DSC 10, Spring 2018
Lecture 18

Normal Distributions

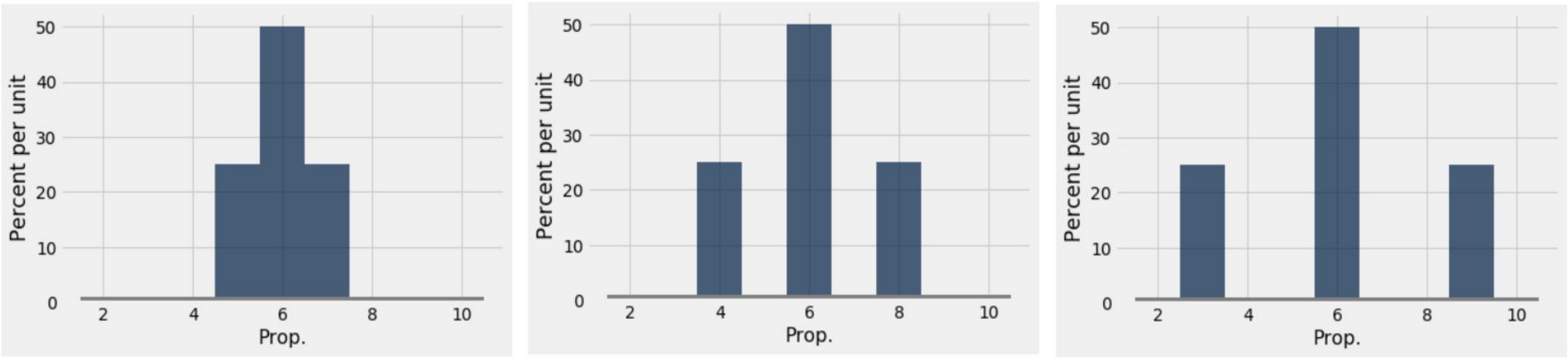
sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Measuring Variability

Center and Spread

- The mean is a measure of **center**.
    - An alternative measure of center is the median.
- Different data sets can have the same mean, but different **spread** or **variability** around that mean.



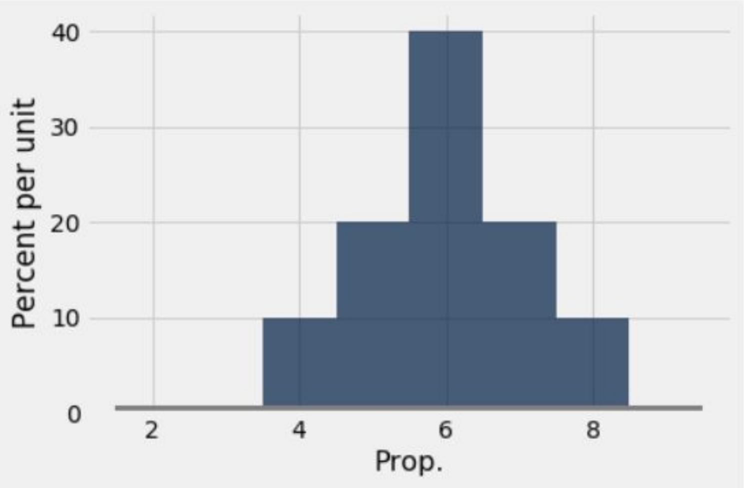
Defining Variability

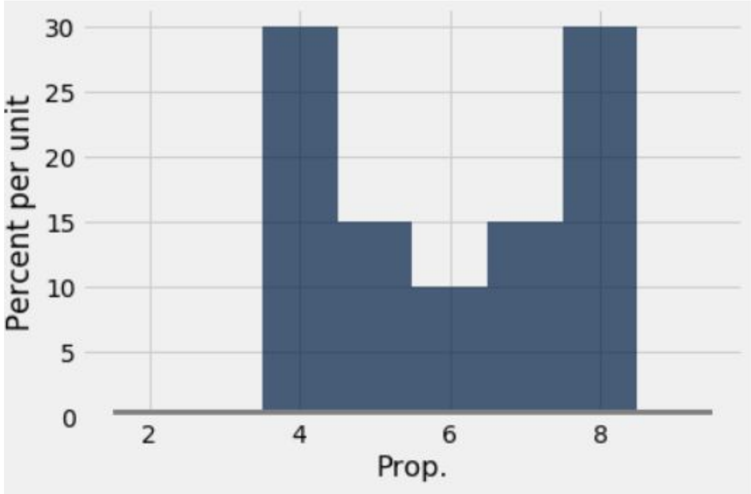
Plan A: “largest value - smallest value”

Defining Variability

Plan A: “largest value - smallest value”

- Doesn’t provide information about the shape of the distribution





Defining Variability

Plan A: “largest value - smallest value”

- Doesn’t provide information about the shape of the distribution

Plan B:

- Measure how far the data is from the mean
- Need a precise way to quantify this

(Demo)

How Far from the Average?

- Standard deviation (SD) measures roughly how far the data are from their average

- SD = root mean square of deviations from average
  5 4 3 2 1

- SD has the same units as the data; hence OK to say "average plus or minus a few SDs"

Discussion Question

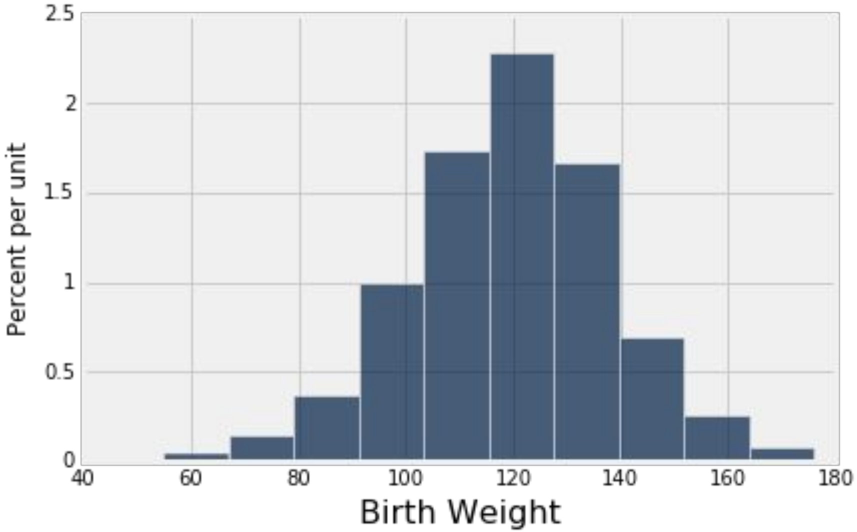
Three gorilla siblings are 2, 3, and 4 years old.

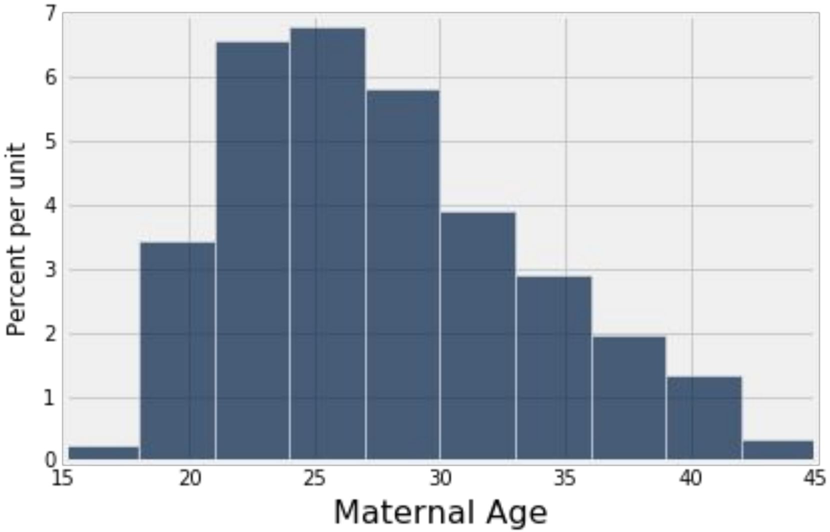
What is the standard deviation of gorilla ages?

A. 1
B. 2
C. sqrt(2)
D. sqrt(⅔)
E. None of the above.

SD = root mean square of deviations from average

Which Has Larger SD?





A. Birth Weight (Left)
B. Maternal Age (Right)
C. Cannot tell from the histograms

(Demo)

Standard Units

Standard Units

- How many SDs above average?
- ***z* = (value - mean)/SD**
  - Negative z: value below average
  - Positive z: value above average
  - z = 0: value equal to average
- When values are in standard units: average = 0, SD = 1
- Most values of *z* are between -5 and 5 (later)

(Demo)

Chebyshev's Inequality

How Big are Most of the Values?

No matter what the shape of the distribution,
<u>the bulk</u> of the data falls in the range "average ± <u>a few</u> SDs"

Chebyshev's Inequality

No matter what the shape of the distribution,
the proportion of values in the range "average ± z SDs" is

at least $1 - 1/z^2$

Chebyshev's Bounds

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)
average ± 5 SDs	at least 1 - 1/25 (96%)

No matter what the distribution looks like

(Demo)

The Normal Distribution

The SD and the Histogram

- Usually, it's not easy to estimate the SD by looking at a histogram

- But if the histogram has a bell shape, then you can

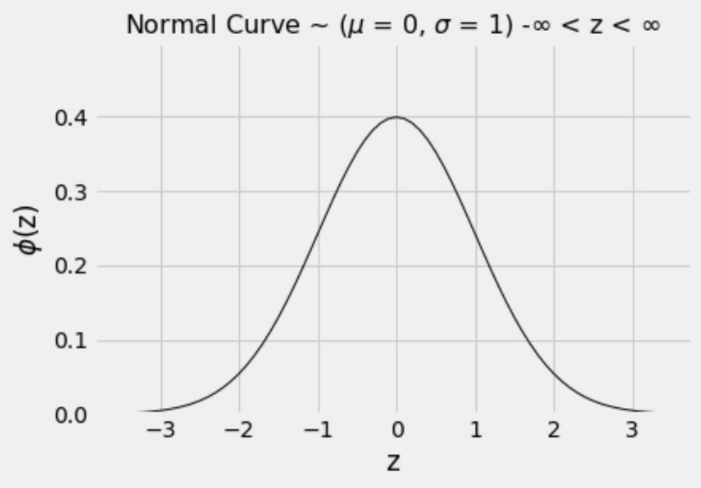
The SD and Bell-Shaped Curves

If a histogram is bell-shaped, then

- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

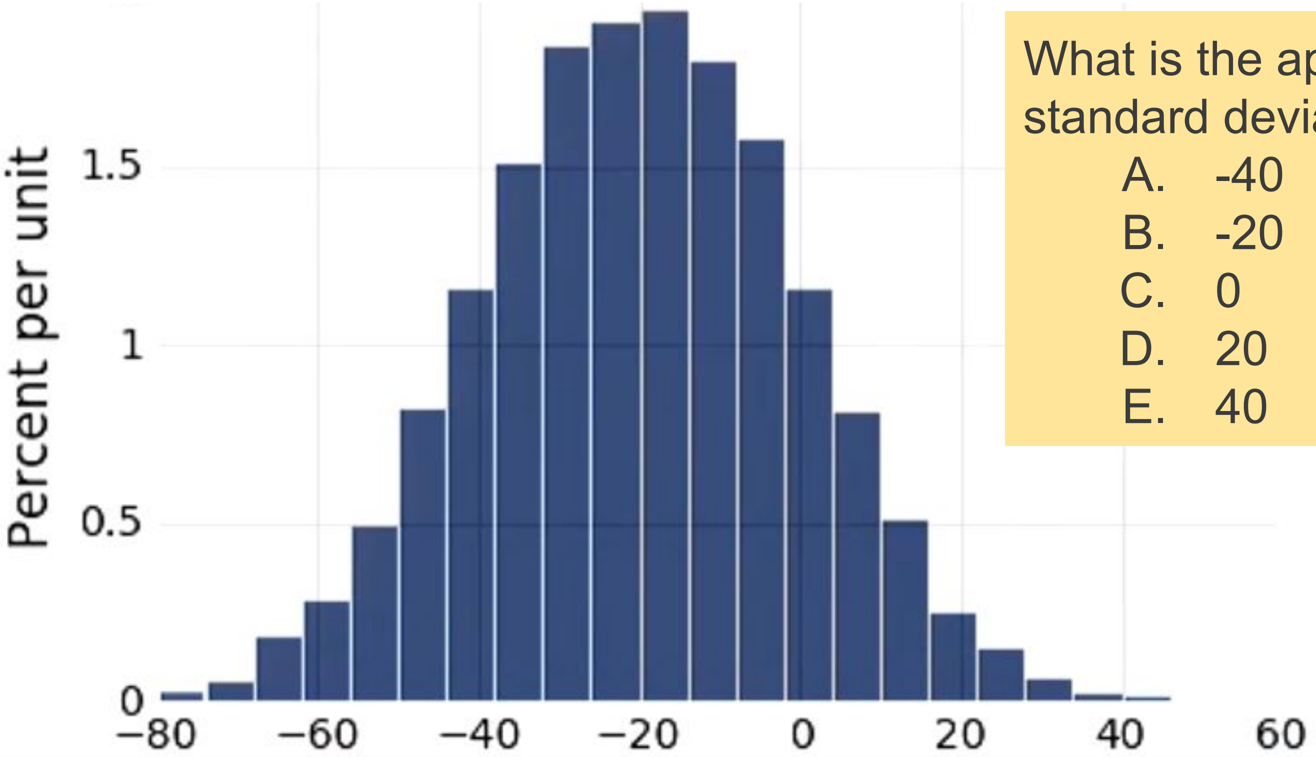
The Standard Normal Curve

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty$$



(Demo)

Question



What is the approximate standard deviation?

A. -40
B. -20
C. 0
D. 20
E. 40

Normal Proportions

How Big are Most of the Values?

No matter what the shape of the distribution,
the bulk of the data falls in the range “average ± a few SDs”

If a histogram is bell-shaped,
almost all of the data falls in the range “average ± 3 SDs”

Bounds and Normal Approximations

Percent in Range	All Distributions	Normal Distribution
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

(Demo)

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the probability distribution of the sample sum (or of the sample average) is roughly bell-shaped

(Demo)