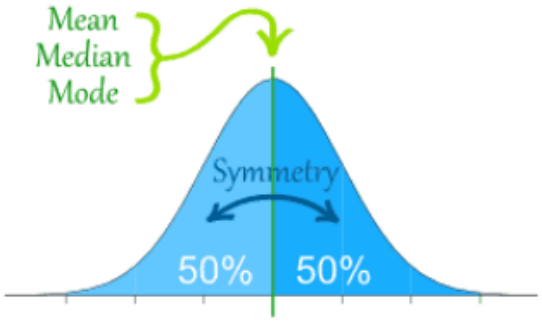
Putting it all together: Descriptive stats and the normal distribution



(from mathisfun.com)

Normal Distributions

Putting it all together: Descriptive stats and the normal distribution



(from mathisfun.com)

Central Tendency

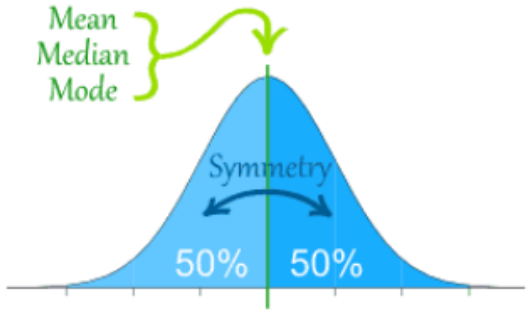
In a real normal distribution …

-the mean, mode, median, and midpoint will all be the same.

→Most of the time, however, you have something approximating normal distribution. If it is approximating a normal distribution, then these measures of central tendencies should be similar.

Normal Distributions

Normal distributions



(from mathisfun.com)

Dispersion

-In a real normal distribution ...

-the lowest and highest values should be exactly the same in distance from the mean.

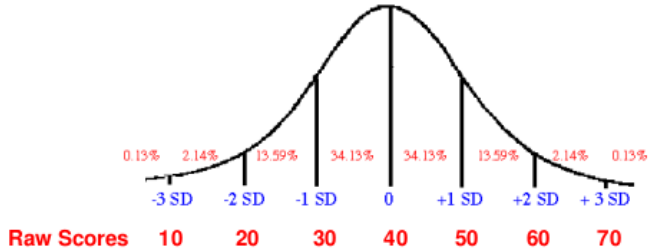
i.e., the range is symmetrical.

-also, standard deviations (SDs) conform to a certain pattern.

Example of Dispersion

Percentiles

The total percentage of students who scored equal to or below a given point in the normal distribution.



(from gifted.uconn.edu)

Let's say the mean is 40, and the SD is 10.

→What percentile is the score of 40? … 50th percentile

→What percentile is the score of 30? … 0.13+2.14+13.59 = 15.86 → 16th

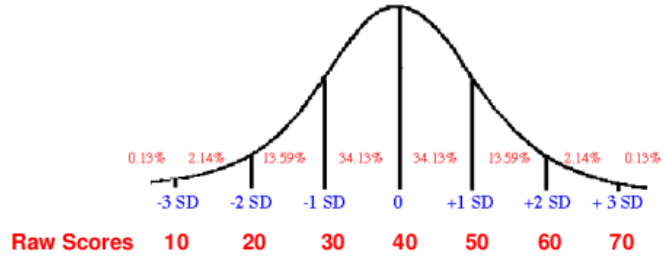
→What percentile is the score of 50?

→What percentile is the score of 60?

Example of Dispersion

Percentiles

The total percentage of students who scored equal to or below a given point in the normal distribution.



(from gifted.uconn.edu)

Let's say the mean is 40, and the SD is 10.

→What percentile is the score of 40? … 15.86+34.13 = 49.99 → 50th

→What percentile is the score of 30? … 0.13+2.14+13.59 = 15.86 → 16th

→What percentile is the score of 50? … 50+34.13 = 84.13 → 84th

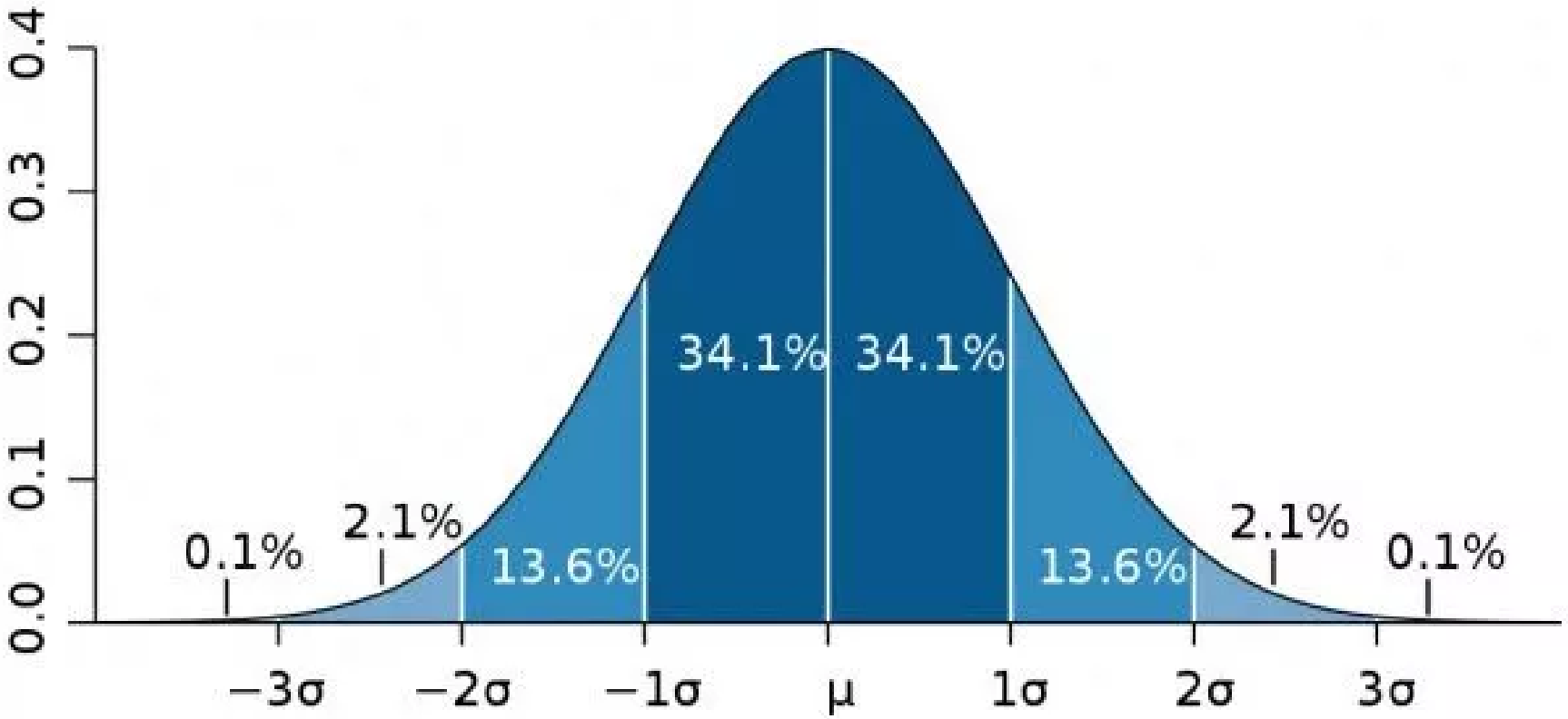
→What percentile is the score of 60? … 84.13+13.59 = 97.72 → 98th

Normal Distributions

When a dataset that is normally distributed, the normal distribution will follow the below rules:

- The center of the bell curve is the mean of the data point (also the highest point in the bell curve).
- 68.2% of the total data points lie in the range (Mean – Standard Deviation to Mean + Standard Deviation).
- 95.5% of the total data points lie in the range (Mean – 2*Standard Deviation to Mean + 2*Standard Deviation)
- 99.7% of the total data points lie in the range (Mean – 3*Standard Deviation to Mean + 3*Standard Deviation)

Normal Distributions



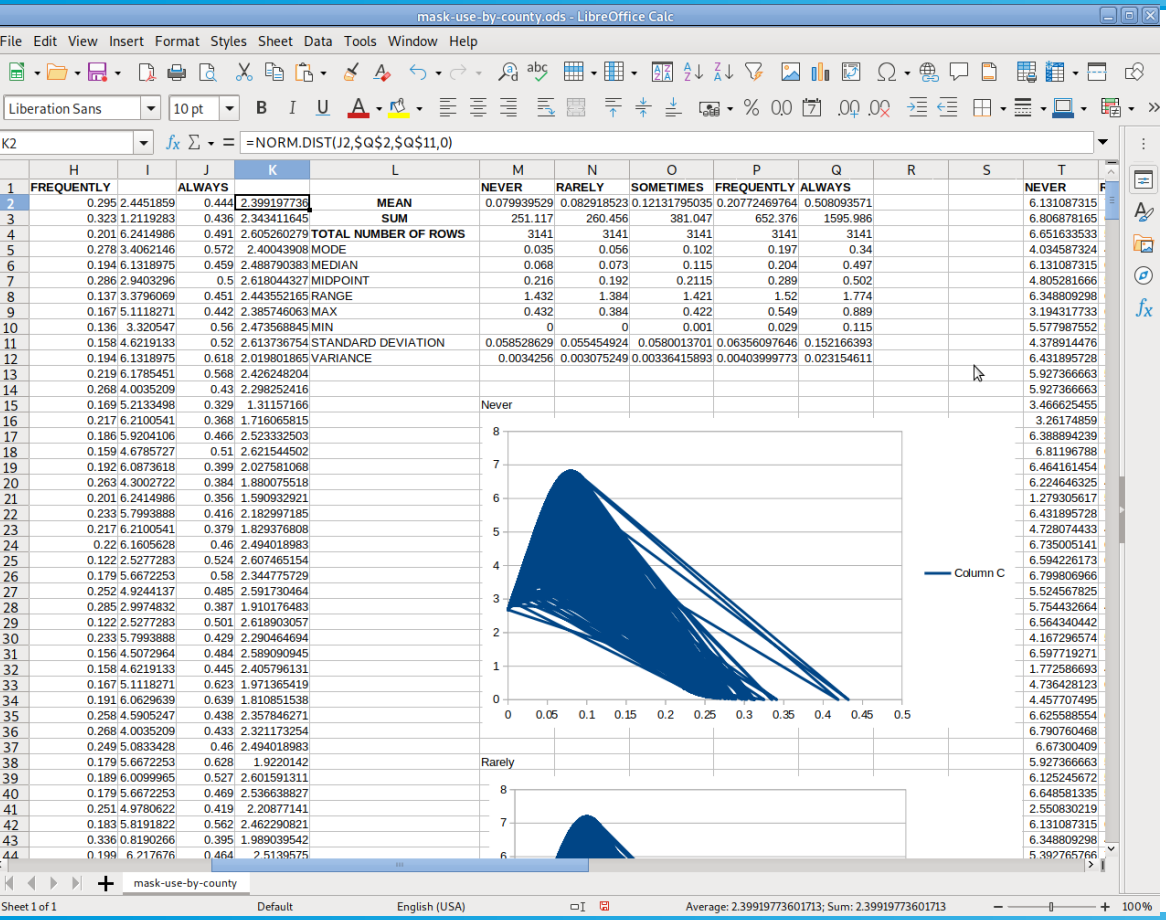
Normal Distributions

Use the formula **Mean – 3* Standard Deviation** to develop the y-axis. The mean and standard deviation should be created from the column that will be the x-axis.

- In Excel or Libre Calc, use the function **=NORM.DIST(the cell from the spread sheet, the mean of the column that contains that cell, the standard deviation of the column that contains that cell, False)**
- For example, **=NORM.DIST(J2,Q2,Q11,0)**

The fourth input can be False or 0, the same thing. Use the dollar signs to prevent Excel or Libre Calc from altering the cell if dragging or copying. Lastly, do not forget using the equal sign before the function; otherwise, the spreadsheet will not recognize this as a function.

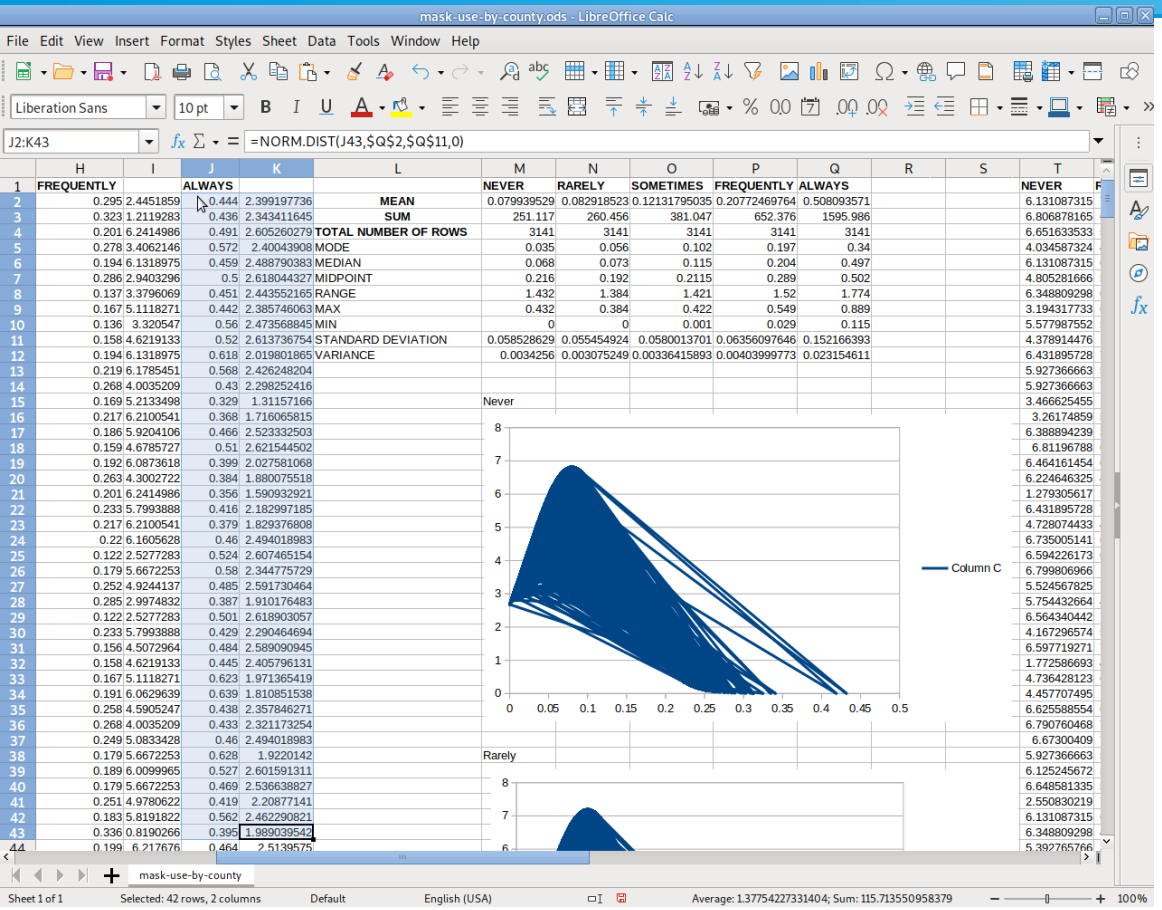
Normal Distributions



	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	FREQUENTLY		ALWAYS			NEVER	RARELY	SOMETIMES	FREQUENTLY	ALWAYS			NEVER
2	0.295	2.4451859	0.444	2.399197736	MEAN	0.079939529	0.082918523	0.12131795035	0.20772469764	0.508093571			6.131087315
3	0.323	1.2119283	0.436	2.343411645	SUM	251.117	260.456	381.047	652.376	1595.986			6.806878165
4	0.201	6.2414986	0.491	2.605260279	TOTAL NUMBER OF ROWS	3141	3141	3141	3141	3141			6.651633533
5	0.278	3.4062146	0.572	2.40043908	MODE	0.035	0.056	0.102	0.197	0.34			4.034587324
6	0.194	6.1318975	0.459	2.488790383	MEDIAN	0.068	0.073	0.115	0.204	0.497			6.131087315
7	0.286	2.9403296	0.5	2.618044327	MIDPOINT	0.216	0.192	0.2115	0.289	0.502			4.805281666
8	0.137	3.3796069	0.451	2.443552165	RANGE	1.432	1.384	1.421	1.52	1.774			6.348809298
9	0.167	5.1118271	0.442	2.385746063	MAX	0.432	0.384	0.422	0.549	0.889			3.194317733
10	0.136	3.320547	0.56	2.473568845	MIN	0	0	0.001	0.029	0.115			5.577987552
11	0.158	4.6219133	0.52	2.613736754	STANDARD DEVIATION	0.058528629	0.055454924	0.0580013701	0.06356097646	0.152166393			4.378914476
12	0.194	6.1318975	0.618	2.019801865	VARIANCE	0.0034256	0.003075249	0.00336415893	0.00403999773	0.023154611			6.431895728
13	0.219	6.1785451	0.568	2.426248204									5.927366663
14	0.268	4.0035209	0.43	2.298252416									5.927366663
15	0.169	5.2133498	0.329	1.31157166									3.466625455
16	0.217	6.2100541	0.368	1.716065815									3.26174859
17	0.186	5.9204106	0.466	2.523332503									6.388894239
18	0.159	4.6785727	0.51	2.621544502									6.81196788
19	0.192	6.0873618	0.399	2.027581068									6.464161454
20	0.263	4.3002722	0.384	1.880075518									6.224646325
21	0.201	6.2414986	0.356	1.590932921									1.279305617
22	0.233	5.7993888	0.416	2.182997185									6.431895728
23	0.217	6.2100541	0.379	1.829376808									4.728074433
24	0.22	6.1605628	0.46	2.494018983									6.735005141
25	0.122	2.5277283	0.524	2.607465154									6.594226173
26	0.179	5.6672253	0.58	2.344775729									6.799806966
27	0.252	4.9244137	0.485	2.591730464									5.524567825
28	0.285	2.9974832	0.387	1.910176483									5.754432664
29	0.122	2.5277283	0.501	2.618903057									6.564340442
30	0.233	5.7993888	0.429	2.290464694									4.167296574
31	0.156	4.5072964	0.484	2.589090945									6.597719271
32	0.158	4.6219133	0.445	2.405796131									1.772586693
33	0.167	5.1118271	0.623	1.971365419									4.736428123
34	0.191	6.0629639	0.639	1.810851538									4.457707495
35	0.258	4.5905247	0.438	2.357846271									6.625588554
36	0.268	4.0035209	0.433	2.321173254									6.790760468
37	0.249	5.0833428	0.46	2.494018983									6.67300409
38	0.179	5.6672253	0.628	1.9220142									5.927366663
39	0.189	6.0099965	0.527	2.601591311									6.125245672
40	0.179	5.6672253	0.469	2.536638827									6.648581335
41	0.251	4.9780622	0.419	2.20877141									2.550830219
42	0.183	5.8191822	0.562	2.462290821									6.131087315
43	0.336	0.8190266	0.395	1.989039542									6.348809298
44	0.199	6.217676	0.464	2.5139575									5.392765766

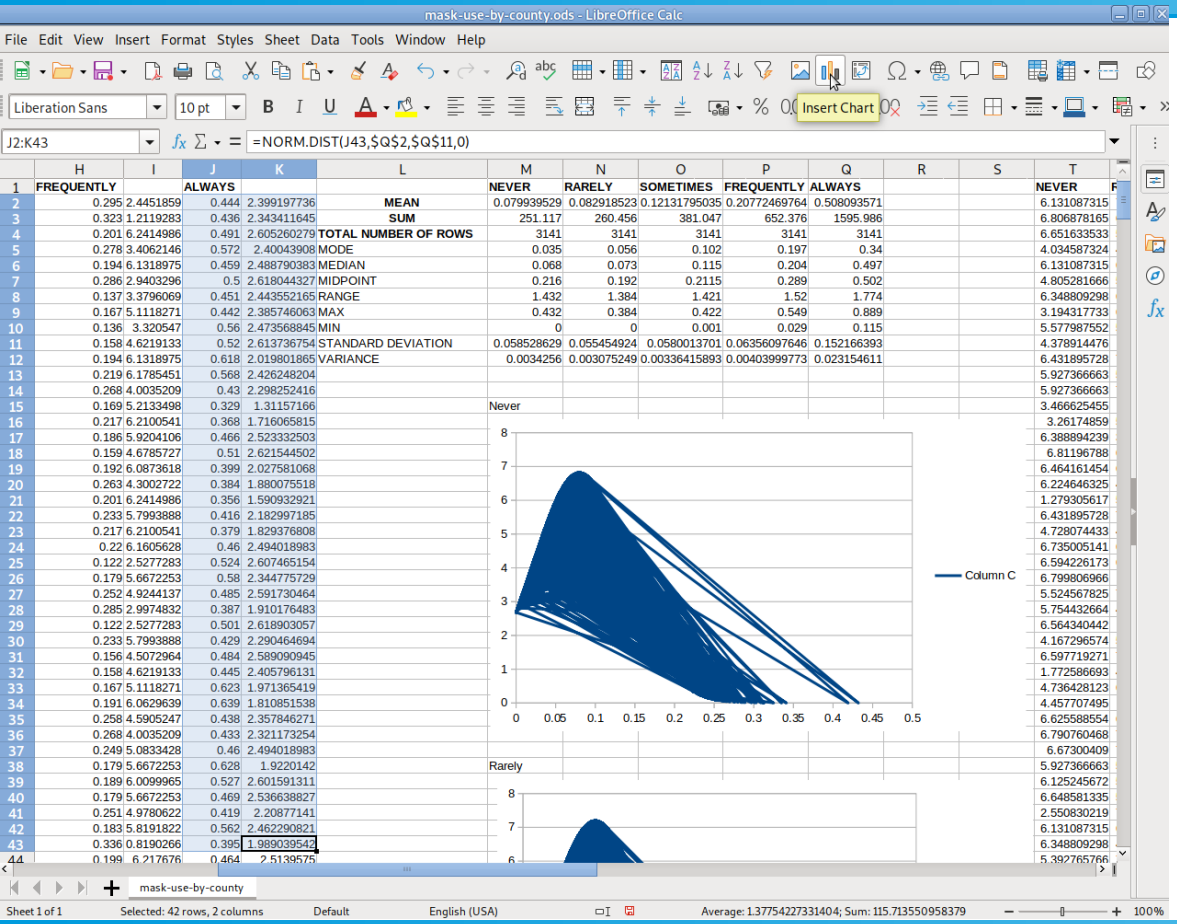
For example, use the norm.dist function in the cell k2, applying to j2. Then this same function can be applied to all the cells in k as shown in the picture to the left.

Normal Distributions



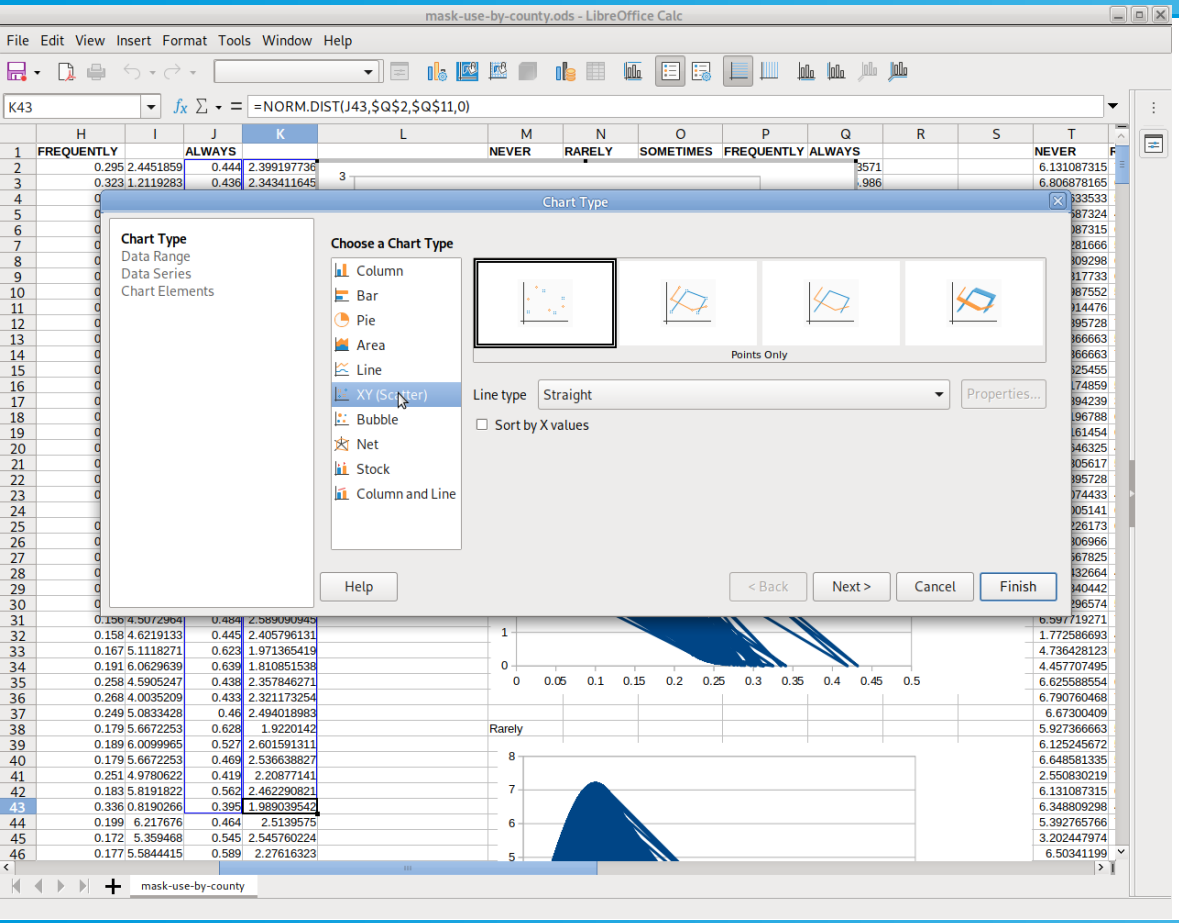
Then select all cells in both j and k columns.

Normal Distributions



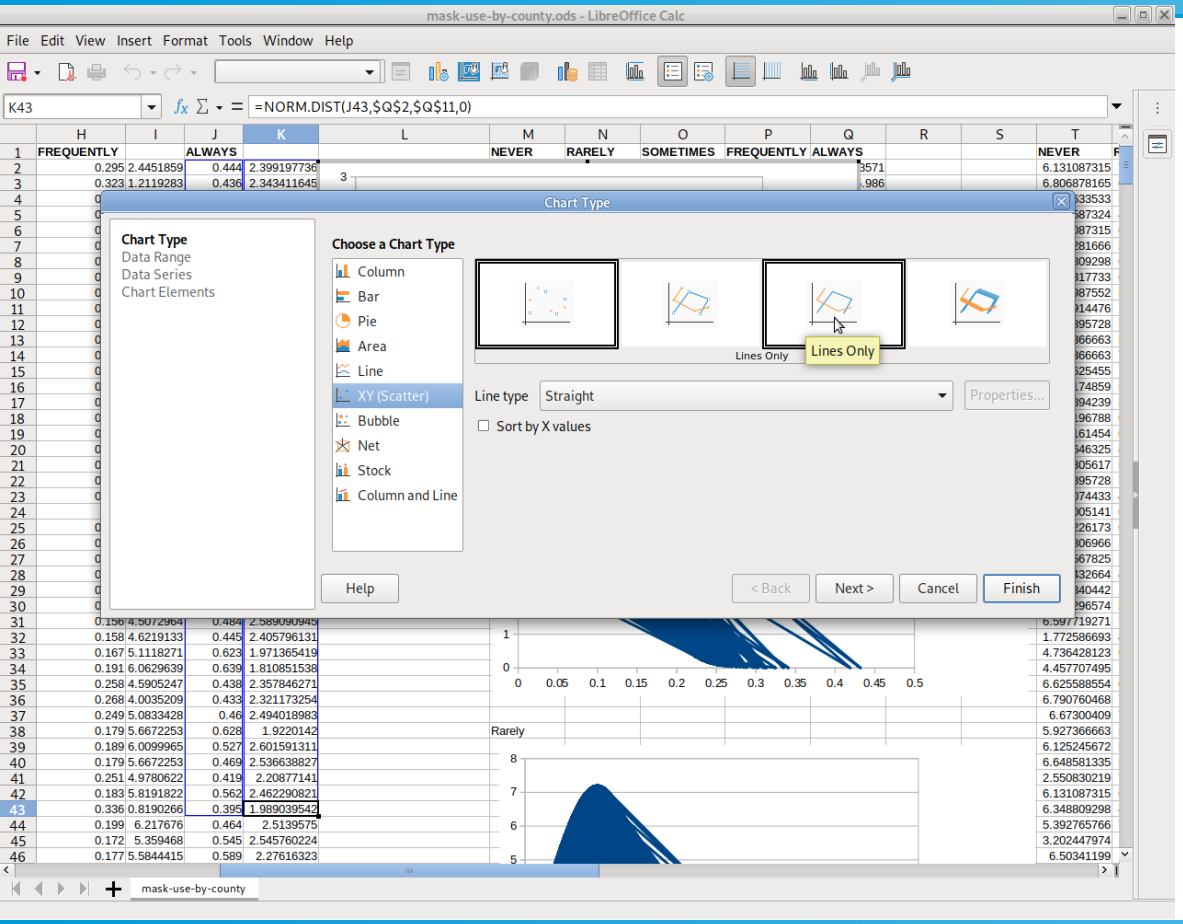
Click on Insert Chart as shown in the picture to the left as columns j and k are selected, indicated by j and k being highlighted blue.

Normal Distributions



After clicking on Insert Chart, select XY (Scatter), or Scatter, depending on the office program used for opening spreadsheets like Excel.

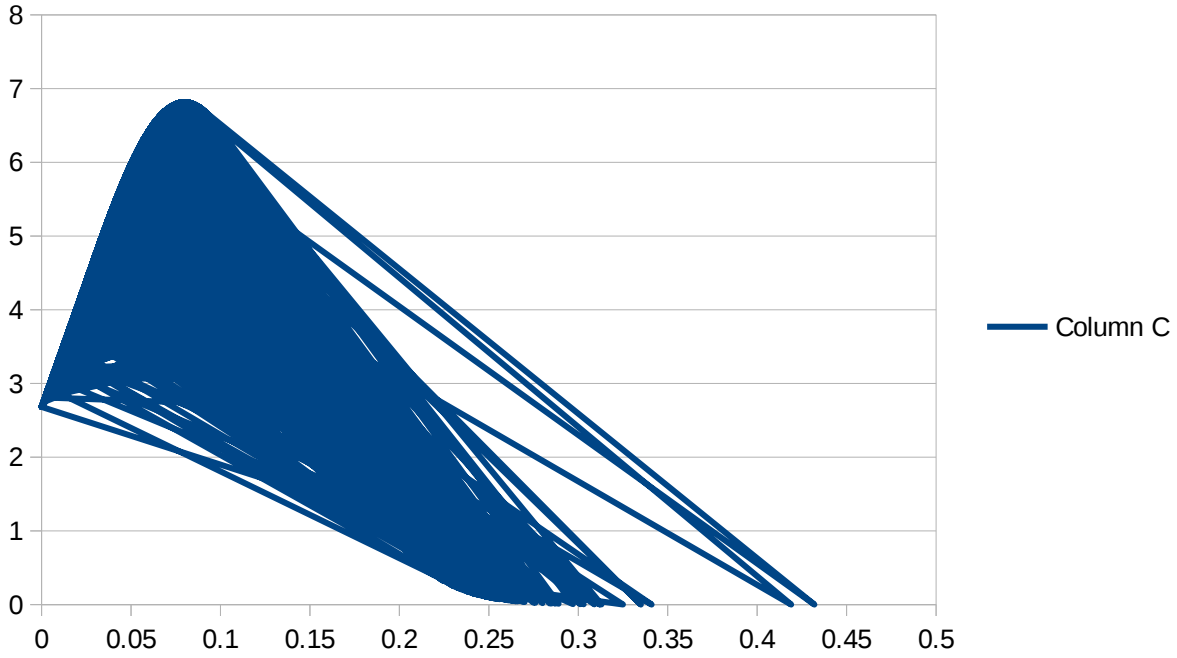
Normal Distributions



Lastly, select Lines Only, or Smooth Lines if using Excel, before clicking on Finish to create a graph or figure for a normal distribution as shown of the skewed distribution that has been shown in the pictures illustrating these steps.

Normal Distributions

Then the result should look like the figure or graph shown to the right, illustrating a normal distribution varying according to how skewed it might be.



Summary of Functions

Excel cheat sheet

Number of students (*N*)	`=COUNT(range)`
Mean	`=AVERAGE(range)`
Mode	`=MODE(range)`
Median	`=MEDIAN(range)`
Midpoint	`=(MAX(range)+MIN(range))/2`
High	`=MAX(range)`
Low	`=MIN(range)`
Range	`=MAX(range)-MIN(range)+1`
Variance (*N* formula)	`=VARP(range)`
Variance (*N*-1 formula)	`=VAR(range)`
SD (*N* formula)	`=STDEVP(range)`
SD (*N*-1 formula)	`=STDEV(range)`

Example table

Table 4.7: Fall 1986 (First Administration) ELIPT Results

	Subtests			
Statistics	**Listening**	**Reading**	**Vocabulary**	**Writing**
N	153.00	153.00	154.00	153.00
Total items (*k*)	55.00	60.00	100.00	100.00
Mean ($\bar{X}$)	34.76	40.64	69.34	75.08
Mode	32.00	43.00	86.00	77.00
Median	34.45	41.00	71.67	75.50
Midpoint	34.50	39.00	59.50	69.00
Low–High	17–52	21–57	20–99	44–94
Range	36.00	37.00	80.00	51.00
S	7.29	7.48	16.08	8.94

(Brown, 1996)

Example table

Table 4.8: Fall 1986 ELIPT Results

			Central Tendency				Dispersion		
Subtest	*N*	*k*	$\bar{X}$	Mode	Median	Midpoint	Low–High	Range	*S*
Listening	153	55	34.76	32	34.45	34.50	17–52	36	7.29
Reading	153	60	40.64	43	41.00	39.00	21–57	37	7.48
Vocabulary	154	100	69.34	86	71.67	59.50	20–99	80	16.68
Writing	153	100	75.08	77	75.50	69.00	44–94	51	8.94

(Brown, 1996)