Math 140
Introductory Statistics

Math 140 tutoring:
LIVE OAK 1319

MW 11:30 - 4:30 TTh 3:30-5:30 F 10:30-12:30

General Hours:

M - Th 10:00 - 5:30 F 10:00 - 3:00

Later: Saturday from 11 to 2.

Last time

Westvaco laid off 3 people at round 2

55, 55, 64

Their average age was 58

The average age of all workers was 46.4

The average age of retained workers was 41.4

Was it by chance?

Let's do a simulation - group of 3

Model of a chance process

Consider all ages of workers at Round 2 whose ages are

25 33 35 38 48 55 55 55 56 64

Let's do a simulation - group of 3

Model of a chance process

Person ONE will label them at random

25	33	35	38	48	55	55	55	56	64
F	J	D	G	A	H	I	E	C	B

(example)

Let's do a simulation - group of 3

Model of a chance process

Persons TWO and THREE will simulate cases of of three people to be laid off from A to J

example 1 - Layoffs: B, F, A
example 2 - Layoffs: C, A, I
…
example 5 - Layoffs: F, G, A

5 per person, so that you have a sample of 10 simulation events

Let's do a simulation - group of 3

Model of a chance process

Match layoff letters with ages
Calculate average age of laid off people

You should end up with TEN estimates
Put them in a dot plot

Questions:

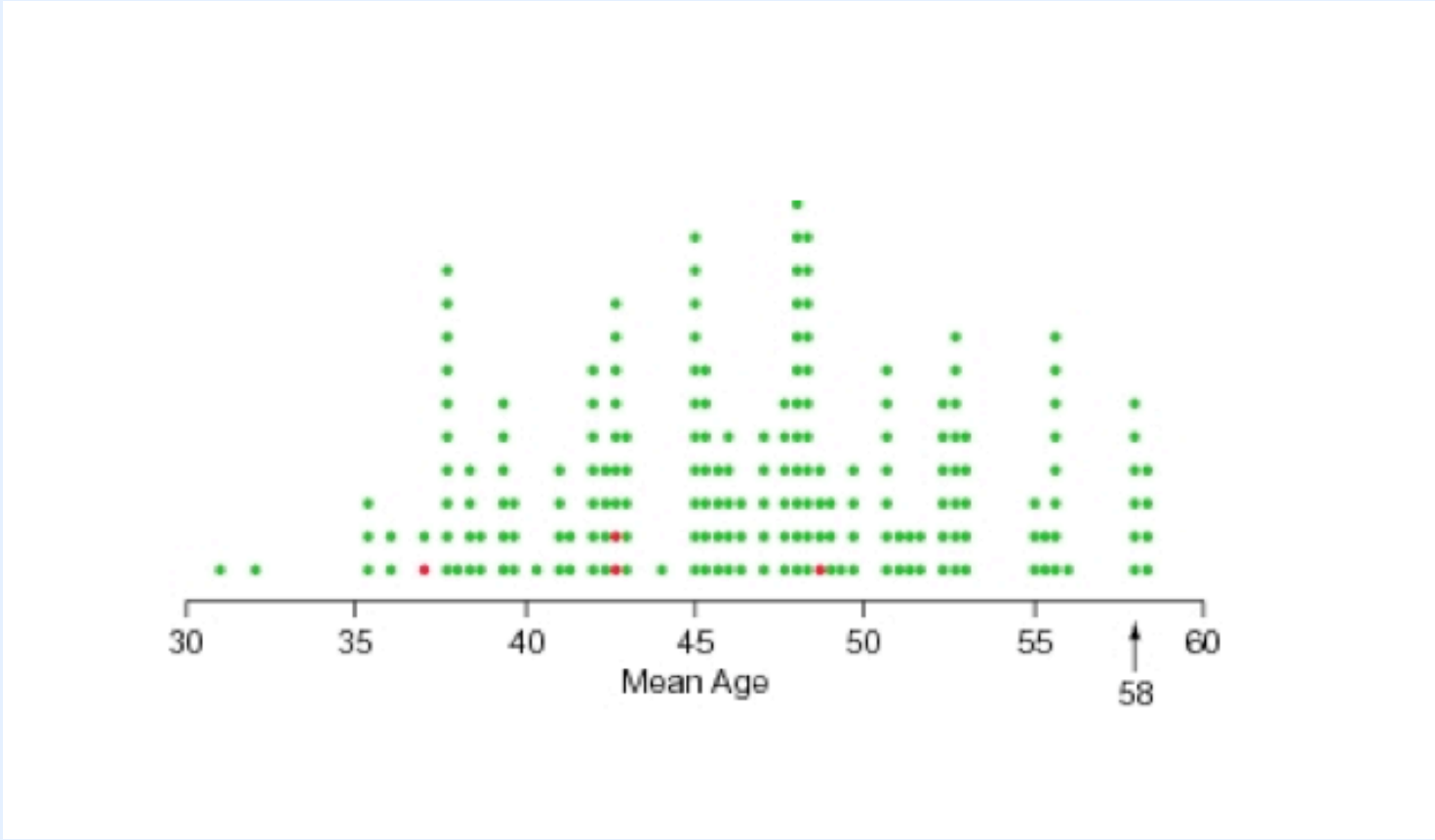
What is the largest possible average?
What is the lowest possible average?

What to do with this?

How many times did we find that the average layoff age was equal or greater than 58?

We should have roughly 400 estimates in total in this class

In this case only 5% of outcomes were higher or equal to 58 years



some results were 42.7, 48.0, 42.7, 37.0

Inference

Getting an average of 58 or more is not impossible

It COULD HAVE BEEN BY CHANCE!

But it is extremely unlikely

The lay off being by chance is

POSSIBLE but UNLIKELY

Other reasons?

Martin vs. Westvaco

The statistician hired by Mr. Martin argued that the probability of getting an average layoff age larger than Westvaco's average layoff age, was only 5%.

The layoffs were most likely not randomly distributed in age.

There could have been other reasons though.

The law sets the limit at 2.5%

The case was settled out of court

Simulation steps

Define your random model

Calculate summary statistic (for us average age)

Repeat many times

Display the distribution

Estimate the probability (for us about 5%)

Come to some conclusion

How do we make sense of it?

We compared the age of 58

to that of the rest of Westvaco's workers

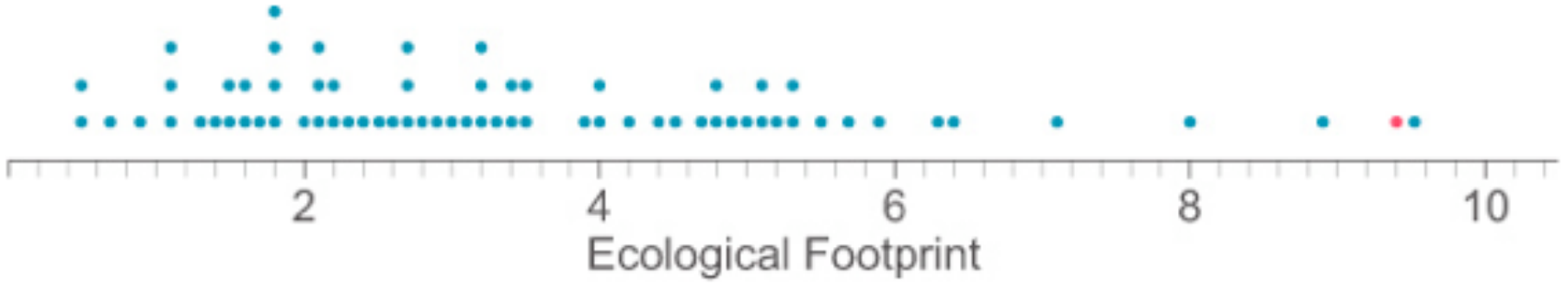
The average number of hectares of the Earth

we use in the USA per person

for our basic needs is 9.7.

Is that a lot or a little?

Visualize the distribution



Number of hectares of earth used per person
to satisfy our basic needs.

The red dot is us - the USA

Shapes of distribution

4 main possibilities:

Uniform or rectangular

Normal

Skewed

Bimodal or multimodal

Uniform distribution

Each outcome occurs roughly the same number of times if we repeat measurements over and over

Uniform distribution

Each outcome occurs roughly the same number of times if we repeat measurements over and over

Examples:

Number of people born per month

Randomly distributed numbers

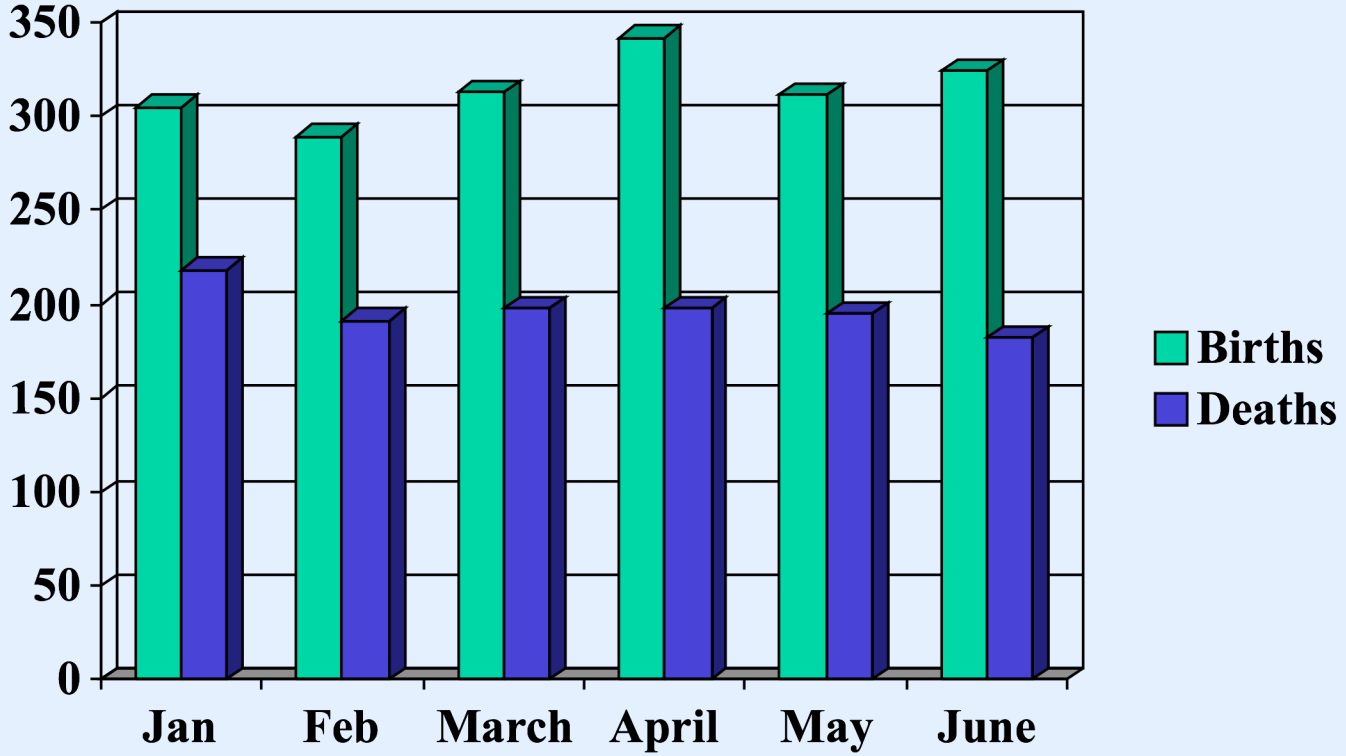
Number of times you get head/tail from coin toss

Number of times you get any number on the roll of dice

Number
of births and deaths
in the USA
for the year 1997
(in thousands)

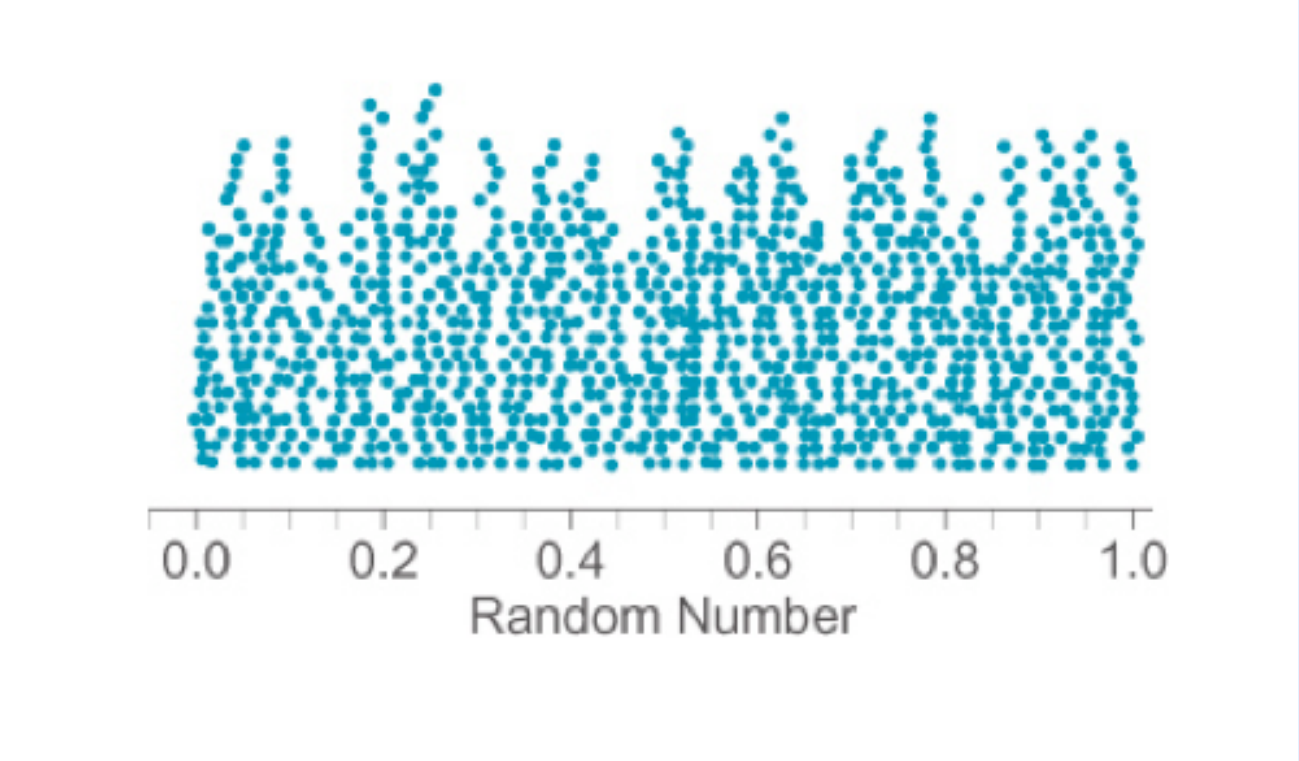
Months	Births	Deaths
January	305	218
February	289	191
March	313	198
April	342	189
May	311	195
June	323	182
July	345	192
August	341	178
September	353	176
October	329	193
November	304	189
December	324	192

Plot the distributions



Uniform or rectangular distribution

1000 random numbers



generated by a computer

Where is center? What is the spread like?
What does it mean?

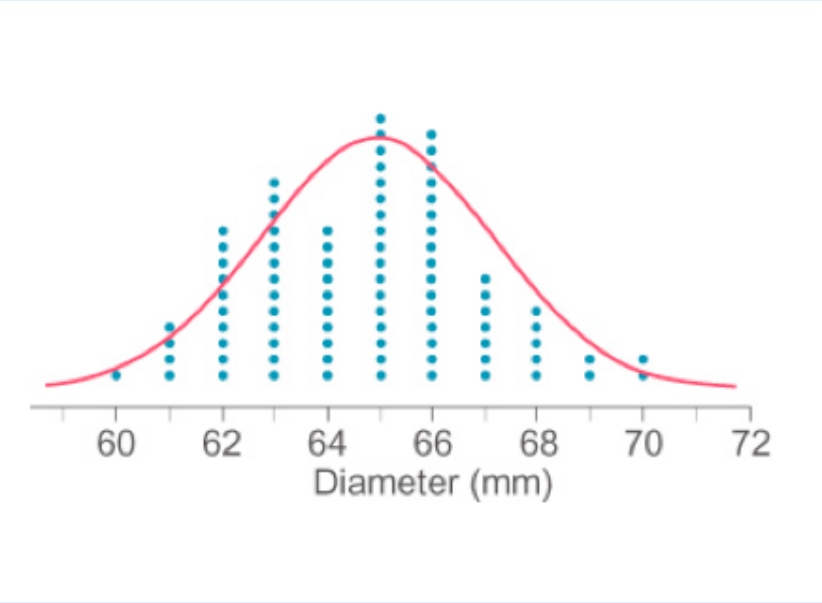
Other examples?

What should not be uniformly distributed?

Normal distributions

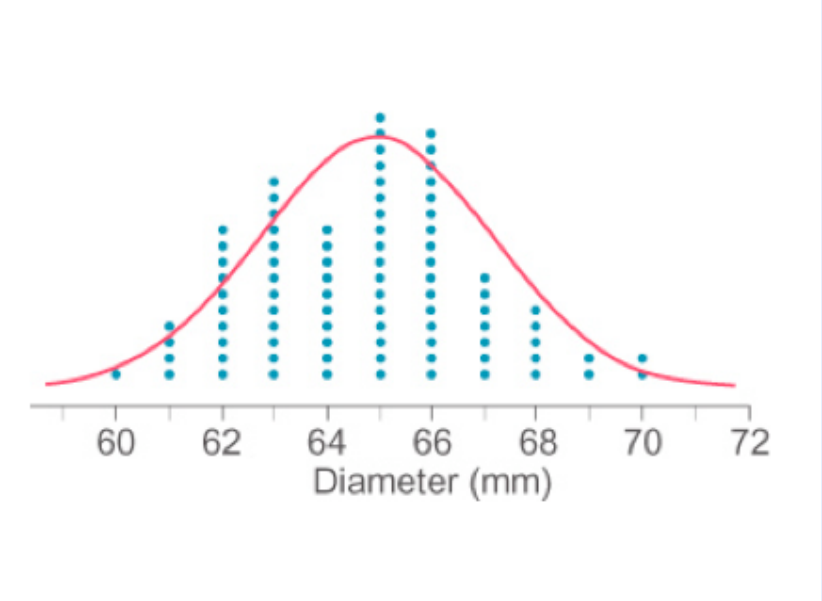
Objects that are manufactured

(diameter of pennies, tennis balls)



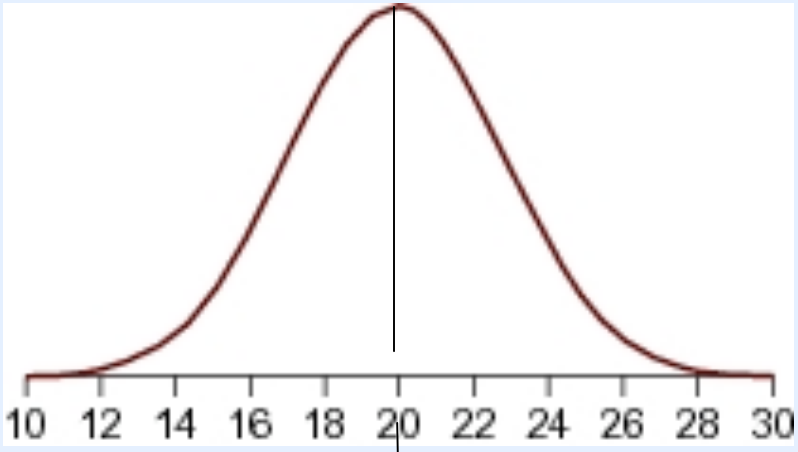
Tennis ball diameter - what do you observe?

Normal distributions

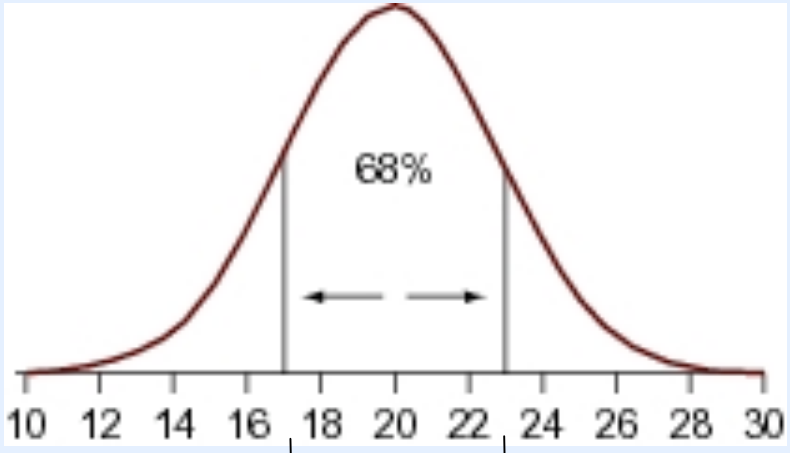


Bell shaped around a maximum
Symmetric left and right

Idealized normal distributions

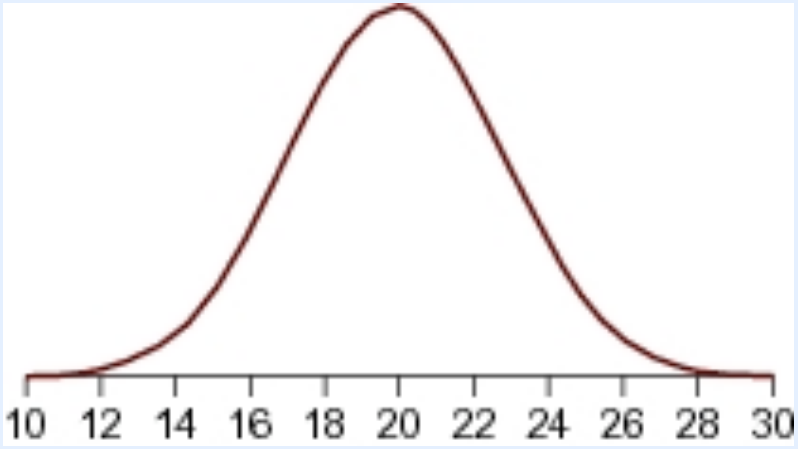


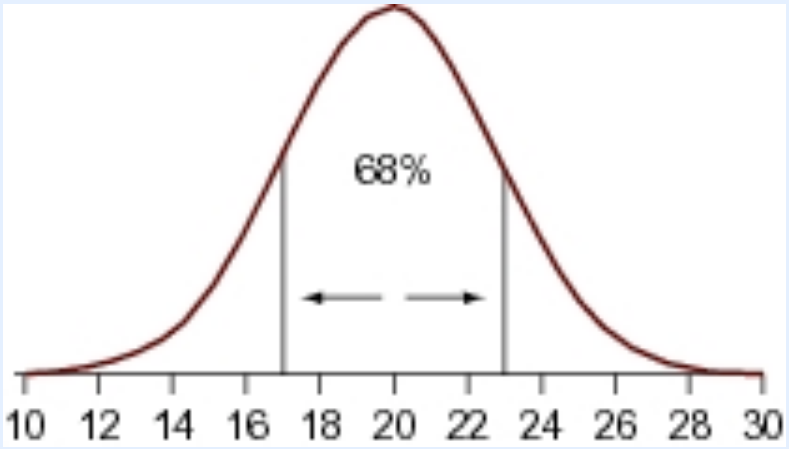
mean



inflection points

Idealized normal distributions



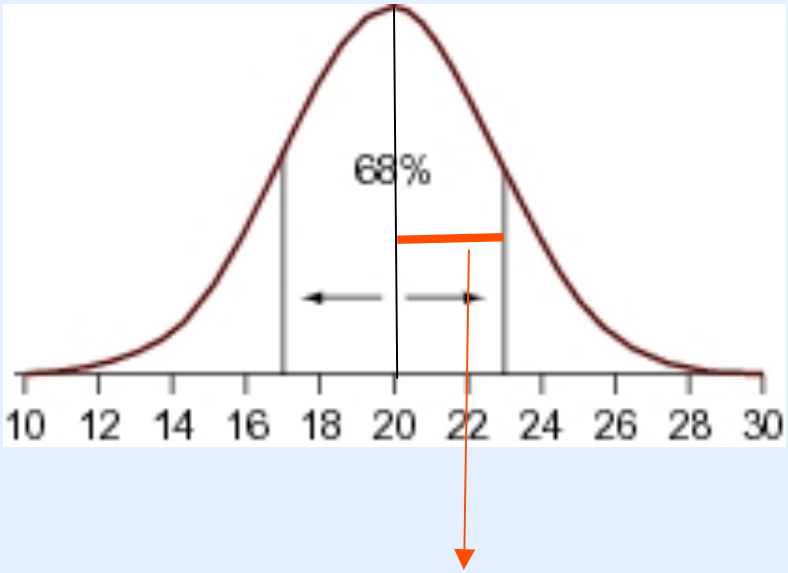


Center value is called the "mean".

The distribution is symmetric
with respect to the mean.

The concavity changes at the "inflection points"
Roughly 2/3 of the area below
the curve is between the inflection points.

Idealized normal distributions

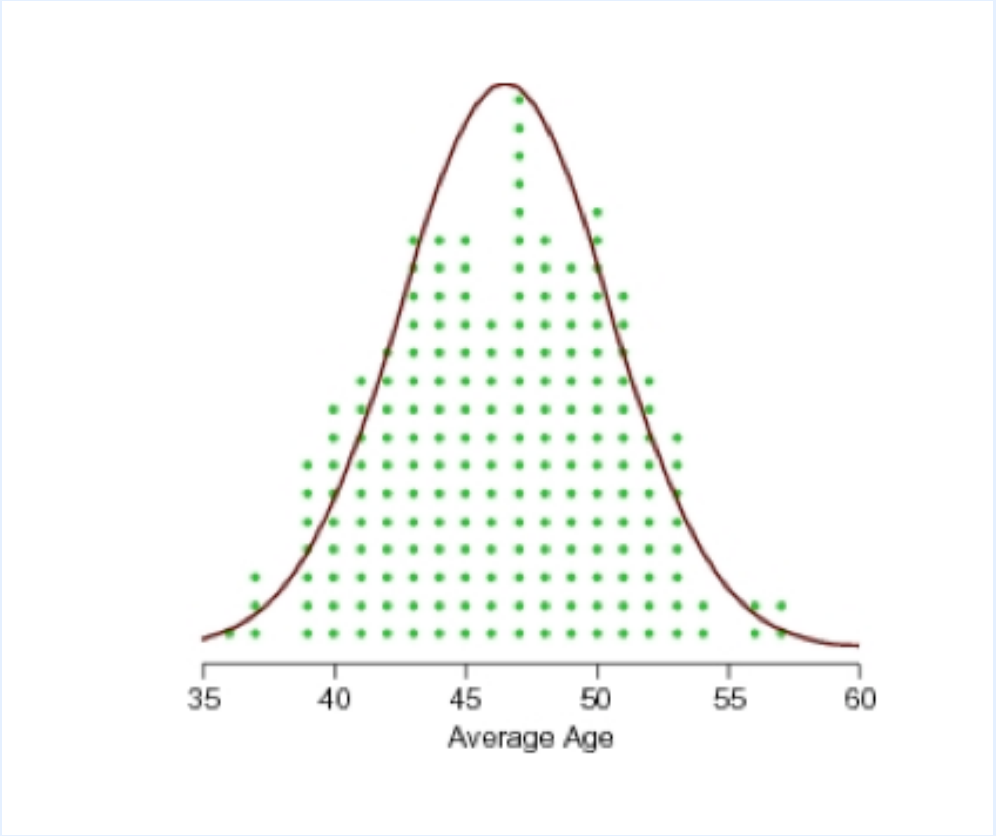


The distance between the mean and either of the inflection points is called

Standard Deviation (SD)

and measures how spread the distribution is

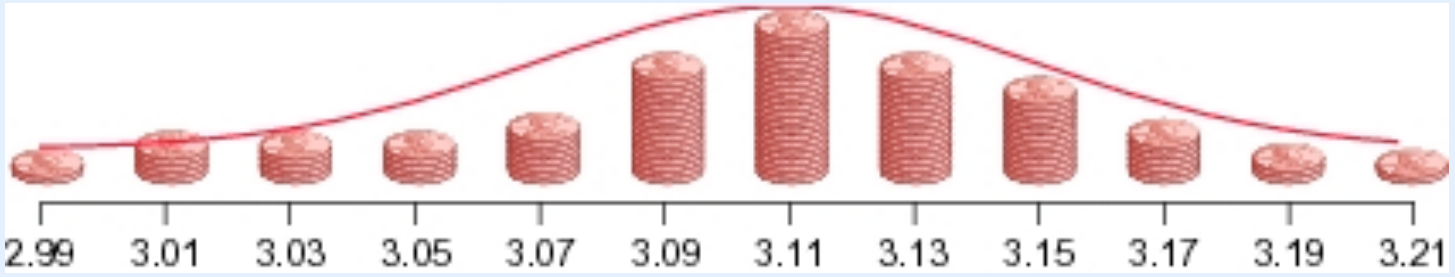
An example



center is at 47, about 68% of ages fall between 43 and 51

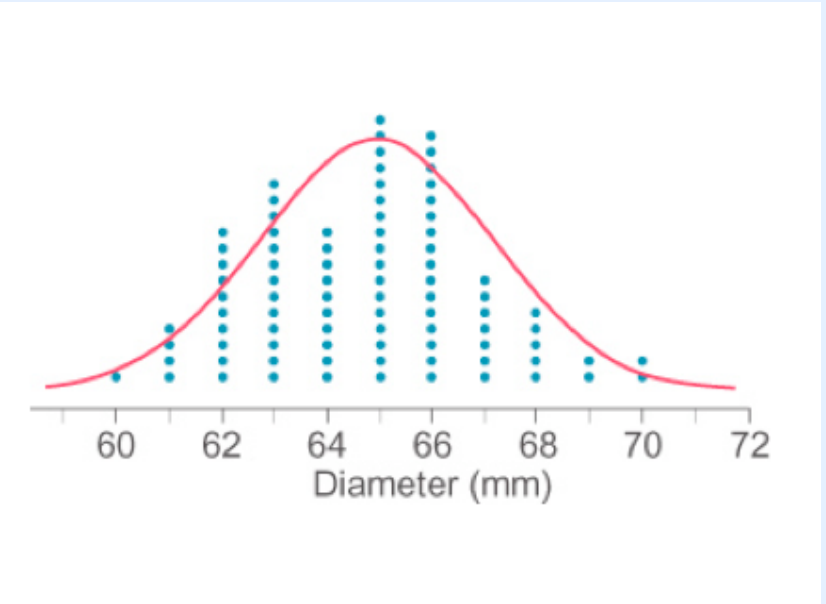
We say the mean is 47 with SD of 4
The average age is 47 give or take 4.

Let's try



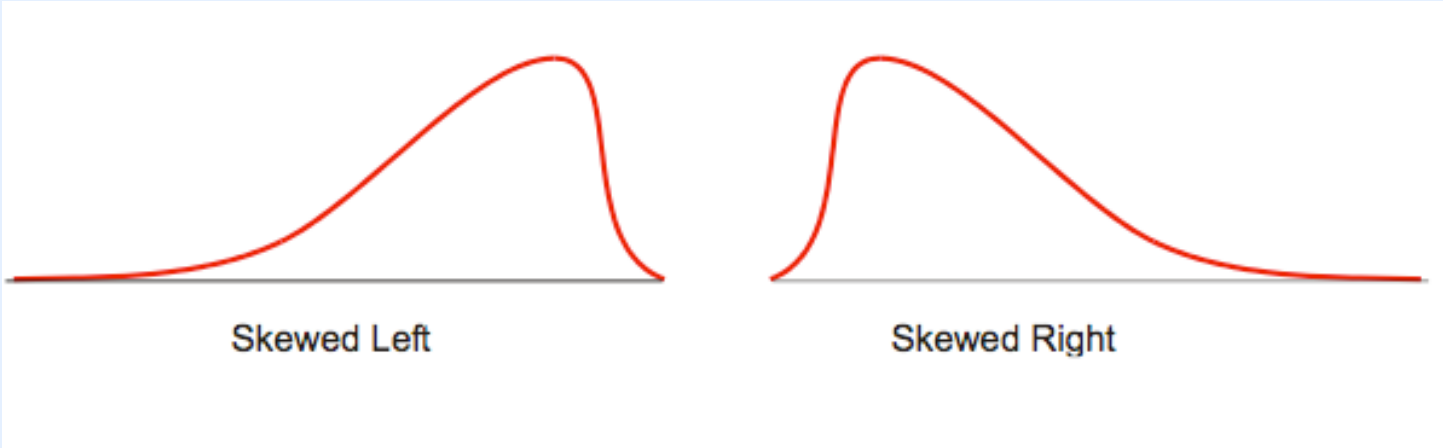
Weight of pennies (grams)

Diameter of tennis balls (mm)



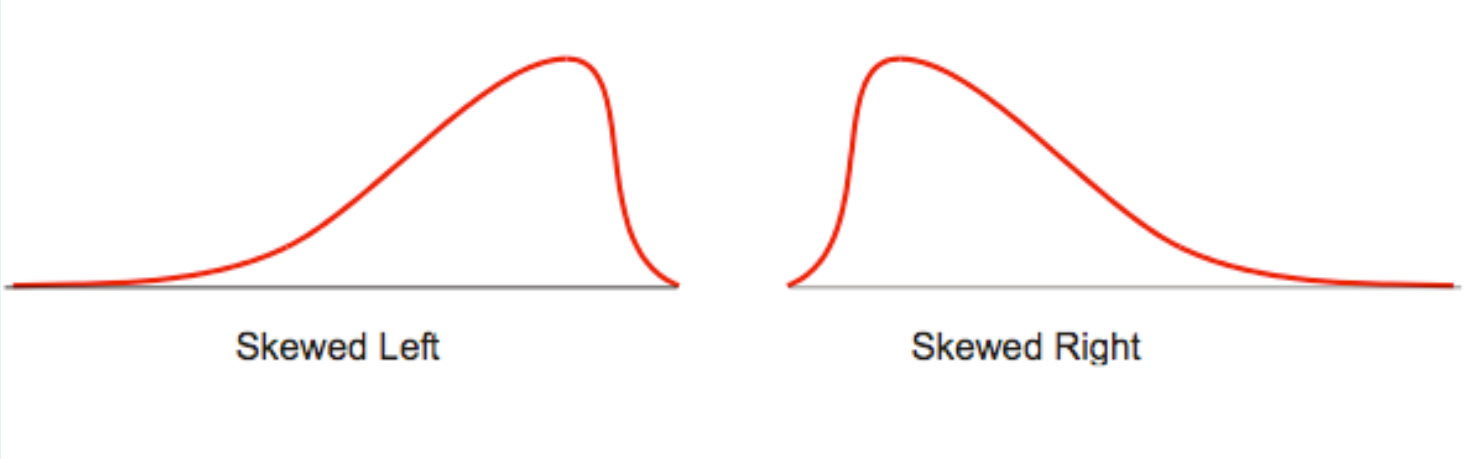
Estimate mean and SD

Skewed distributions



Not symmetric curves
Data is bunched on one end
and a tail appears on the other side

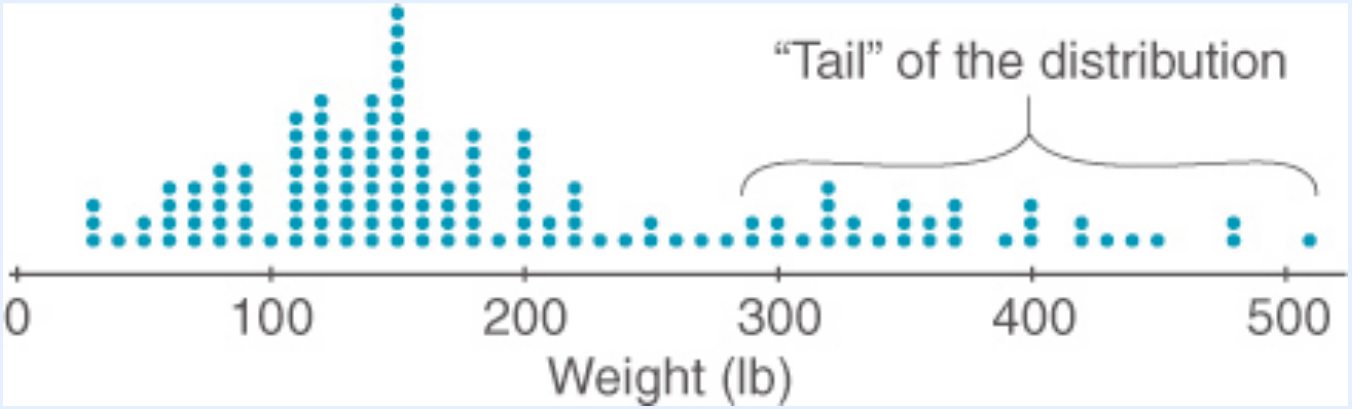
Skewed distributions



This happens because there is a ‘wall’
A value of data you cannot go beyond.

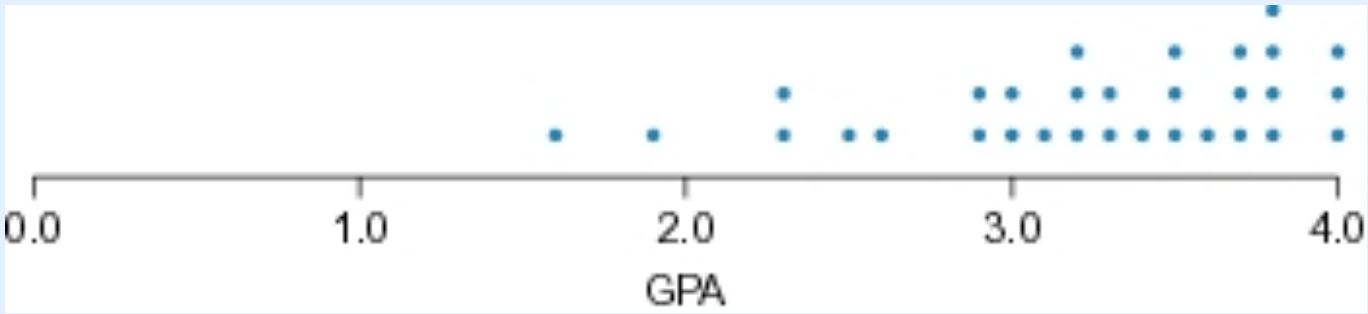
Maybe 0 (for things like counts or measurements)
or 100 (for percentiles)

Skewed distributions



The weight of bears in pounds

Skewed distributions



Your GPAs - can't go beyond 4!

We should look at quartiles and medians to investigate better

Do in class and homework

Problems E13, E14, E17, E21 (page 16)

E1, E4, E5, E6, E10, E15 (page 42)

For problems E13 and E21

Number of ways to pick 3 objects from group of 4

$$\binom{4}{3} = 4$$

$$\binom{6}{2} = 15$$

Number of ways to pick 2 objects from group of 6