DATA 8

Summer 2018

Lecture 24

The Normal Distribution

Contributions by Vinitra Swamy (vinitra@berkeley.edu) and Fahad Kamran (fhdkmrn@berkeley.edu)
Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)

Announcements

Questions for This Week

- How can we quantify natural concepts like “center” and “variability”?

- Why do many of the empirical distributions that we generate come out bell shaped?

- How is sample size related to the accuracy of an estimate?

Standard Deviation (Review)

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

Why Use the SD?

There are two main reasons.

- **The first reason:**

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

- **The second reason:**

Coming at the end of lecture.

Chebyshev's Inequality

The Mathematician’s Name

- Chebyshev
- Chebychev
- Chebishov
- Čebyšev
- Tchebichev
- Tchebicheff
- Tschebyscheff
- Tschebyschew
- **Чебышёв**

Pafnuty ChebyshevPafnuty C

How Big are Most of the Values?

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

Standard Units

Standard Units

- How many SDs above average?
- ***z* = (value - average)/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
- Chebyshev: At least 96% of the values of *z* are between -5 and 5

(Demo)

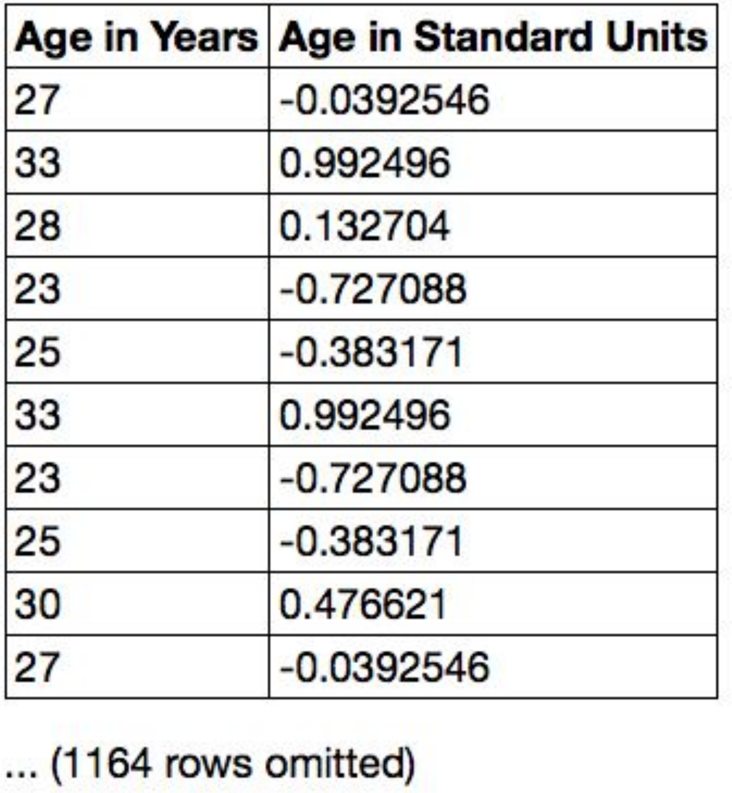
Discussion Question

Find whole numbers that are close to:

(a) the average age

(a) the SD of the ages

(Demo)



Age in Years	Age in Standard Units
27	-0.0392546
33	0.992496
28	0.132704
23	-0.727088
25	-0.383171
33	0.992496
23	-0.727088
25	-0.383171
30	0.476621
27	-0.0392546

... (1164 rows omitted)

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

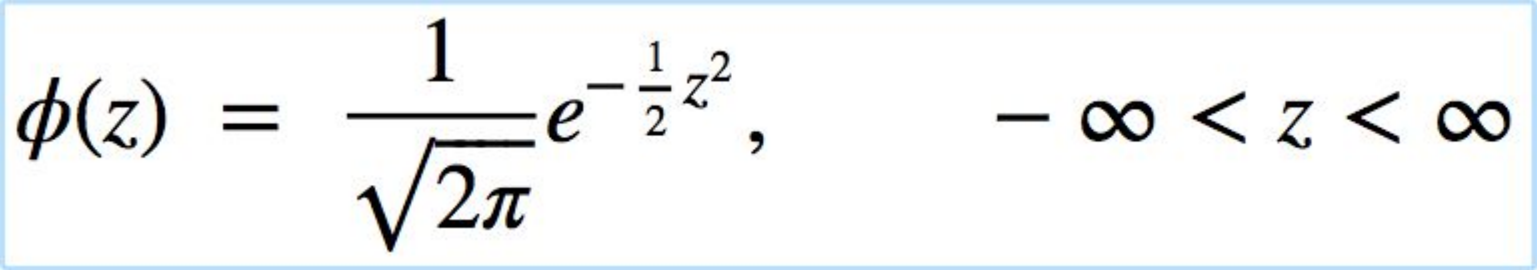
- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

(Demo)

The Normal Distribution

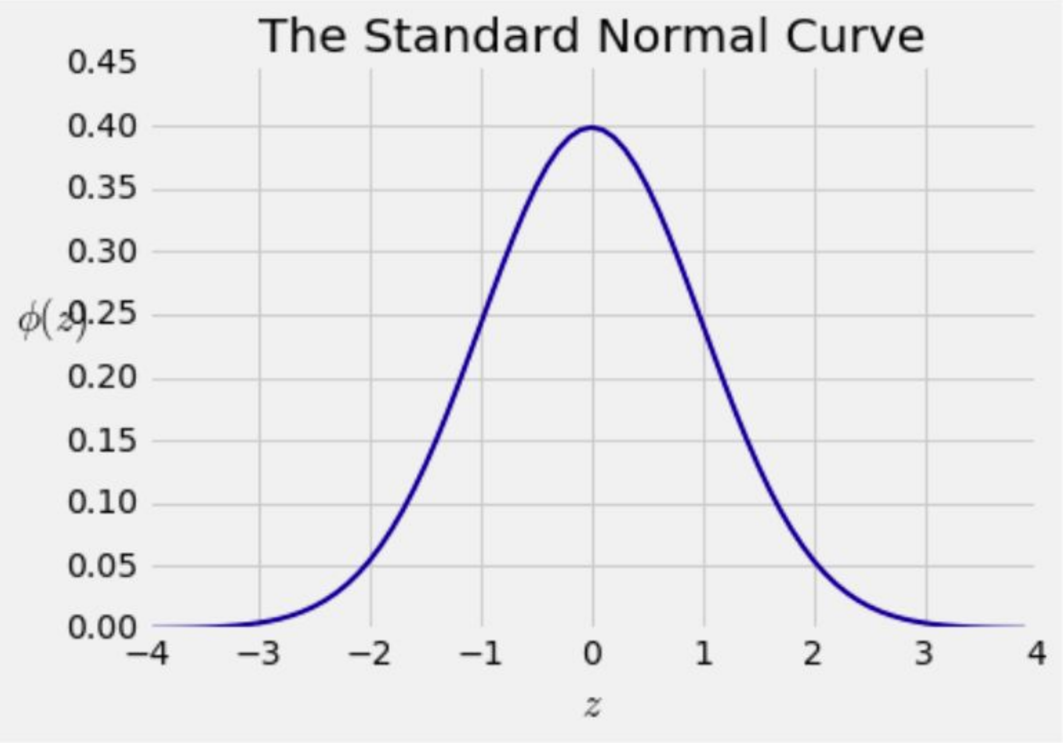
The Standard Normal Curve

A beautiful formula that we won’t use at all:



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

Bell Curve


The Standard Normal Curve
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$\phi(z)$
−4
−3
−2
−1
0
1
2
3
4
z

Normal Proportions

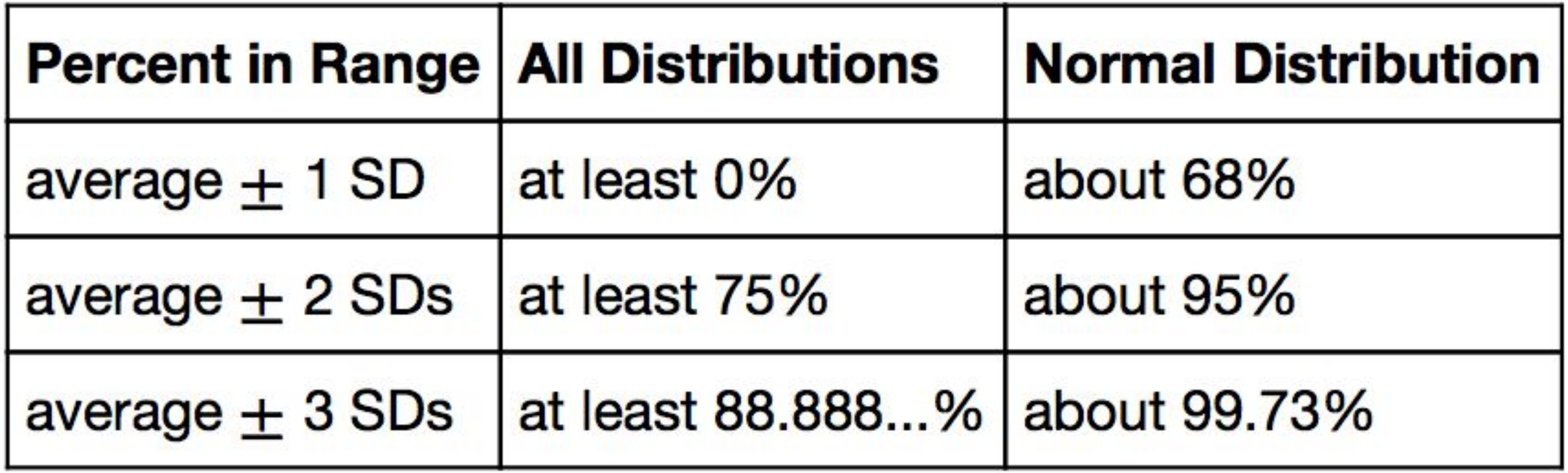
How Big are Most of the Values?

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

If a histogram is bell-shaped, then

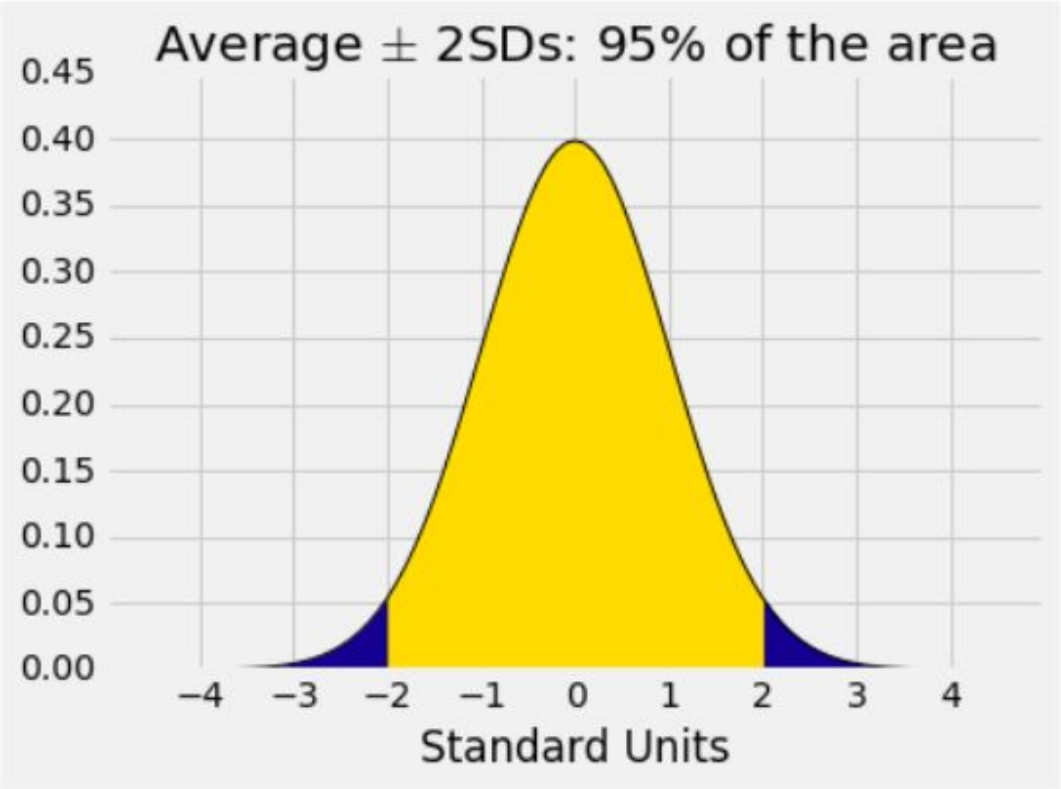
- Almost all of the data are in the range “average ± 3 SDs”

Bounds and Normal Approximations



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

A “Central” Area



Central Limit Theorem

Second Reason for Using the SD

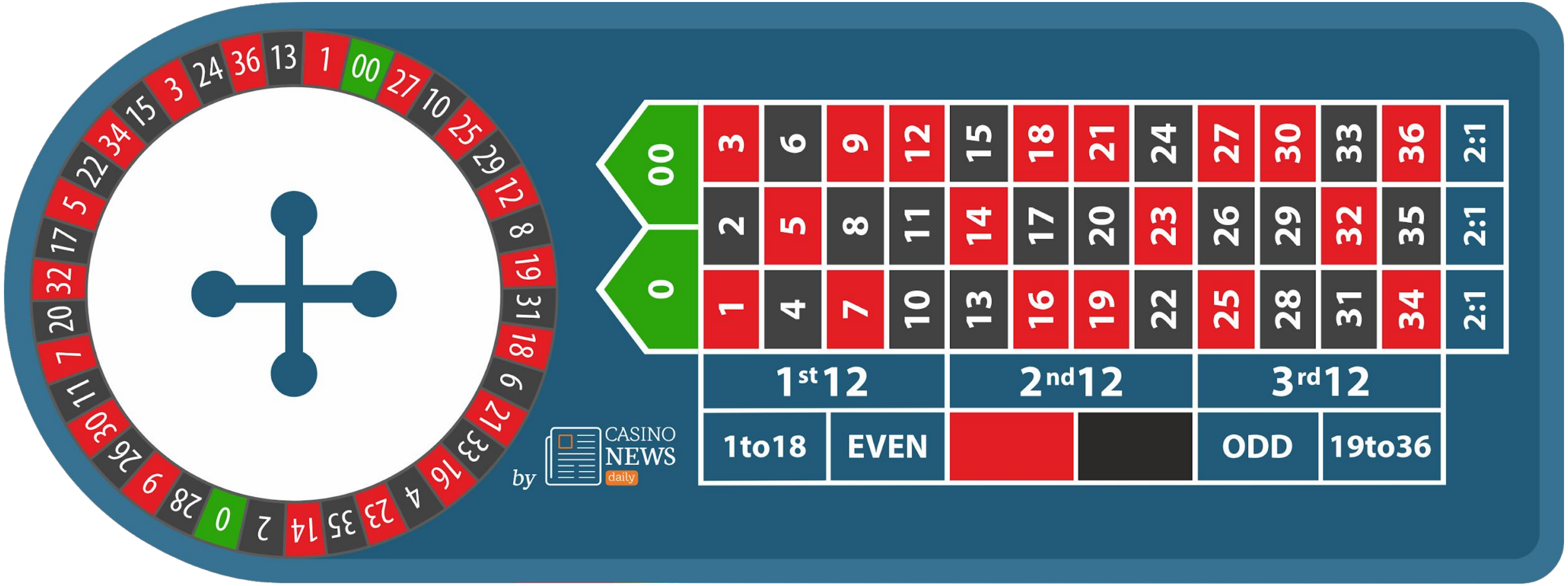
If the sample is

- 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 normal

Roulette



(Demo)