The Normal Distribution

DEPARTMENT OF STATISTICS
RAMJAS COLLEGE,
UNIVERSITY OF DELHI

Presented By:

Shubham Mehta

BSc. (H) Statistics

Examples of continuous probability distributions:

The normal and standard normal

Order of Study

- Introduction
- Functions
- Graphs
- Data Analysis and Problems
- Case Study and Questions
- Binomial Approximations
- Conclusion

Normal Distribution

Normal Distribution, also called Gaussian Distribution, is one of the widely used continuous distributions existing which is used to model a number of scenarios such as marks of students, heights of people, salaries of working people etc.

It is given utmost importance due to **Central Limit Theorem** which shall not be dealt with here.

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

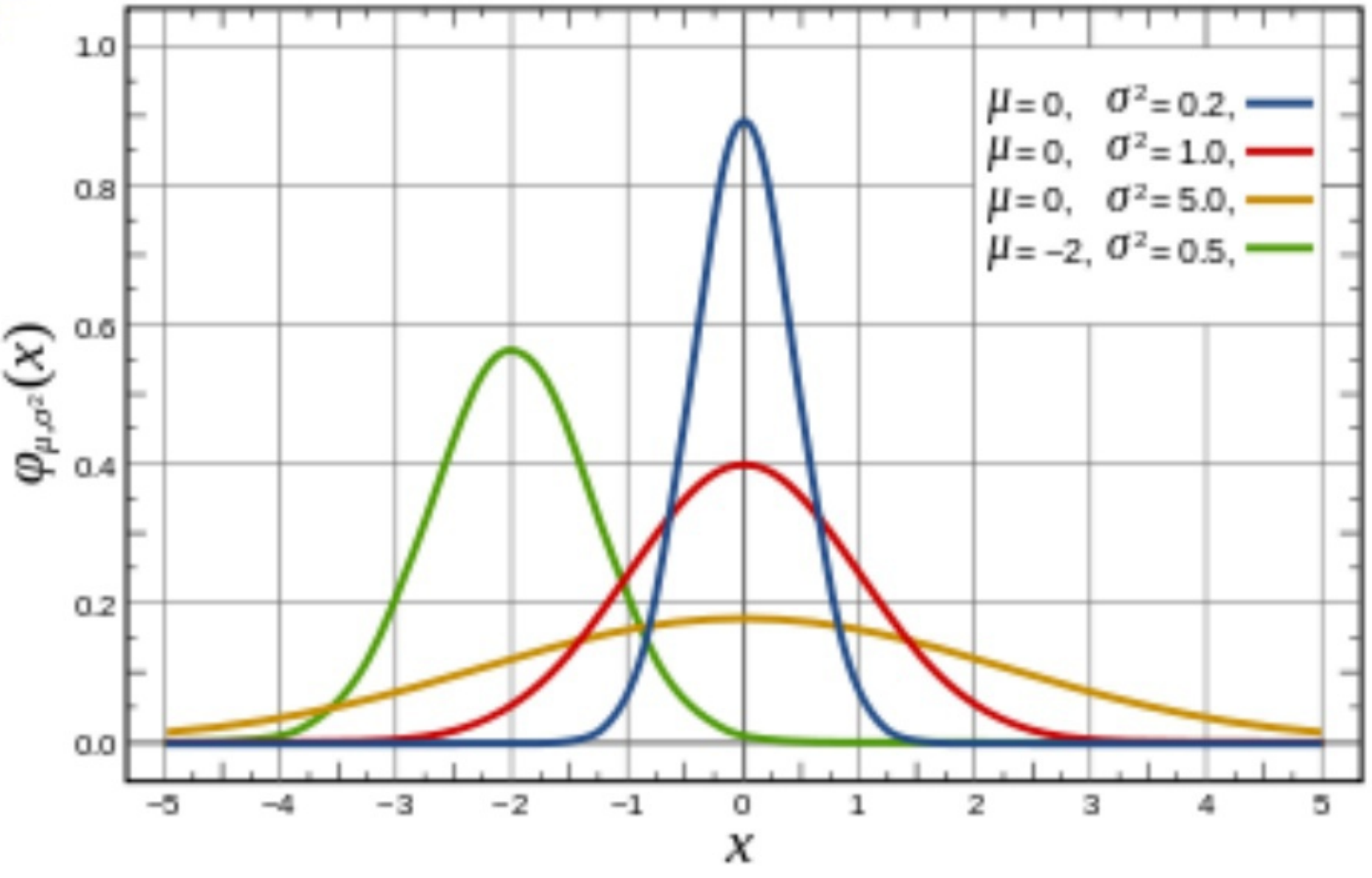
e=2.71828

The Normal PDF

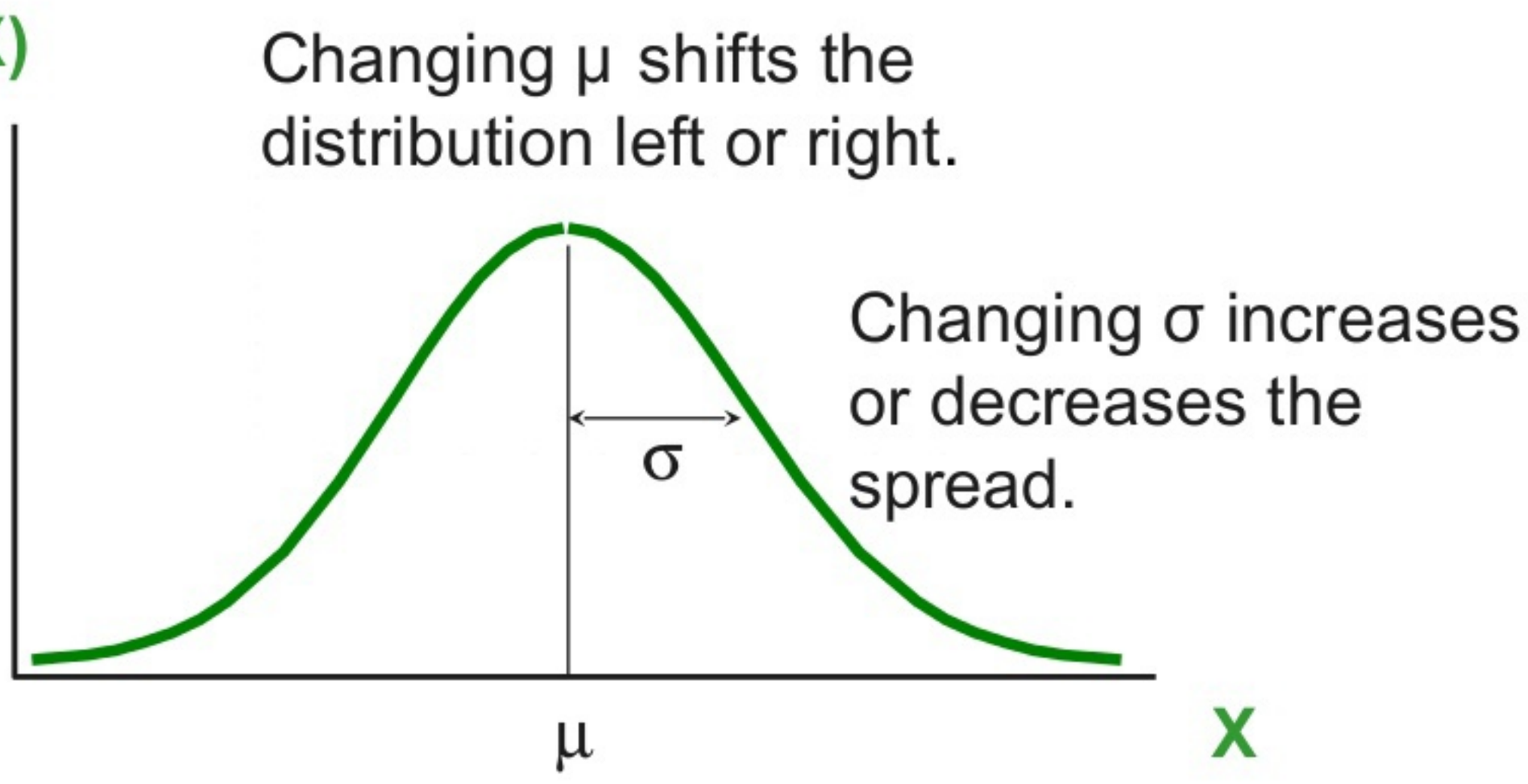
It's a probability function, so no matter what the values of μ and σ, must integrate to 1!

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

Normal Distribution (pdf)



The Normal Distribution



Normal distribution is defined by its mean and standard dev.

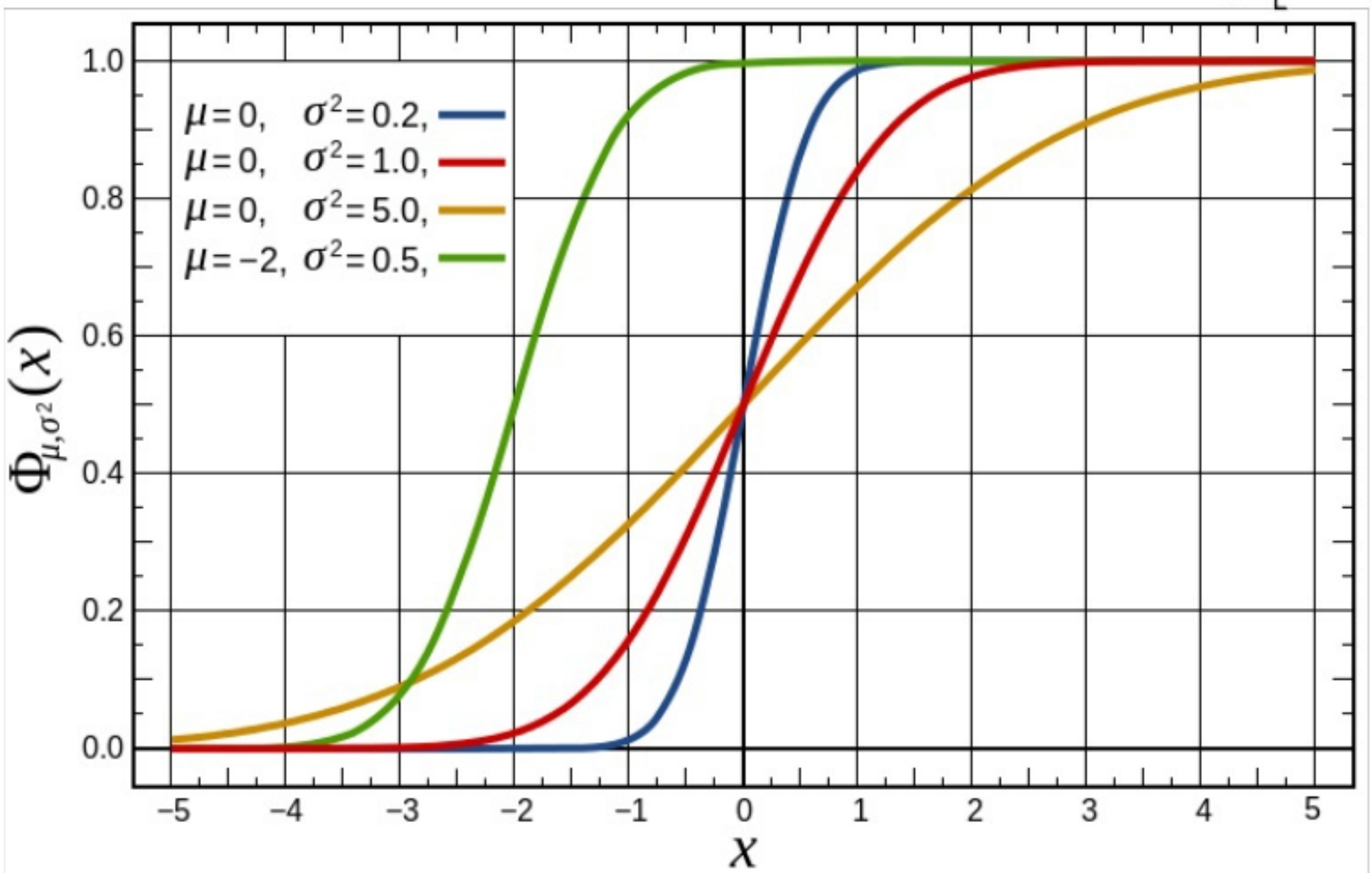
$$E(X)=\mu = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

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

Standard Deviation(X)=σ

Normal Distribution (cdf)

$$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$



Properties of Normal Distribution

- Mean=Median=Mode=μ
- Skewness = 0
- Variance=σ^2
- All odd order central moments are zero.
- Moment Generating Function

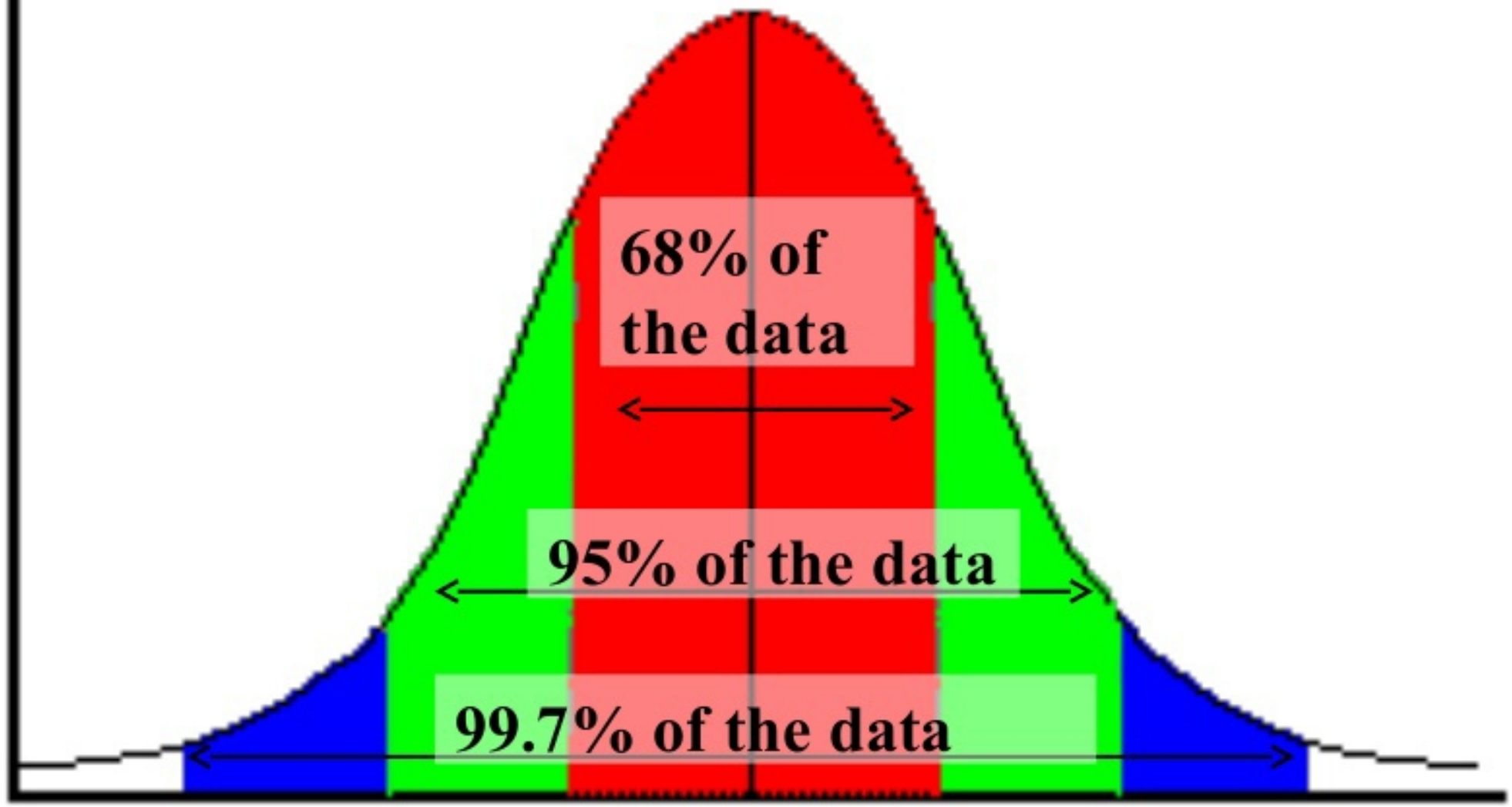
$$
\begin{aligned}
M_X(t) &= \mathrm{E}\left[\exp\left(tX\right)\right] \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \exp\left(tx\right) dx \\
&= \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)
\end{aligned}
$$

*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 Mathematical terms…

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \, dx = 0.68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \, dx = 0.95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \, dx = 0.997$$

How good is rule for real data?

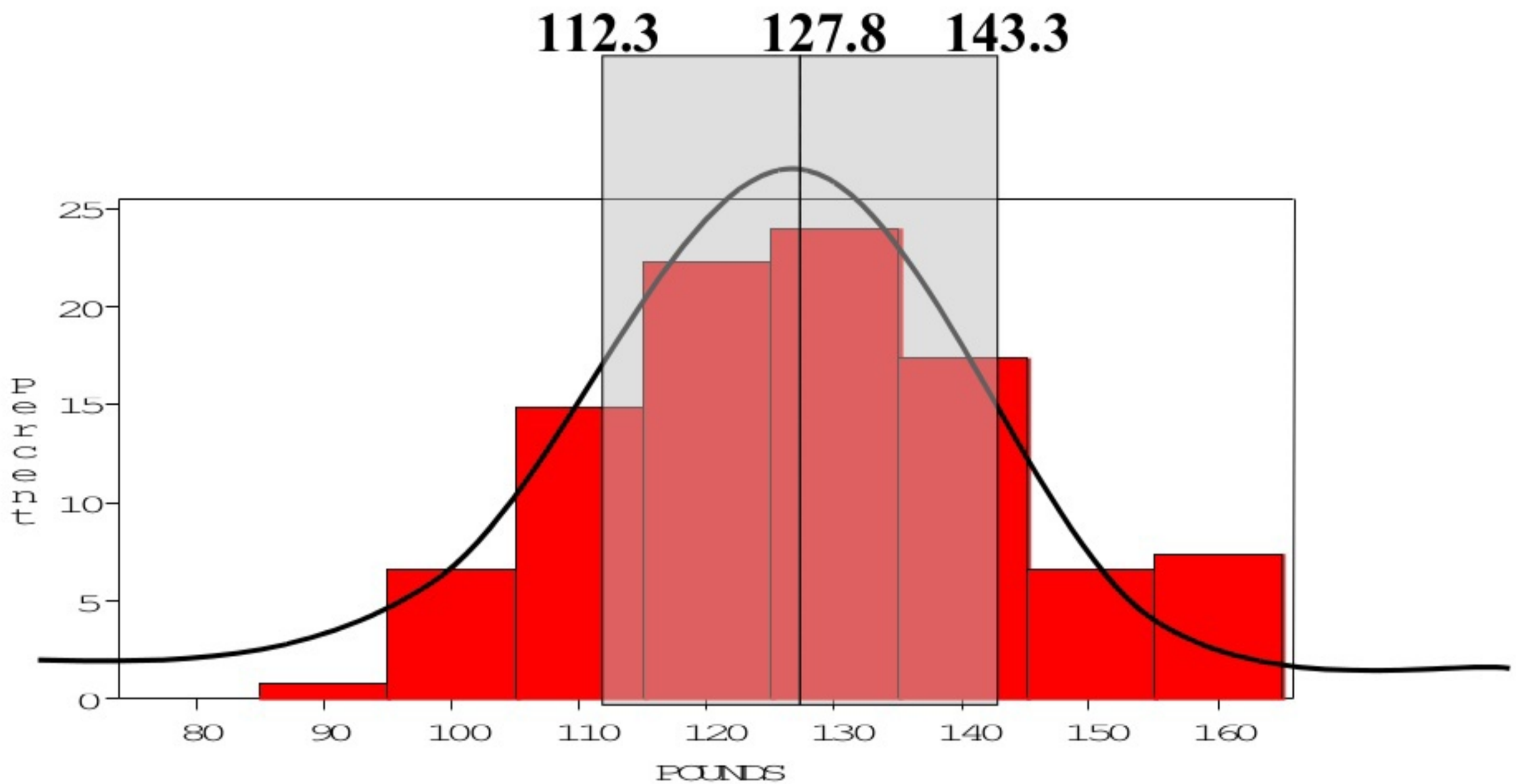
Check some example data:

The mean of the weight of the women = 127.8

The standard deviation (SD) = 15.5

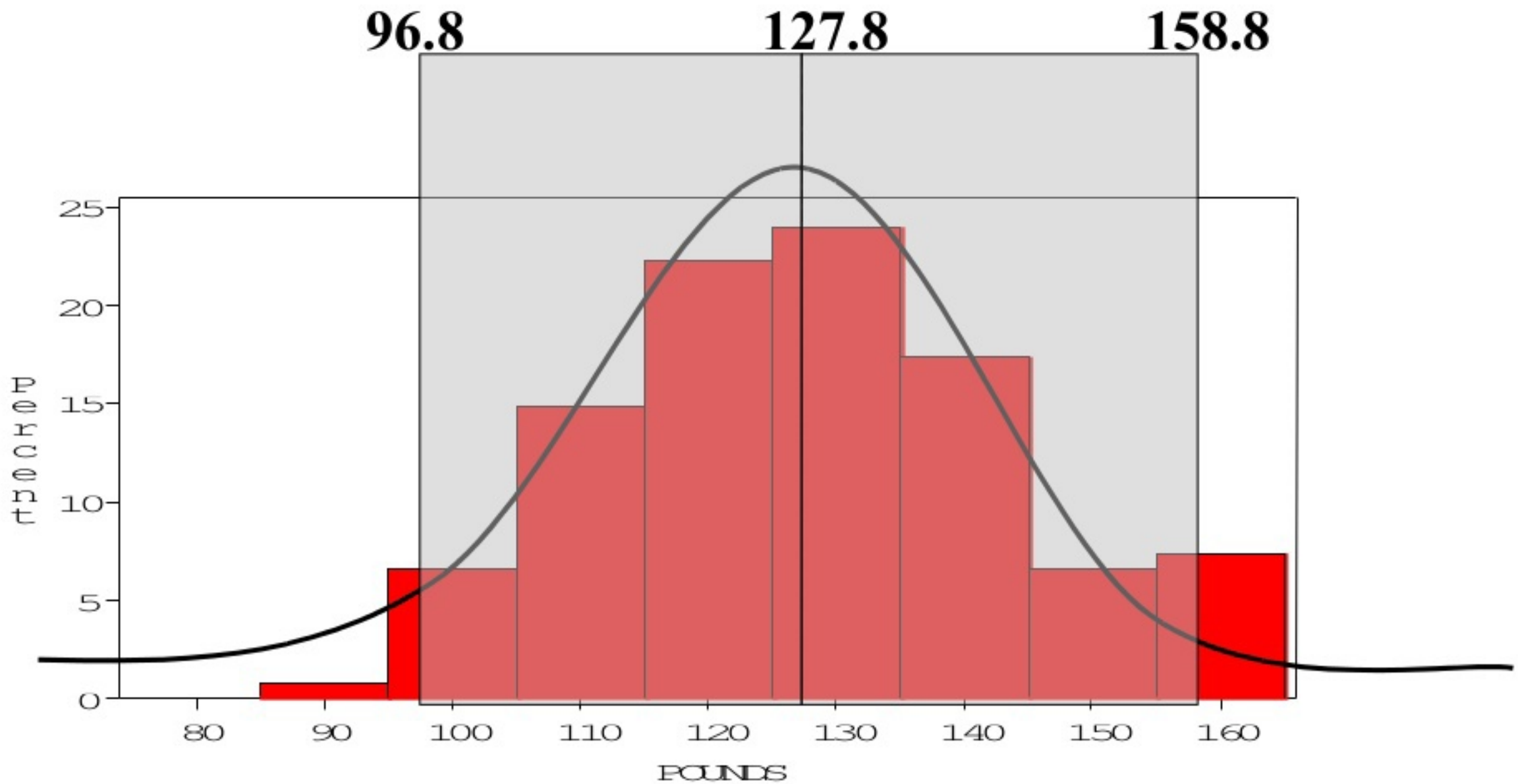
68% of 120 = .68x120 = ~ 82 runners

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



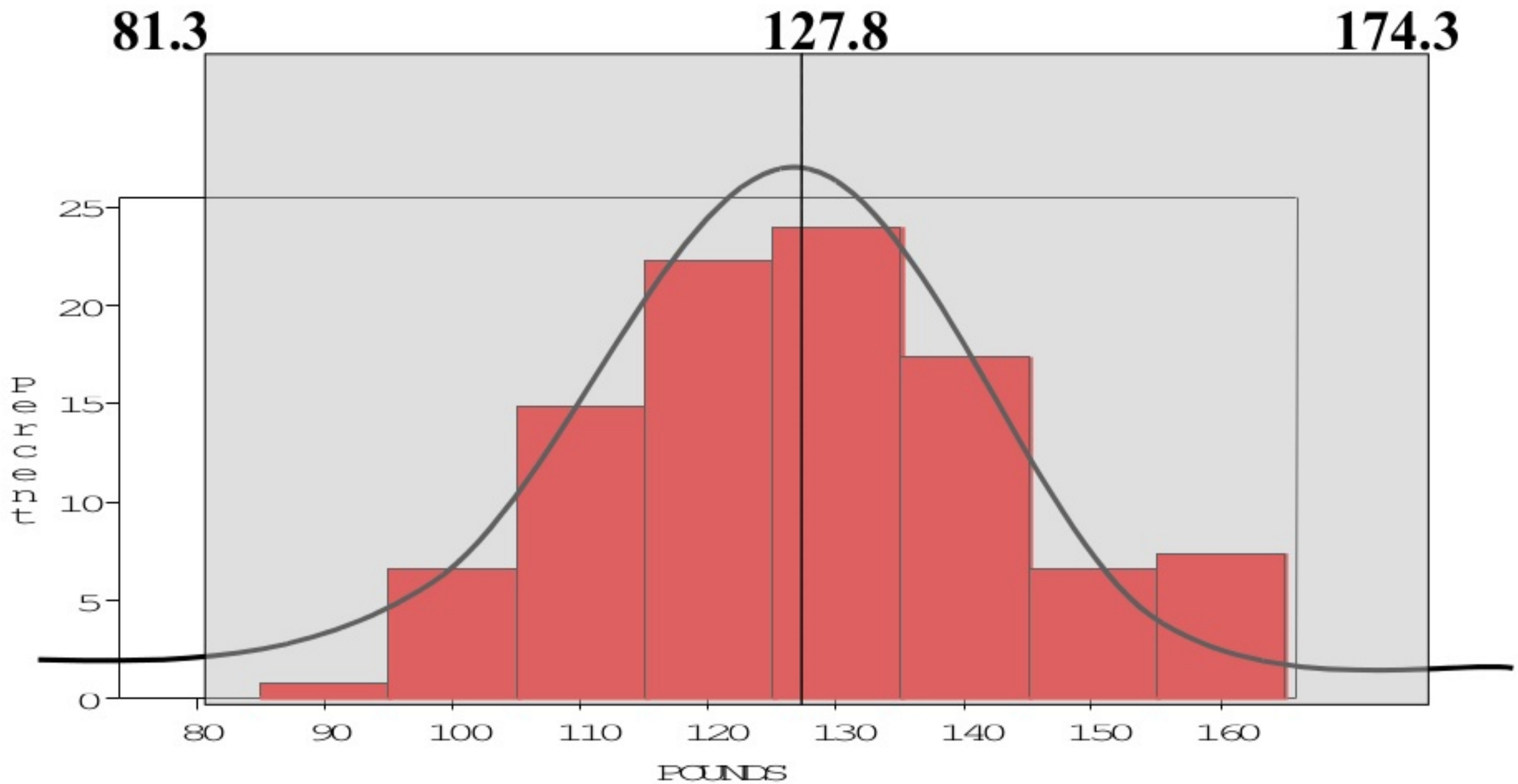
95% of 120 = .95 x 120 = ~ 114 runners

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 JAM scores roughly follows a normal distribution in the Indian population of college-bound students (with range restricted to 200-800), and the average JAM Score be 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 JAM score corresponding to the 90th percentile (=90% of students are lower)?

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

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

Solve for Q?…. by looking at tables.

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!

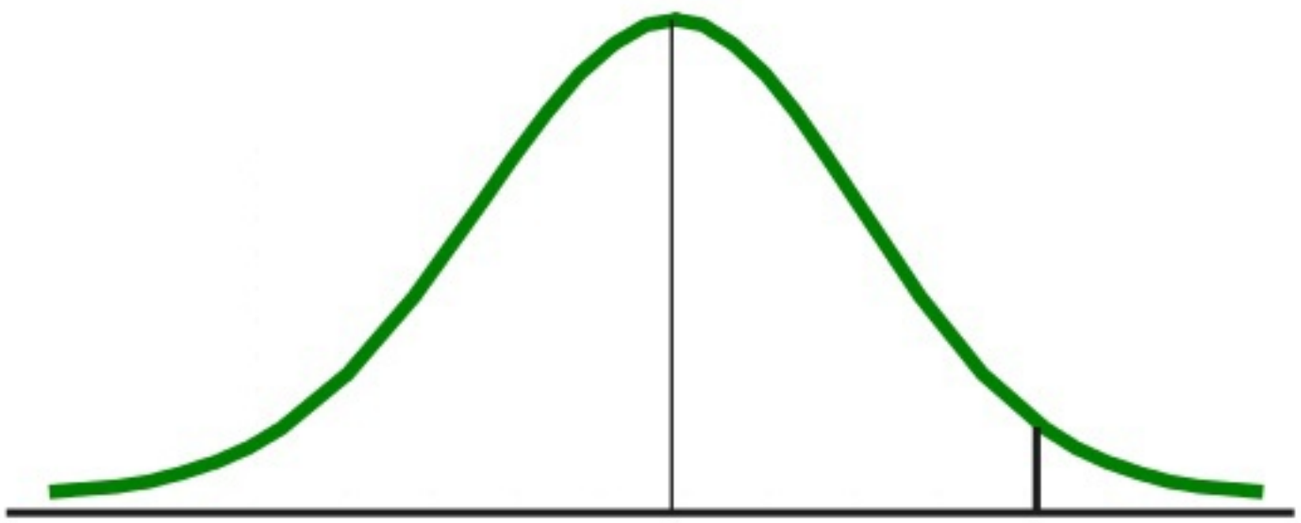
Even better, computers now do all the integration.

The Standard Normal (Z): "Universal Currency"

The formula for the standardized normal probability density function is

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

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 JAM score of 575 or less, $\mu = 500$ and $\sigma = 50$?

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

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

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

But to look up Z = 1.5 in standard normal chart = 0.9332

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?

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

From the chart → Z of 2.46 corresponds to a right tail (greater than) area of: $P(Z \geq 2.46) = 1 - P(Z \leq 2.46) = 1 - (0.9931) = 0.0069$ or 0.69 %

Practice problem

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

$$Z = \frac{120 - 109}{13} = 0.85$$

From the chart → Z of 0.85 corresponds to a left tail area of:

$P(Z \leq 0.85) = 0.8023 = 80.23\%$

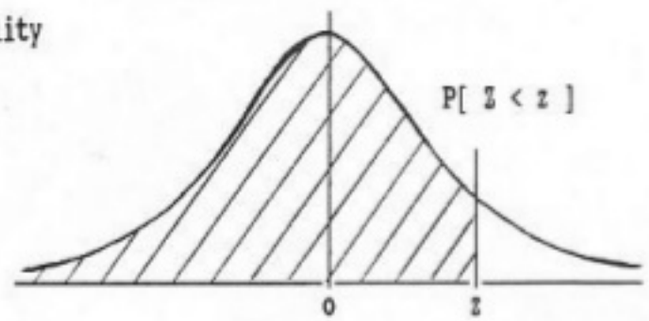
Looking up probabilities in the standard normal table

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

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

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



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

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
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Z=1.51

Z=1.51

Area is 93.45%

Is my data "normal"?

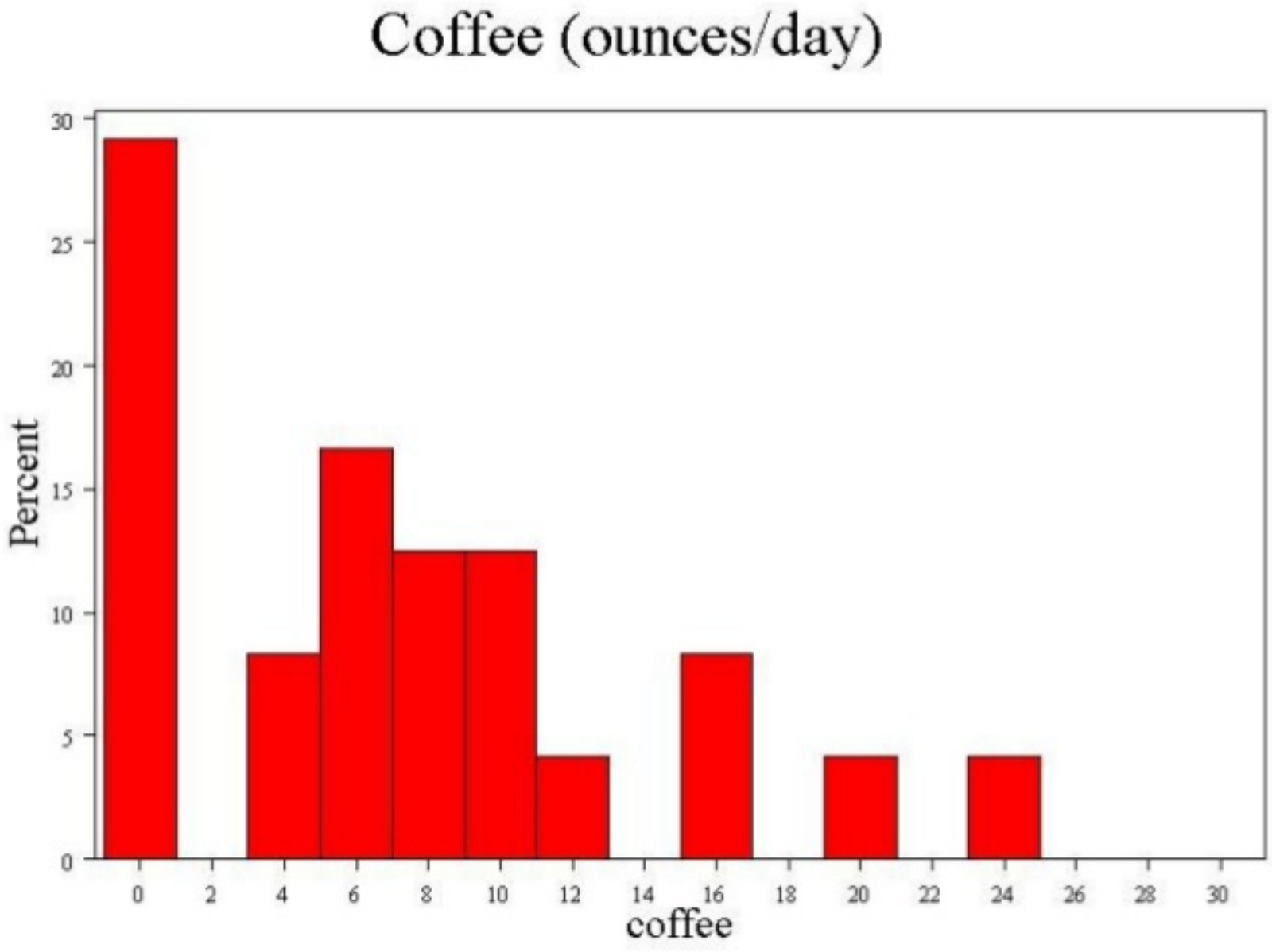
- Not all continuous random variables are normally distributed!!
- It is important to evaluate how well the data are approximated by a normal distribution

Note: At this level, we cannot state whether the Normal Distribution is the best fit for the given data or NOT.

Is 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 1st std dev of the mean? Do 95% of observations lie within 2nd std dev of the mean?
4. Look at a normal probability plot—is it approximately linear?

Data from a class…



Median = 6

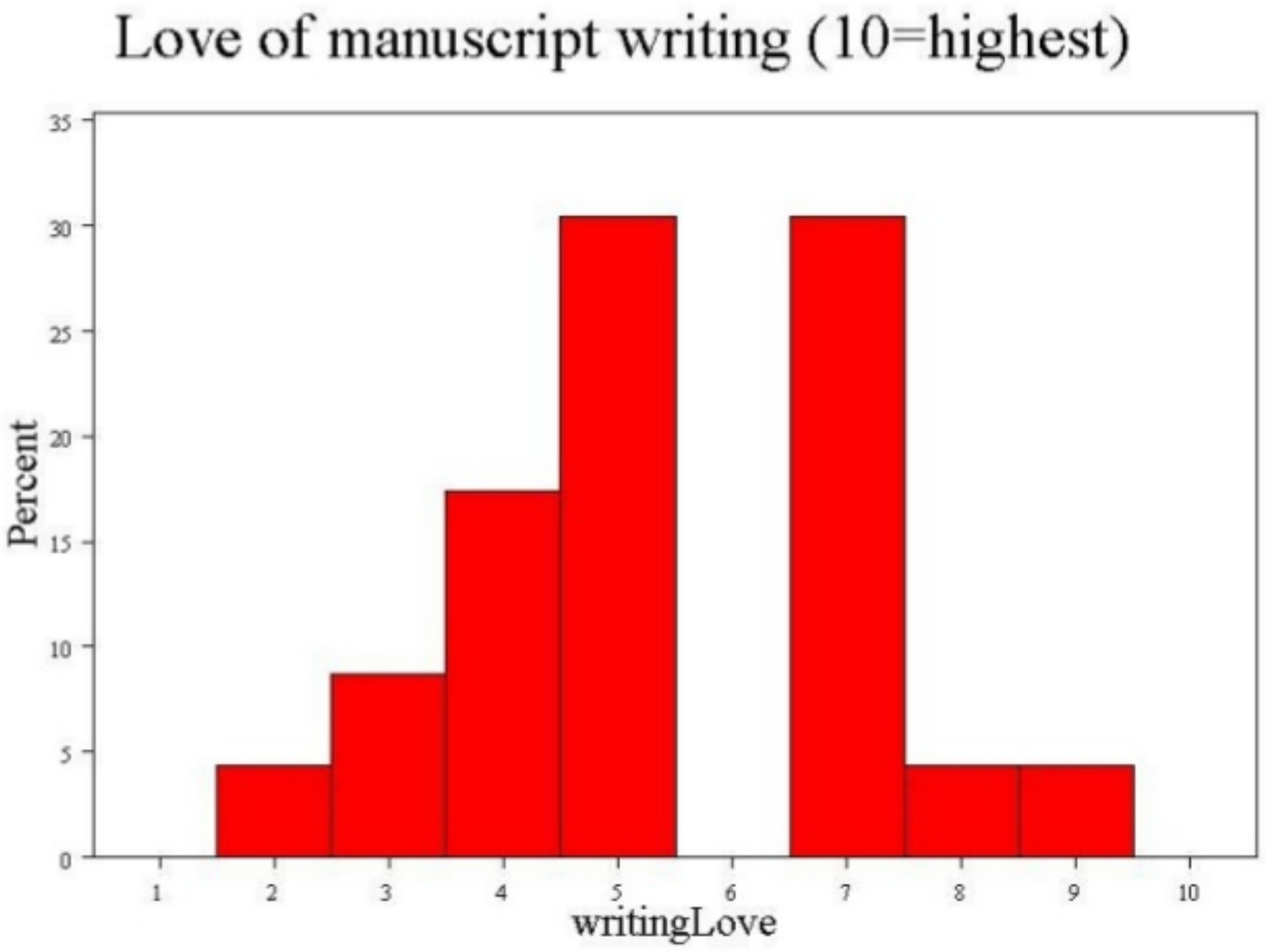
Mean = 7.1

Mode = none

SD = 6.8

Range = 0 to 24
(= 3.5 σ)

Data from a class…



Median = 5

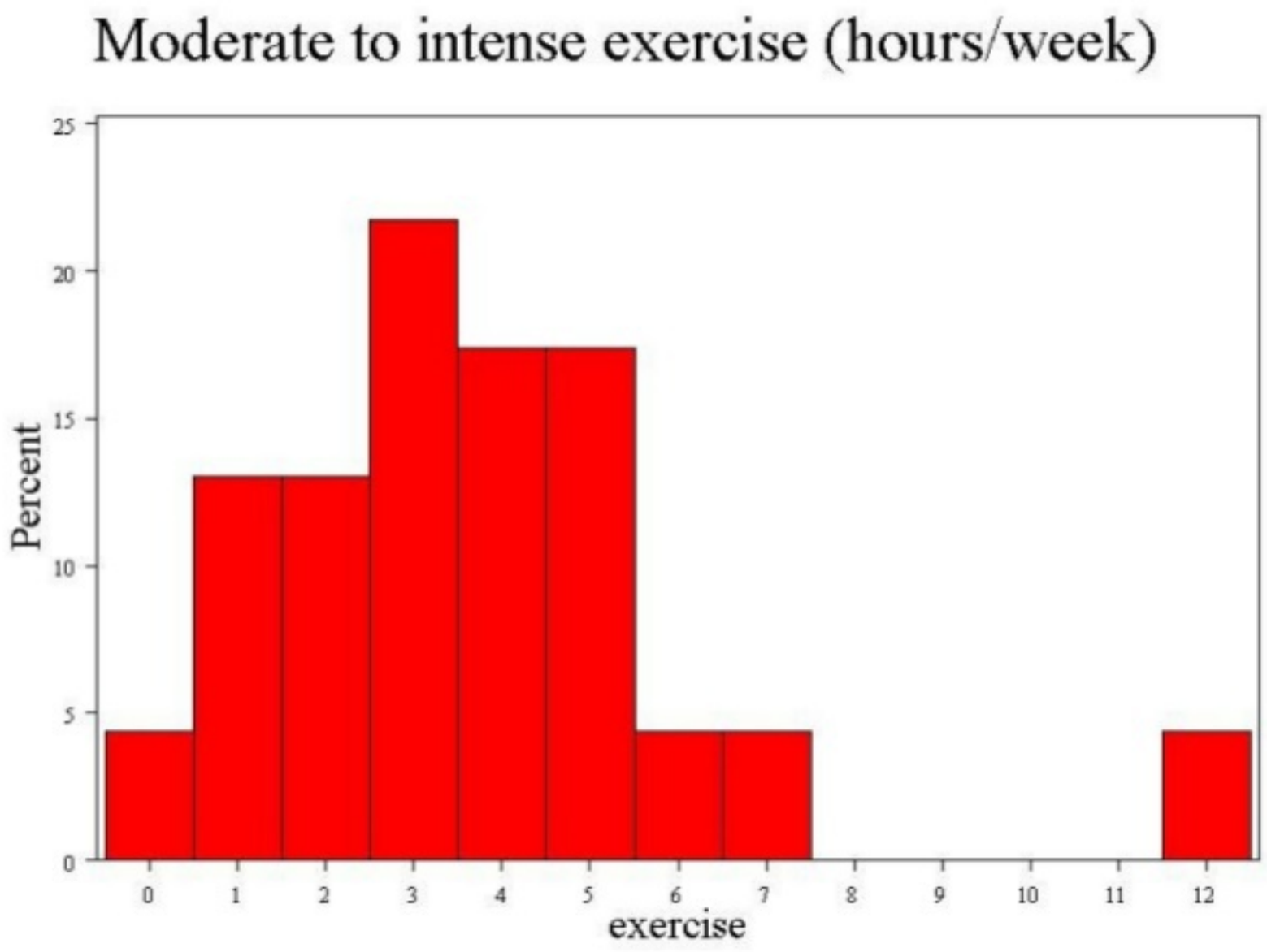
Mean = 5.4

Mode = none

SD = 1.8

Range = 2 to 9
(~ 4 σ)

Data from a class…



Median = 3

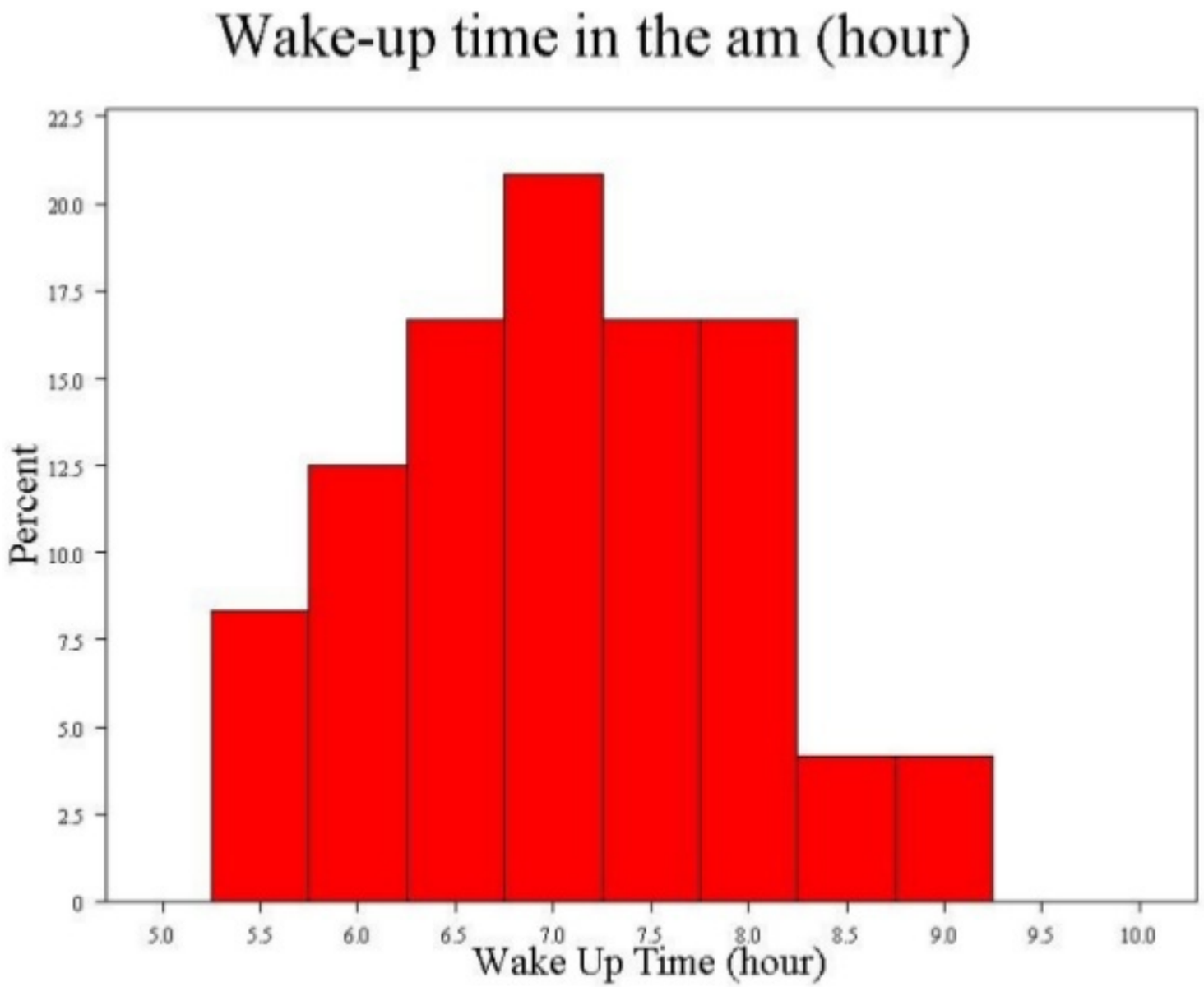
Mean = 3.4

Mode = 3

SD = 2.5

Range = 0 to 12
(~ 5 σ)

Data from a class…



Median = 7:00

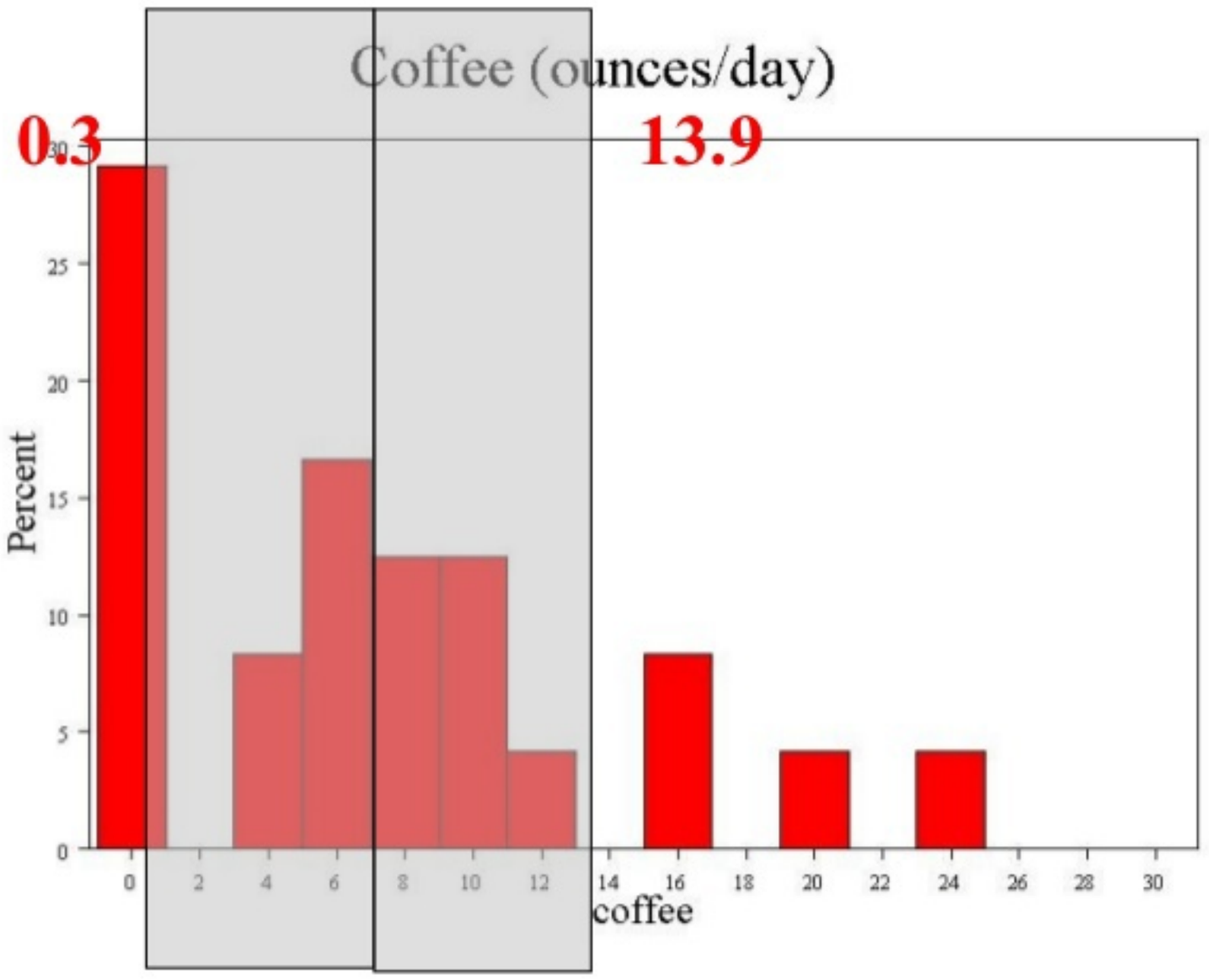
Mean = 7:04

Mode = 7:00

SD = :55

Range = 5:30 to 9:00 (~4 σ)

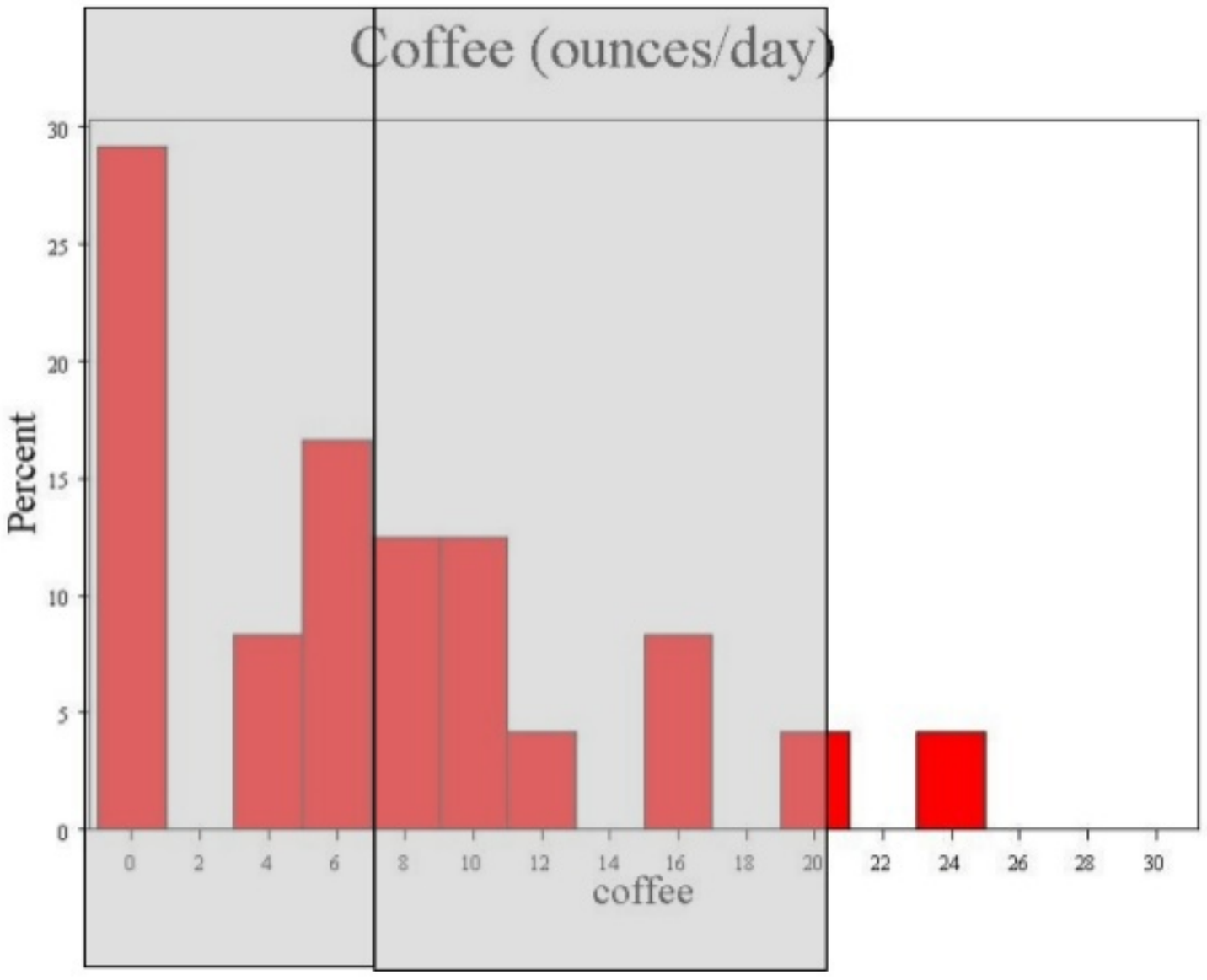
Data from a class…



7.1 +/- 6.8 =

0.3 – 13.9

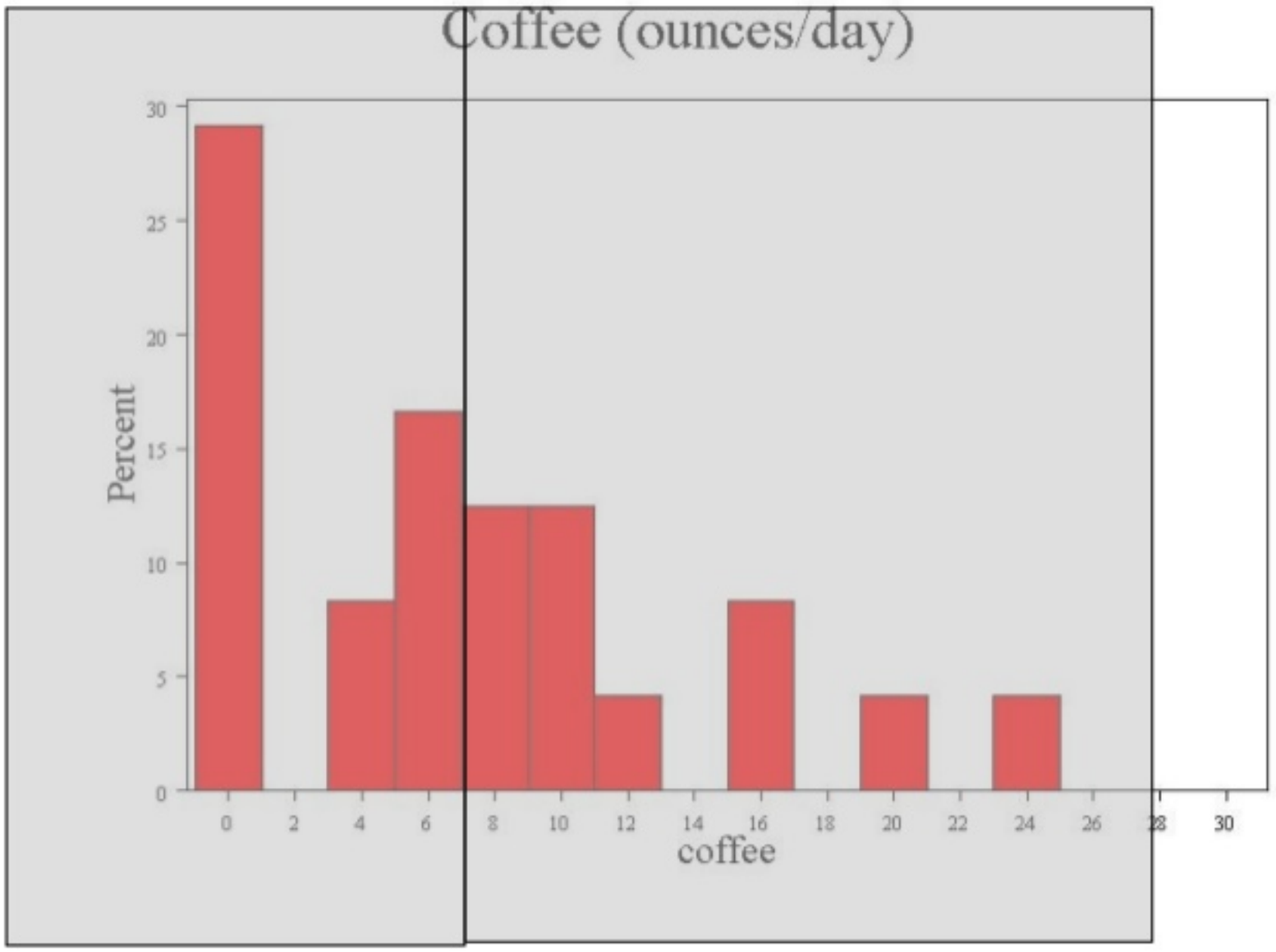
Data from a class…



7.1 +/- 2*6.8 =

0 – 20.7

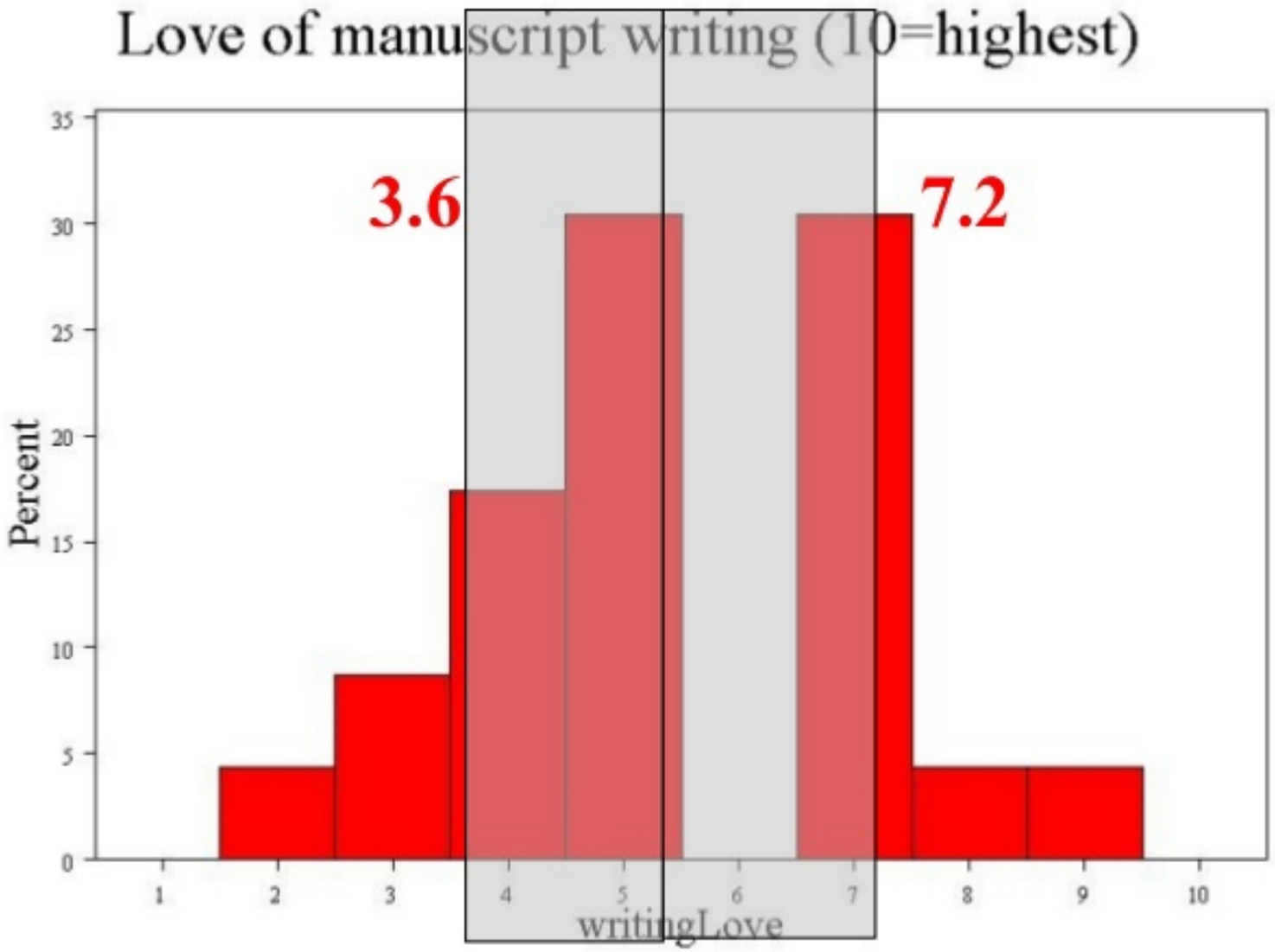
Data from a class…



7.1 +/- 3*6.8 =

0 – 27.5

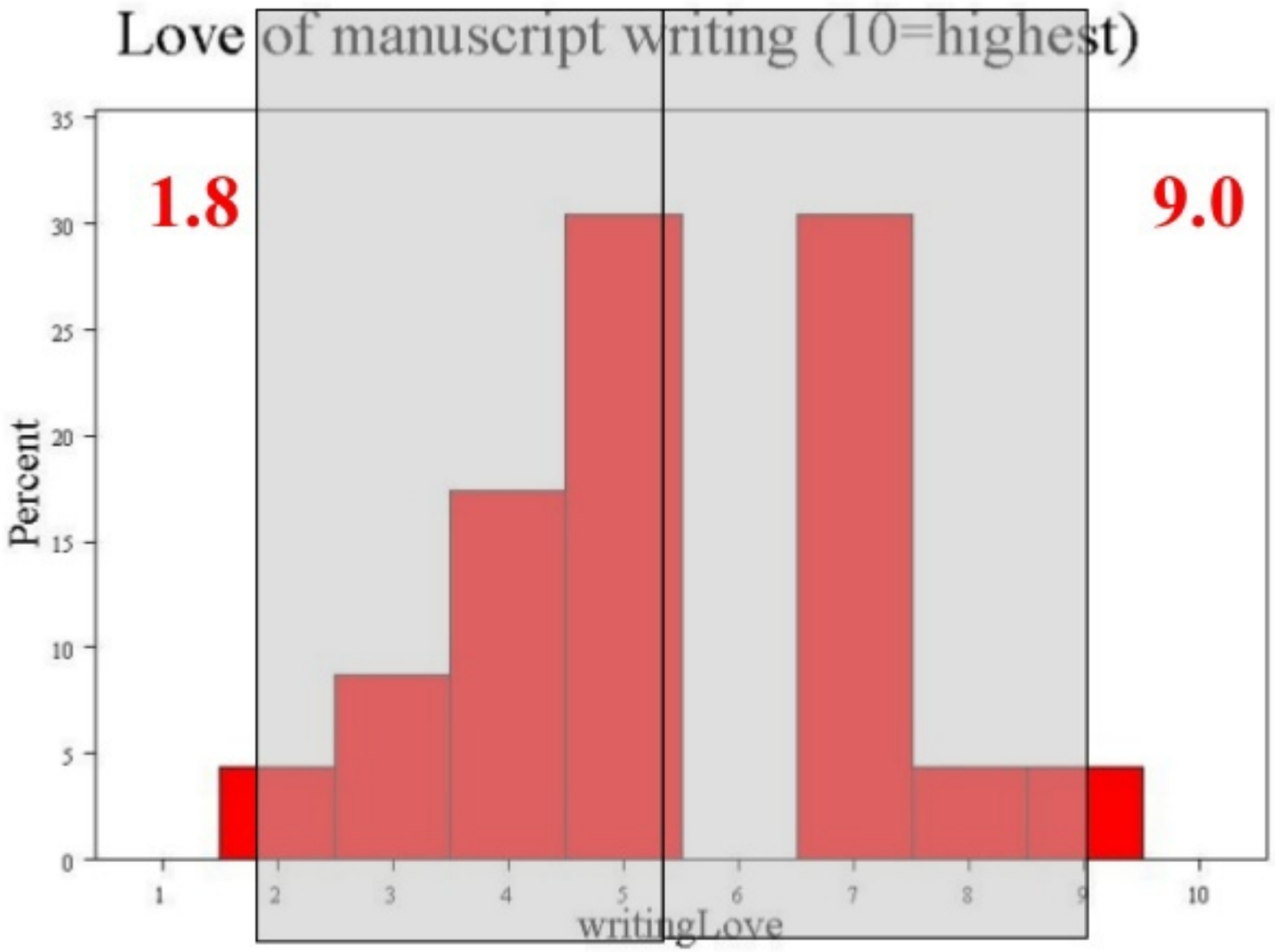
Data from a class…



5.4 +/- 1.8 =

3.6 – 7.2

Data from a class…



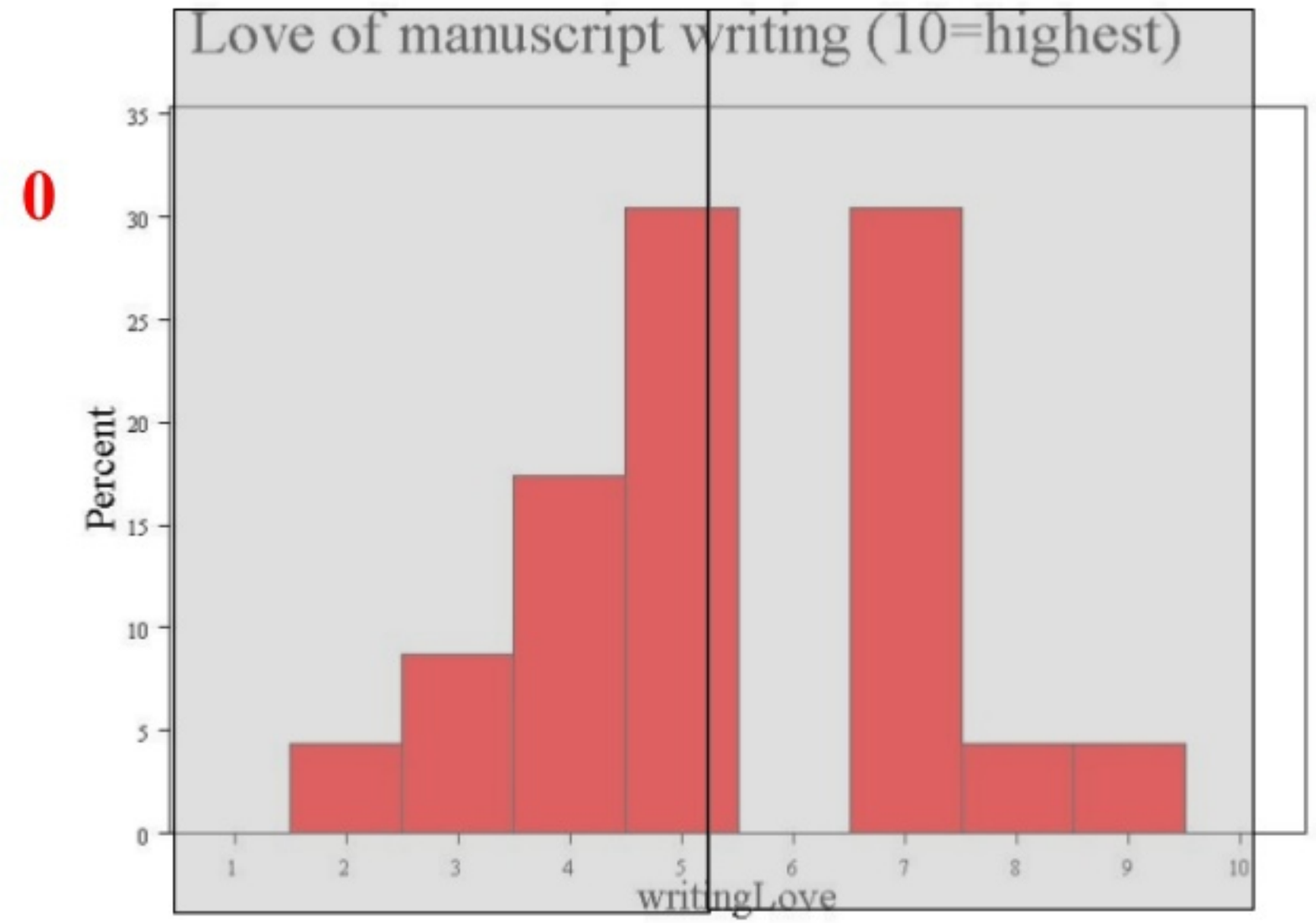
5.4 +/- 2*1.8 =

1.8 – 9.0

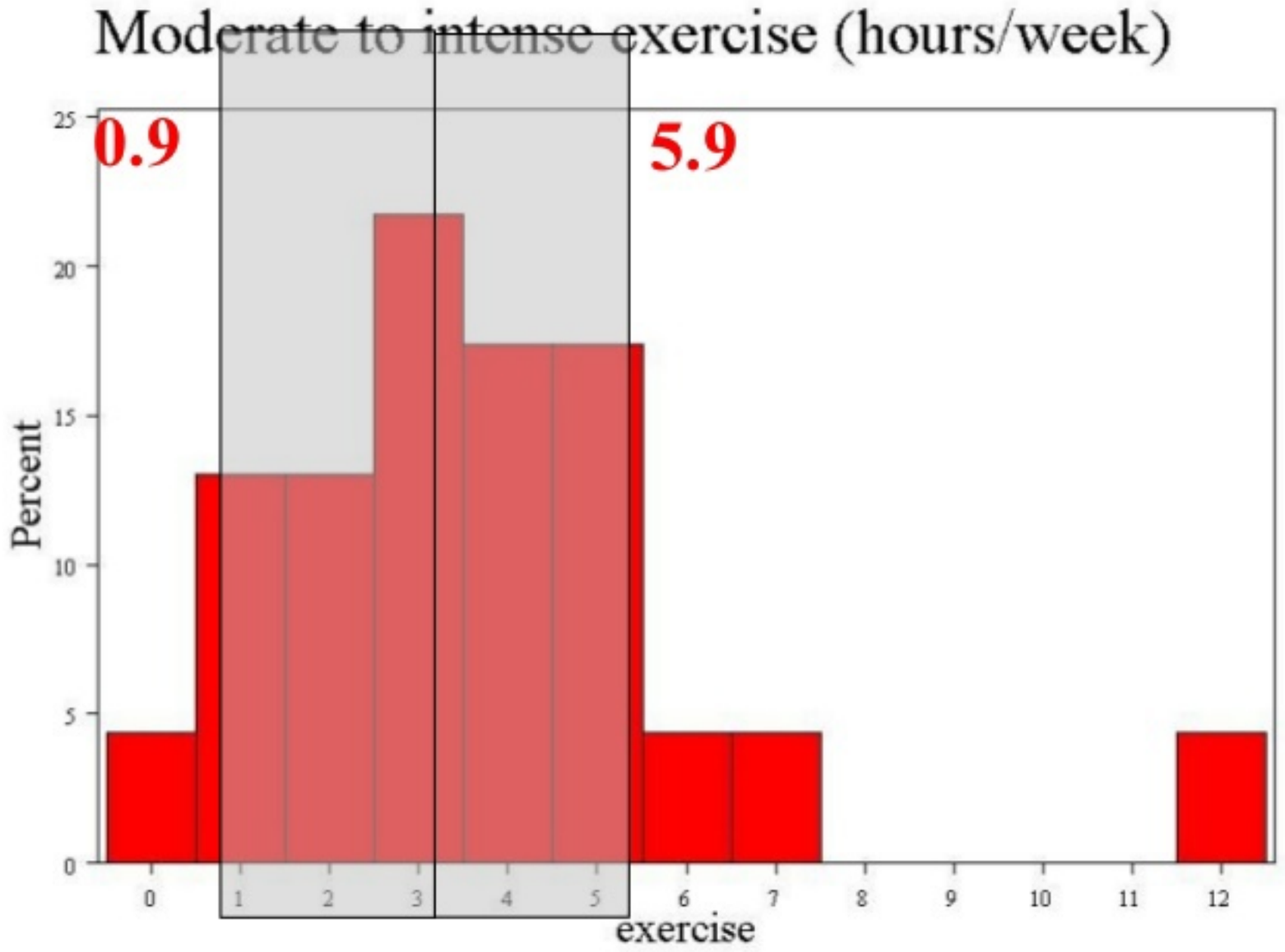
Data from a class…

5.4 +/- 3*1.8 =

0– 10



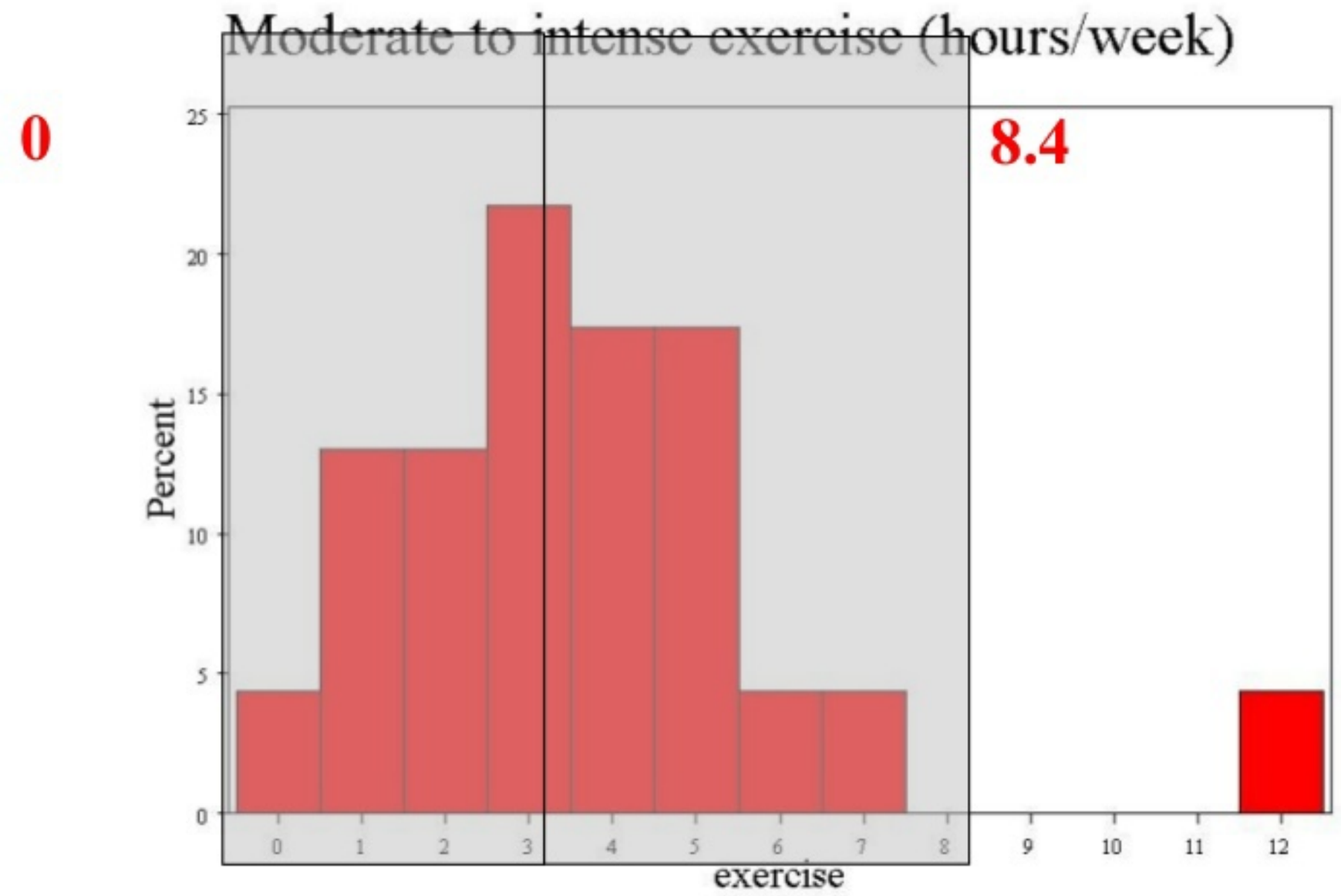
Data from a class…



3.4 +/- 2.5=

0.9 – 7.9

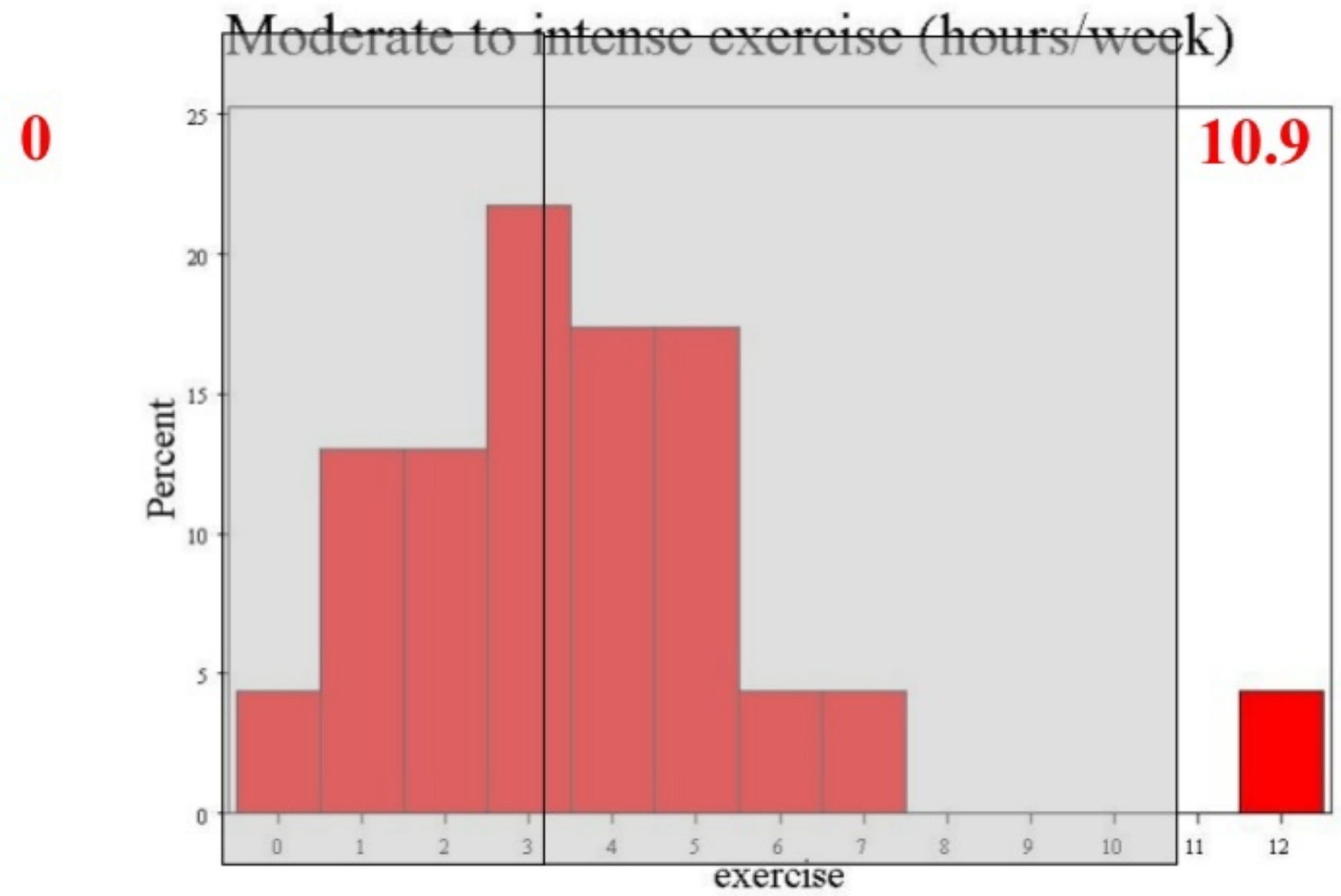
Data from a class…



3.4 +/- 2*2.5=

0 – 8.4

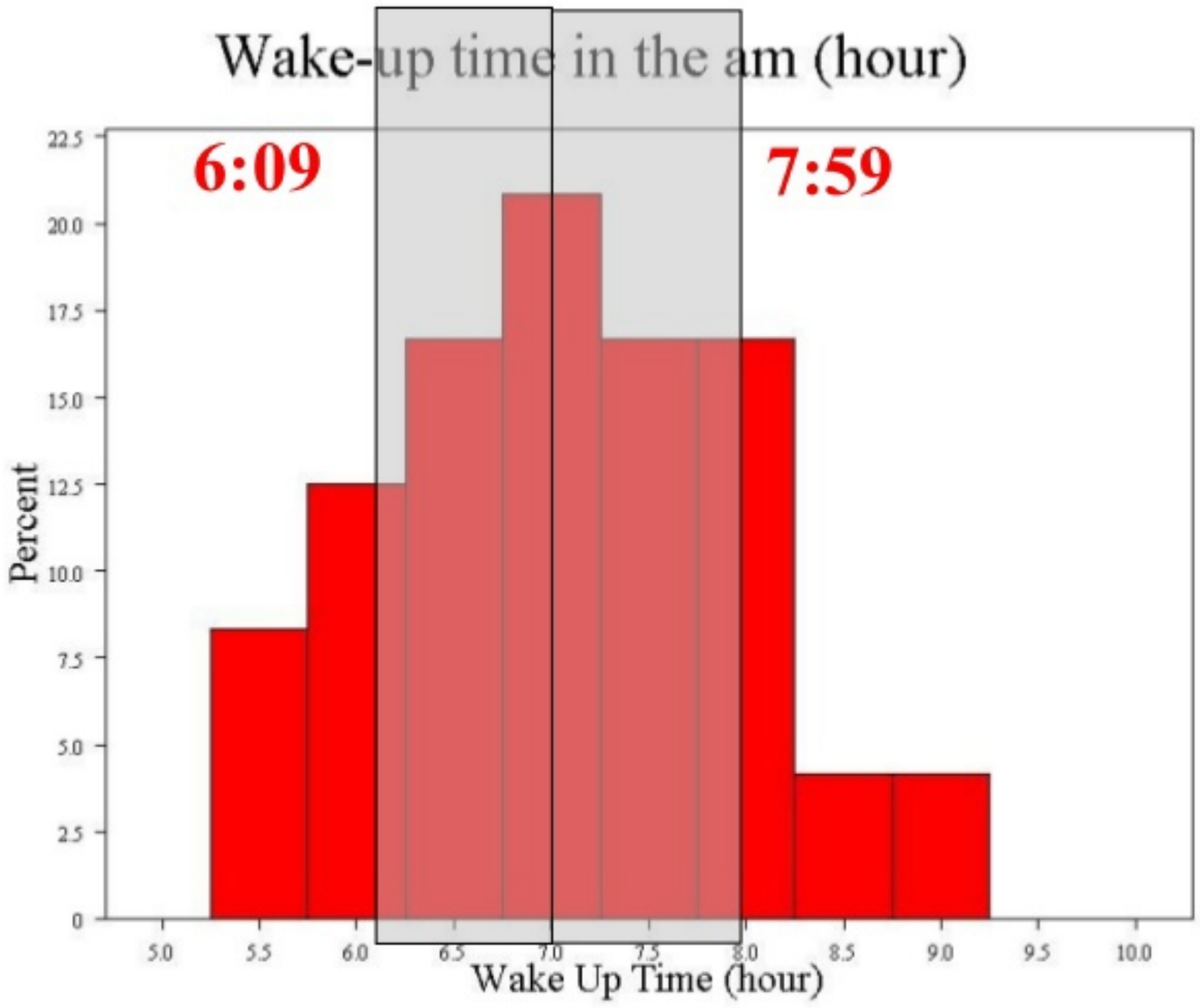
Data from a class…



3.4 +/- 3*2.5=

0 – 10.9

Data from a class…

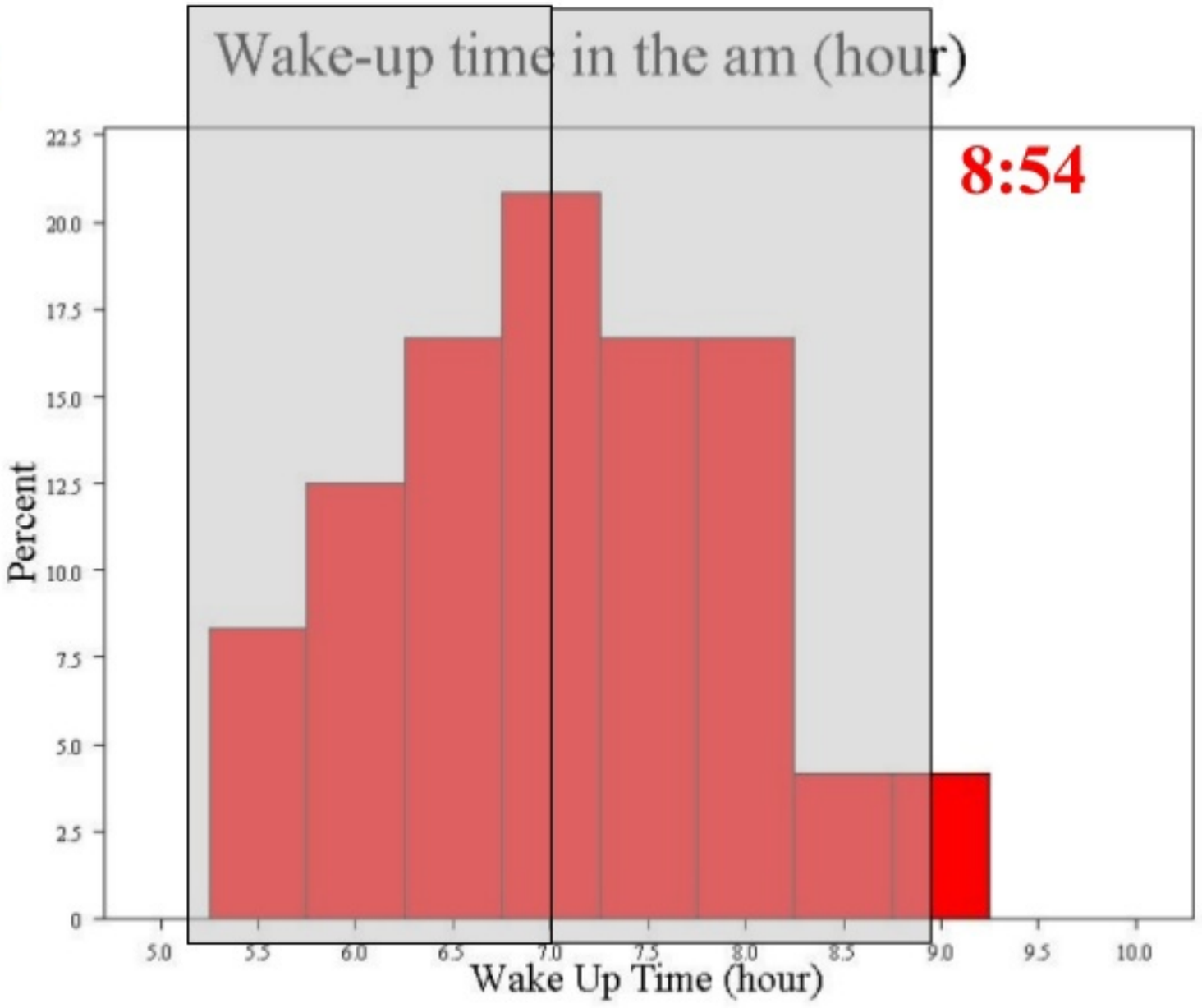


7:04+/- 0:55 =

6:09 – 7:59

Data from a class…

5:14



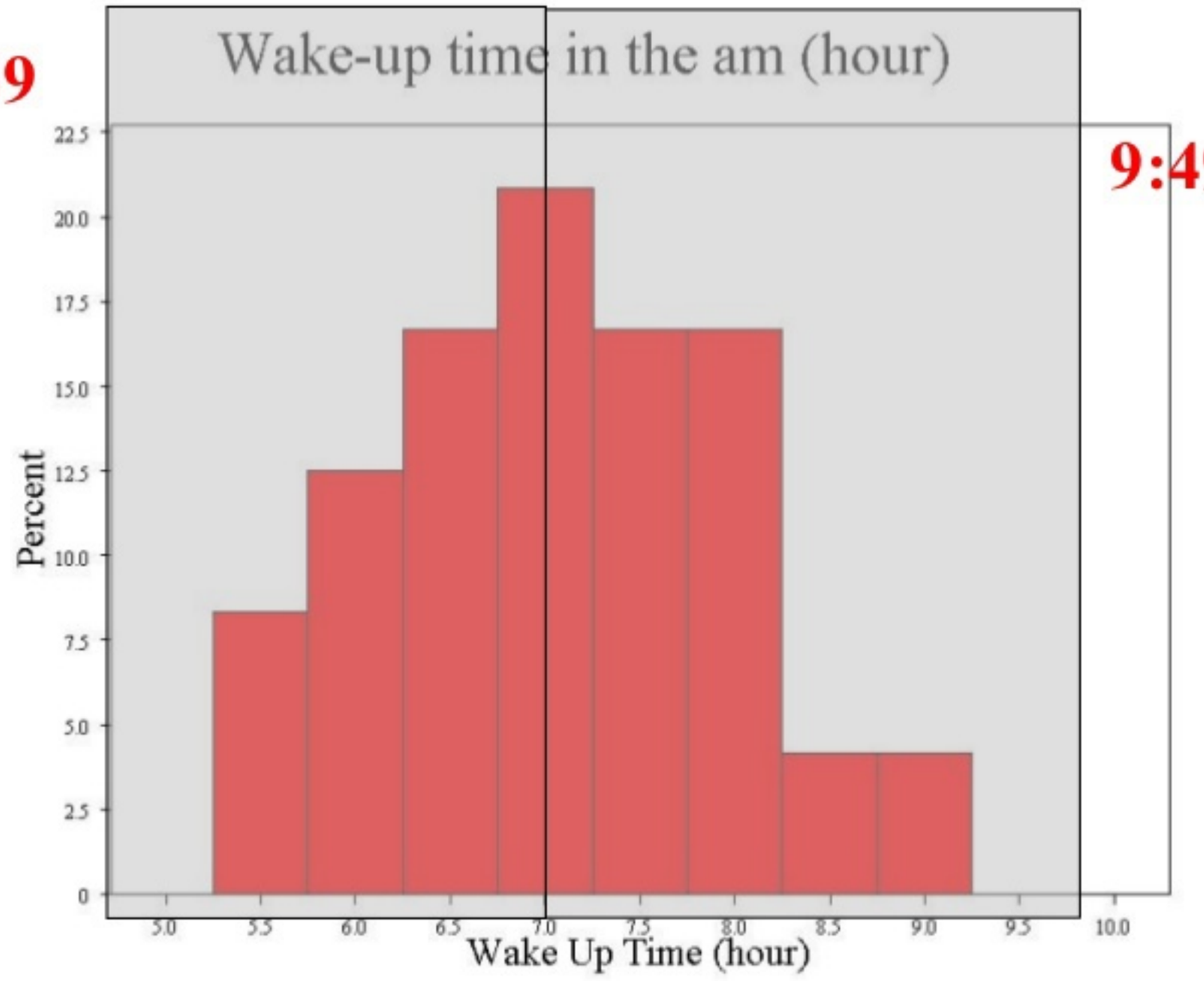
8:54

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

5:14 – 8:54

Data from a class…

4:59



9:49

7:04+/- 3*0:55 =

4:59 – 9:49

The Normal Probability Plot

- The normal probability plot is a graphical technique for normality testing: assessing whether or not a data set is approximately normally distributed.
- The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.
- The Probit function is the quantile function associated with the standard normal distribution.

$$\Phi(-1.96) = 0.025 = 1 - \Phi(1.96).$$

$$\text{probit}(0.025) = -1.96 = -\text{probit}(0.975)$$

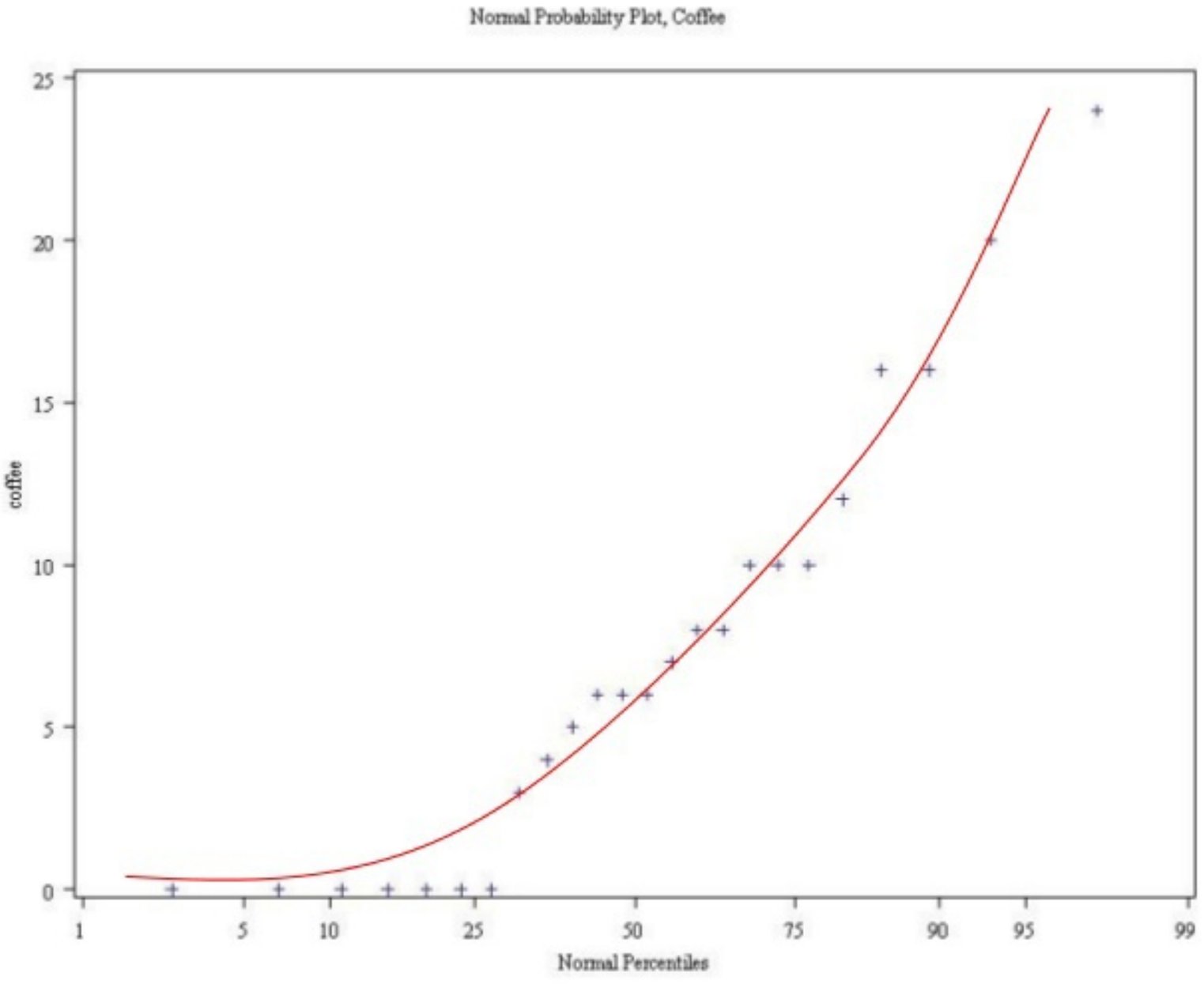
$$\Phi(\text{probit}(p)) = p \qquad \text{probit}(\Phi(z)) = z.$$

The Normal Probability Plot

- Normal probability plot
  - Order the data.
  - Find corresponding standardized normal quantile values: $i^{th}\ \text{quantile} = \phi(\frac{\text{i}}{\text{n}+1})$

    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.

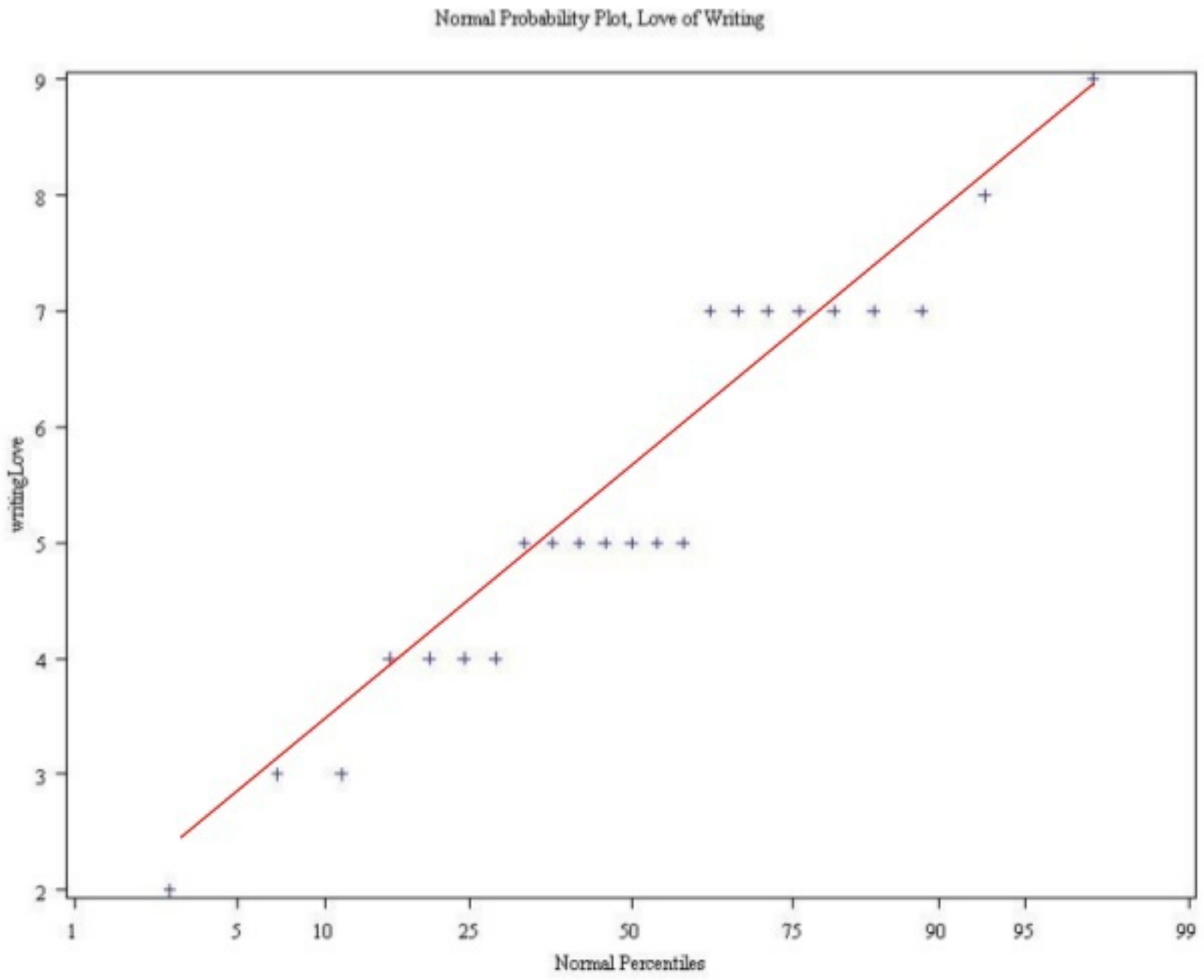
Normal probability plot coffee…



Right-Skewed

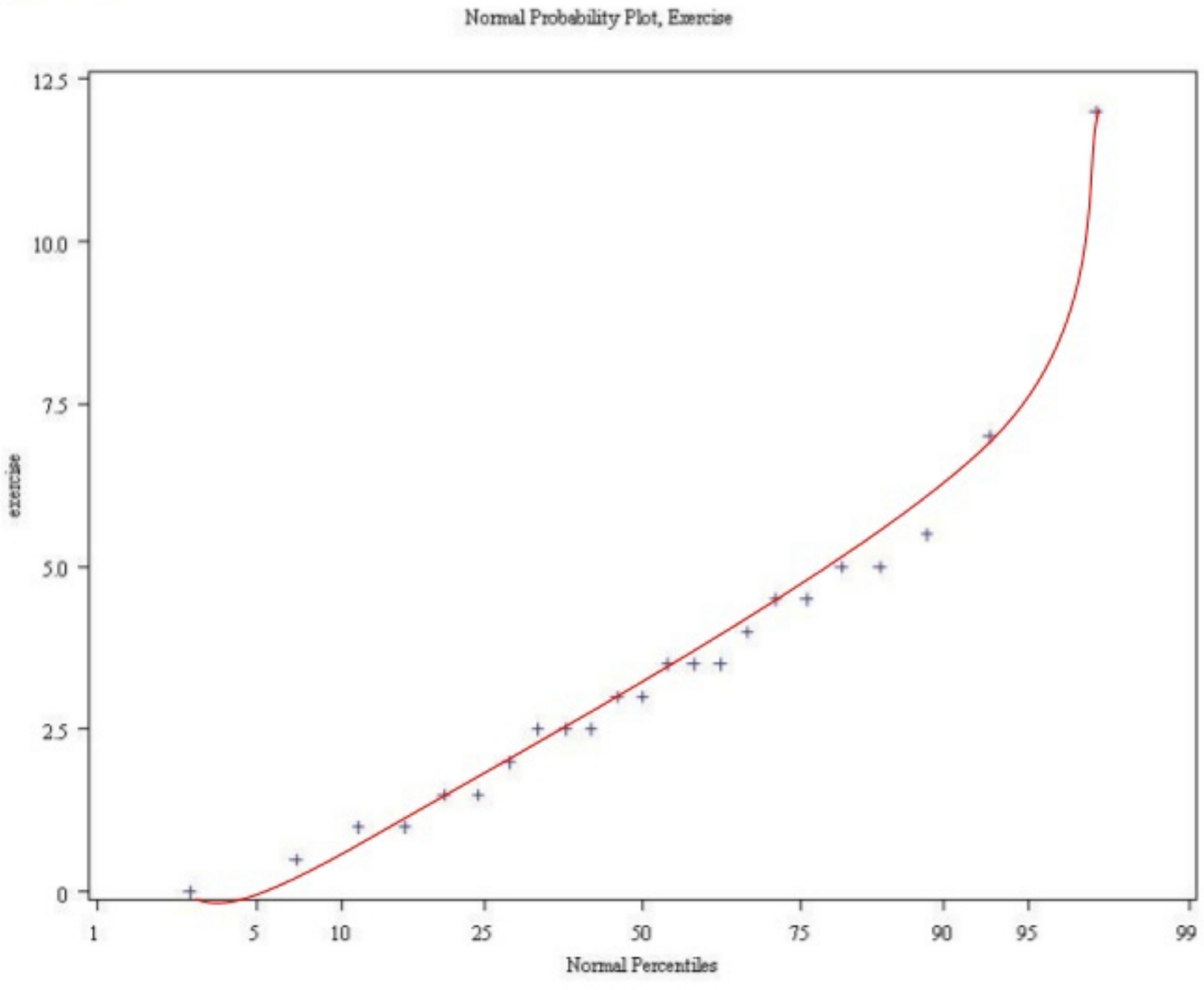
(concave up)

Normal probability plot love of writing…



Neither right-skewed nor left-skewed, but big gap at 6.

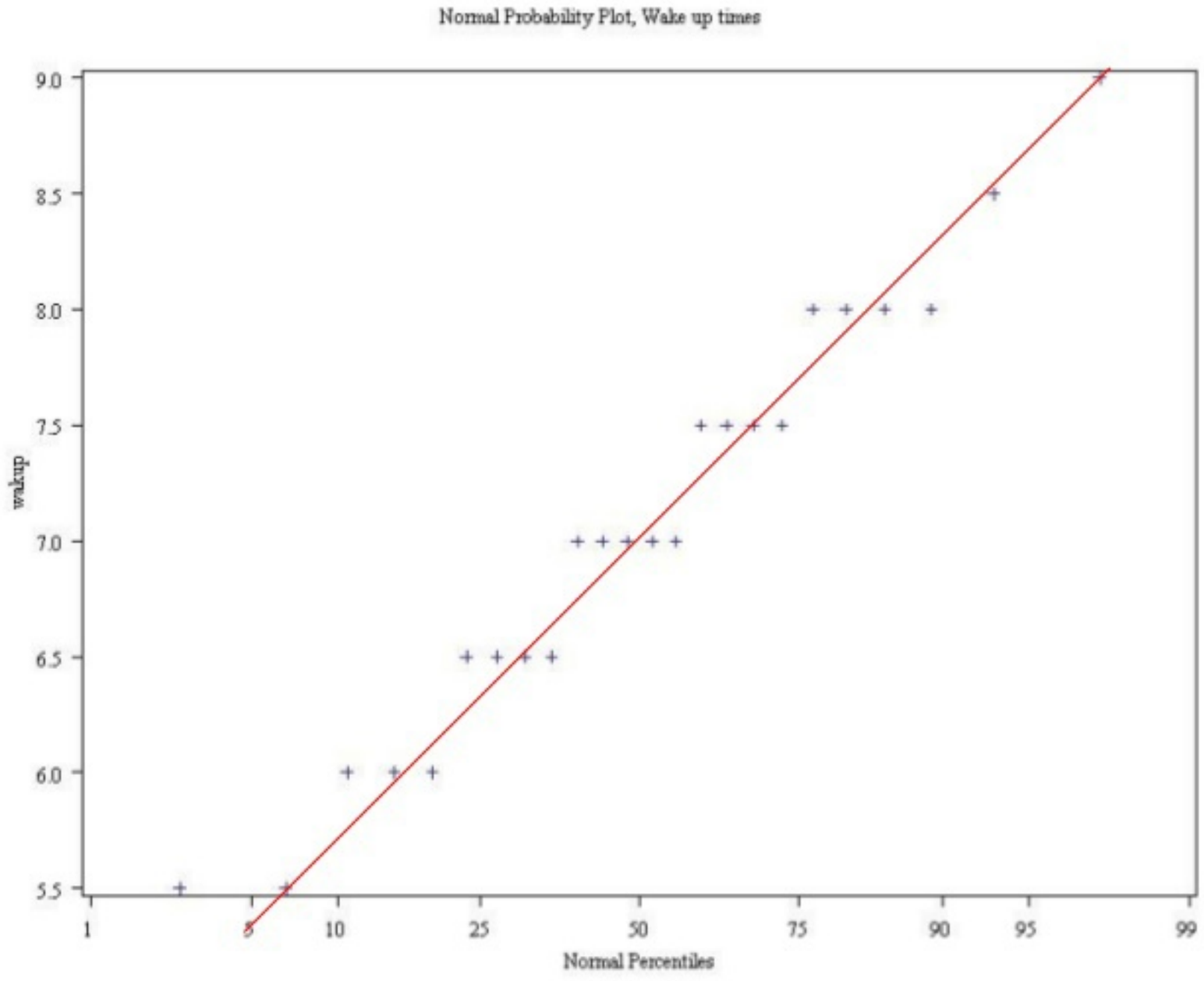
Norm prob. plot Exercise…



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 ($p<0.01$)
- Writing love: Moderate evidence of non-normality ($p = 0.01$)
- Exercise: Weak to no evidence of non-normality ($p>0.10$)
- Wakeup time: No evidence of non-normality ($p>0.25$)

Normal approximation to the binomial

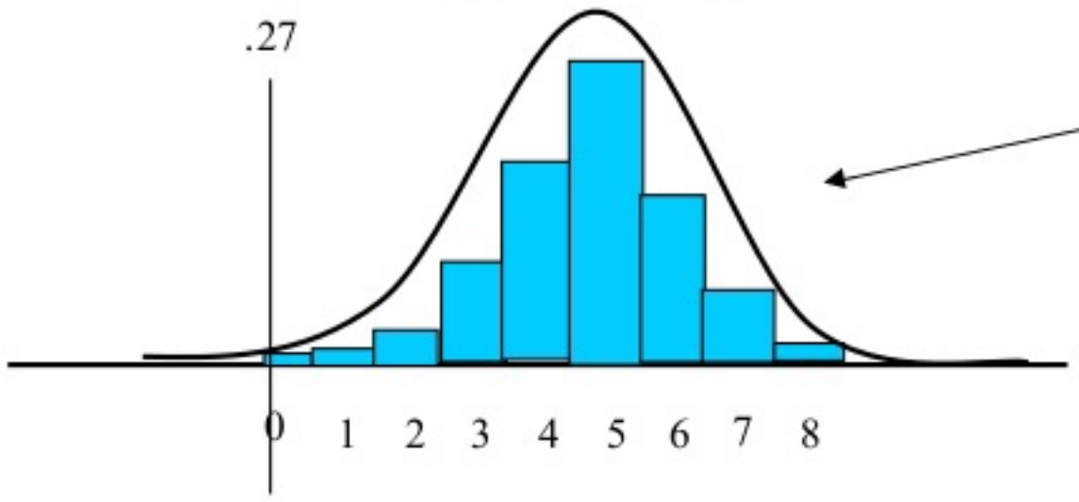
When you have a binomial distribution where n is large and p is middle-of-the road (not too small, not too big, closer to 0.5), then the binomial starts to look like a normal distribution→ In fact, this doesn't even take a particularly large n.

<u>Recall:</u> What is the probability of being a smoker among a group of cases with lung cancer is 0.6; What's the probability that in a group of 8 cases you have less than 2 smokers?

Normal approximation to the binomial

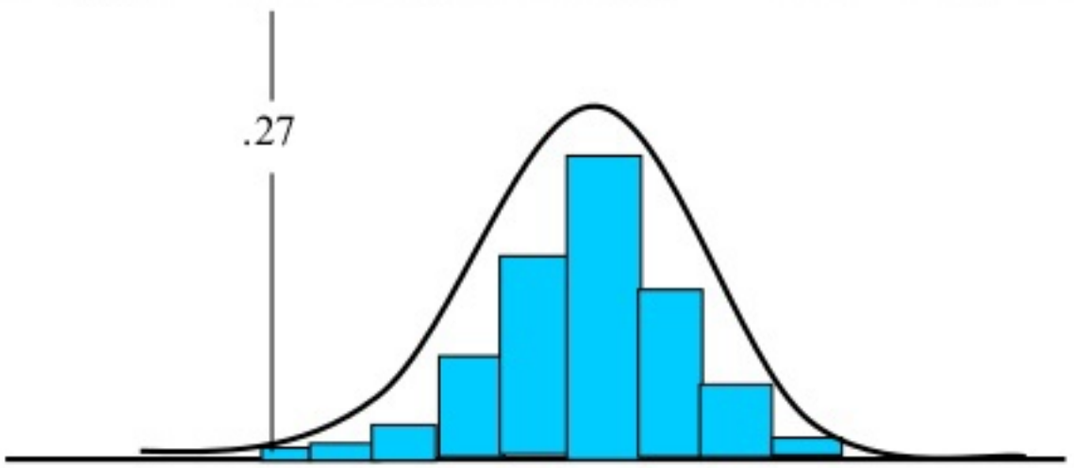
When you have a binomial distribution where *n* is large and p isn't too small (rule of thumb: mean>5), then the binomial starts to look like a normal distribution.

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) = 0.00065 + 0.008 = 0.00865

Normal approximation probability:

$\mu=4.8$

$\sigma=1.39$

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

$P(Z<2) = 0.022$

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")…

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

$P(Z \leq -2.37) = 0.0069$

A fairly good approximation of the exact probability, 0.00865.

Practice problem

1. You are performing a cohort study. If the probability of developing disease in the exposed group is 0.25 for the study duration, then if you sample (randomly) 500 exposed people, what's the probability that **<u>at most</u>** 120 people develop the disease?

Answer

By hand:

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

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

OR USE normal approximation:

$\mu = np = 500(0.25) = 125$ and $\sigma^2 = np(1-p) = 93.75$; $\sigma = 9.68$

$$Z = \frac{120 - 125}{9.68} = -0.52 \qquad P(Z < -0.52) = 0.3015$$

Case Study

A company supplies pins in bulk to a customer. The company uses an automatic lathe to produce the pins. Due to many causes - vibration, temperature, wear and tear, and the like – the lengths of the pins made by the machine are normally distributed with a mean of 1.012 inches and a standard deviation of 0.018 inch. The customer will buy only those pins with lengths in the interval 1.00±0.02 inch. In other words, the customer wants the length to be 1.00 inch but will accept up to 0.02 inch deviation on either side. This 0.02 inch is known as the **tolerance.**

Questions

Q. What percentage of pins will be accepted by the customer?

A. Required Percentage =
(NORM.DIST(1.02,1.012,0.018,TRUE)-
NORM.DIST(0.98,1.012,0.018,TRUE))

= 0.633919177 [Using MS Excel]

It means that around 63.4% of the nails will be accepted by the customer.

Q. In order to improve percentage accepted, the production manager and the engineers discuss adjusting the population mean and standard deviation of the length of the pins. If the lathe can be adjusted to have the mean of the lengths to any desired value, what should it be adjusted to? Why?

A. The mean length should be 1 because that is what customer wants with a tolerance of 0.02

Q. Suppose mean cannot be adjusted but standard deviation can be reduced, what maximum standard deviation would make the customer accept -

1. 95% of the sample
2. 99% of the sample

Which, according to you, is easier to adjust mean or standard deviation?

A. We know that 95% of the values lie between ±1.96. Therefore,

=(1.02-1.012)/1.96

=0.004082 New Standard Deviation

99% of the values lie between ±2.58. Therefore

=(1.02-1.012)/2.58

=0.003101 New Standard Deviation

Q. The production manager then considers the costs involved. The cost of resetting the machine to adjust the population mean involves the engineers' time and the cost of production time lost.

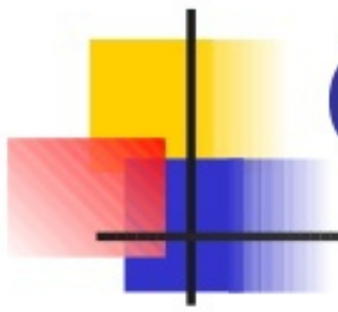
The cost of reducing the population standard deviation involves, in addition to these cost, the cost of overhauling the machine and reengineering the process.

Assume it costs $\$150000x^2$ to decrease the standard deviation by x inch.

Find the cost of reducing the standard deviation to the values found in previous question?

- Amount needed to decrease standard deviation will be
  = ((0.018-0.004082)^2)*150000
  = $29.06

- Amount needed to decrease standard deviation will be
  = ((0.018-0.003101)^2)*150000
  = $33.30



Question

Q. A survey showed that 95% of the users use Internet explorer as their browser. You randomly select 200 people and ask them their browser. What is the probability that exactly 194 say yes?

A. The problem can be approximately as normal having mean np=200*0.95=190 and variance npq=200*0.95*0.05=9.5 which are greater than 5.

P(X=194)=P(193.5<X<194.5)=0.055926456 (approx.)

P(X=194)=0.061401 exact

% error = 8.9%

The example shows the application. It tries to find approximate probability and actual probability and then percentage error.

Conclusion

Normal distributions can be used to describe many real-life situations and are widely used in the fields of science, business, and psychology. They are the most important probability distributions in statistics and can be used to approximate other distributions, such as discrete binomial distributions. The most incredible application of the normal distribution lies in the Central Limit Theorem. This theorem states that no matter what type of distribution a population may have, as long as the sample size is at least 30, the distribution of sample means will be normal. If the population is itself normal, then the distribution of sample means will be normal no matter how small the sample is.