DATA 8

Fall 2016

# Lecture 26, October 26

## SDs and Bell Shaped Curves

Slides created by Ani Adhikari and John DeNero

# Announcements

- Project 2 will be released today!
- Homework due as usual
- I’ve posted on Piazza about courses to consider if you are interested in data science. I have no further info yet. I’ll post on Piazza as soon as I do.

# Standard Deviation

**Standard deviation** (SD)

=

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

Measures roughly how far off the values are from average

# Chebychev'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

“average ± *z* SDs”

- *z* measures “how many SDs above average”
- If *z* is negative, the value is below average
- *z* is called **standard units**
- Almost all standard units are in the range (-5, 5)
- To convert a value to standard units:

$$z = \frac{\text{value - average}}{\text{SD}}$$

# The SD and the histogram

- Usually not easy to estimate the SD by looking at a histogram

- But if the histogram has a special shape, then maybe

(Demo)

# 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 standard normal curve

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

(Demo)

# How big are most of the values?

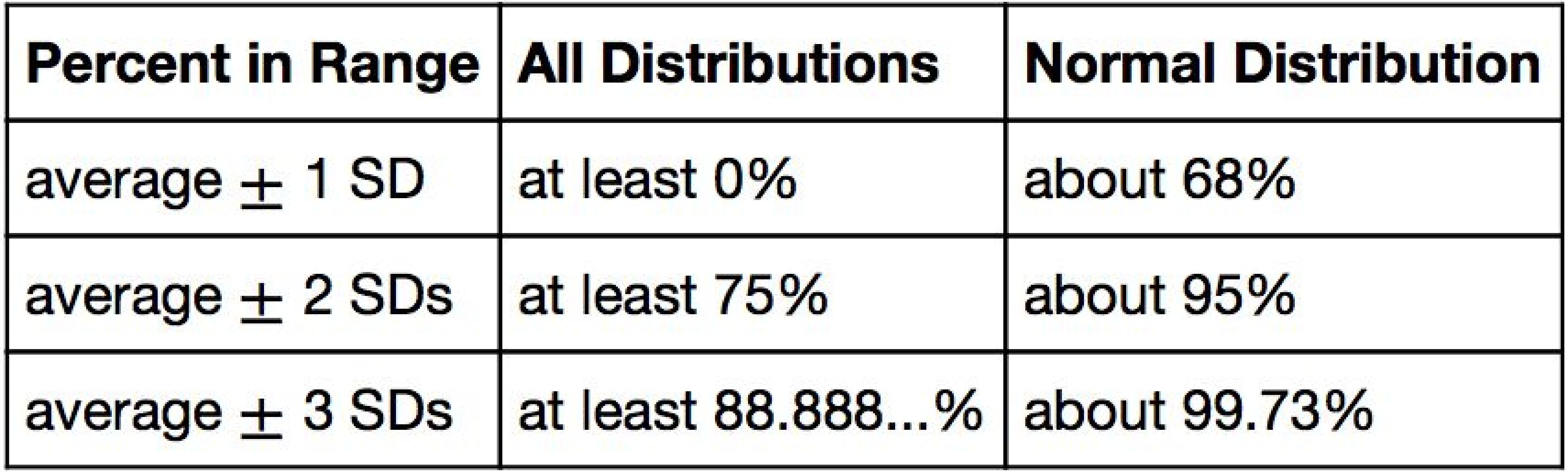
***No matter what the shape of the distribution,***
the bulk of the data are in the range "average ± a few SDs"

***If a histogram is bell-shaped***, then

- the SD is the distance between the average and the points of inflection on either side
- Almost all of the data are in the range "average ± 3 SDs"

(Demo)

# 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% |

(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**