Normal Distribution contd..

Dr.Mamatha.H.R
Professor
Department of Computer Science and Engineering
PES University
Bangalore

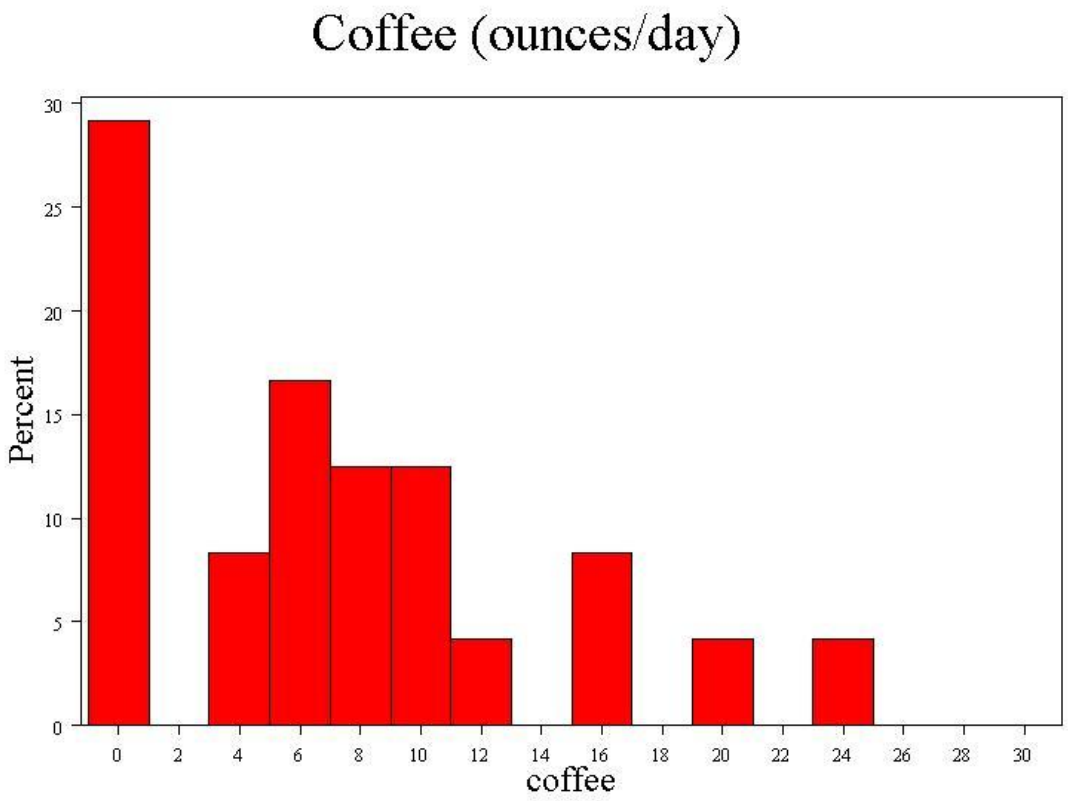
Course material created using various Internet resources and text book

Are my data “normal”?

- Not all continuous random variables are normally distributed!!
- It is important to evaluate how well the data are approximated by a normal distribution

Are my data normally distributed?

1. Look at the histogram! Does it appear bell shaped?
2. Compute descriptive summary measures—are mean, median, and mode similar?
3. Do 2/3 of observations lie within 1 std dev of the mean? Do 95% of observations lie within 2 std dev of the mean?
4. Look at a normal probability plot—is it approximately linear?
5. Run tests of normality (such as Kolmogorov-Smirnov). But, be cautious, highly influenced by sample size!



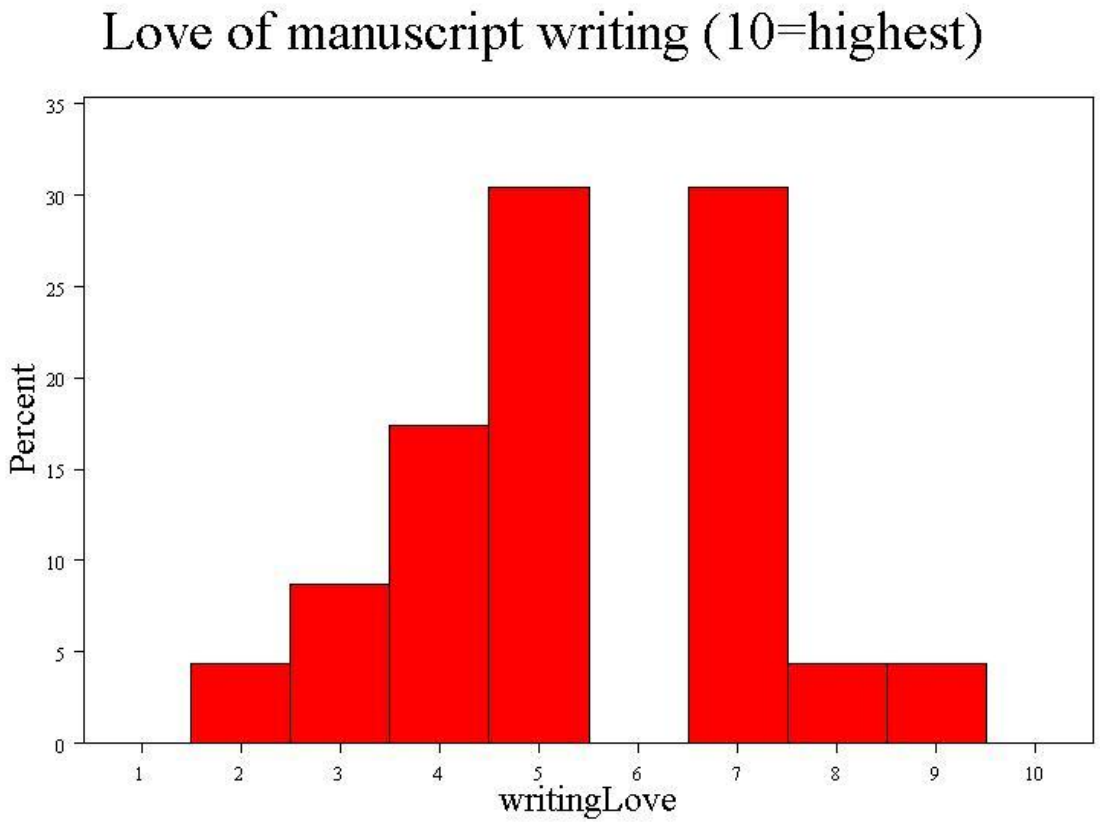
Median = 6

Mean = 7.1

Mode = 0

SD = 6.8

Range = 0 to 24
(= 3.5 ∫)



Median = 5

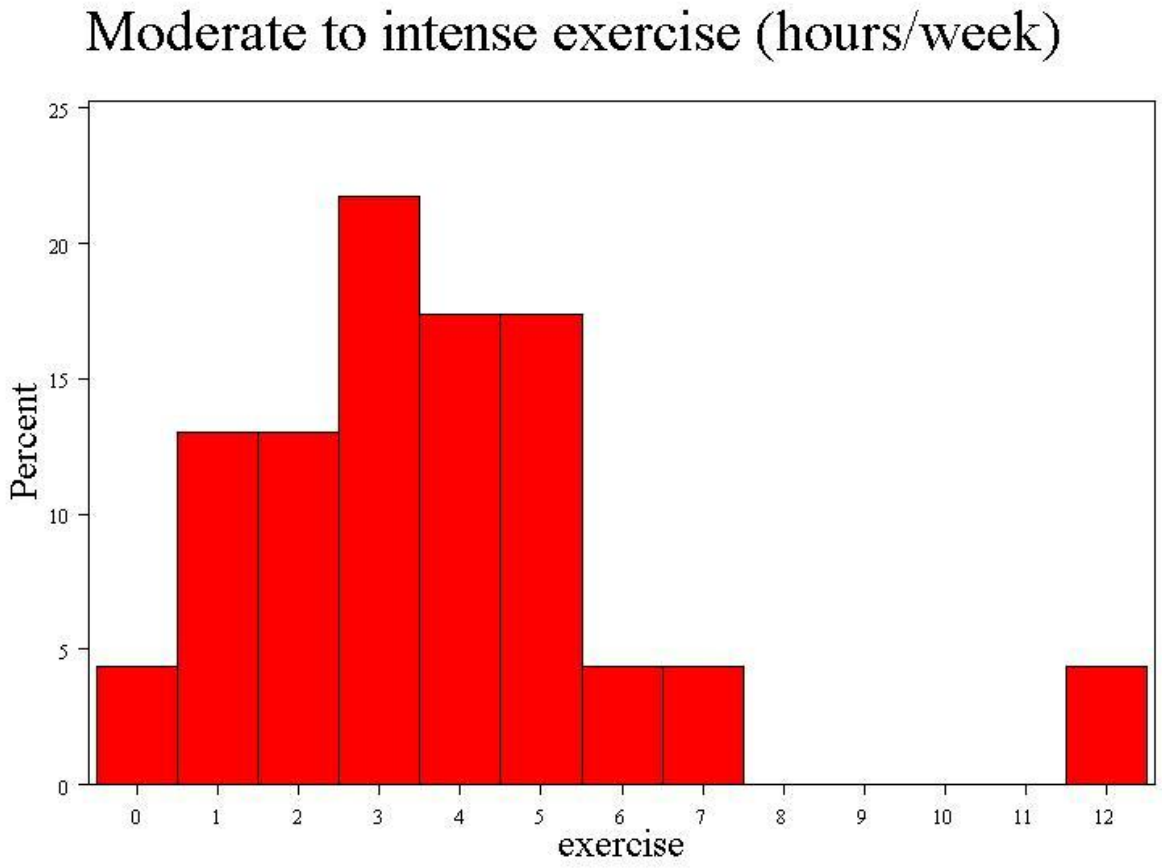
Mean = 5.4

Mode = none

SD = 1.8

Range = 2 to 9

(~ 4 ∫)



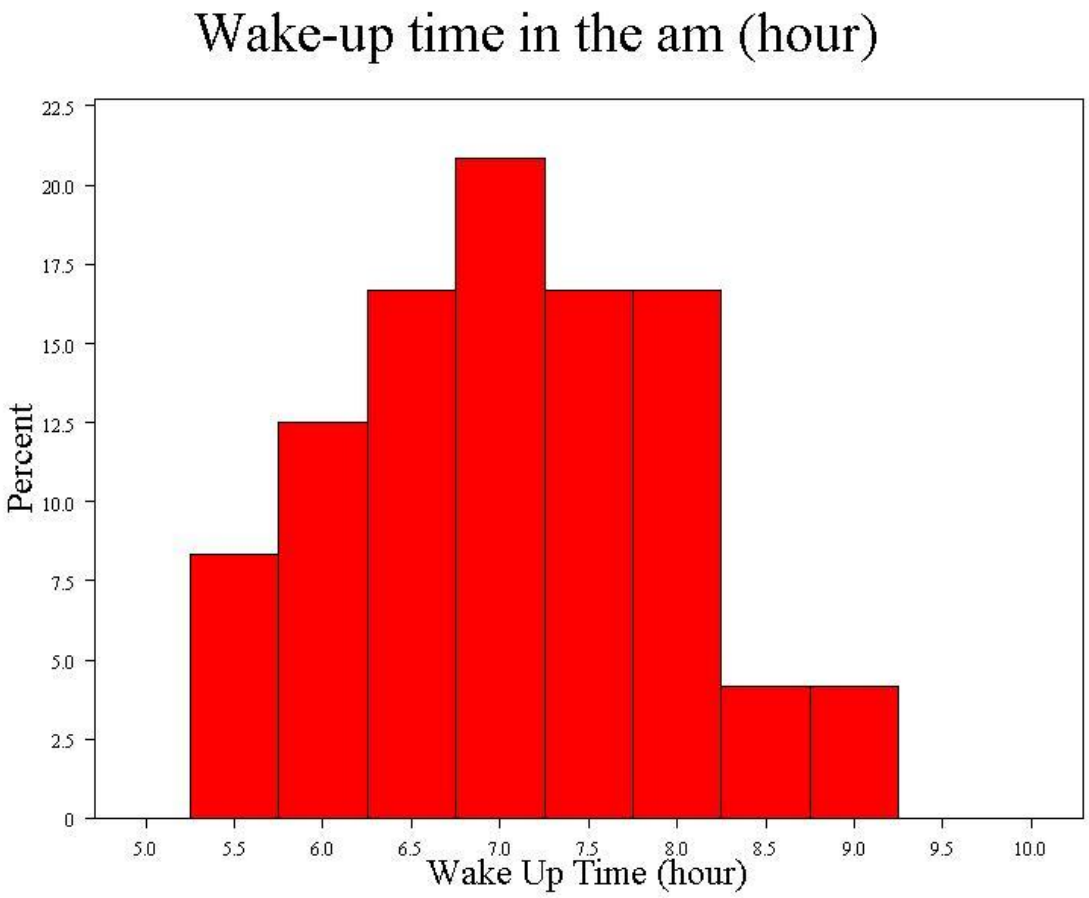
Median = 3

Mean = 3.4

Mode = 3

SD = 2.5

Range = 0 to 12
(~ 5 ∫)



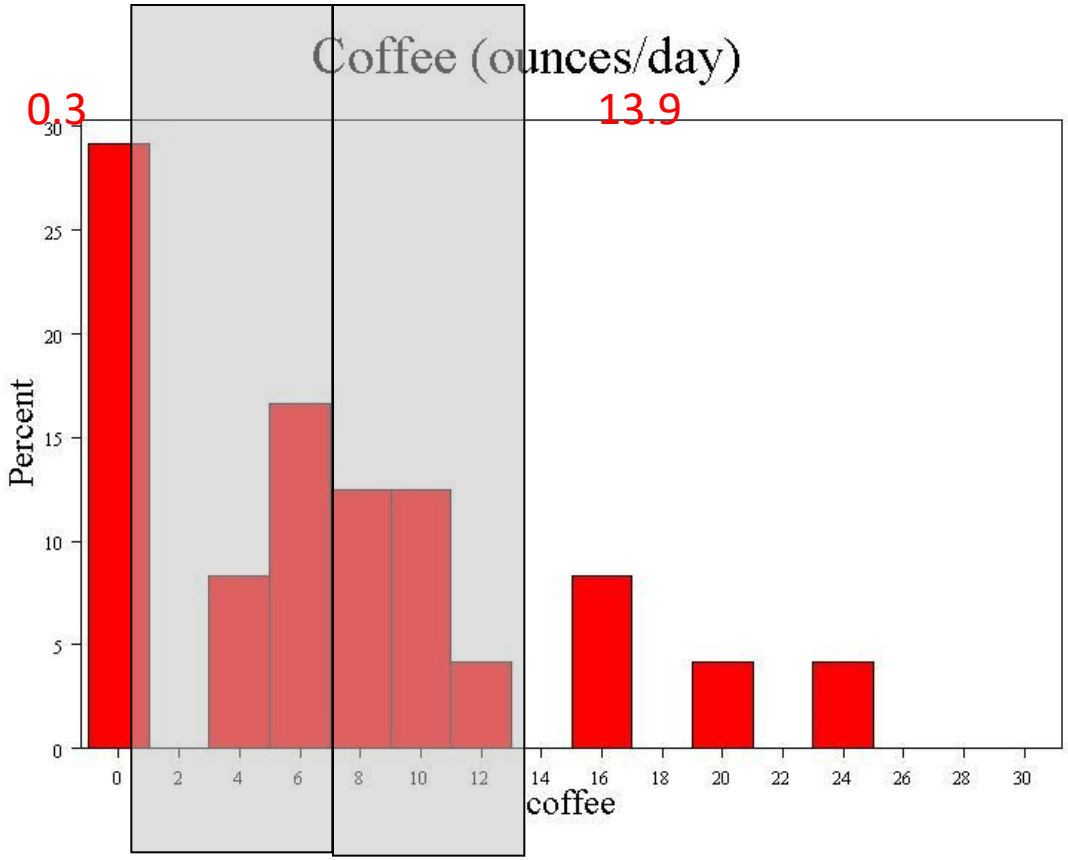
Median = 7:00

Mean = 7:04

Mode = 7:00

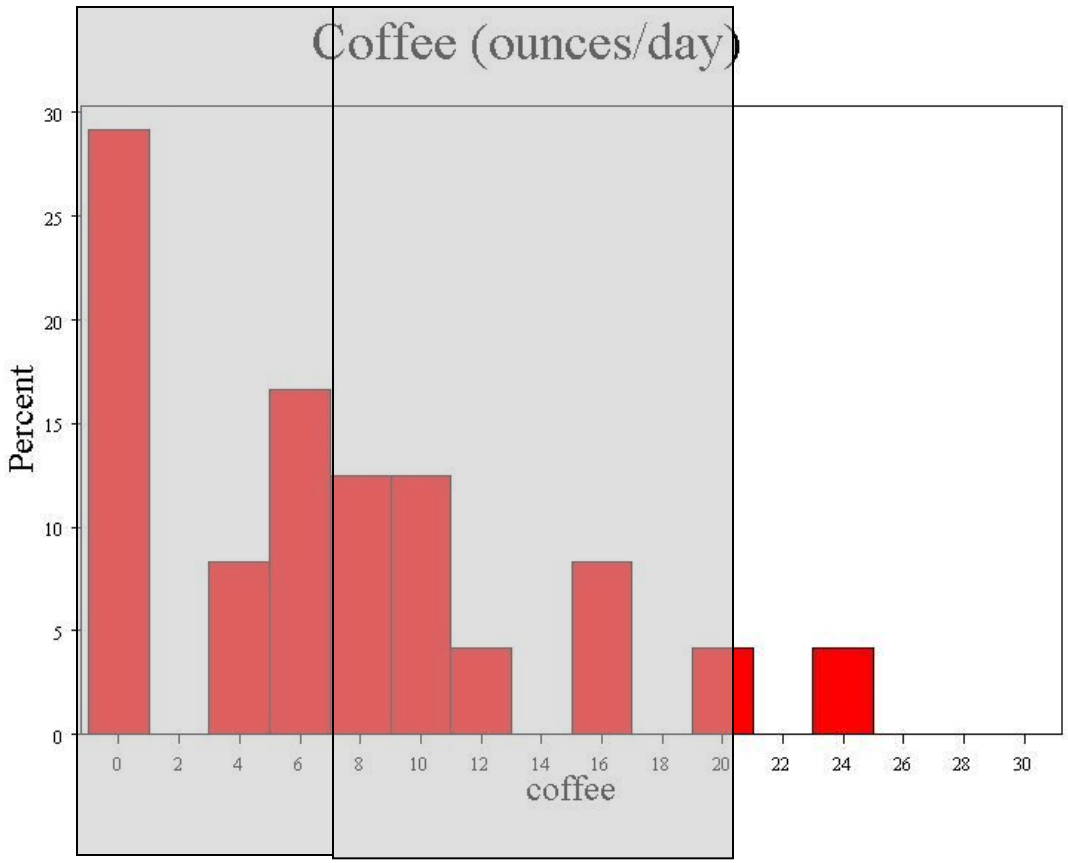
SD = :55

Range = 5:30 to 9:00 (~4 ∫)


Coffee (ounces/day)
0.3
13.9
Percent
0
5
10
15
20
25
30
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
coffee

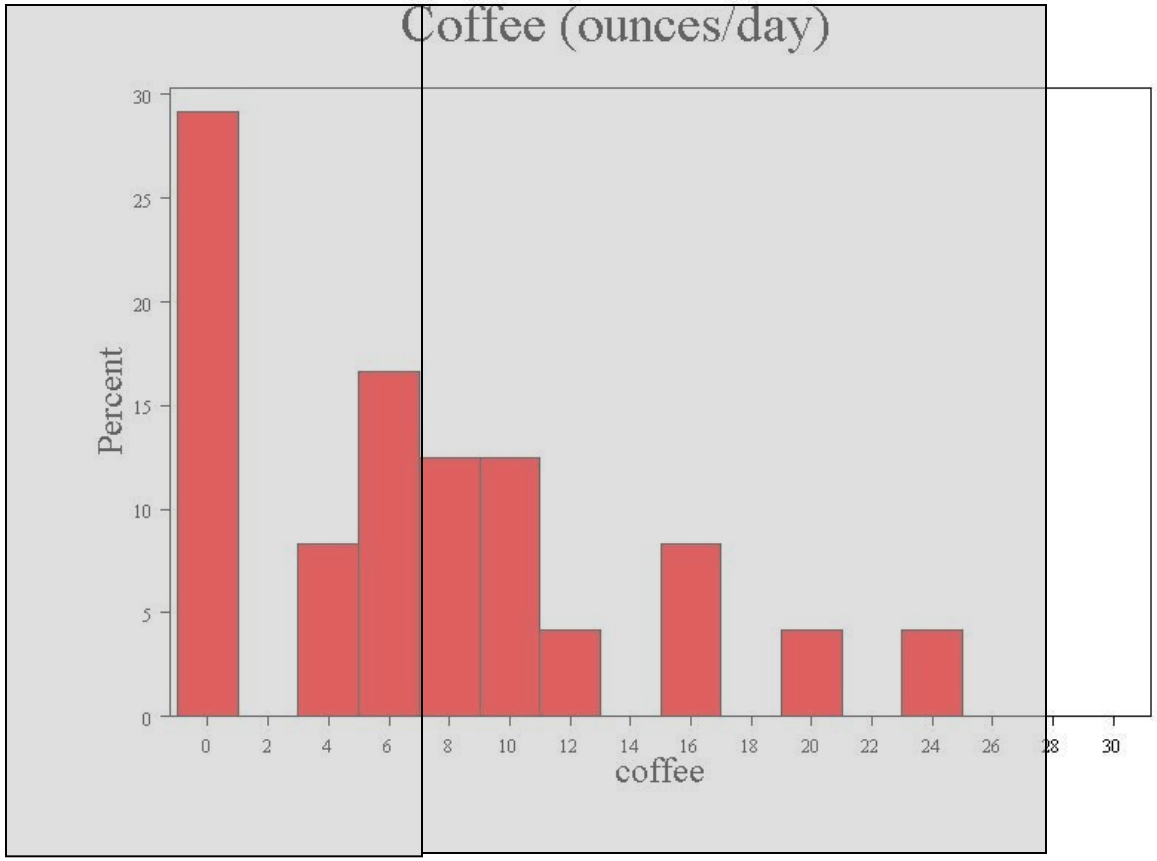
7.1 +/- 6.8 =

0.3 – 13.9


Coffee (ounces/day)
Percent
30
25
20
15
10
5
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
coffee

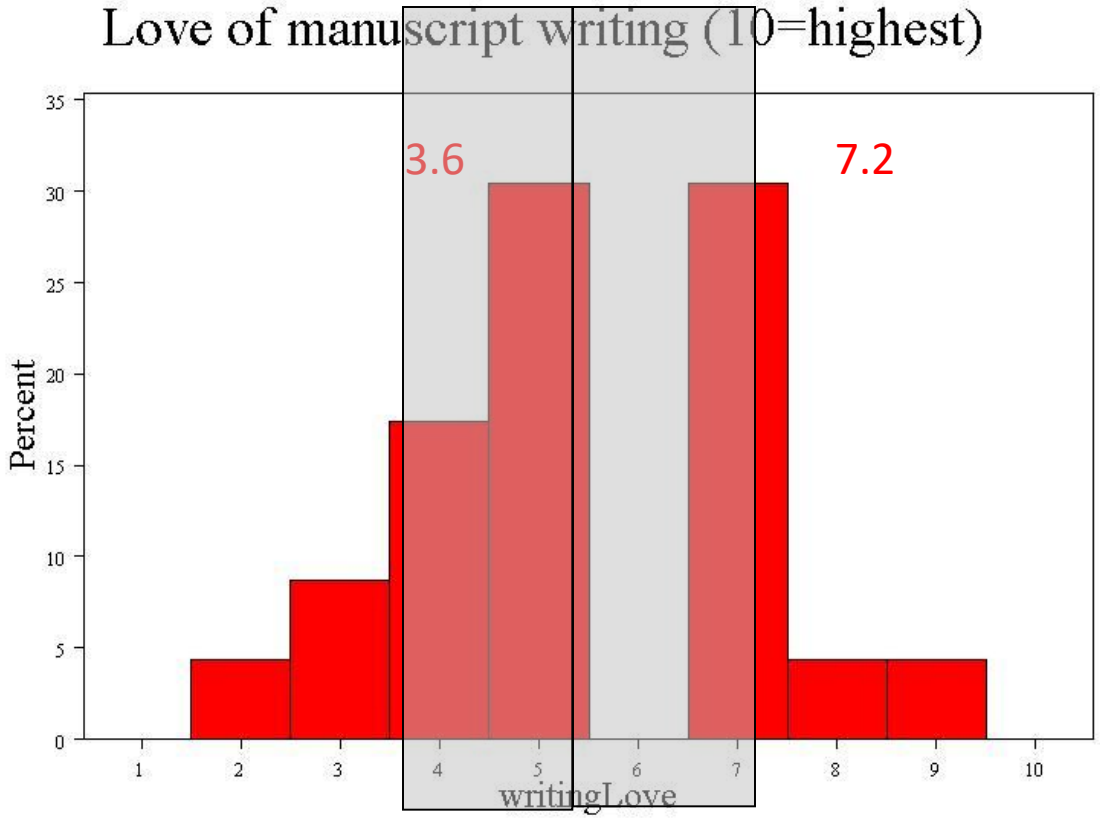
7.1 +/- 2*6.8 =

0 – 20.7


Coffee (ounces/day)
Percent
30
25
20
15
10
5
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
coffee

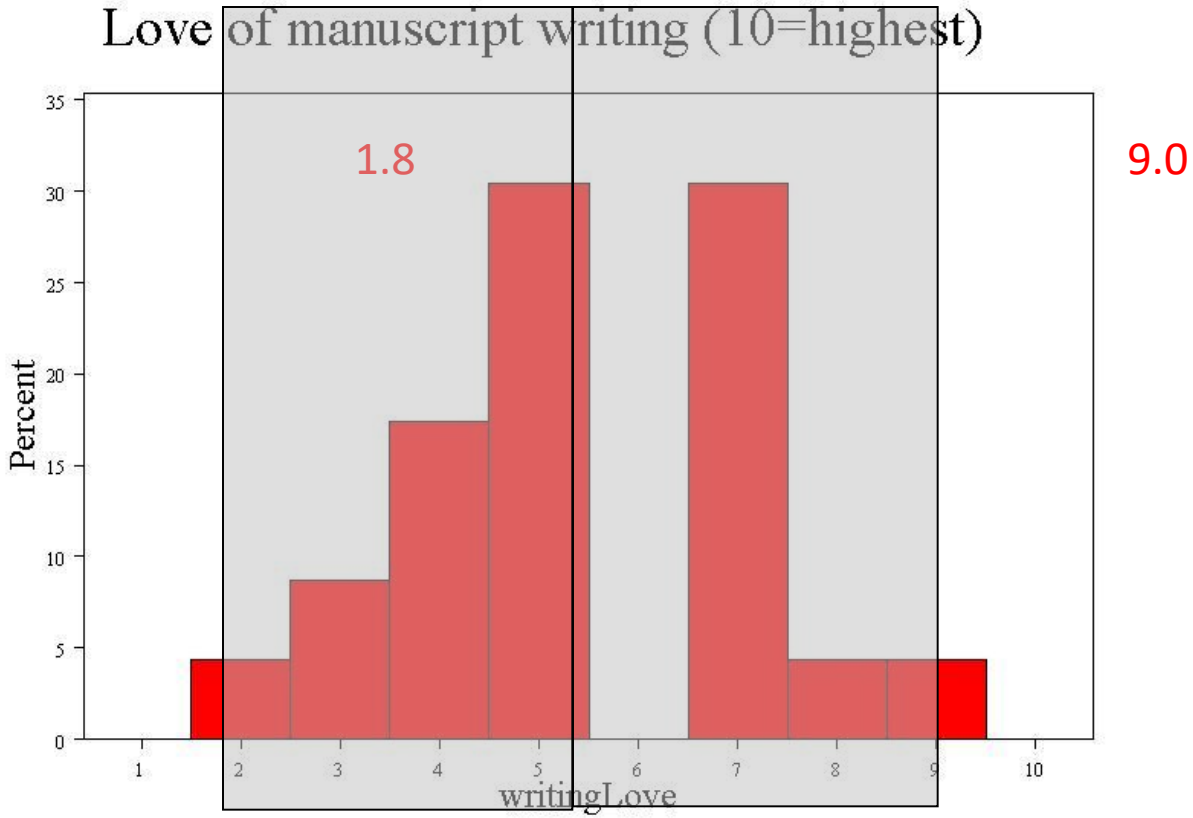
7.1 +/- 3*6.8 =

0 – 27.5


Love of manuscript writing (10=highest)
3.6
7.2
Percent
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
writingLove

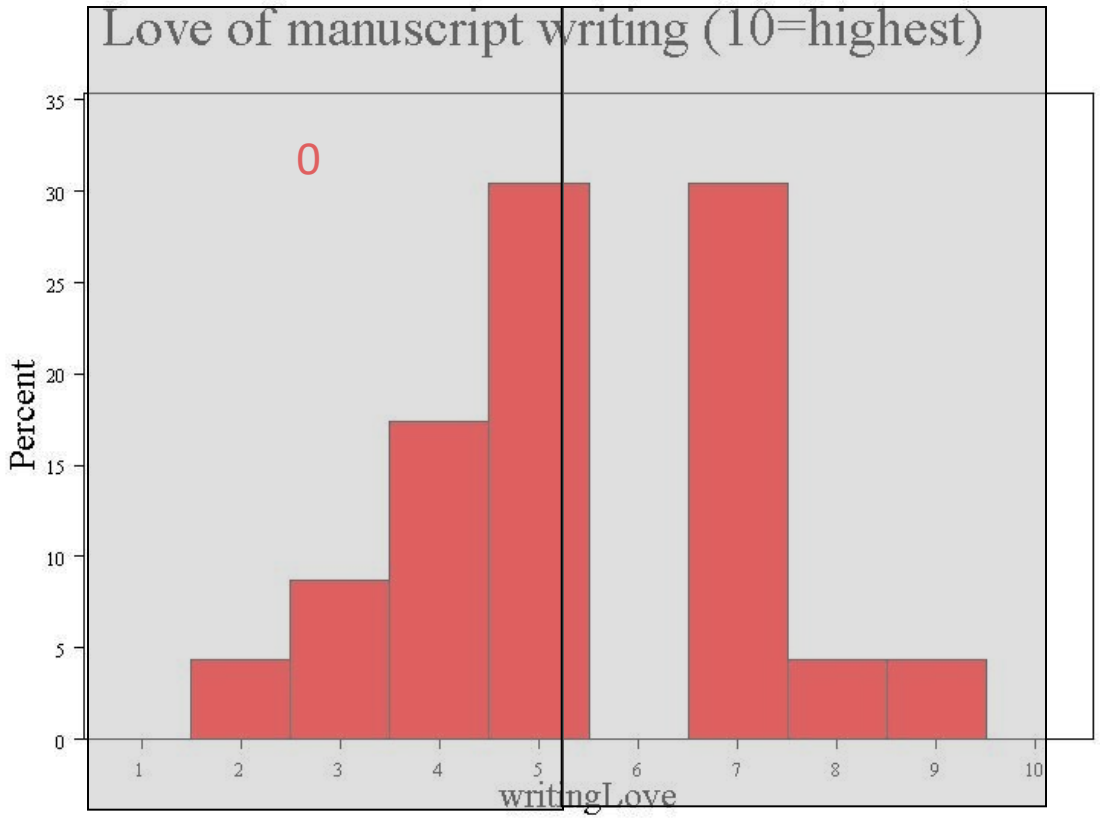
5.4 +/- 1.8 =

3.6 – 7.2


Love of manuscript writing (10=highest)
1.8
9.0
Percent
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
writingLove

5.4 +/- 2*1.8 =

1.8 – 9.0

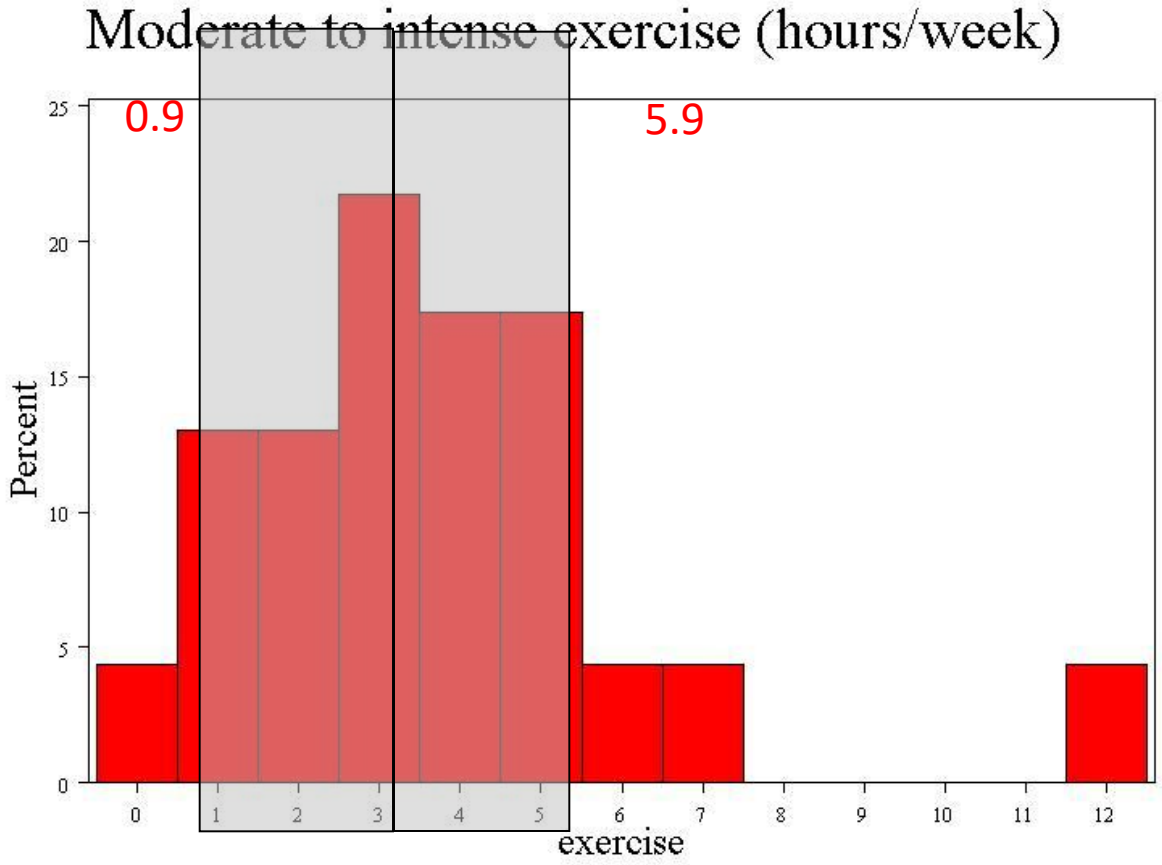


0

10

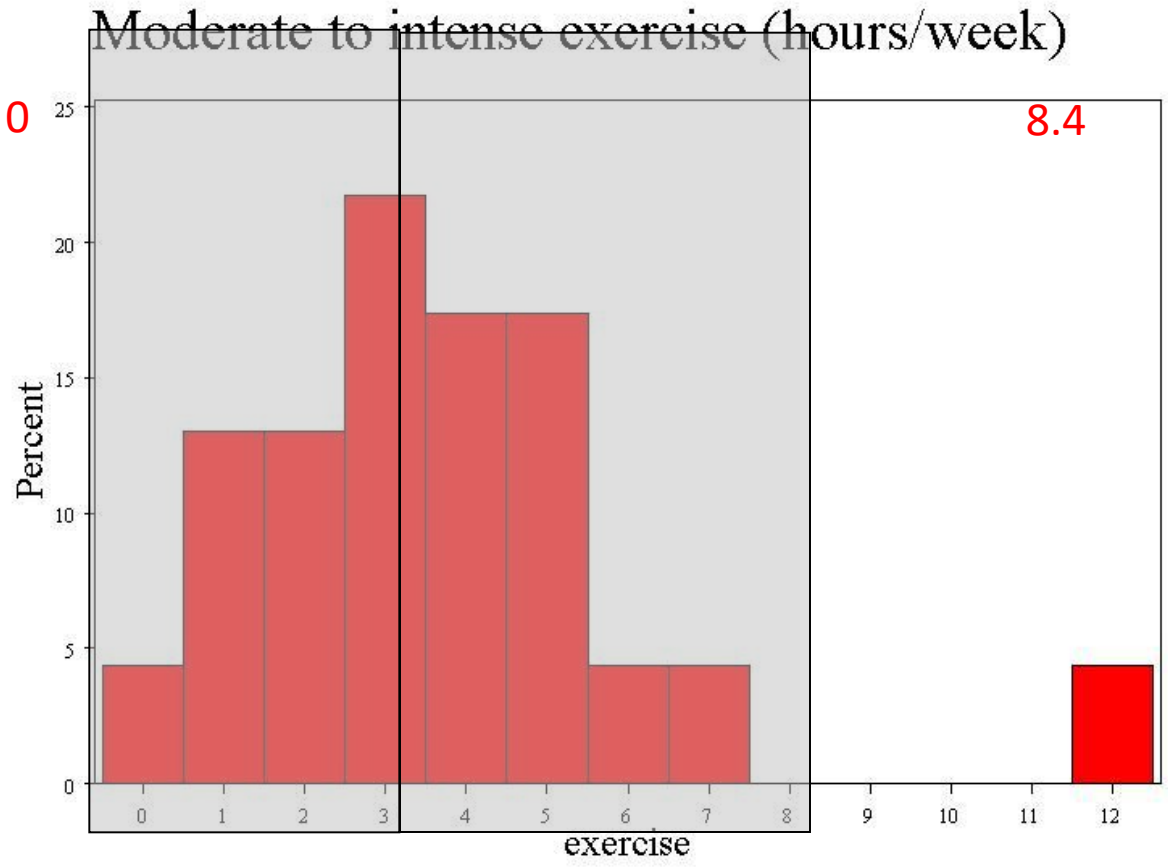
5.4 +/- 3*1.8 =

0– 10


Moderate to intense exercise (hours/week)
0.9
5.9
Percent
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
11
12
exercise

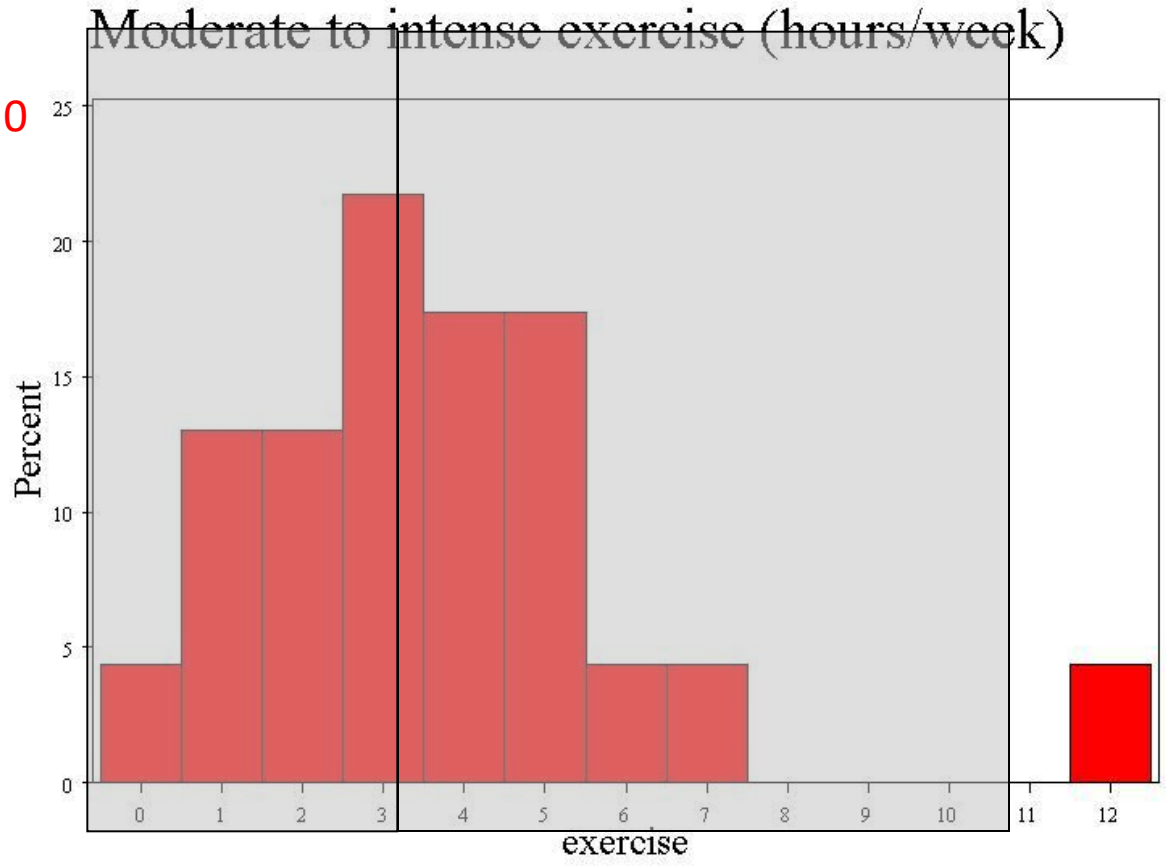
3.4 +/- 2.5=

0.9 – 7.9


Moderate to intense exercise (hours/week)
0
8.4
Percent
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
11
12
exercise

3.4 +/- 2*2.5=

0 – 8.4

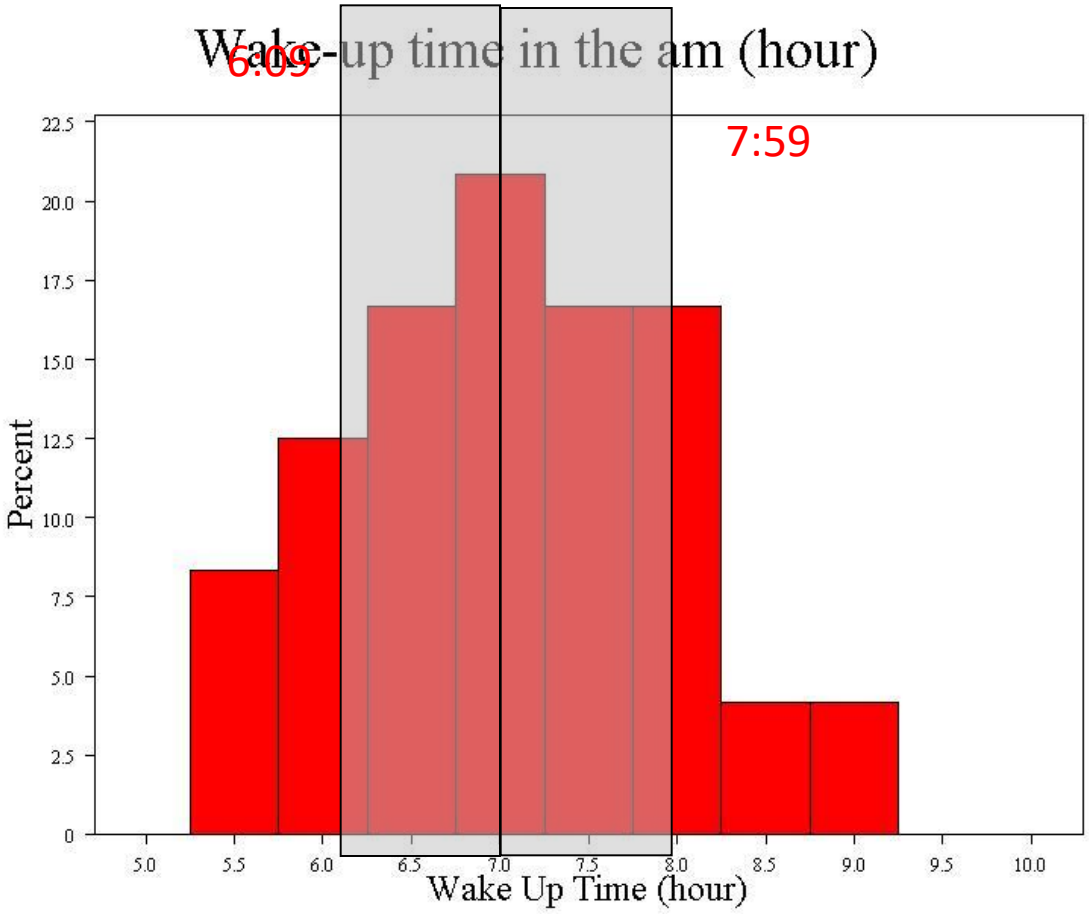


0

10.9

3.4 +/- 3*2.5=

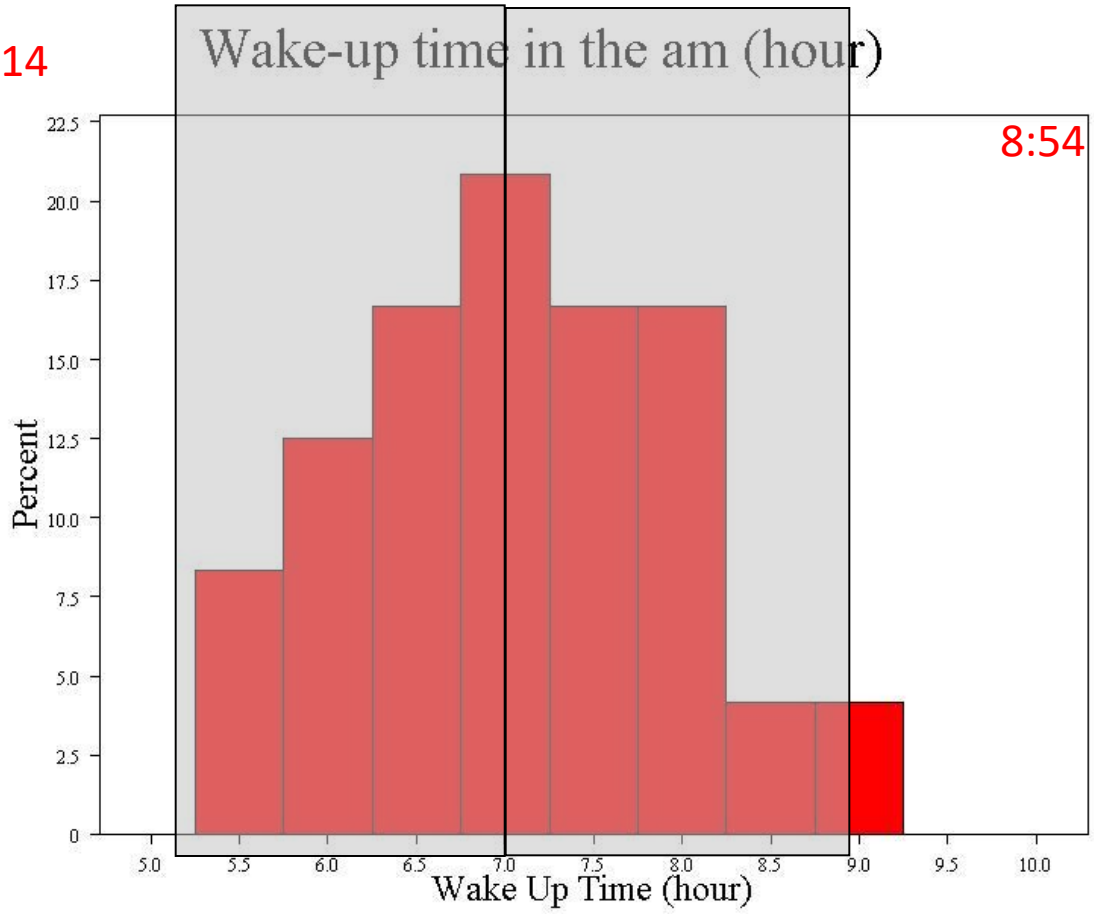
0 – 10.9



7:04+/- 0:55 =

6:09 – 7:59

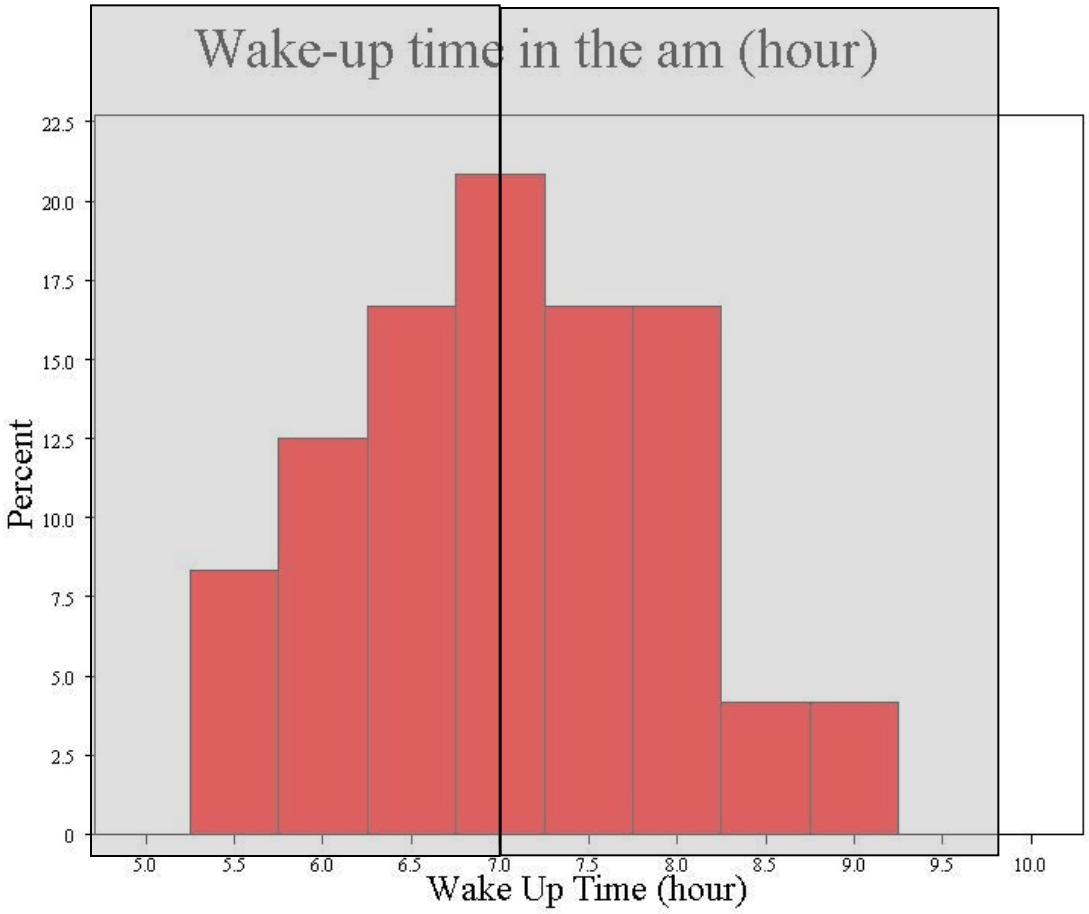
5:14



7:04+/- 2*0:55
=

5:14 – 8:54

4:19



9:49

7:04+/- 2*0:55
=

4:19 – 9:49

The Normal Probability Plot

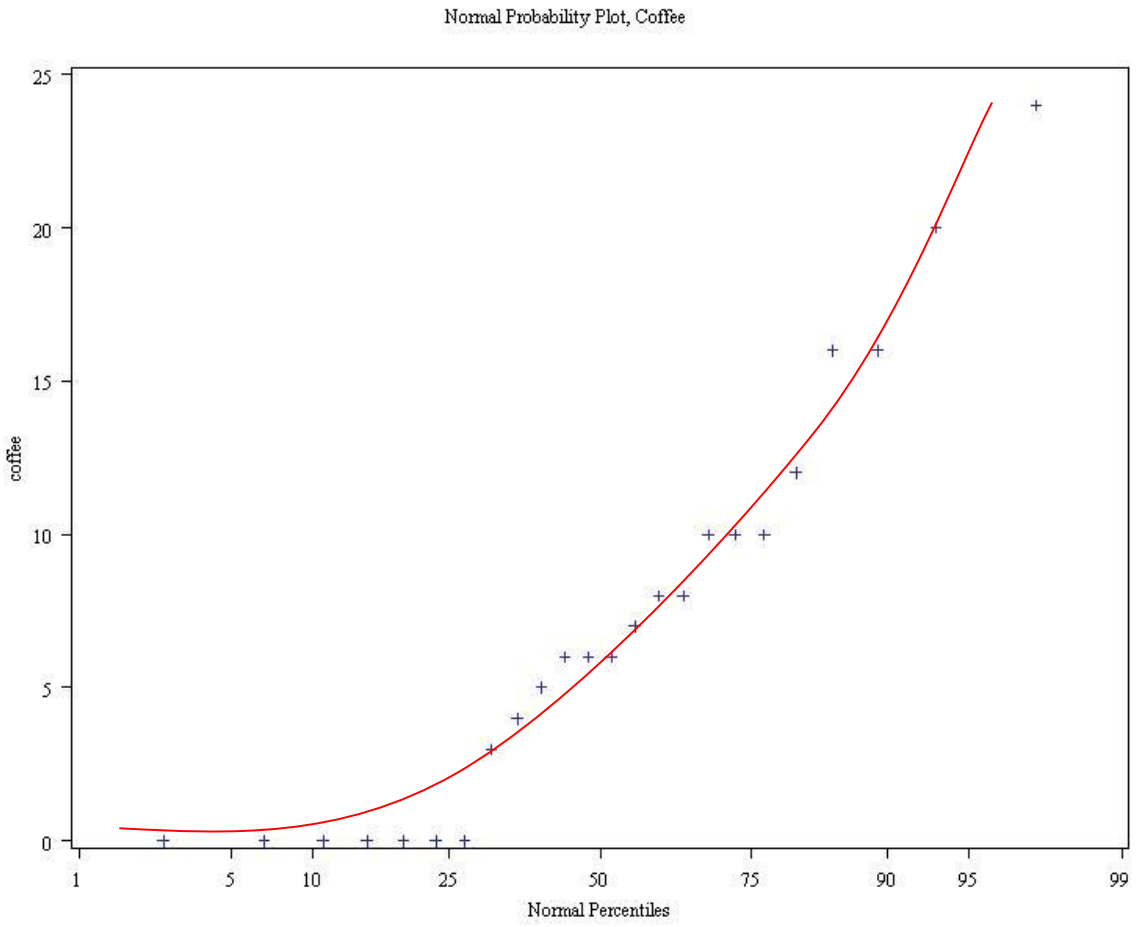
- **Normal probability plot**
  - **Order the data.**
  - **Find corresponding standardized normal quantile values:**

$$i^{th}\ \text{quantile} = \phi\left(\frac{i}{n+1}\right)$$

where ϕ is the probit function, which gives the Z value that corresponds to a particular left-tail area

  - **Plot the observed data values against normal quantile values.**
  - **Evaluate the plot for evidence of linearity.**

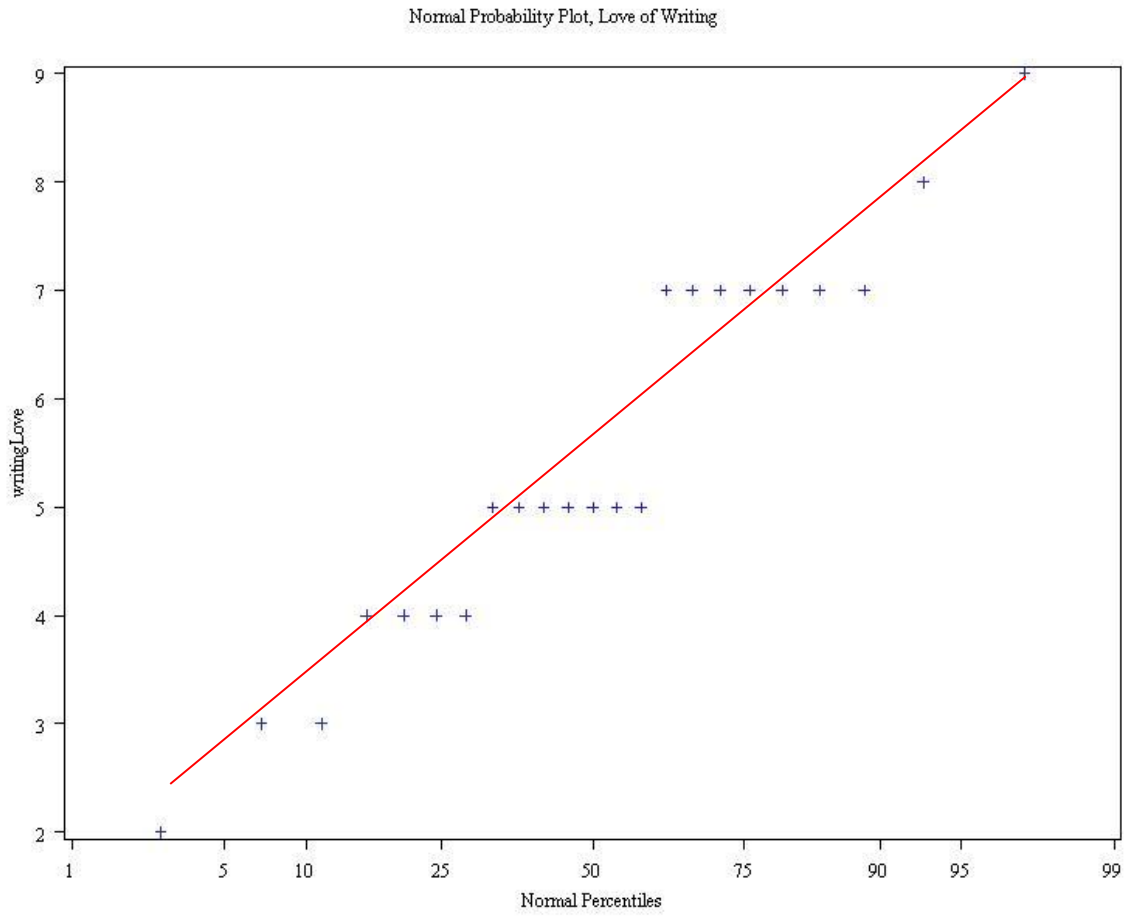
Normal probability plot coffee…



Right-Skewed!

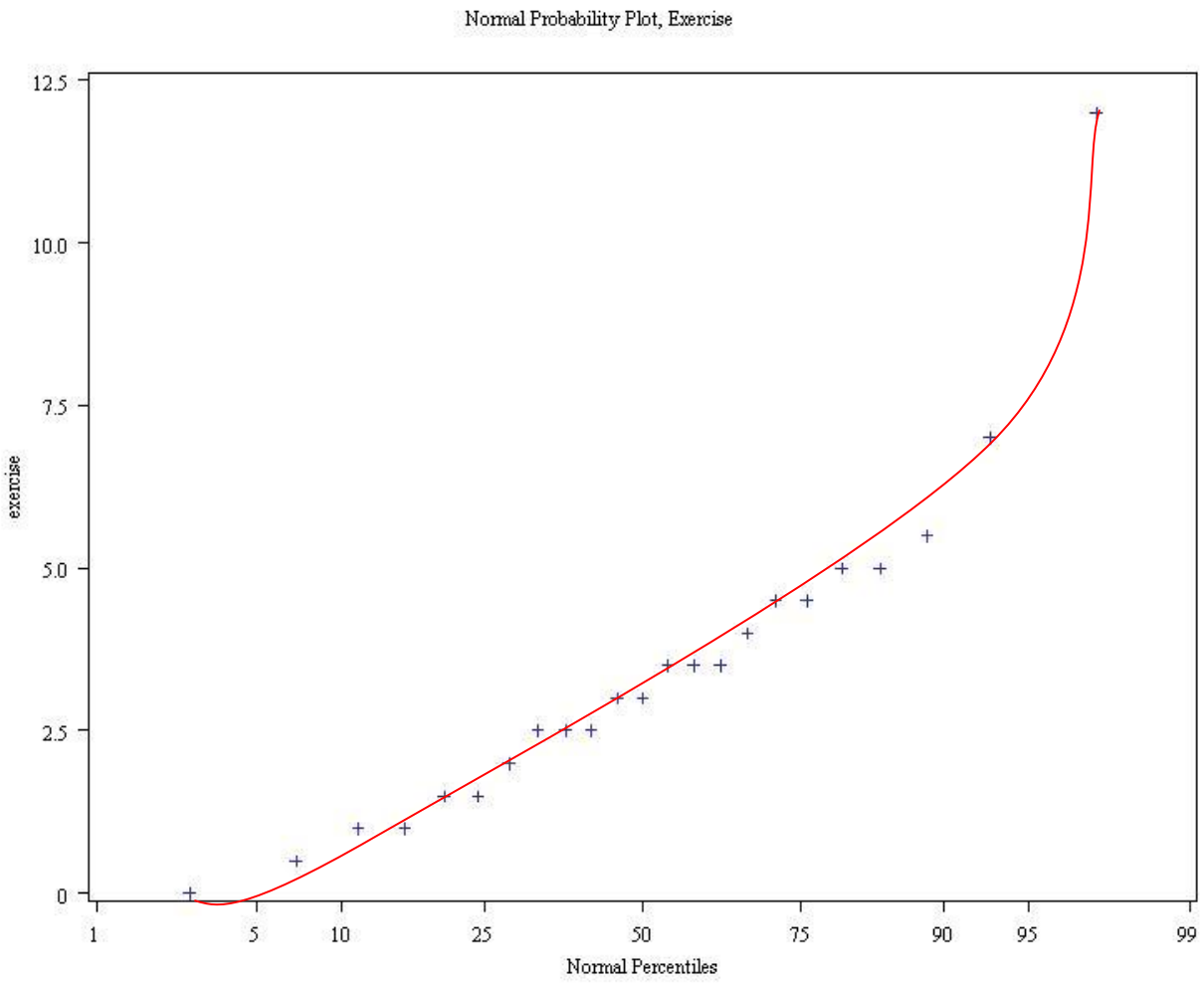
(concave up)

Normal probability plot writing…



Neither right-skewed or left-skewed, but big gap at 6.

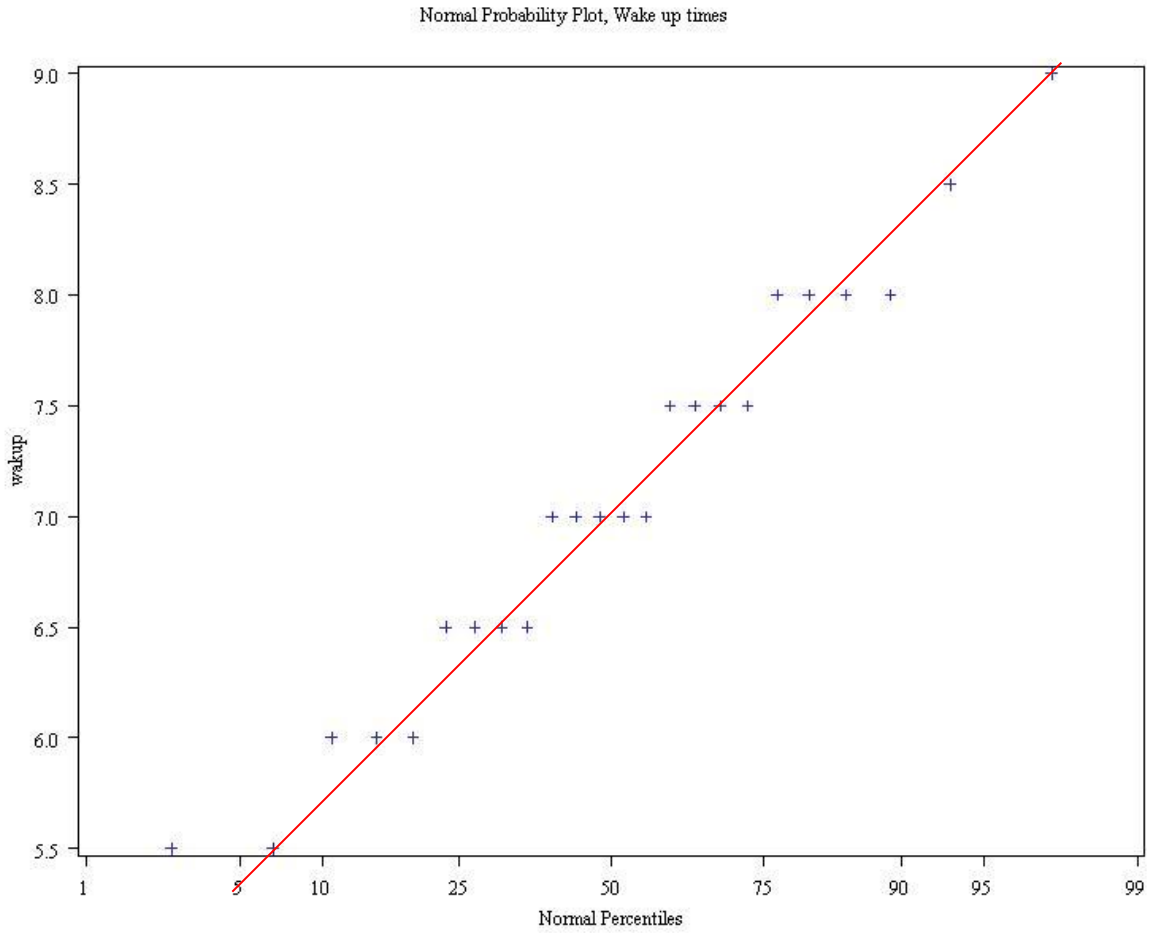
Norm prob. plot Exercise…



Right-Skewed!

(concave up)

Norm prob. plot Wake up time



Closest to a straight line…

Formal tests for normality

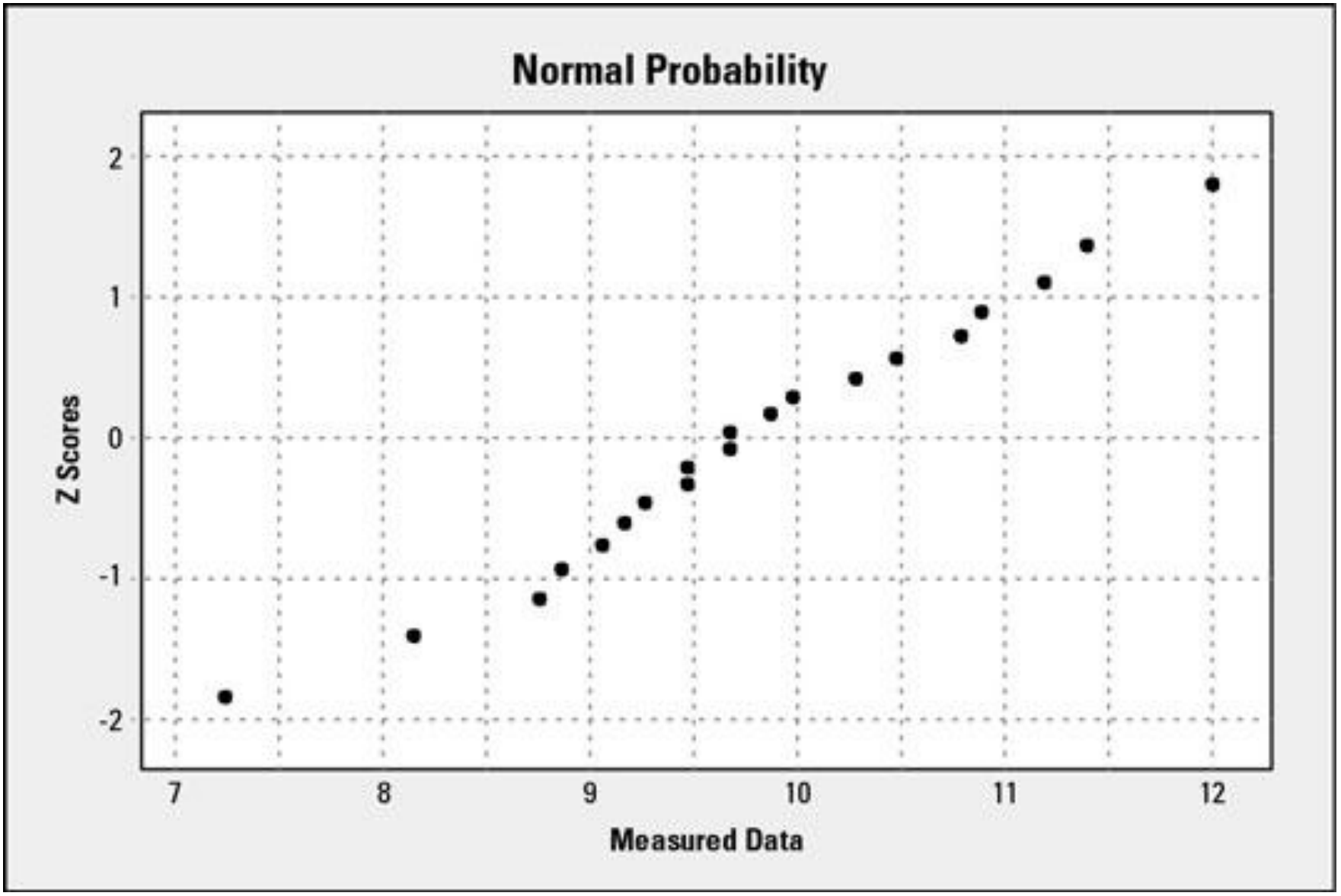
- Results:
- Coffee: Strong evidence of non-normality ($p<.01$)
- Writing love: Moderate evidence of non-normality ($p=.01$)
- Exercise: Weak to no evidence of non-normality ($p>.10$)
- Wakeup time: No evidence of non-normality ($p>.25$)

- Order your *n* number of points of raw data from the minimum value to the maximum observed values.
- Assign a rank order number (*i*) to each of the *n* points of data.That is, from minimum to maximum, is the point of data the 1st, 7th, or 98th?
- Calculate the cumulative probability (p_i) associated with each rank-ordered point of data.
- Use the followi

$$p_i = \frac{i - 0.5}{n}$$

- Use the standard normal table found in Z Table to calculate the z_i value for each of your *n* points of data.
- Create an *x*-*y* scatter plot of your measured data points versus their determined *z* values.
- The measured data go on the *x*-axis, and the *z* values go on the *y*-axis.

Rank-Ordered Data	I	p_i	z_i
7.3	**1**	**0.025**	**–1.96**
8.2	**2**	**0.075**	**–1.44**
8.8	**3**	**0.125**	**–1.15**
8.9	**4**	**0.175**	**–0.93**
9.1	**5**	**0.225**	**–0.76**
9.2	**6**	**0.275**	**–0.60**
9.3	**7**	**0.325**	**–0.45**
9.5	**8**	**0.375**	**–0.32**
9.5	**9**	**0.425**	**–0.19**
9.7	**10**	**0.475**	**–0.06**
9.7	**11**	**0.525**	**0.06**
9.9	**12**	**0.575**	**0.19**
10.0	**13**	**0.625**	**0.32**
10.3	**14**	**0.675**	**0.45**
10.5	**15**	**0.725**	**0.60**
10.8	**16**	**0.775**	**0.76**
10.9	**17**	**0.825**	**0.93**
11.2	**18**	**0.875**	**1.15**
11.4	**19**	**0.925**	**1.44**


Normal Probability
Z Scores
2
1
0
-1
-2
7
8
9
10
11
12
Measured Data

- The closer the points are to forming a single line, the more normal your data are;
- the more scattered the points are, the less normal your data are.