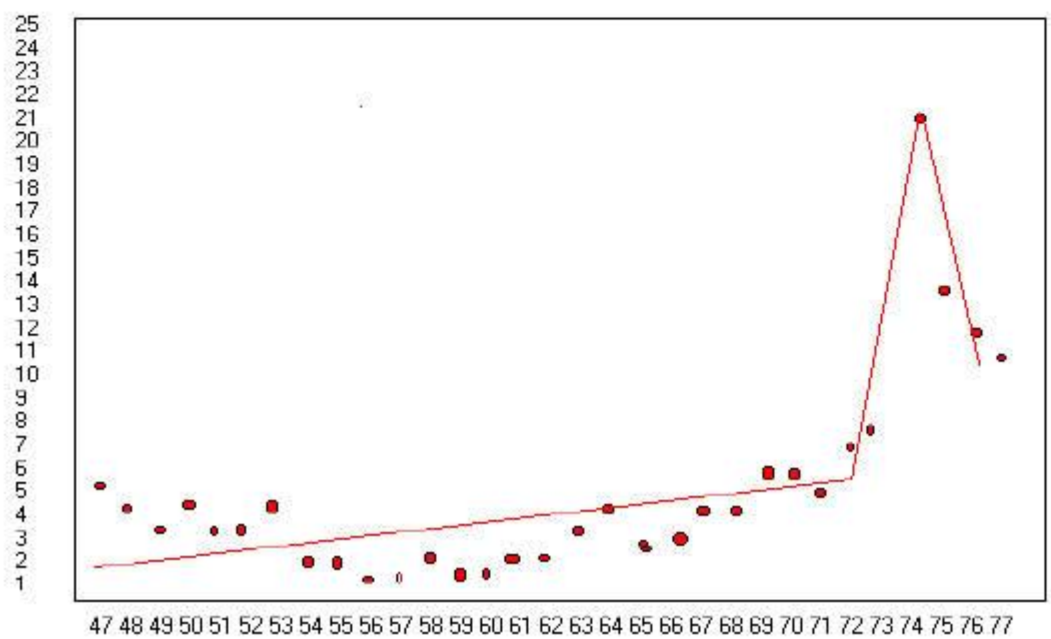
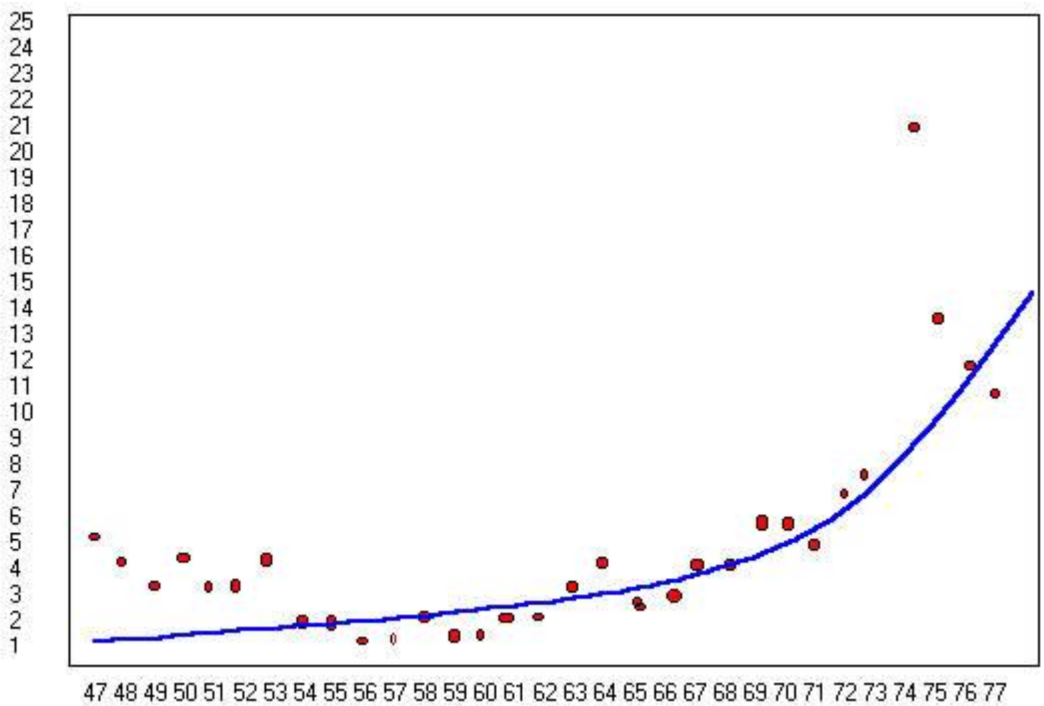
Non-Linear Data

Topics of this chapter are: Curve Fitting, Curvilinearity, Normal Curve, and Skewness

I. Curve Fitting



Divorce rates in Australia in years from 1947 to 1977. In 1974 divorce became legal. The above lines seem to capture the relationship of divorce rates over time and also the change in the law in 1974 when divorce became legal. There was a clear spike when the law changed and a decline after the "pent up" divorces were accomplished. The lines as drawn seem to pass very close to all of the points.



The curve in the second graph presents a different interpretation. It indicates that the divorce rate has been on the rise since the middle sixties and possibly the law caught up with the practice. Further, the line may pass about as close to the data points as the above graph. The lower line is also more simple. It could be generated with three functions (starting point, slope and accelerating) while it may take as many as six functions to generate the top graph. The bottom graph is more parsimonious.

Curvilinearity

A positively accelerating curve

As indicated in the graphs above trends may be linear or non-linear. The data in Table I is hypothetical trend for 10% interest on $1000.00 for 30 years. Figure 1 shows the 10% Interest at 5 year intervals. Note the non-linear trend. When the numbers are squared the trend becomes more linear but it still remains non-linear in Figure 2. When the natural log is used to transform the data it becomes linear. This difference can be tested by using multiple regression and correcting skewed data.

Table 1. Ten percent interest on $1000.00 for 30 years

Year	10% Interest	Square Root	Natural Log
1960	1000	31.6228	6.90775528
1961	1100	33.1662	7.00306546
1962	1210	34.7851	7.09837564
1963	1331	36.4829	7.19368582
1964	1464	38.2623	7.28892769
1965	1610	40.1248	7.38398946
1966	1771	42.0833	7.47929964
1967	1948	44.1362	7.57455848
1968	2143	46.2925	7.66996200
1969	2357	48.5489	7.76514490
1970	2593	50.9215	7.86057079
1971	2852	53.4041	7.95577578
1972	3137	56.0089	8.05102221
1973	3451	58.7452	8.14641932
1974	3796	61.6117	8.24170316
1975	4176	64.6220	8.33710913
1976	4594	67.7791	8.43250638
1977	5053	71.0845	8.52773741
1978	5558	74.5520	8.62299361
1979	6114	78.1921	8.71833650
1980	6725	82.0061	8.81358720
1981	7398	86.0116	8.90896497
1982	8138	90.2109	9.00429973
1983	8952	94.6150	9.09963225
1984	9847	99.2321	9.19492212
1985	10832	104.0769	9.29026000
1986	11915	109.1559	9.38555339
1987	13107	114.4858	9.48090172
1988	14418	120.0750	9.57623270
1989	15860	125.9365	9.67155550
1990	17446	132.0833	9.76686567

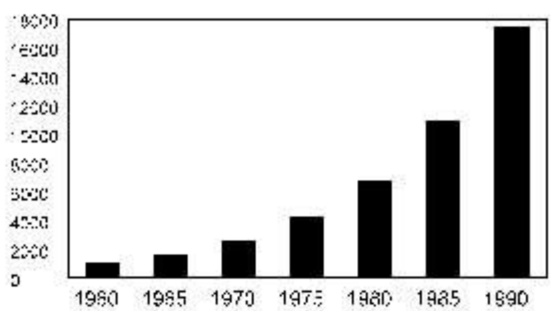


Figure 1. 10% Interest.

The 10% Interest produces a positive accelerating curve as seen in Figure 1. Notice the each year the gap widens between the amount of change.

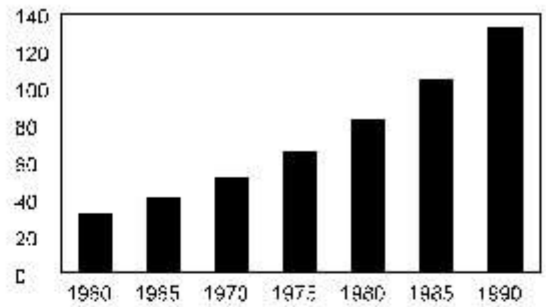


Figure 2. Square root of 10% Interest.

The square root of the 10% Interest reduces the amount of acceleration of the curve but it still exists. There continues to be a positive accelerating curve.

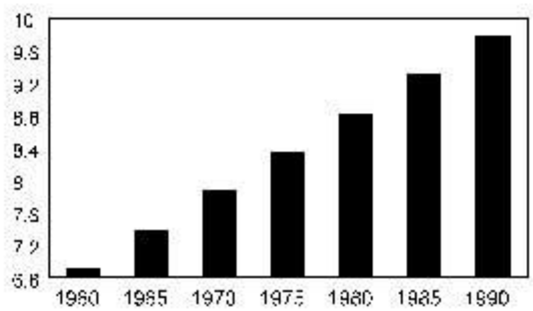


Figure 3. Natural log of 10% Interest.

Notice in both Figures 1 & 2 that the trend is a positively accelerating even though it is less so in Figure 2 where the transformation is accomplished by the square root method. However, in Figure 3 the trend is a straight line (linear). In Figure 3 the natural log is computed and the result is a straight line.

Things that grow throughout their life will follow that same non-linear pattern as inflation. For example many organisms like trees and whales do, and their weight will increase in a non-linear fashion. Many other living things will increase in a non-linear fashion until maturity. Many psychological phenomenon will follow the same pattern. The year from 1 year old to 2 years old seems much longer than the year 50 to 51 years old. A day spent in a psychiatric hospital has a much longer phenomonologically for a person who spends a week in a psychiatric hospital than for a person who spends a year in a psychiatric hospital. As noted above curvilinearity can be corrected by computing a natural log.

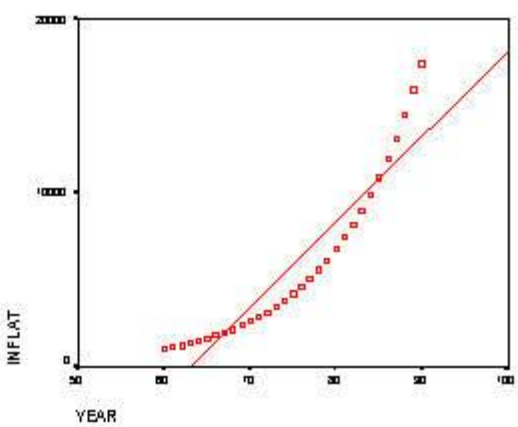
Curvilinearity and Correlation

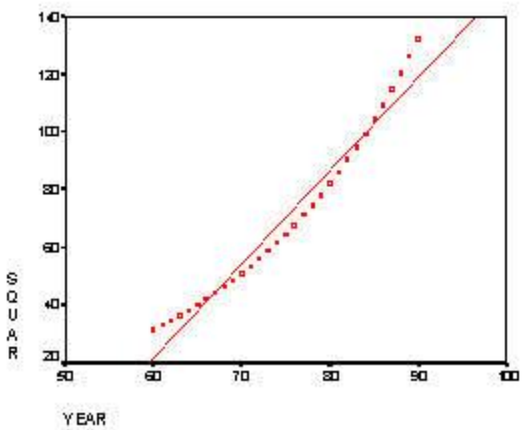
The data presented above also demonstrates how curvilinearity can be assessed in a correlation or regression problem. The table below contains the correlations of YEAR with INFLATION, INFLATION SQUARED, and the NATURAL LOG OF INFLATION. It shows that when YEAR is correlated with inflation it is .938 but when the Natural Log is computed the relationship becomes 1.00. This indicates that there is a curvilinear relationship between INFLATION and YEAR. Consequently, such a relationship can be tested (as was done here) by first assessing a relationship between two variables and then computing the natural log on the variable that is suspected of being curvilinear and testing that new variable. If the correlation improves as in the example one can conclude that there is a curvilinear relationship. (It must be determined that there is a significant difference between the two correlations.) The curvilinear relationship is exemplified in the scattergrams below.

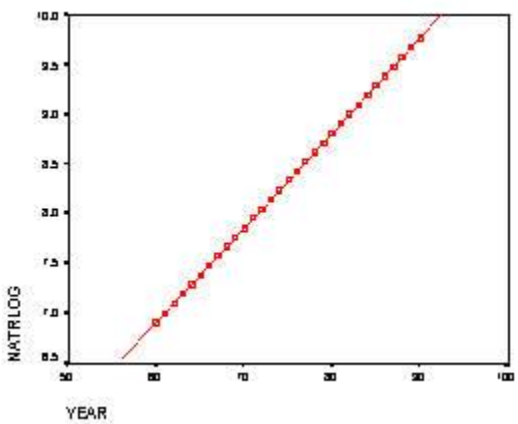
Correlations

		YEAR	INFLAT	SQUAROOT	NATRLOG
YEAR	Pearson Correlation	1.000	.938**	.983**	1.000**
	Sig. (2-tailed)	.	.000	.000	.000
	N	31	31	31	31
INFLAT	Pearson Correlation	.938**	1.000	.986**	.938**
	Sig. (2-tailed)	.000	.	.000	.000
	N	31	31	31	31
SQUAROOT	Pearson Correlation	.983**	.986**	1.000	.983**
	Sig. (2-tailed)	.000	.000	.	.000
	N	31	31	31	31
NATRLOG	Pearson Correlation	1.000**	.938**	.983**	1.000
	Sig. (2-tailed)	.000	.000	.000	.
	N	31	31	31	31

**. Correlation is significant at the 0.01 level (2-tailed).







Curvilinearity and Skewness

Skewness of frequency data presents the problem as curvilinearity data above and computing the Natural Log solves the problem in the same manner. In this next example the variable SHOULD and CONFUSE are skewed. The reason for the skewed data is that most feel that they do what they should and most people don't feel confused. Consequently, most of the responses will be toward the end of the scale.

File Name = lsqfre6.sps

```
get file = 'e:\dape\lsq.sav'
fre variables=should confuse
   /  statistics = all
   / barchart freq.
```

Statistics

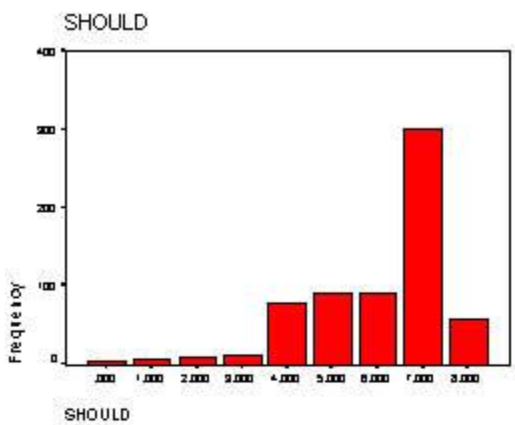
		SHOULD	CONFUSE
N	Valid	655	649
	Missing	7	13
Mean		6.05038	1.78120
Std. Error of Mean		5.94E-02	7.021E-02
Median		7.00000	1.00000
Mode		7.000	1.000
Std. Deviation		1.52017	1.78857
Variance		2.31091	3.19897
Skewness		-1.220	1.239
Std. Error of Skewness		.095	.096
Kurtosis		1.515	1.067
Std. Error of Kurtosis		.191	.192
Range		8.000	8.000
Minimum		.000	.000
Maximum		8.000	8.000
Sum		3963.000	1156.000

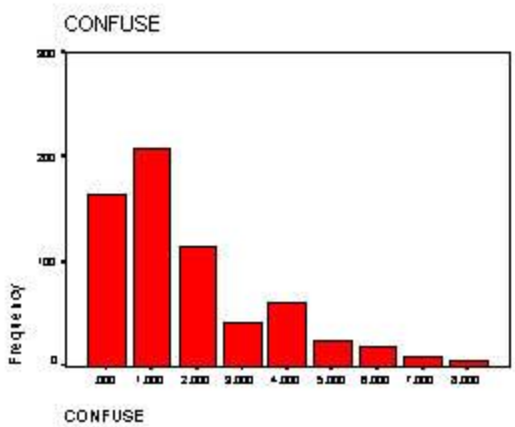
SHOULD

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.000	4	.6	.6	.6
	1.000	6	.9	.9	1.5
	2.000	10	1.5	1.5	3.1
	3.000	13	2.0	2.0	5.0
	4.000	79	11.9	12.1	17.1
	5.000	93	14.0	14.2	31.3
	6.000	91	13.7	13.9	45.2
	7.000	301	45.5	46.0	91.1
	8.000	58	8.8	8.9	100.0
	Total	655	98.9	100.0	
Missing	9.000	7	1.1		
Total		662	100.0		

CONFUSE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.000	164	24.8	25.3	25.3
	1.000	208	31.4	32.0	57.3
	2.000	115	17.4	17.7	75.0
	3.000	42	6.3	6.5	81.5
	4.000	62	9.4	9.6	91.1
	5.000	24	3.6	3.7	94.8
	6.000	19	2.9	2.9	97.7
	7.000	10	1.5	1.5	99.2
	8.000	5	.8	.8	100.0
	Total	649	98.0	100.0	
Missing	9.000	13	2.0		
Total		662	100.0		





Both of the variables are skewed (skewness greater than one). Further, the SHOULD variable has another problem in that it is negatively skewed. The computing the natural log function will correct skewness only when it is positive. In order to correct the negative skewness the item must first be reversed. The log function cannot be computed on 0 (zero) and 1 (one) must be added to all numbers. The syntax file above performs all of the necessary functions.

File = lsqfre7.sps

```
get file = 'e:\dape\lsq.sav'.
compute shouldln=ln(9-should).
compute confusln=ln(1+confuse).
missing value shouldln confusln (9).
fre variables=shouldln confusln
   /  statistics = all
   / barchart freq.
```

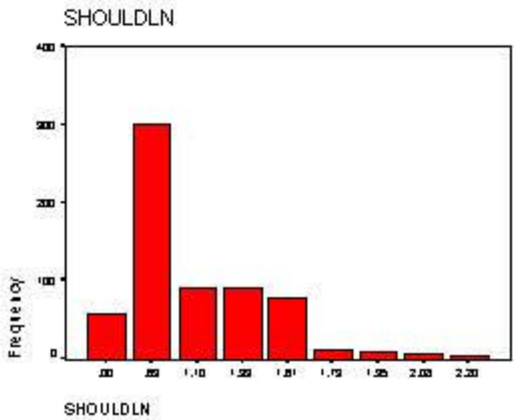
		SHOULD LN	CONFUS LN
N	Valid	655	649
	Missing	7	13
Mean		.9598	.8325
Std. Error of Mean		1.928E-02	2.422E-02
Median		.6931	.6931
Mode		.69	.69
Std. Deviation		.4934	.6170
Variance		.2434	.3806
Skewness		.066	.150
Std. Error of Skewness		.095	.096
Kurtosis		-.347	-.896
Std. Error of Kurtosis		.191	.192
Range		2.20	2.20
Minimum		.00	.00
Maximum		2.20	2.20
Sum		628.70	540.28

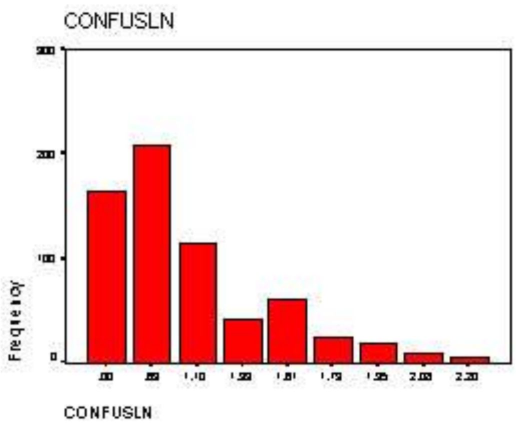
SHOULDLN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	58	8.8	8.9	8.9
	.69	301	45.5	46.0	54.8
	1.10	91	13.7	13.9	68.7
	1.39	93	14.0	14.2	82.9
	1.61	79	11.9	12.1	95.0
	1.79	13	2.0	2.0	96.9
	1.95	10	1.5	1.5	98.5
	2.08	6	.9	.9	99.4
	2.20	4	.6	.6	100.0
	Total	655	98.9	100.0	
Missing	System	7	1.1		
Total		662	100.0		

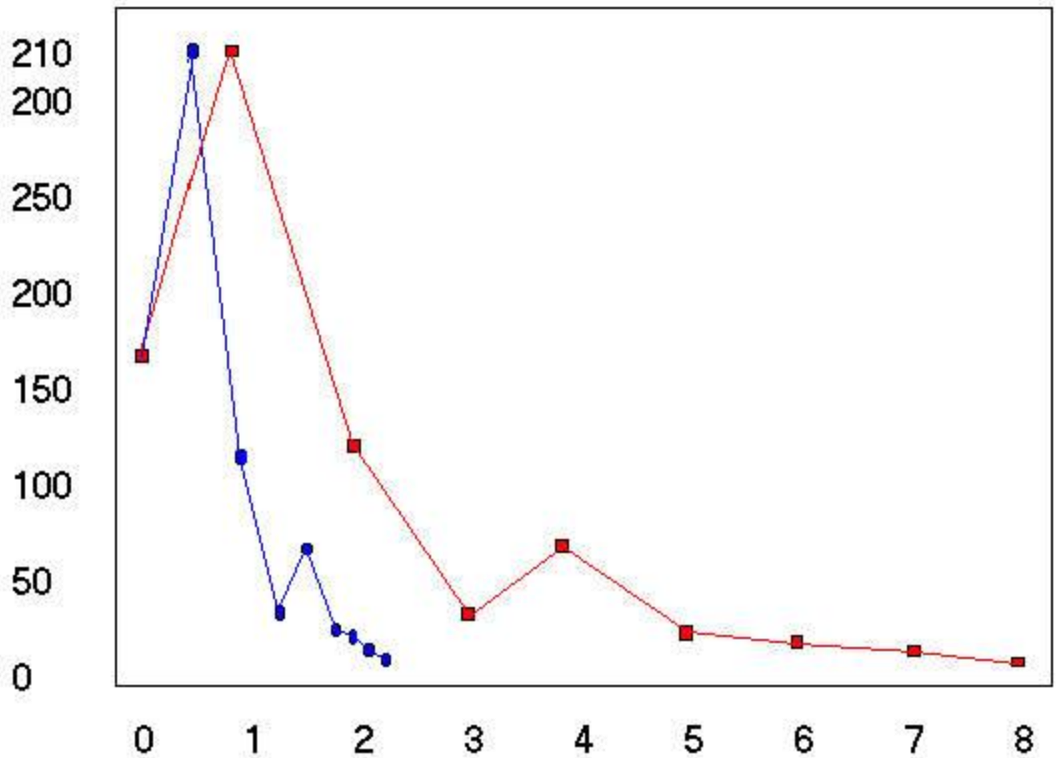
CONFUSLN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	164	24.8	25.3	25.3
	.69	208	31.4	32.0	57.3
	1.10	115	17.4	17.7	75.0
	1.39	42	6.3	6.5	81.5
	1.61	62	9.4	9.6	91.1
	1.79	24	3.6	3.7	94.8
	1.95	19	2.9	2.9	97.7
	2.08	10	1.5	1.5	99.2
	2.20	5	.8	.8	100.0
	Total	649	98.0	100.0	
Missing	System	13	2.0		
Total		662	100.0		

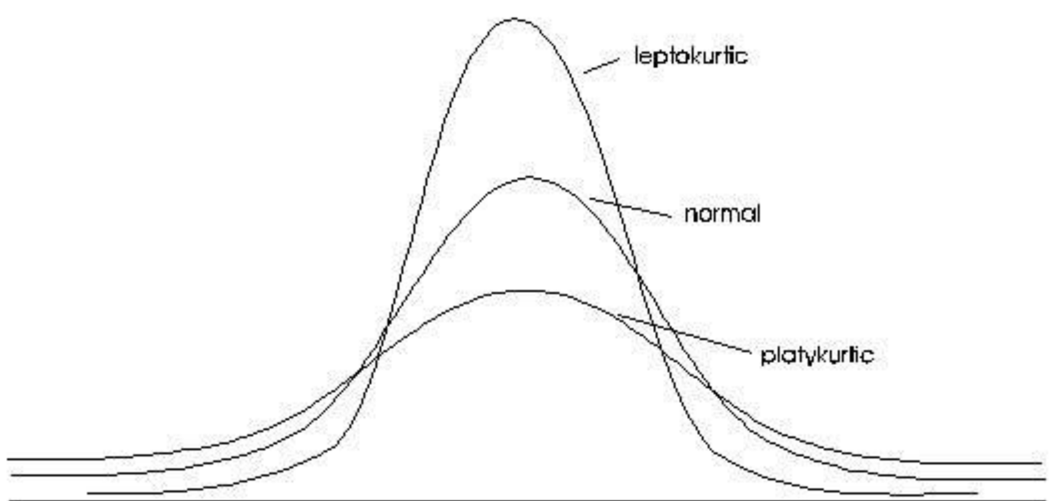


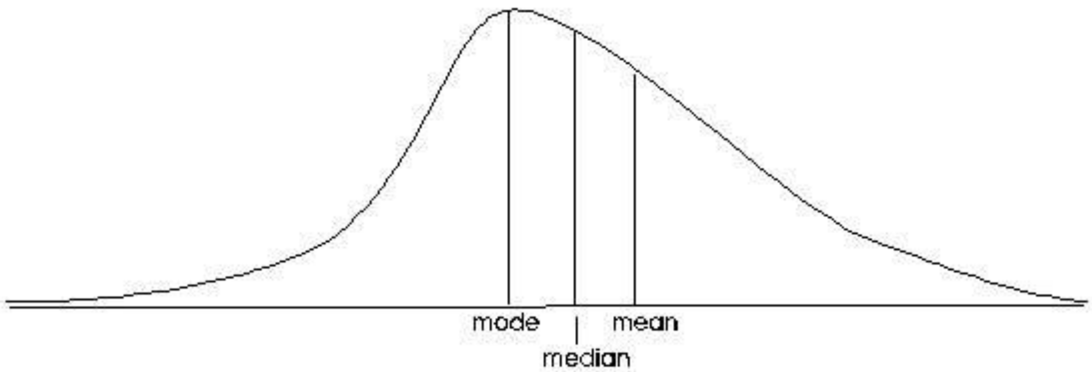


Notice that skewness has been corrected for both variables. Although it is not intuitive when comparing the bar charts. Consequently, the overlay chart has been drawn below to show the correction. In the chart below both the original and natural log of the variable CONFUSE is plotted to show how the skewed variable has become normal.

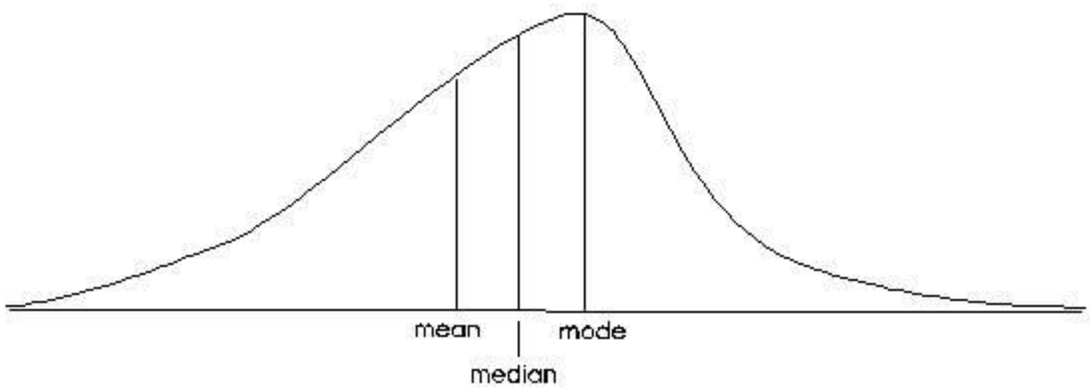


Notice the original variable (in red) has a much longer "tail" indicating skewness.





Positively skewed distribution



Negatively skewed distribution

An Example of Generating Standard Scores

This next set of SPSS computer runs demonstrates how to generate Standard Scores. Two different Standard Scores are generated (1) DEPRES and (2) TENSE. The means of the two variables are taken from the above analysis. The z-score (which is a Standard Score) is

File Name = crslst2.sps

```
get file = '\rdda\crsleq1.sav'/keep=
    GROUP LEISUR FEAR DEPRES FEELG
    ANGRY CONFUS WORTH TENSE USELES SATISF.
compute ltly=depres - 3.45.
compute ltlx=tense  - 3.40.
compute ltly2=ltly * ltly.
compute ltlx2=ltlx * ltlx.
compute crspdts=ltly * ltlx.
list ltly, ltlx.
fre var = ltly2 ltlx2
  /statistics=all.
```

LTLY	LTLX
-3.45	-2.40
.55	.60
-2.45	-1.40
-2.45	-3.40
-3.45	-2.40
-3.45	-1.40
.55	.60
-2.45	-2.40
-1.45	-1.40
.55	-.40
2.55	2.60
4.55	4.60
.55	-1.40
4.55	4.60
3.55	1.60
-3.45	-3.40
-.45	-.40
2.55	-.40
2.55	1.60
.55	4.60

Number of cases read: 20 Number of cases listed: 20

LTLY2

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	.20	1	5.0	5.0	5.0
	.30	5	25.0	25.0	30.0
	2.10	1	5.0	5.0	35.0
	6.00	3	15.0	15.0	50.0
	6.50	3	15.0	15.0	65.0
	11.90	4	20.0	20.0	85.0
	12.60	1	5.0	5.0	90.0
	20.70	2	10.0	10.0	100.0
	Total	20	100.0	100.0	

Mean	7.148	Std err	1.452	Median	6.253
Mode	.302	Std dev	6.492	Variance	42.148
Kurtosis	-.107	S E Kurt	.992	Skewness	.748
S E Skew	.512	Range	20.500	Minimum	.203
Maximum	20.702	Sum	142.950		

Valid cases 20 Missing cases 0

LTLX2

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	.16	3	15.0	15.0	15.0
	.36	2	10.0	10.0	25.0
	1.96	4	20.0	20.0	45.0
	2.56	2	10.0	10.0	55.0
	5.76	3	15.0	15.0	70.0
	6.76	1	5.0	5.0	75.0
	11.56	2	10.0	10.0	85.0
	21.16	3	15.0	15.0	100.0
	Total	20	100.0	100.0	

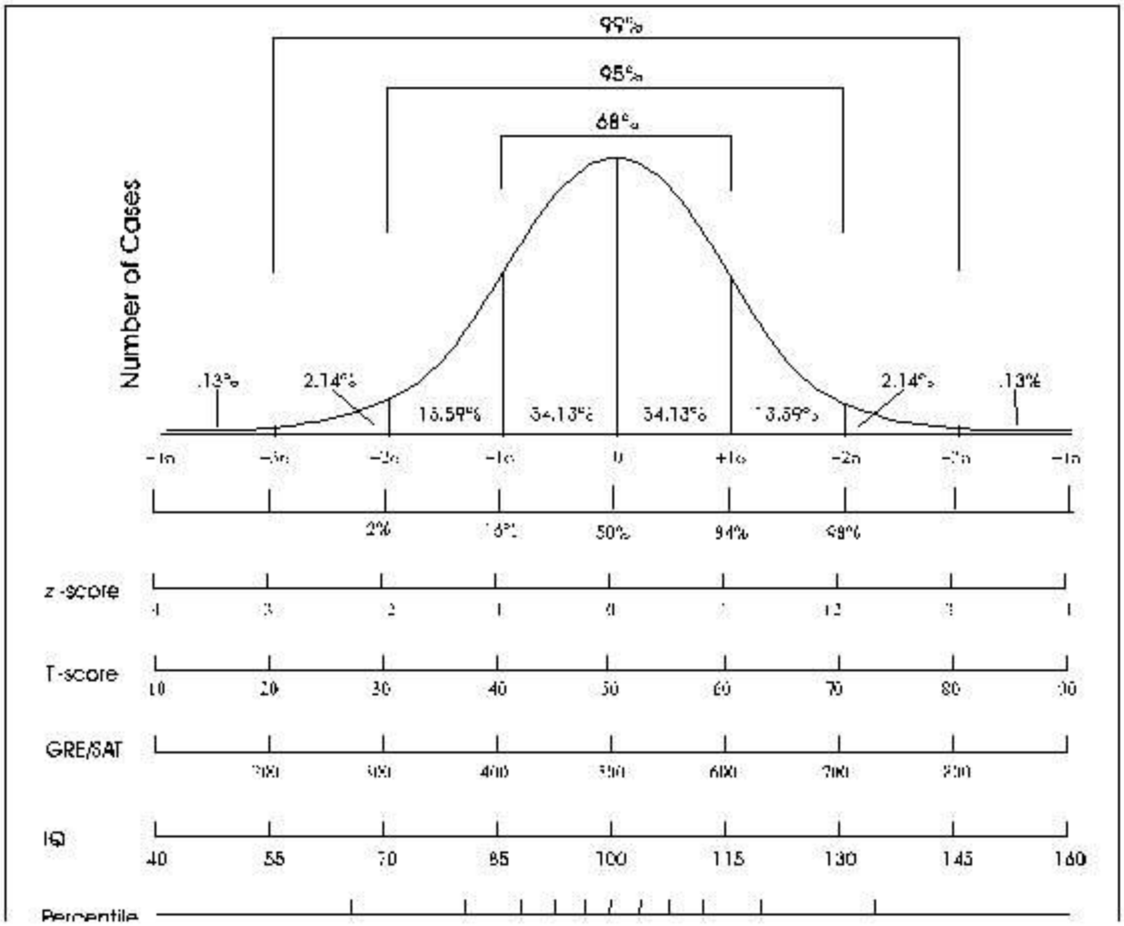
Mean	6.240	Std err	1.627	Median	2.560
Mode	1.960	Std dev	7.276	Variance	52.936
Kurtosis	.546	S E Kurt	.992	Skewness	1.332
S E Skew	.512	Range	21.000	Minimum	.160
Maximum	21.160	Sum	124.800		

The square root of 6.240 = 2.498.
The square root of 7.148 = 2.674.
These values are the population standard ard deviation for TENSE and DEPRES respectively.
The z-score (from chapter 3) is X minus the Mean dividied by the Standard Deviation.

file name = crslst3.sps

```
get file = "\rdda\crsleq1.sav'/keep=
    GROUP LEISUR FEAR DEPRES FEELG
    ANGRY CONFUS WORTH TENSE USELES SATISF.
compute ltly=depres - 3.45.
compute ltlx=tense - 3.40.
compute ltly2=ltly * ltly.
compute ltlx2=ltlx * ltlx.
compute crspdts=ltly * ltlx.
compute ltlzx=ltlx/2.498.
compute ltlzy=ltly/2.674.
list ltlzx, ltlzy.
```

LTLZX	LTLZY
-.96	-1.29
.24	.21
-.56	-.92
-1.36	-.92
-.96	-1.29
-.56	-1.29
.24	.21
-.96	-.92
-.56	-.54
-.16	.21
1.04	.95
1.84	1.70
-.56	.21
1.84	1.70
.64	1.33
-1.36	-1.29
-.16	-.17
-.16	.95
.64	.95
1.84	.21



These are refered to as Standard Scores - they have a mean of 0 and a standard deviation of 1. T-scores have a mean of 50 or 100 and a standard deviation of 5 or 10. The WAIS has a mena of 100 and a standard deviation of 15. The following jobstream computes 3 different standard scores (1) the first has a mean of 50 and a standard deviation of 10, (2) the second has a mean of 100 and a standard deviation of 15, and the final (3) has a mean of 37 (possibly the average age that people become depressed -- nobody says this has to make sense) and a standard deviation of 8.3 (some other possibility). A little weird but you can do anything you want as long as it does not have to make sense.

file name = crslst4.sps
get file = "\rdda\crsleq1.sav'/keep= GROUP LEISUR FEAR DEPRES FEELG ANGRY CONFUS WORTH TENSE USELES SATISF. compute ltly=depres - 3.45. compute ltlx=tense - 3.40. compute ltly2=ltly * ltly. compute ltlx2=ltlx * ltlx. compute crspdts=ltly * lt1x. compute ltlzx=lt1x/2.498. compute ltlzy=ltly/2.674. compute ltlzx5= 50 + ((lt1x/2.498)*10). compute ltlzxh= 100 + ((lt1x/2.498)*15). compute ltlzxd= 37 + ((lt1x/2.498)*8.3). list lt1zx, ltlzy, lt1zx5, ltlzxh, lt1zxd.

```
get file =  "\rdda\crsleq1.sav'/keep=
    GROUP LEISUR FEAR DEPRES FEELG
    ANGRY CONFUS WORTH TENSE USELES SATISF.
compute ltly=depres - 3.45.
compute ltlx=tense  - 3.40.
compute ltly2=ltly * ltly.
compute ltlx2=ltlx * ltlx.
compute crspdts=ltly * lt1x.
compute ltlzx=lt1x/2.498.
compute ltlzy=ltly/2.674.
compute ltlzx5= 50 + ((lt1x/2.498)*10).
compute ltlzxh= 100 + ((lt1x/2.498)*15).
compute ltlzxd= 37  + ((lt1x/2.498)*8.3).
list lt1zx, ltlzy, lt1zx5, ltlzxh, lt1zxd.
```

LTLZX	LTLZY	LTLZX5	LTLZXH	LTLZXD
-.96	-1.29	40.39	85.59	29.03
.24	.21	52.40	103.60	38.99
-.56	-.92	44.40	91.59	32.35
-1.36	-.92	36.39	79.58	25.70
-.96	-1.29	40.39	85.59	29.03
-.56	-1.29	44.40	91.59	32.35
.24	.21	52.40	103.60	38.99
-.96	-.92	40.39	85.59	29.03
-.56	-.54	44.40	91.59	32.35
-.16	.21	48.40	97.60	35.67
1.04	.95	60.41	115.61	45.64
1.84	1.70	68.41	127.62	52.28
-.56	.21	44.40	91.59	32.35
1.84	1.70	68.41	127.62	52.28
.64	1.33	56.41	109.61	42.32
-1.36	-1.29	36.39	79.58	25.70
-.16	-.17	48.40	97.60	35.67
-.16	.95	48.40	97.60	35.67
.64	.95	56.41	109.61	42.32
1.84	.21	68.41	127.62	52.28