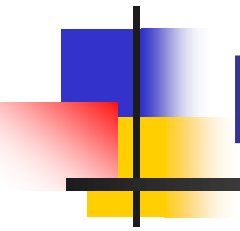
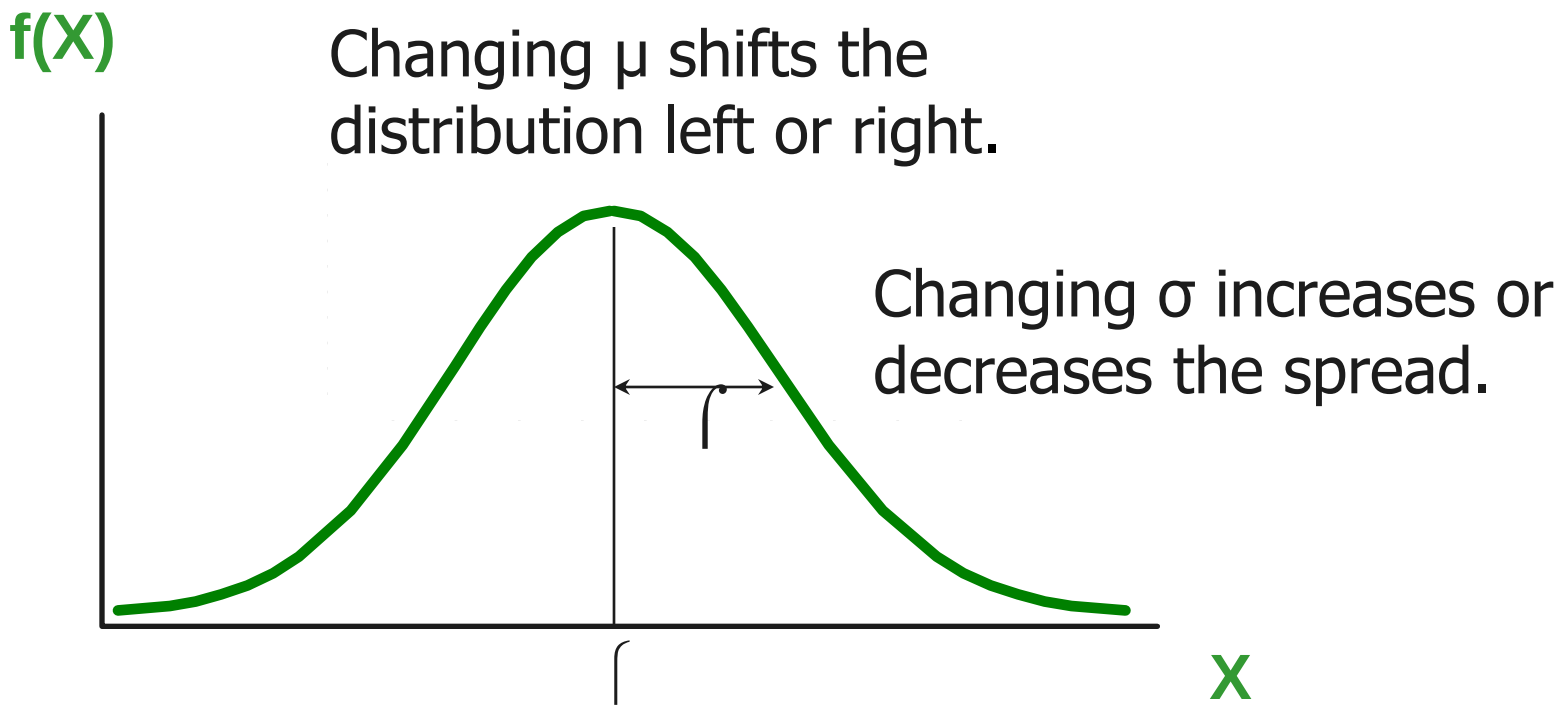


Examples of continuous probability distributions:

The normal and standard normal

The Normal Distribution



The Normal Distribution: as mathematical function (pdf)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

This is a bell shaped curve with different centers and spreads depending on μ and σ

Note constants:

π=3.14159

e=2.71828

The Normal PDF

It's a probability function, so no matter what the values of μ and σ, must integrate to 1!

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

Normal distribution is defined by its mean and standard dev.

$E(X)=\mu = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$

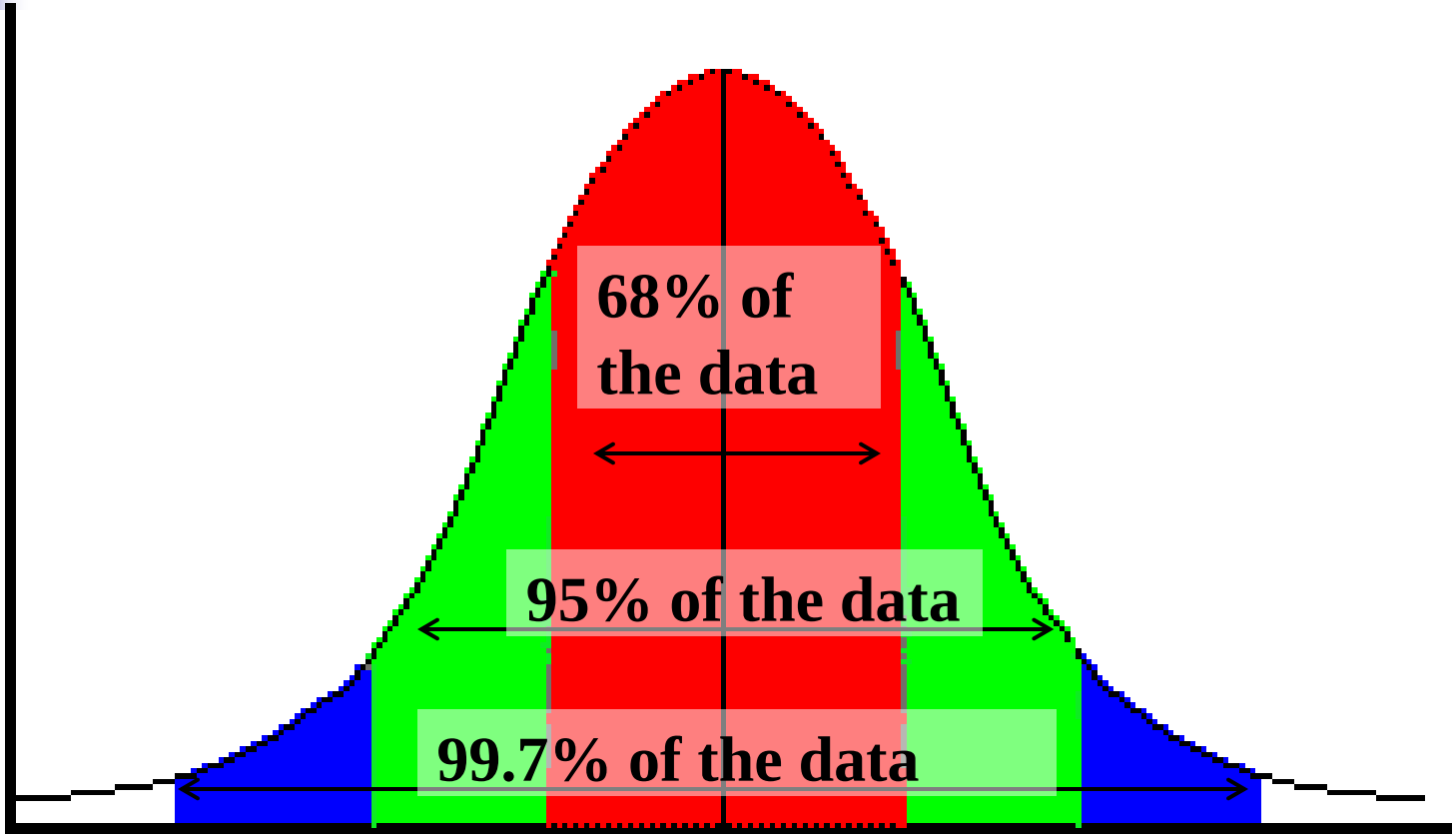
$Var(X)=\sigma^2 = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx) - \mu^2$

Standard Deviation(X)=σ

**The beauty of the normal curve:

No matter what μ and σ are, the area between $\mu-\sigma$ and $\mu+\sigma$ is about 68%; the area between $\mu-2\sigma$ and $\mu+2\sigma$ is about 95%; and the area between $\mu-3\sigma$ and $\mu+3\sigma$ is about 99.7%. Almost all values fall within 3 standard deviations.

68-95-99.7 Rule



68-95-99.7 Rule in Math terms…

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .997$$

How good is rule for real data?

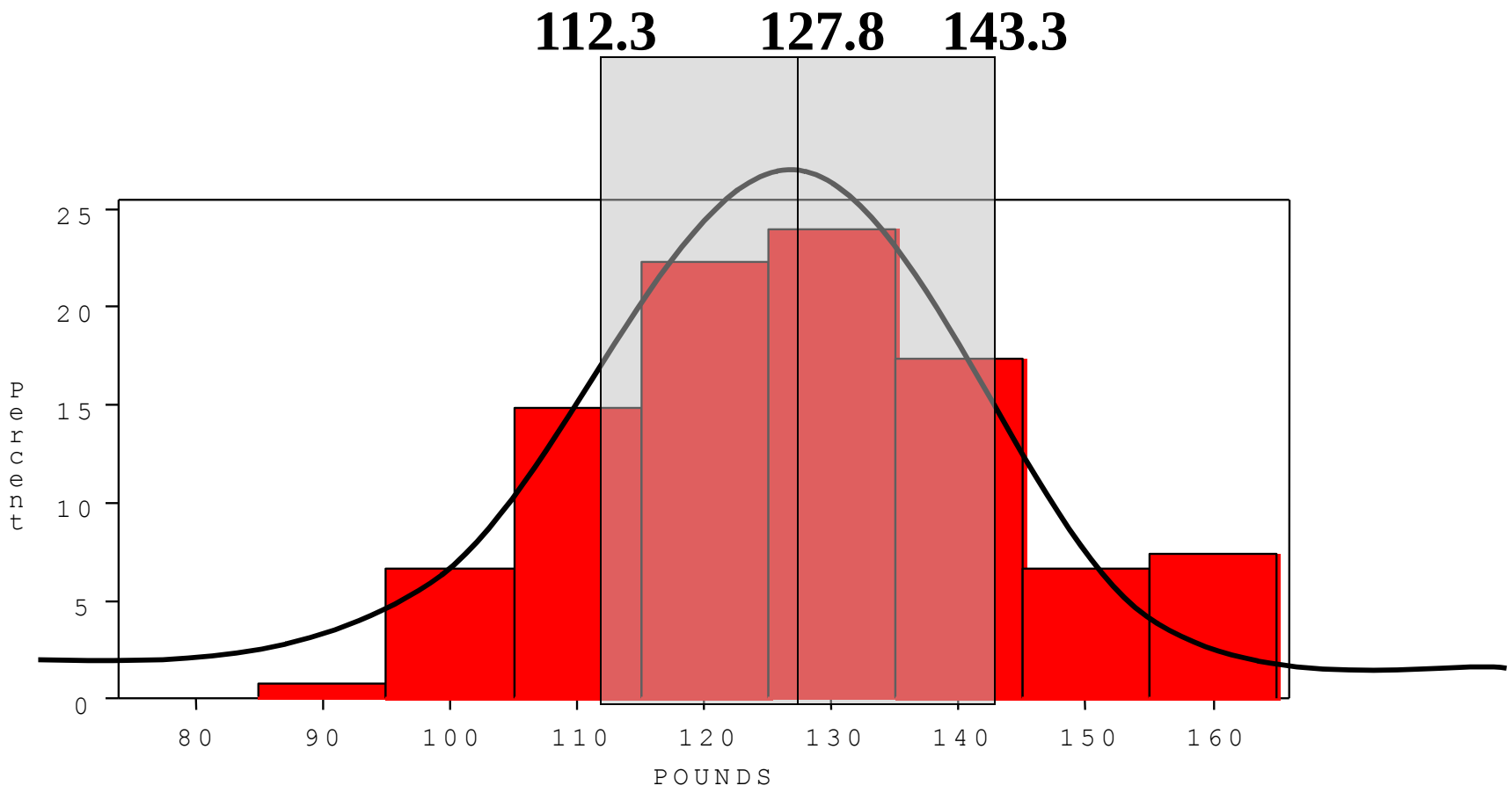
Check some example data:

The mean of the weight of the women = 127.8

The standard deviation (SD) = 15.5

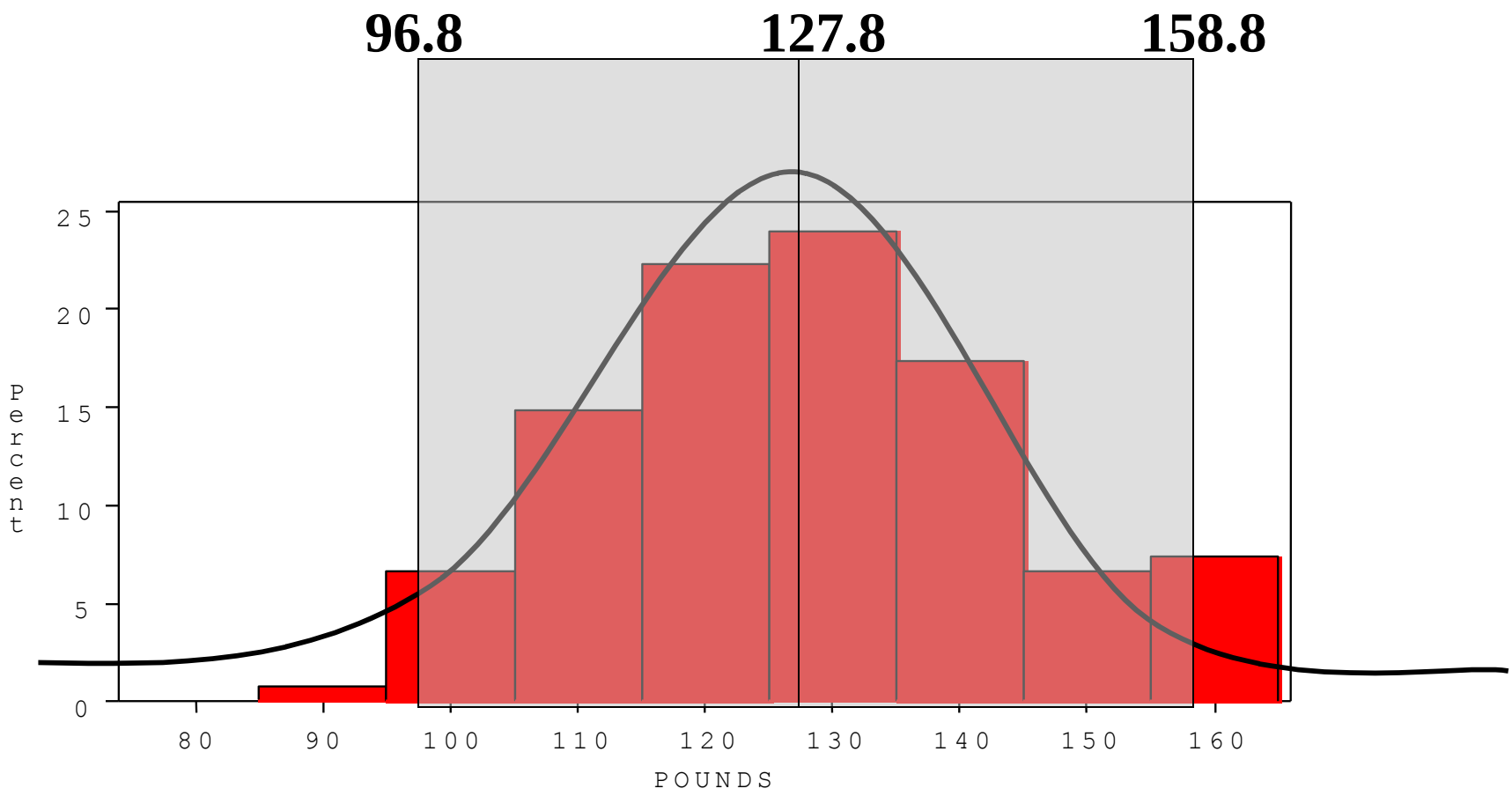
68% of 120 = .68x120 = ~ 82 runners

In fact, 79 runners fall within 1-SD (15.5 lbs) of the mean.



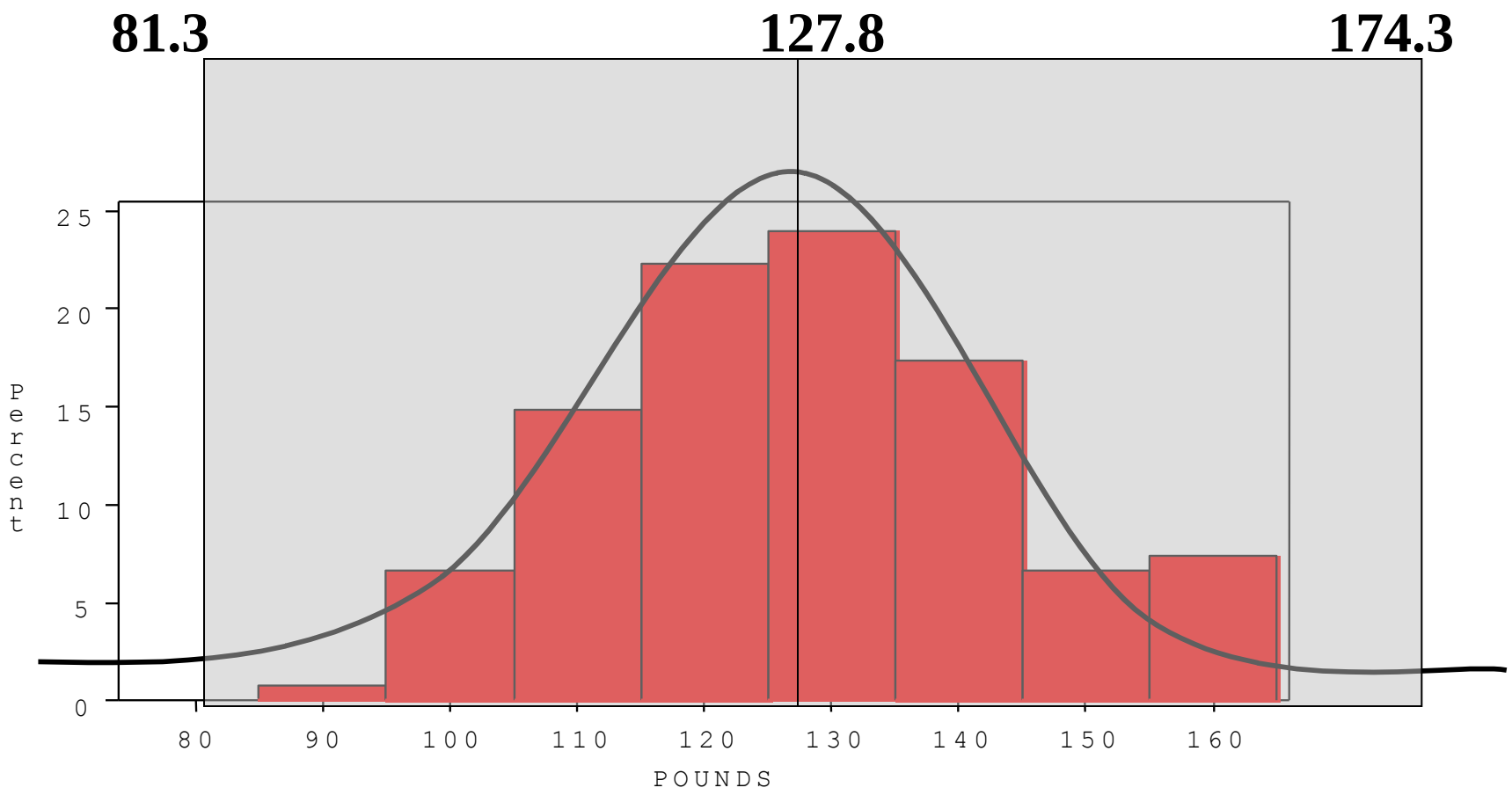
95% of 120 = .95 x 120 = ~ 114 runners

In fact, 115 runners fall within 2-SD's of the mean.



99.7% of 120 = .997 x 120 = 119.6 runners

In fact, all 120 runners fall within 3-SD's of the mean.



Example

- Suppose SAT scores roughly follows a normal distribution in the U.S. population of college-bound students (with range restricted to 200-800), and the average math SAT is 500 with a standard deviation of 50, then:
  - 68% of students will have scores between 450 and 550
  - 95% will be between 400 and 600
  - 99.7% will be between 350 and 650

Example

- BUT…
- What if you wanted to know the math SAT score corresponding to the 90th percentile (=90% of students are lower)?

$P(X \leq Q) = .90 \rightarrow$

$$\int_{200}^{Q} \frac{1}{(50)\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-500}{50}\right)^2} dx = .90$$

Solve for Q?….Yikes!

The Standard Normal (Z): "Universal Currency"

The formula for the standardized normal probability density function is

$$p(Z) = \frac{1}{(1)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{Z-0}{1}\right)^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(Z)^2}$$

The Standard Normal Distribution (Z)

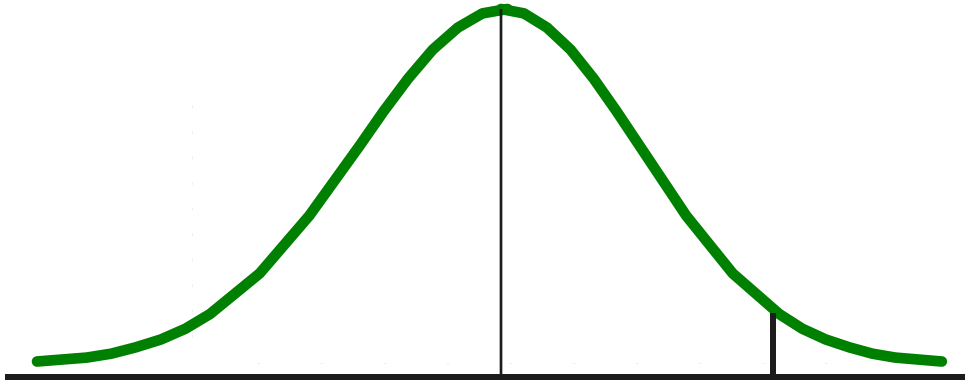
All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

Somebody calculated all the integrals for the standard normal and put them in a table! So we never have to integrate!

Even better, computers now do all the integration.

Comparing X and Z units



100	**200**	**X**	($\mu = 100, \sigma = 50$)
0	**2.0**	**Z**	($\mu = 0, \sigma = 1$)

Example

- For example: What's the probability of getting a math SAT score of 575 or less, μ=500 and σ=50?

$$Z = \frac{575 - 500}{50} = 1.5$$

- i.e., A score of 575 is 1.5 standard deviations above the mean

$$\therefore P(X \leq 575) = \int_{200}^{575} \frac{1}{(50)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-500}{50}\right)^2} dx \longrightarrow \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}Z^2} dz$$

Yikes!

But to look up Z= 1.5 in standard normal chart (or enter into SAS)→ no problem! = .9332

Practice problem

If birth weights in a population are normally distributed with a mean of 109 oz and a standard deviation of 13 oz,

a. What is the chance of obtaining a birth weight of 141 oz *or heavier* when sampling birth records at random?

b. What is the chance of obtaining a birth weight of 120 *or lighter*?

Answer

a. What is the chance of obtaining a birth weight of 141 oz *or heavier* when sampling birth records at random?

$$Z = \frac{141 - 109}{13} = 2.46$$

From the chart or SAS → Z of 2.46 corresponds to a right tail (greater than) area of: $P(Z \geq 2.46) = 1-(.9931) = .0069$ or .69 %

Answer

b. What is the chance of obtaining a birth weight of 120 *or lighter*?

$$Z = \frac{120 - 109}{13} = .85$$

From the chart or SAS → Z of .85 corresponds to a left tail area of:
$P(Z \leq .85) = .8023 = 80.23\%$

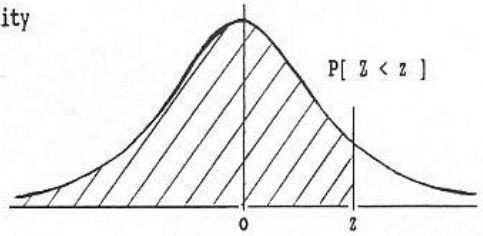
Looking up probabilities in the standard normal table

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}Z^2)\ dZ$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Z=1.51

Z=1.51

What is the area to the left of Z=1.51 in a standard normal curve?

Area is 93.45%

Normal probabilities in SAS

```
data _null_;
  theArea=probnorm(1.5);
  put theArea;
run;
0.9331927987
```

The "probnorm(Z)" function gives you the probability from negative infinity to Z (here 1.5) in a standard normal curve.

And if you wanted to go the other direction (i.e., from the area to the Z score (called the so-called "Probit" function →

```
data _null_;
  theZValue=probit(.93);
  put theZValue;
run;
1.4757910282
```

The "probit(p)" function gives you the Z-value that corresponds to a left-tail area of p (here .93) from a standard normal curve. The probit function is also known as the inverse standard normal function.

Probit function: the inverse

ϕ(area)= Z: gives the Z-value that goes with the probability you want

For example, recall SAT math scores example. What's the score that corresponds to the 90th percentile?

In Table, find the Z-value that corresponds to area of .90 → Z= 1.28

Or use SAS

```
data _null_;
        theZValue=probit(.90);
        put theZValue;
run;
```

1.2815515655

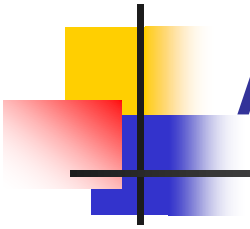
If Z=1.28, convert back to raw SAT score →

$1.28 = \frac{X - 500}{50}$

X – 500 =1.28 (50)

X=1.28(50) + 500 = 564 (1.28 standard deviations above the mean!)

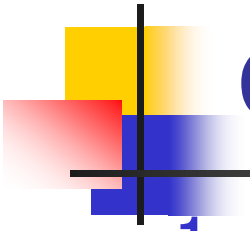
`



Are my data “normal”?

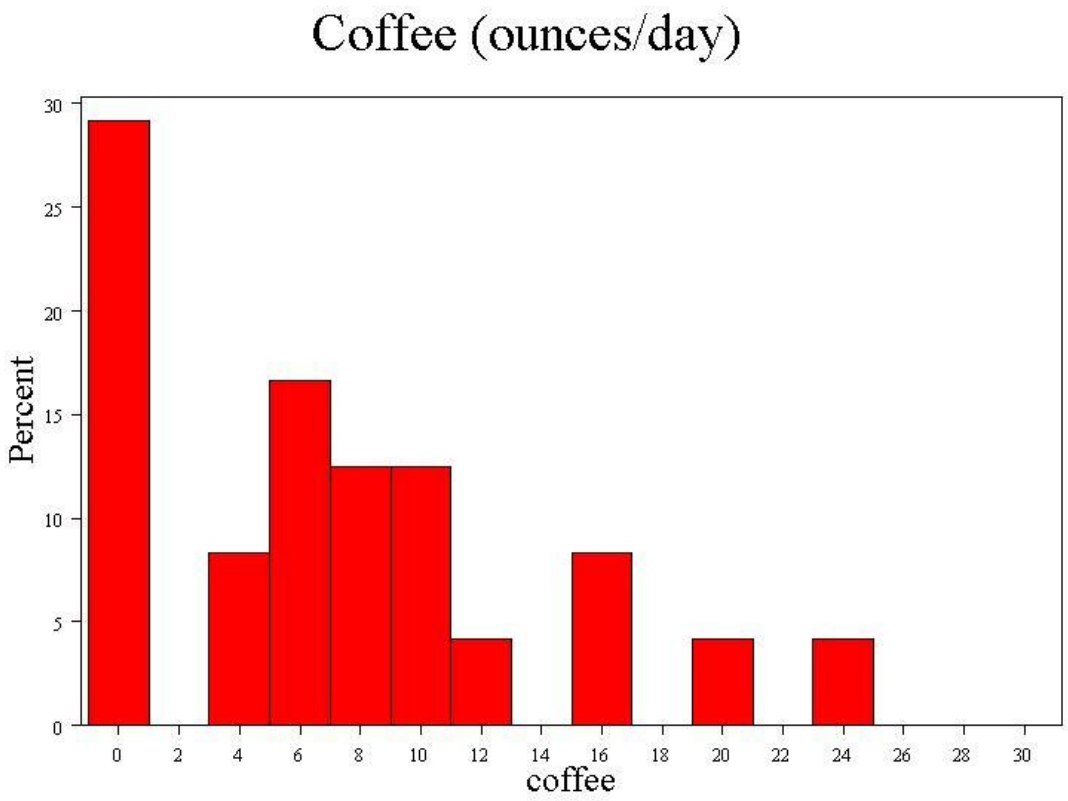
- Not all continuous random variables are normally distributed!!
- It is important to evaluate how well the data are approximated by a normal distribution

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!

Data from our class…



Median = 6

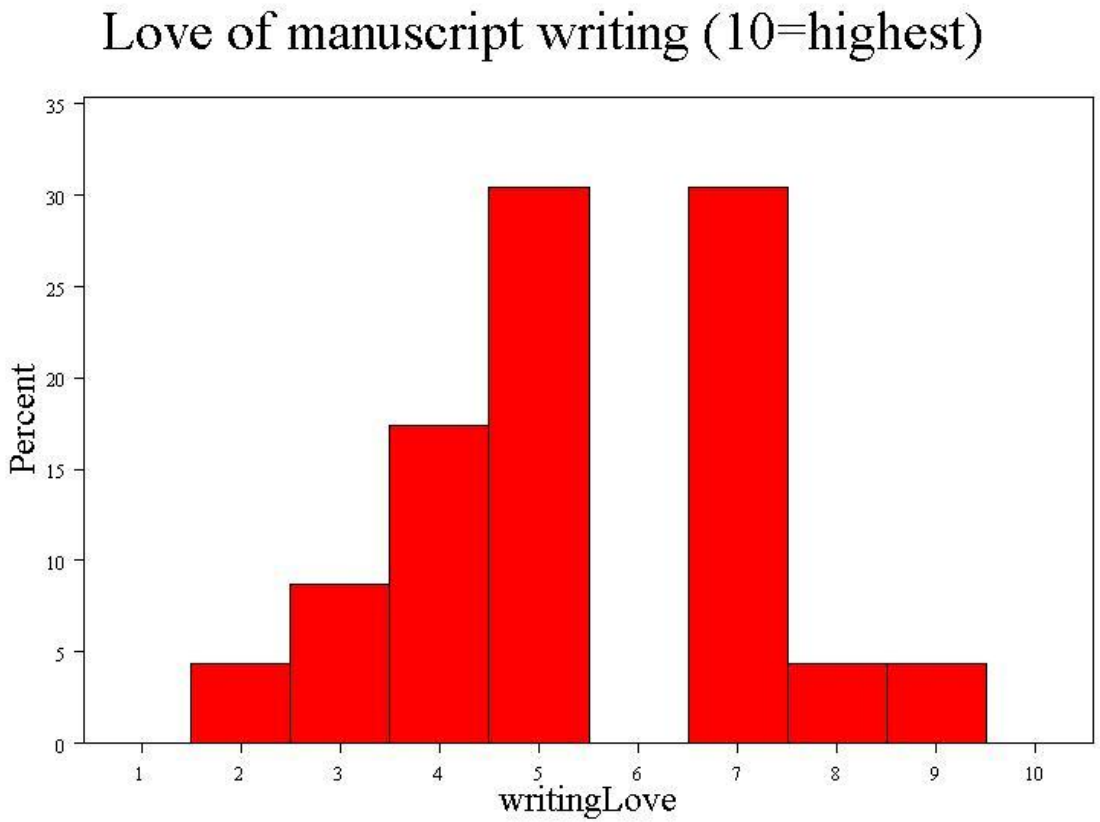
Mean = 7.1

Mode = 0

SD = 6.8

Range = 0 to 24
(= 3.5 ∫)

Data from our class…



Median = 5

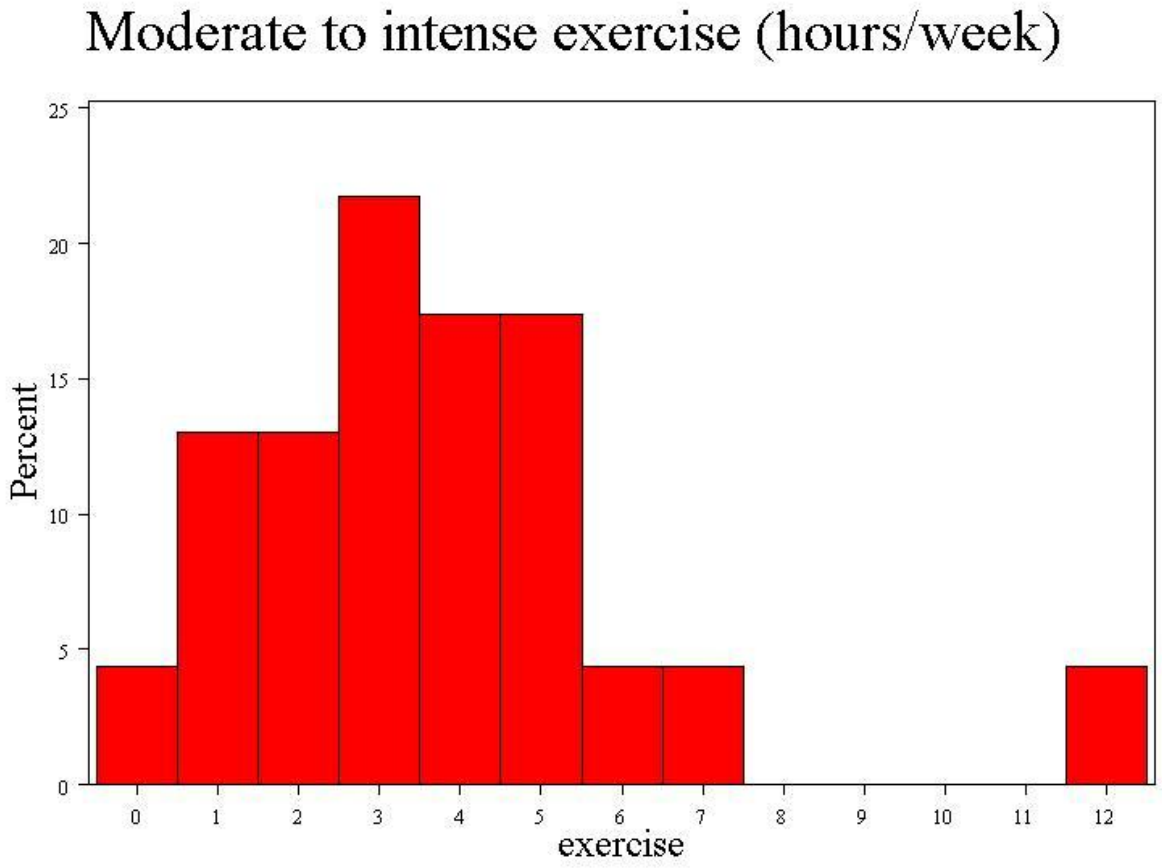
Mean = 5.4

Mode = none

SD = 1.8

Range = 2 to 9
(~ 4 ∫)

Data from our class…



Median = 3

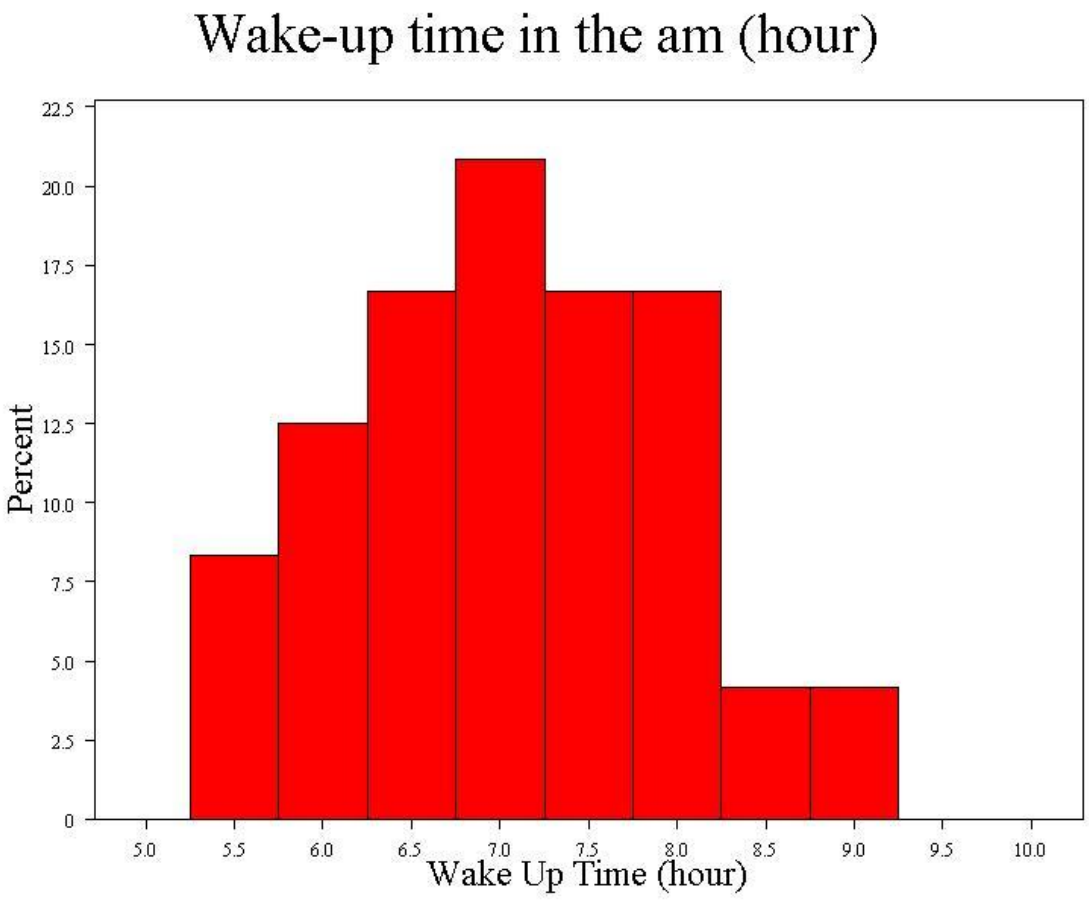
Mean = 3.4

Mode = 3

SD = 2.5

Range = 0 to 12
(~ 5 ∫)

Data from our class…



Median = 7:00

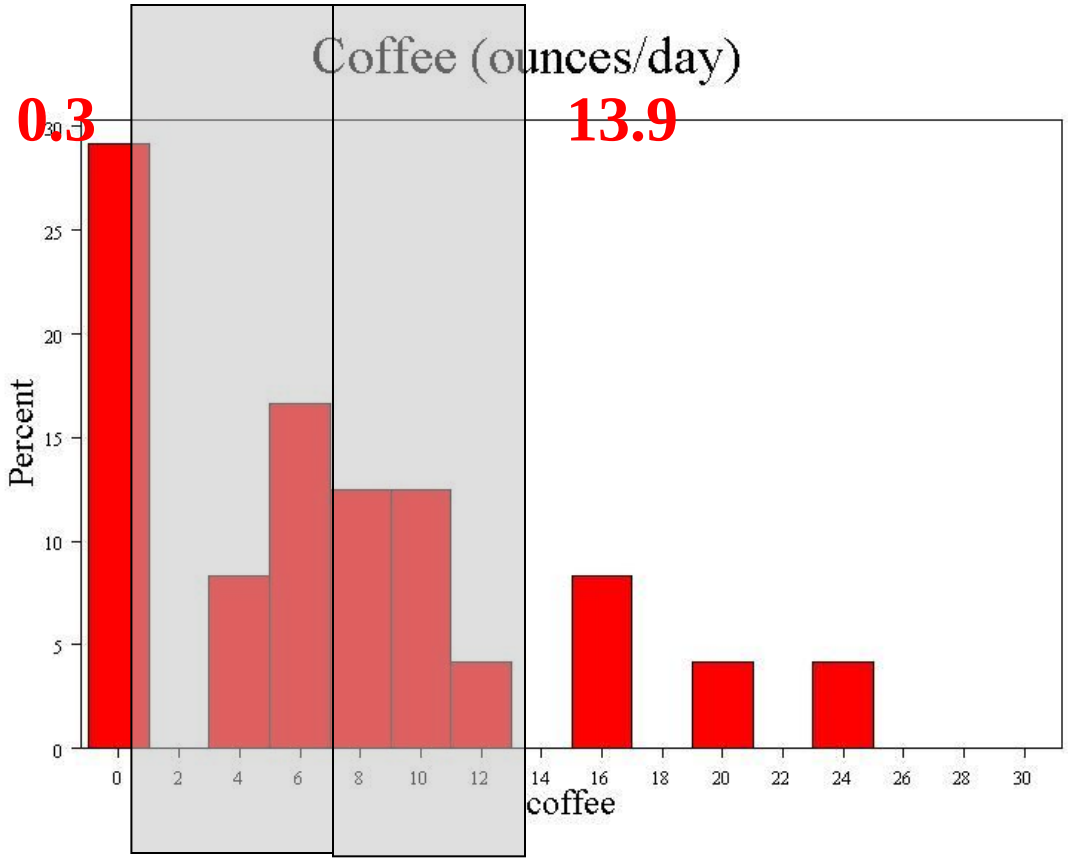
Mean = 7:04

Mode = 7:00

SD = :55

Range = 5:30 to 9:00 (~4 ∫)

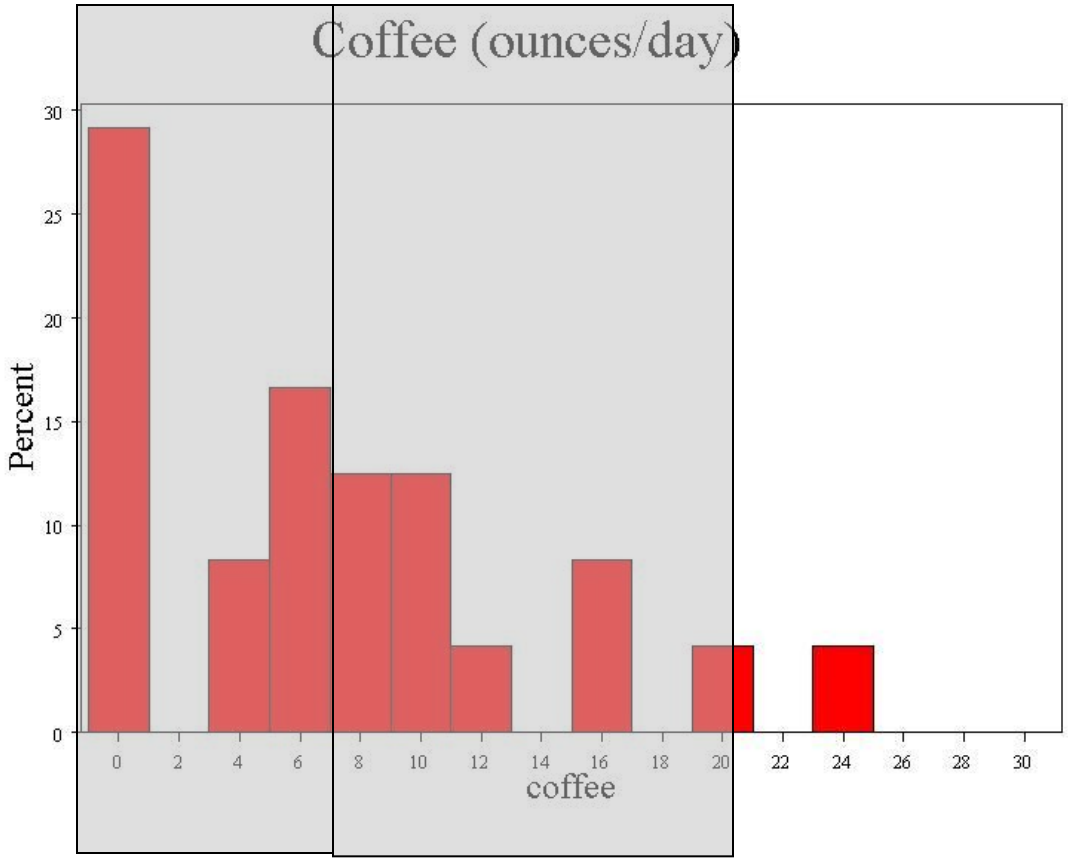
Data from our class…



7.1 +/- 6.8 =

0.3 – 13.9

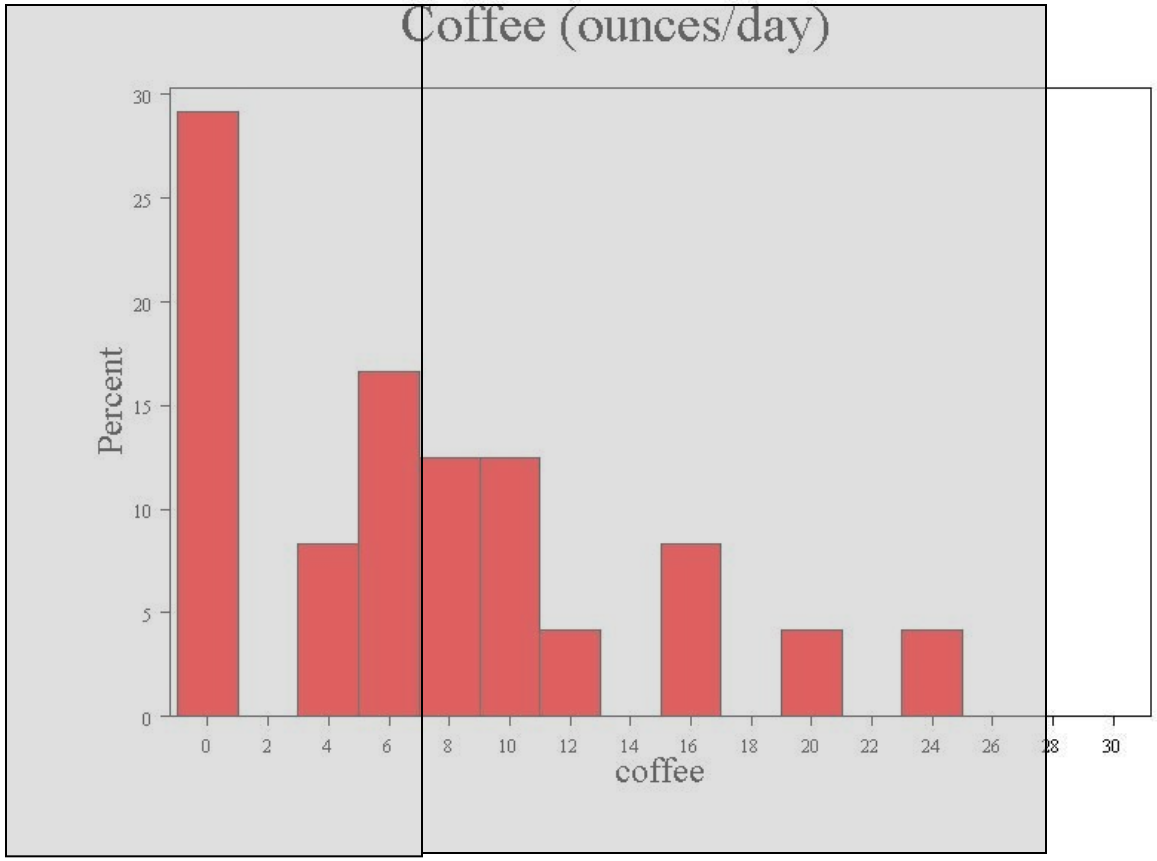
Data from our class…



7.1 +/- 2*6.8 =

0 – 20.7

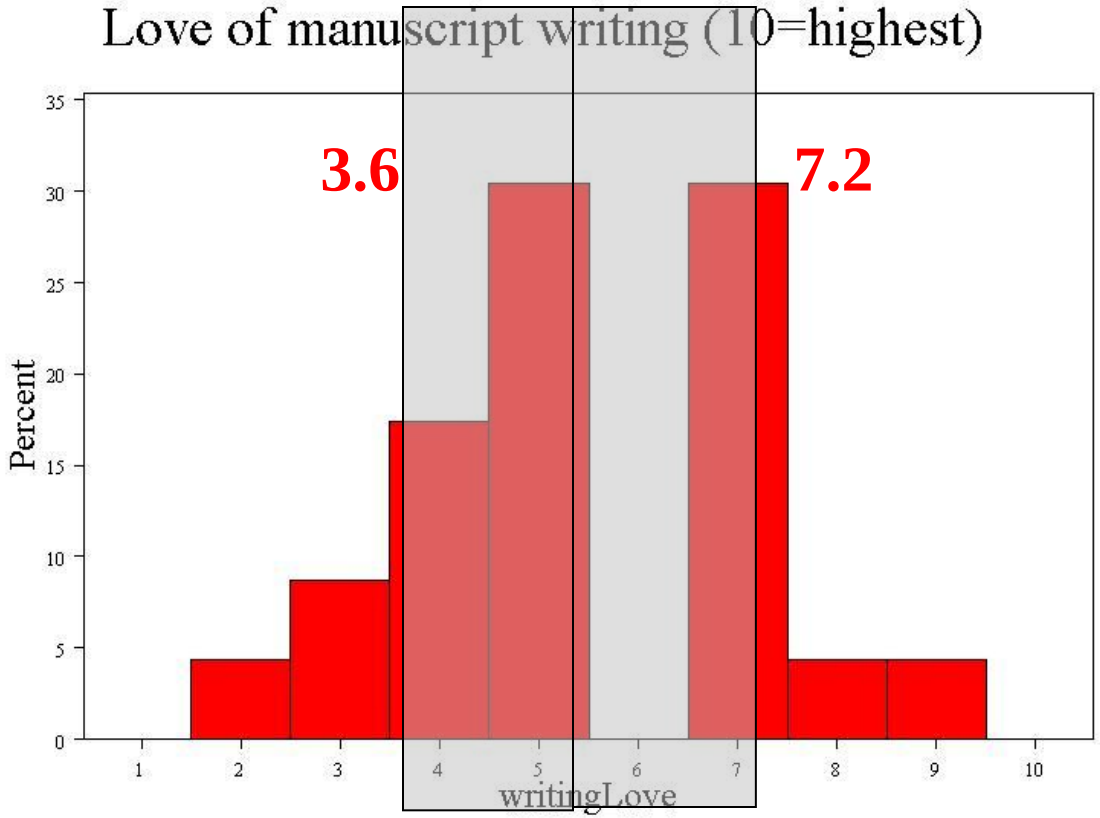
Data from our class…



7.1 +/- 3*6.8 =

0 – 27.5

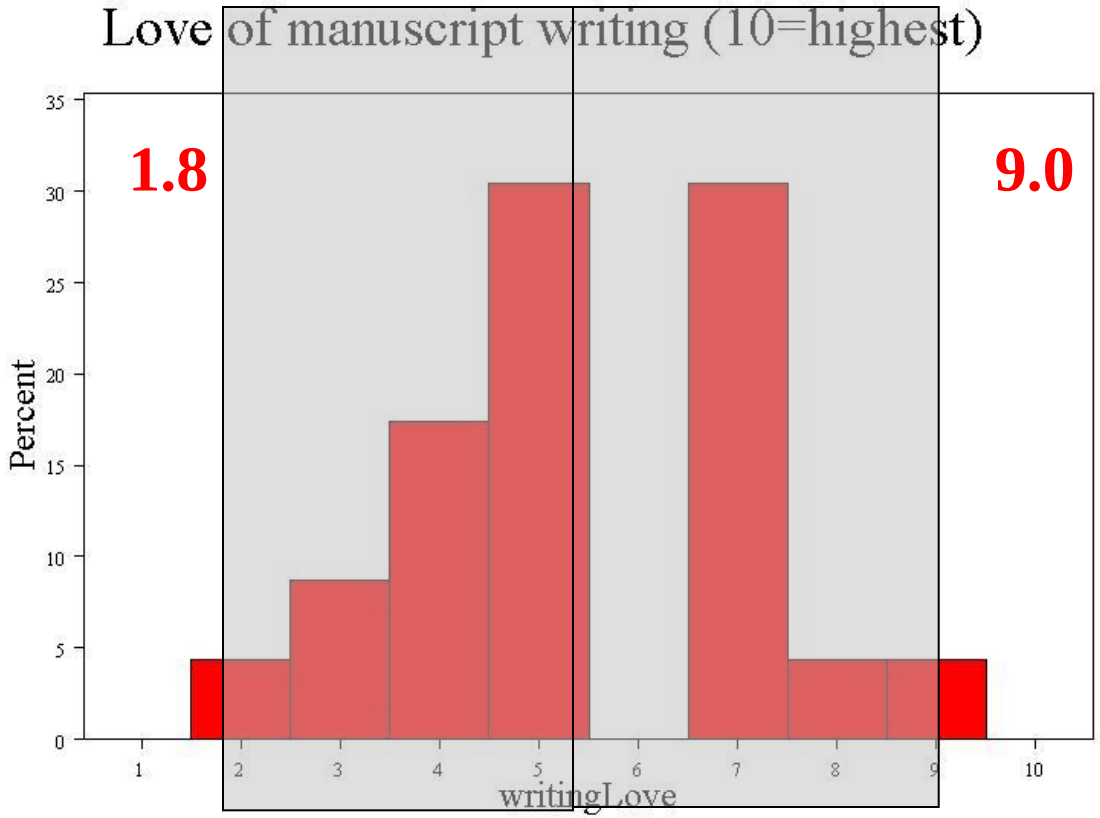
Data from our class…



5.4 +/- 1.8 =

3.6 – 7.2

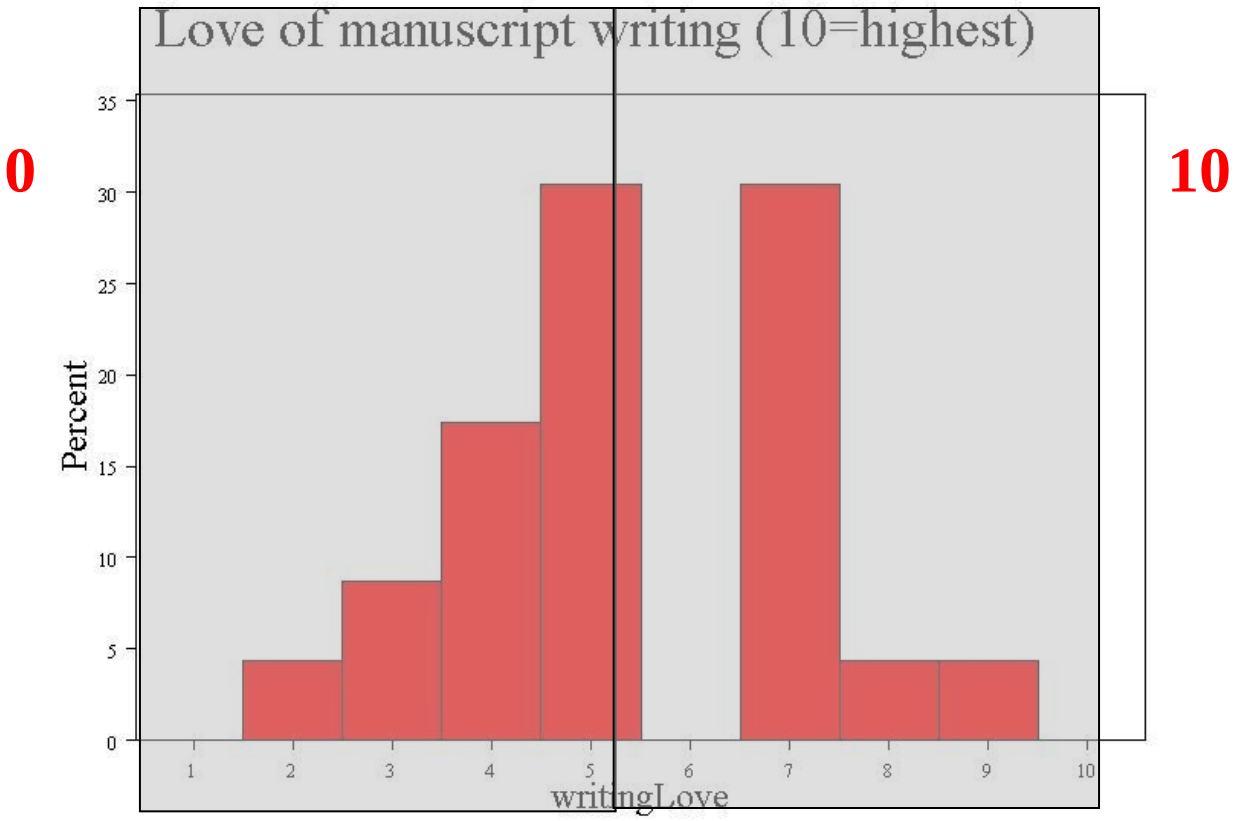
Data from our class…



5.4 +/- 2*1.8 =

1.8 – 9.0

Data from our class…



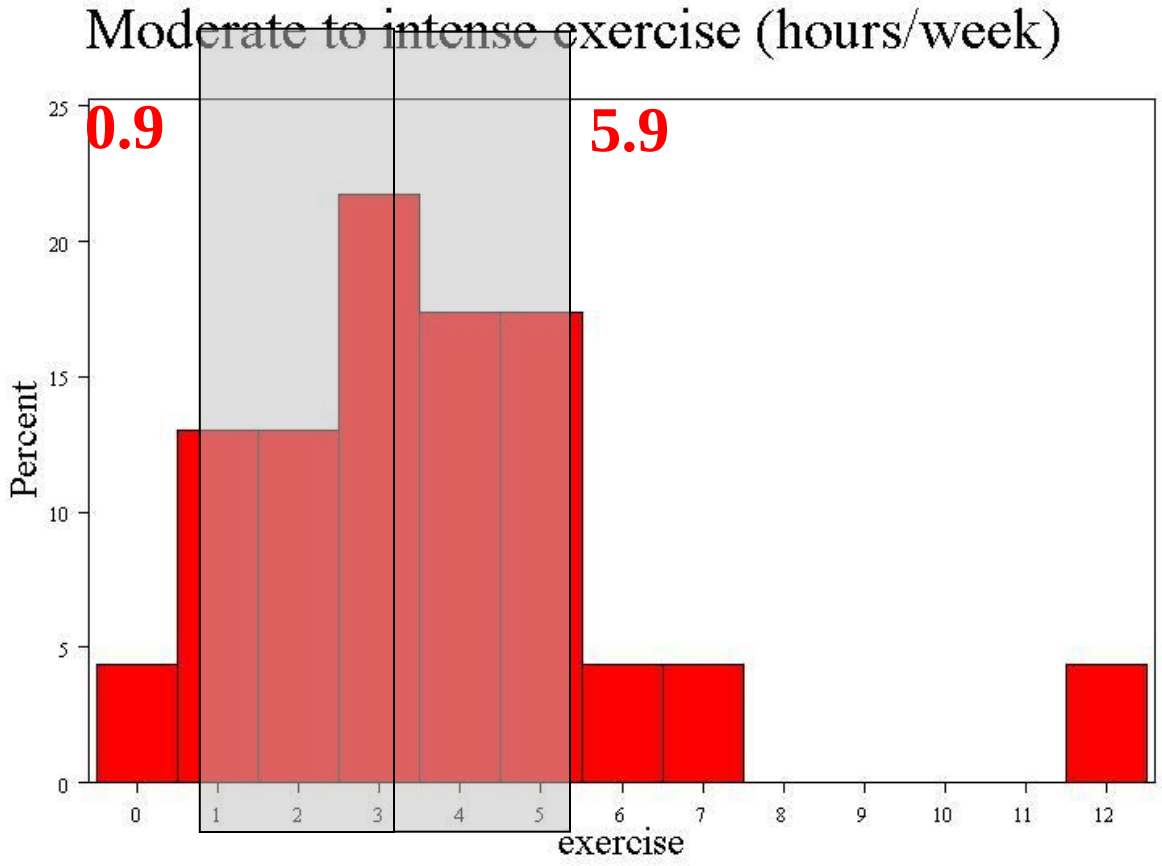
0

10

5.4 +/- 3*1.8 =

0– 10

Data from our class…

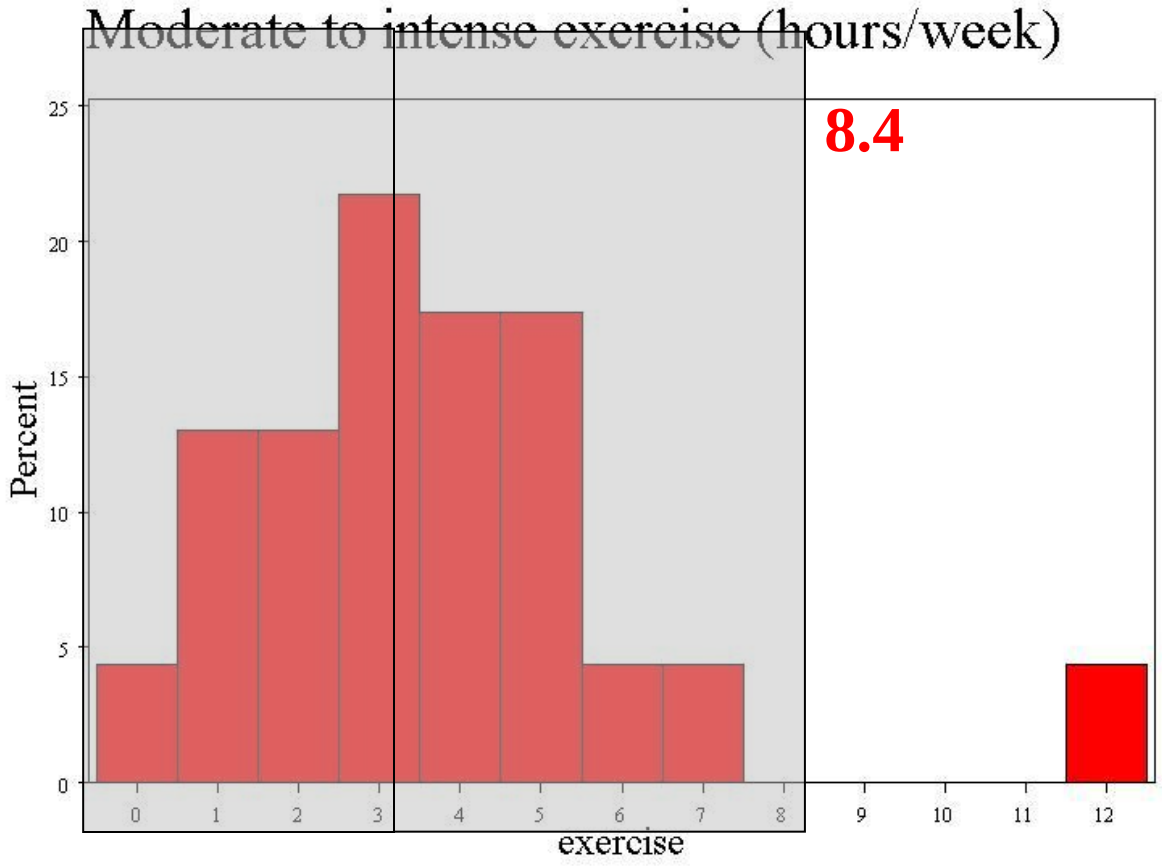


3.4 +/- 2.5=

0.9 – 7.9

Data from our class…

0



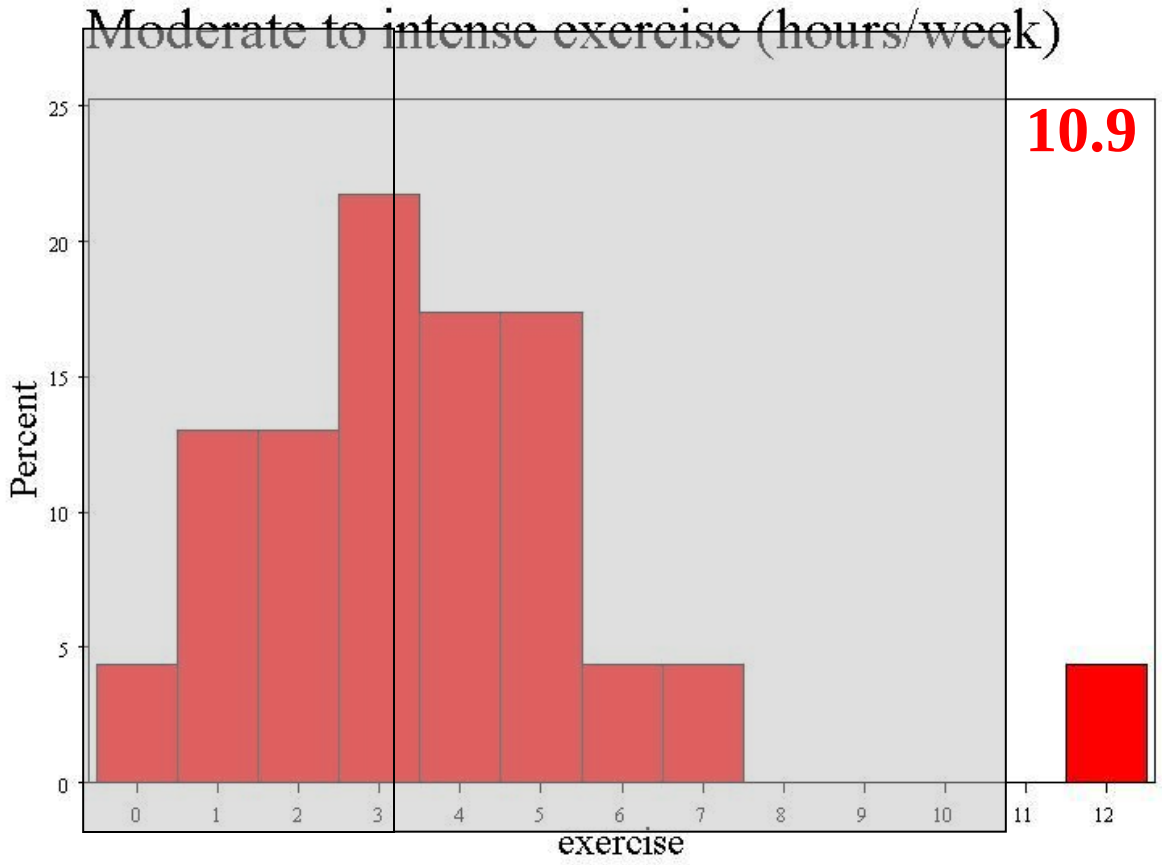
8.4

3.4 +/- 2*2.5=

0 – 8.4

Data from our class…

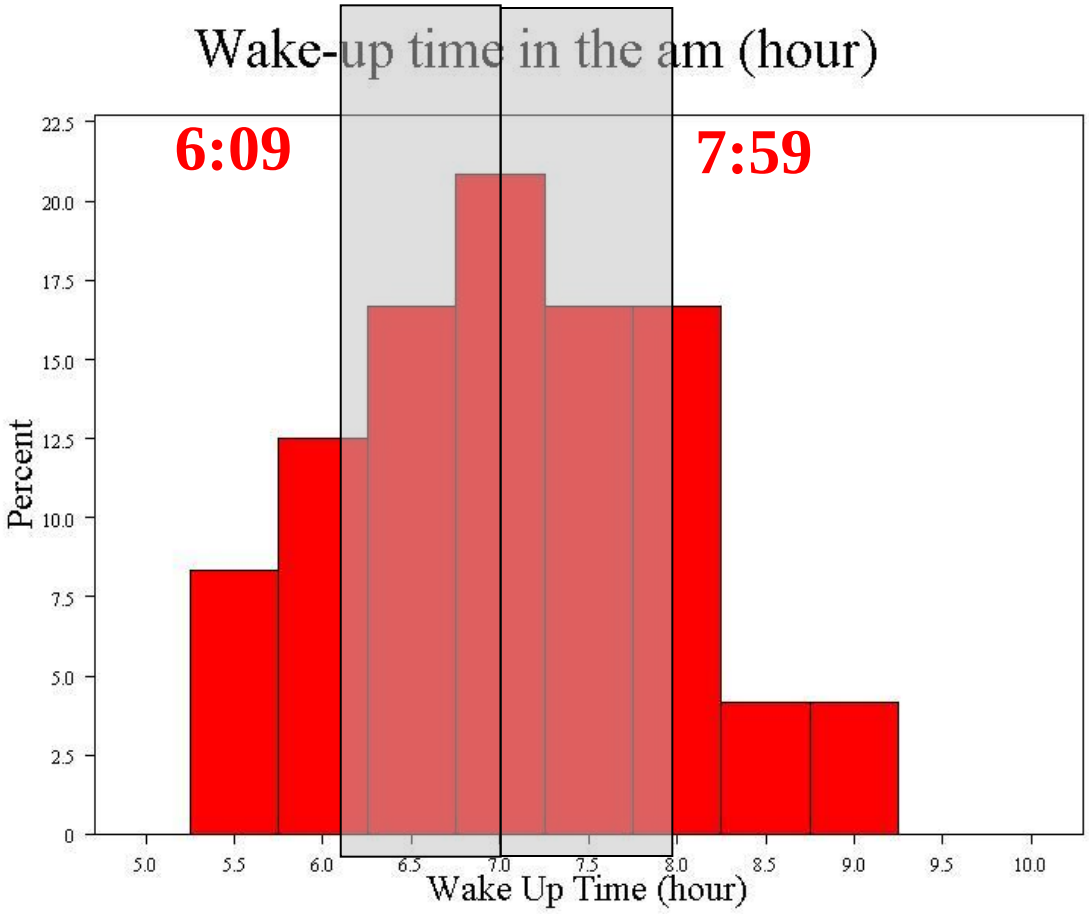
0



3.4 +/- 3*2.5=

0 – 10.9

Data from our class…

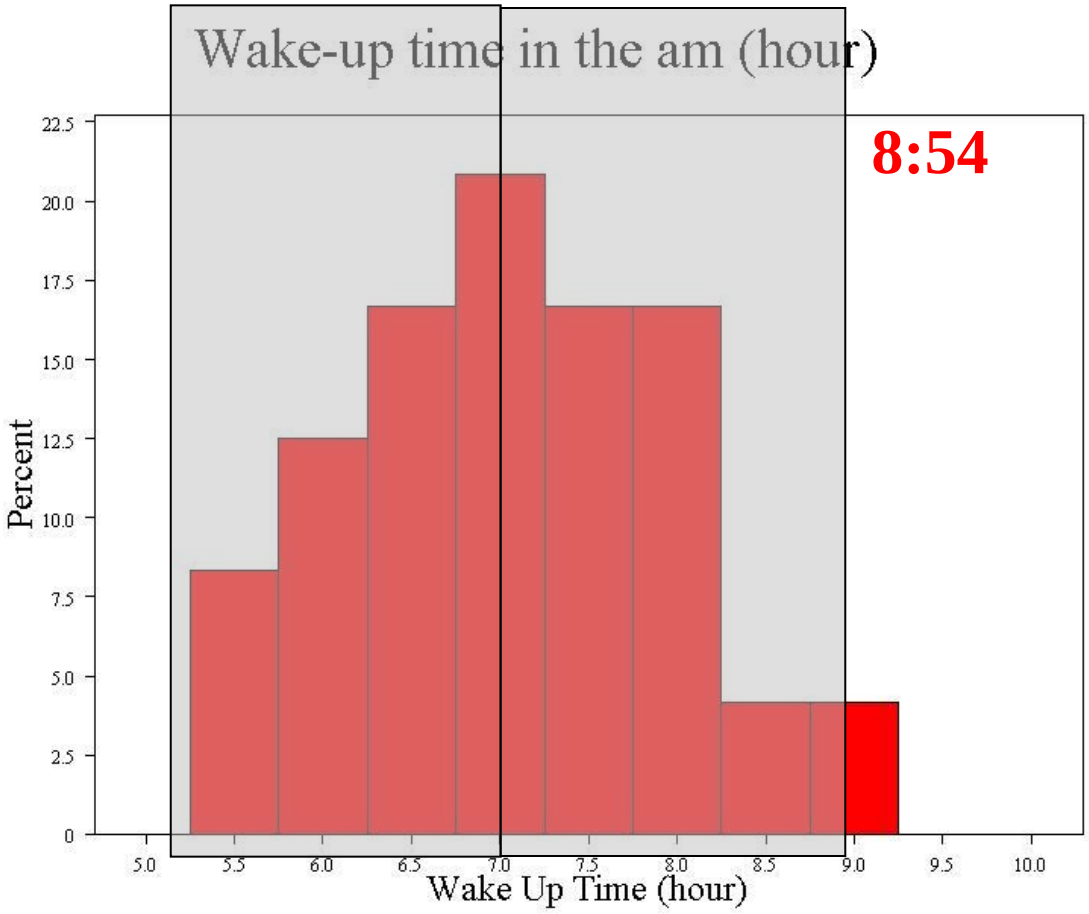


7:04+/- 0:55 =

6:09 – 7:59

Data from our class…

5:14

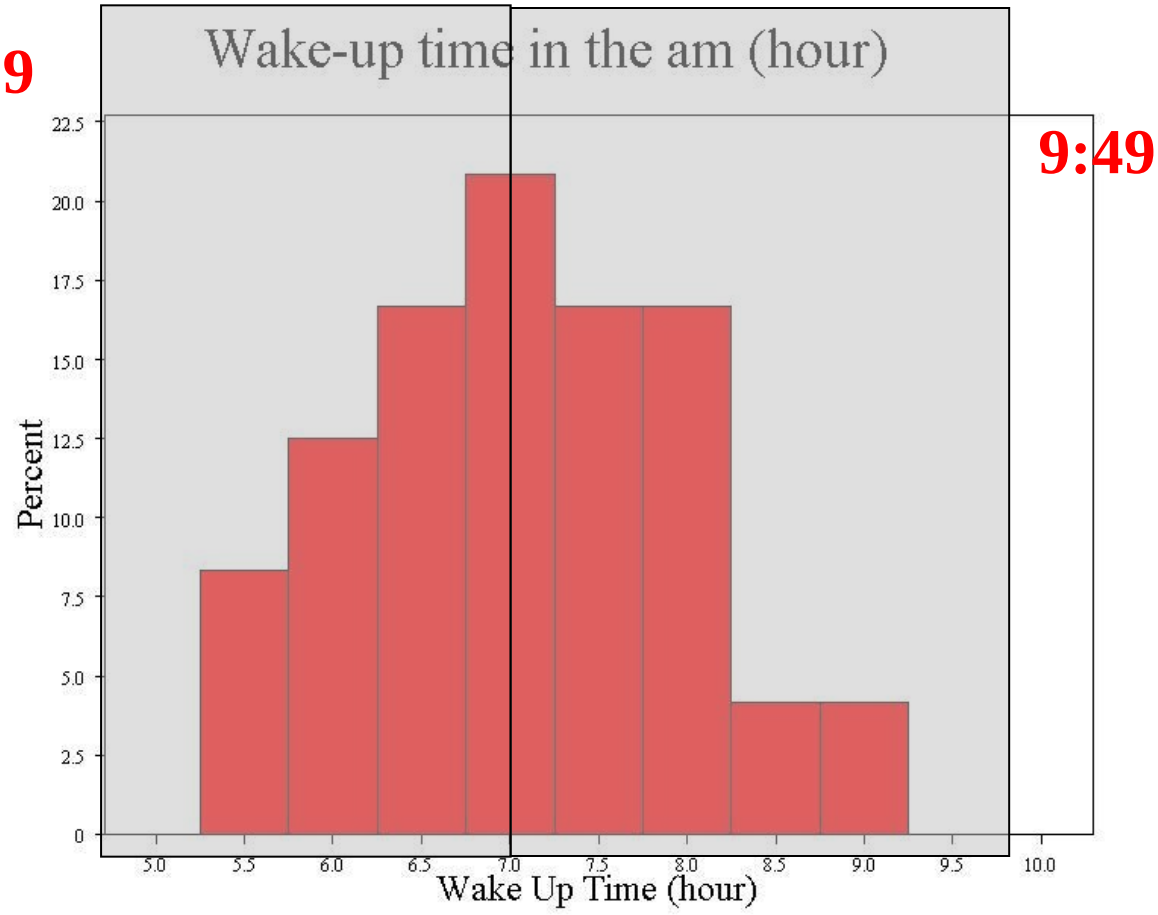


7:04+/- 2*0:55
=

5:14 – 8:54

Data from our class…

4:19



9:49

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

4:19 – 9:49

The Normal Probability Plot

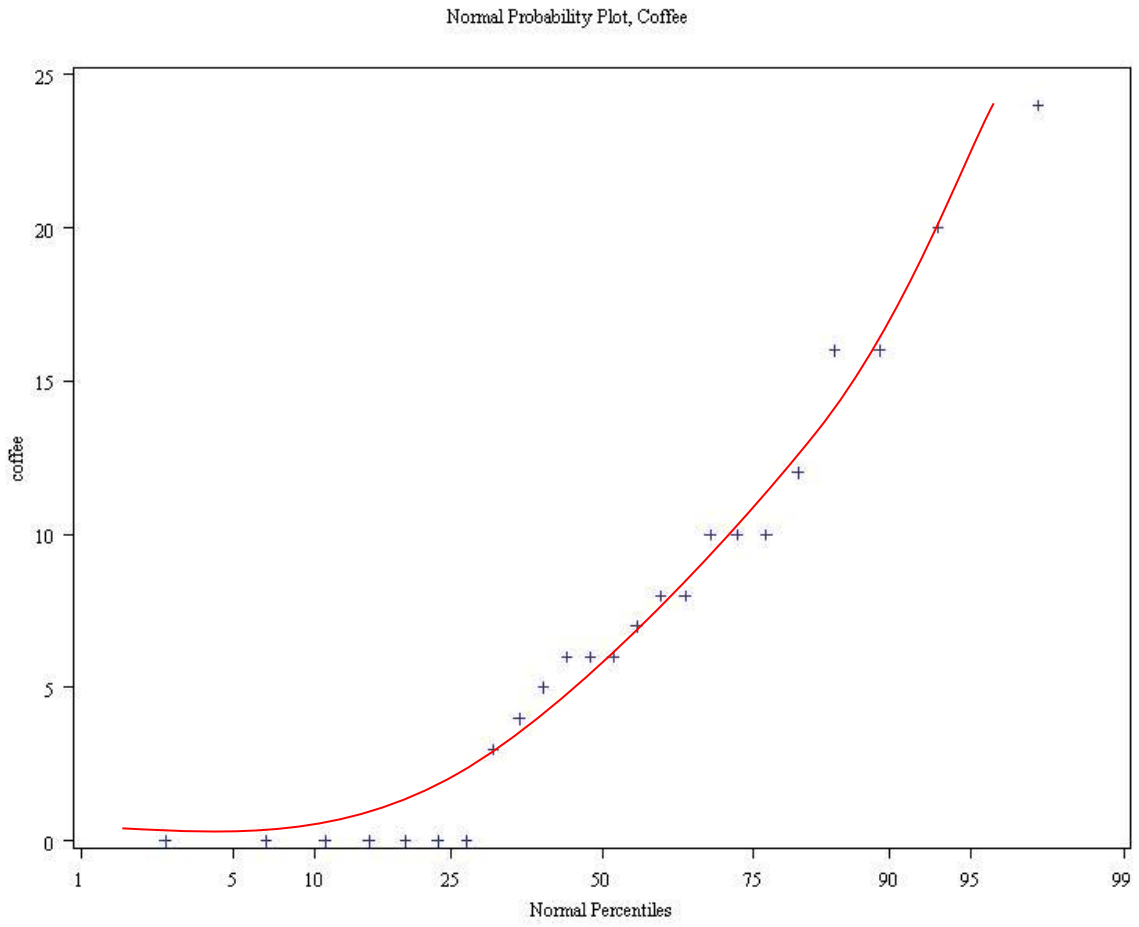
- Normal probability plot
  - Order the data.
  - Find corresponding standardized normal quantile values:

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

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.

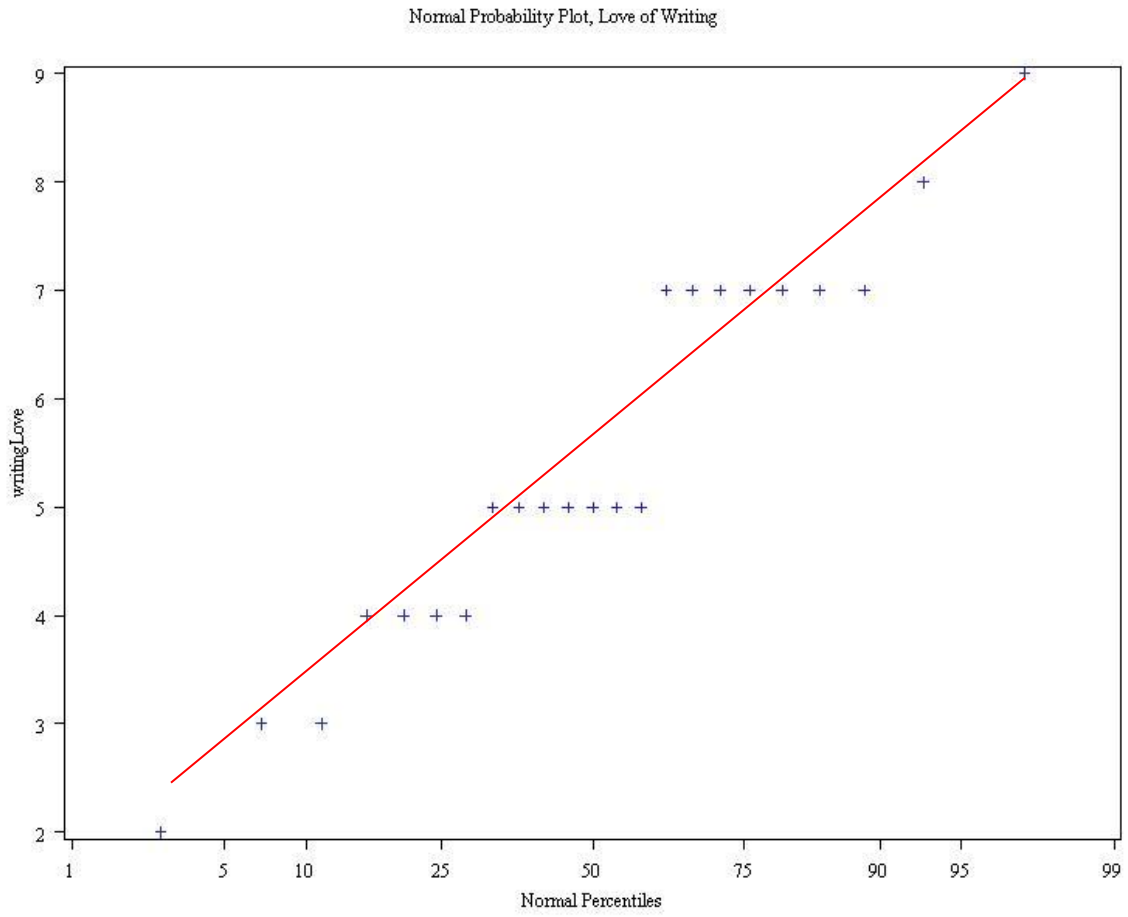
Normal probability plot coffee…



Right-Skewed!

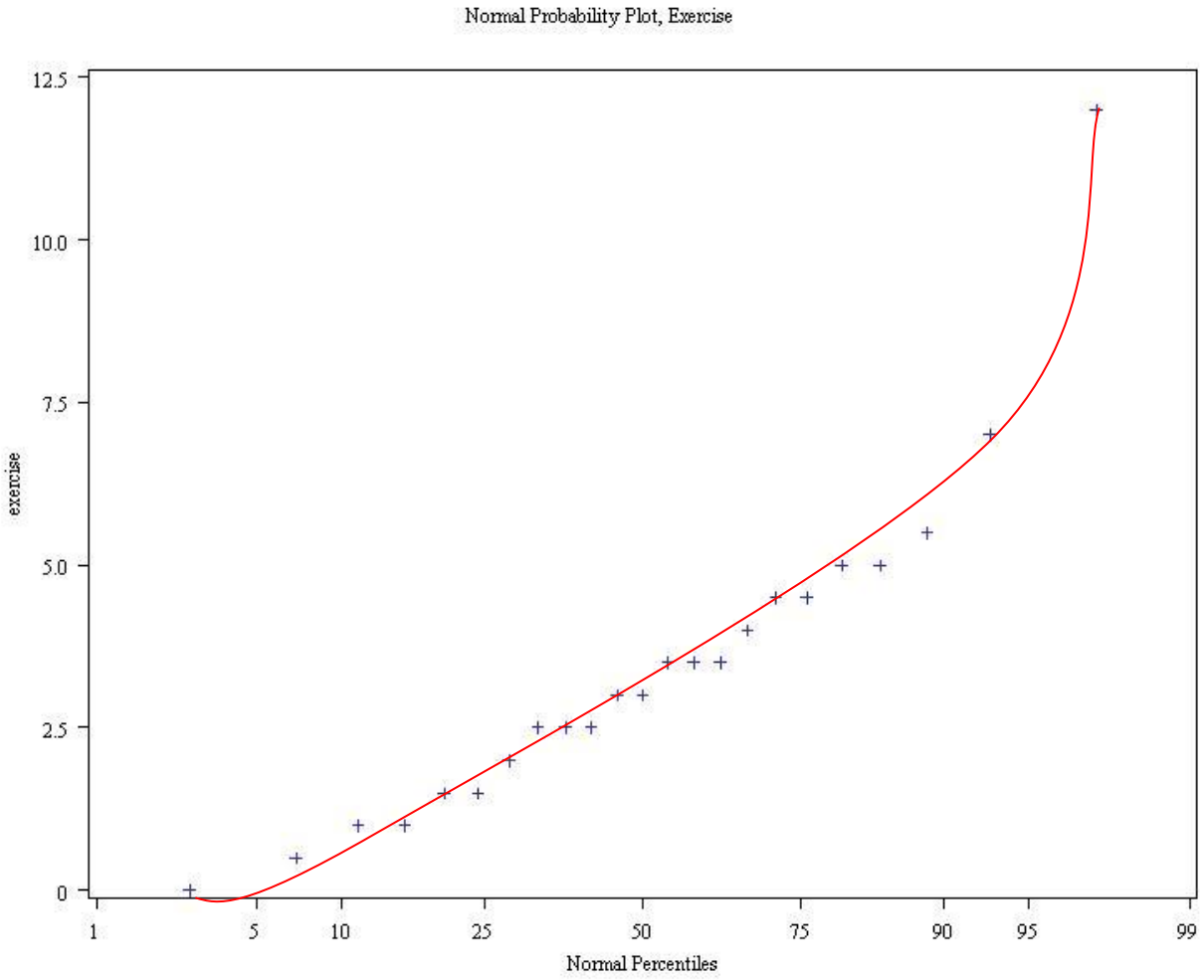
(concave up)

Normal probability plot love of writing…



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

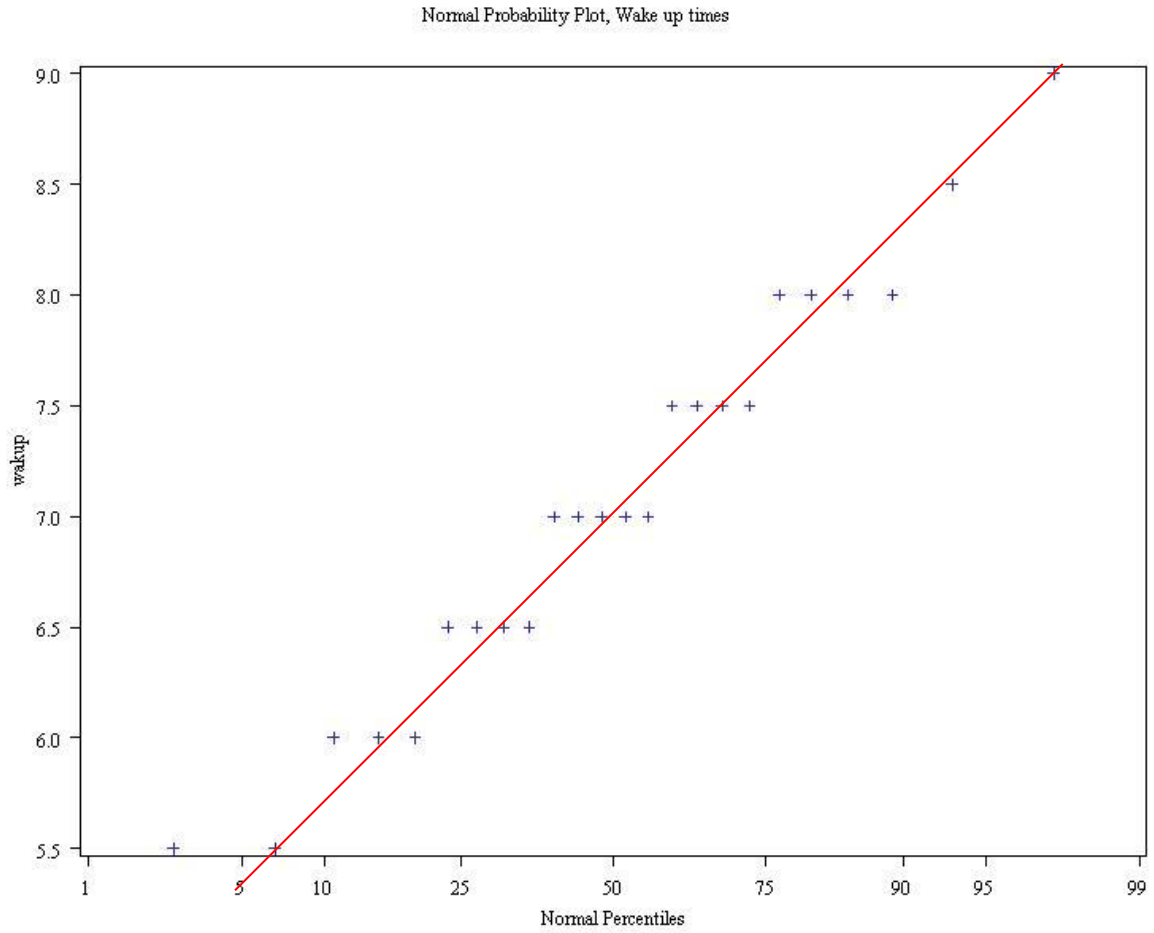
Norm prob. plot Exercise…



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 ($p>.25$)

Normal approximation to the binomial

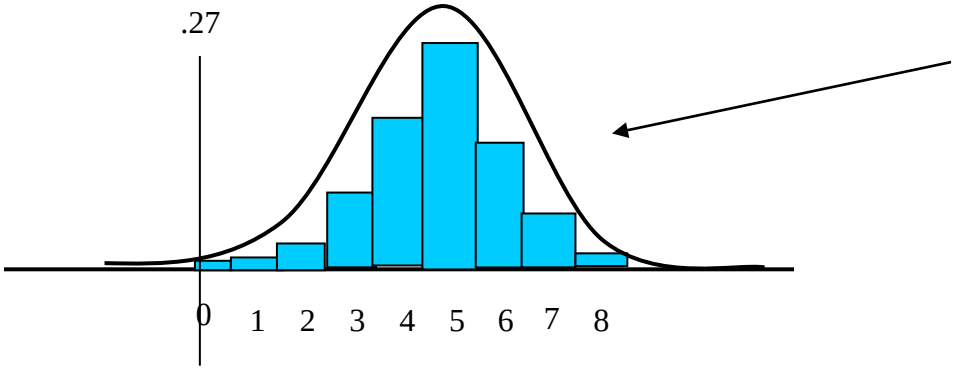
When you have a binomial distribution where *n* is large and *p* is middle-of-the road (not too small, not too big, closer to .5), then the binomial starts to look like a normal distribution→ in fact, this doesn't even take a particularly large *n*→

Recall: What is the probability of being a smoker among a group of cases with lung cancer is .6, what's the probability that in a group of 8 cases you have less than 2 smokers?

Normal approximation to the binomial

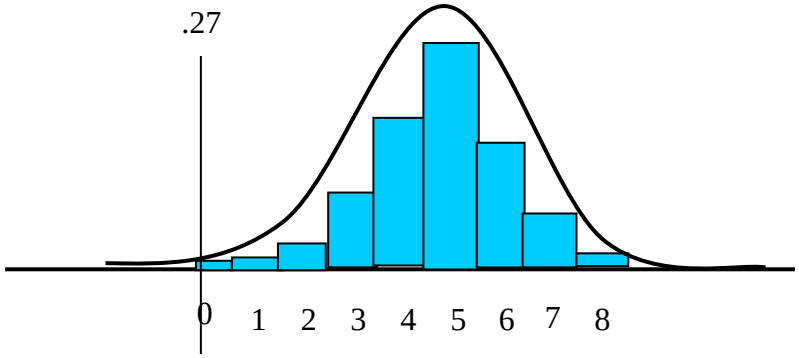
When you have a binomial distribution where *n* is large and p isn't too small (rule of thumb: mean>5), then the binomial starts to look like a normal distribution→

Recall: smoking example…



Starting to have a normal shape even with fairly small *n*. You can imagine that if *n* got larger, the bars would get thinner and thinner and this would look more and more like a continuous function, with a bell curve shape. Here np=4.8.

Normal approximation to binomial



What is the probability of fewer than 2 smokers?

Exact binomial probability (from before) = .00065 + .008 = .00865

Normal approximation probability:

μ=4.8

σ=1.39

$$Z \approx \frac{2-(4.8)}{1.39} = \frac{-2.8}{1.39} = -2$$

P(Z<2)=.022

A little off, but in the right ballpark… we could also use the value to the left of 1.5 (as we really wanted to know less than but not including 2; called the "continuity correction")…

$$Z \approx \frac{1.5-(4.8)}{1.39} = \frac{-3.3}{1.39} = -2.37$$

$P(Z \leq -2.37) = .0069$

A fairly good approximation of the exact probability, .00865.

Practice problem

1. You are performing a cohort study. If the probability of developing disease in the exposed group is .25 for the study duration, then if you sample (randomly) 500 exposed people, What's the probability that **at most** 120 people develop the disease?

Answer

By hand (yikes!):

$$P(X \le 120) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + \ldots + P(X=120) =$$

$$\binom{500}{120}(.25)^{120}(.75)^{380} + \binom{500}{2}(.25)^{2}(.75)^{498} + \binom{500}{1}(.25)^{1}(.75)^{499} + \binom{500}{0}(.25)^{0}(.75)^{500} \ldots$$

OR Use SAS:

```
data _null_;
Cohort=cdf('binomial', 120, .25, 500);
put Cohort;
run;
```

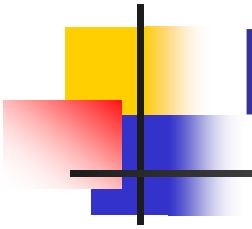
0.323504227

OR use, normal approximation:

$\mu = np = 500(.25) = 125$ and $\sigma^2 = np(1-p) = 93.75$; $\sigma = 9.68$

$$Z = \frac{120 - 125}{9.68} = -.52$$

P(Z<-.52)= .3015



Proportions…

- The binomial distribution forms the basis of statistics for proportions.
- A proportion is just a binomial count divided by n.
  - For example, if we sample 200 cases and find 60 smokers, X=60 but the observed proportion=.30.
- Statistics for proportions are similar to binomial counts, but differ by a factor of n.

Stats for proportions

For binomial:

$$\mu_x = np$$

$$\sigma_x^{\,2} = np(1-p)$$

$$\sigma_x = \sqrt{np(1-p)}$$

Differs by a factor of n.

Differs by a factor of n.

For proportion:

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}}^{\,2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

P-hat stands for "sample proportion."

It all comes back to Z…

- Statistics for proportions are based on a normal distribution, because the binomial can be approximated as normal if np>5