YData: Introduction to Data Science

Lecture 26: center, spread and normal distribution

Overview

Very quick review of the bootstrap

Measures of central tendency and variability

Chebyshev's Inequality

Standardized units

If there is time

- The normal distribution
- The Central Limit Theorem

Measures of central tendency

Measures of central tendency

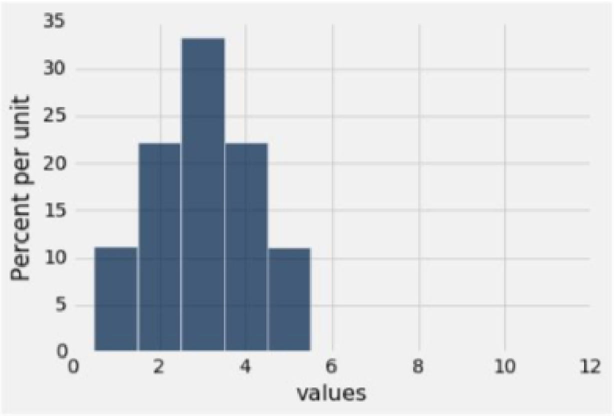
The average (or mean)

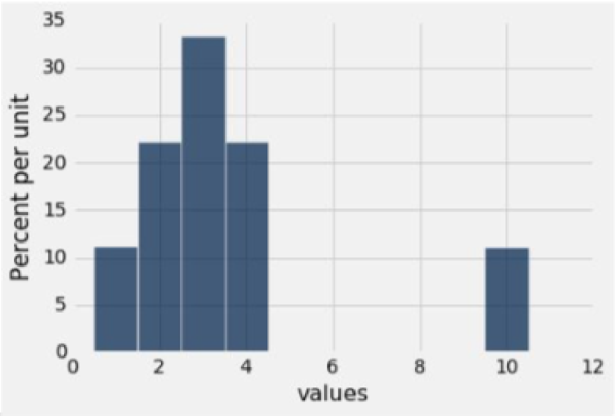
- Data: 2, 3, 3, 9   Average = (2+3+3+9)/4 = 4.25
- Can be heavily influenced by outliers

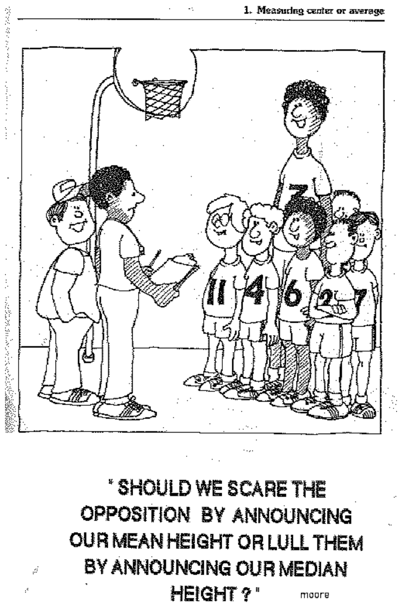
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The median

- Value that splits out data in half
- Resistant to outliers



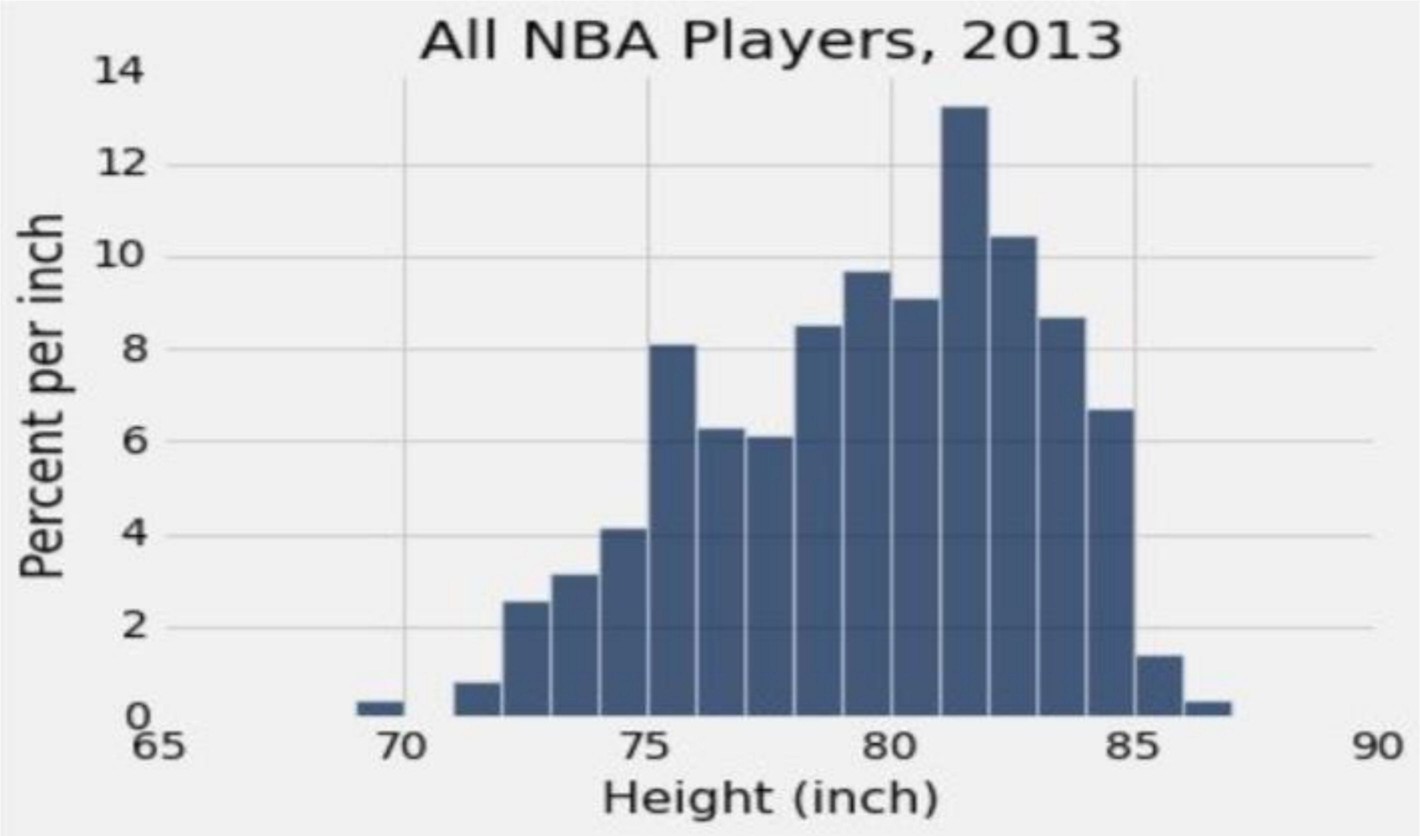




Discussion question

Which is bigger?

- A. The mean
- B. The median



Let's explore this in Jupyter!

The standard deviation

Defining variability

There are many different potential ways to measure variability in our data

Plan A: "biggest value - smallest value"

- Doesn't tell us much about the shape of the distribution

Plan B:

- Measure variability around the mean
- Need to figure out a way to quantify this

How far away from average?

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

SD = root mean square of deviations from average

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

SD has the same units as the data

In Python: np.std(array)

Chebyshev's Inequality

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 ± $z \cdot$ SDs" is at least $1 - 1/z^2$

Range	**Proportion**
Average ± 2 SDs	at least 1 - 1/4 (75%)
Average ± 3 SDs	at least 1 - 1/9 (88.88...%)
Average ± 4 SDs	at least 1 - 1/16 (93.75%)
Average ± 5 SDs	at least 1 - 1/25 (96%)

Let's explore this in Jupyter!

Standardized units

Standardized units

Item in the world are often measured on very different scales

How can we create a standard scale to quantify unusual/large/impressive values?

Z-scores measure how many SDs a value is from average:

$$z = (\text{value} - \text{average})/\text{SD}$$

- Negative z: value below average
- Positive z: value above average
- z = 0: value equal to average

Which Accomplishment is most impressive?

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

The summary statistics of the NBA in 2011 are given below:

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{SD}$$

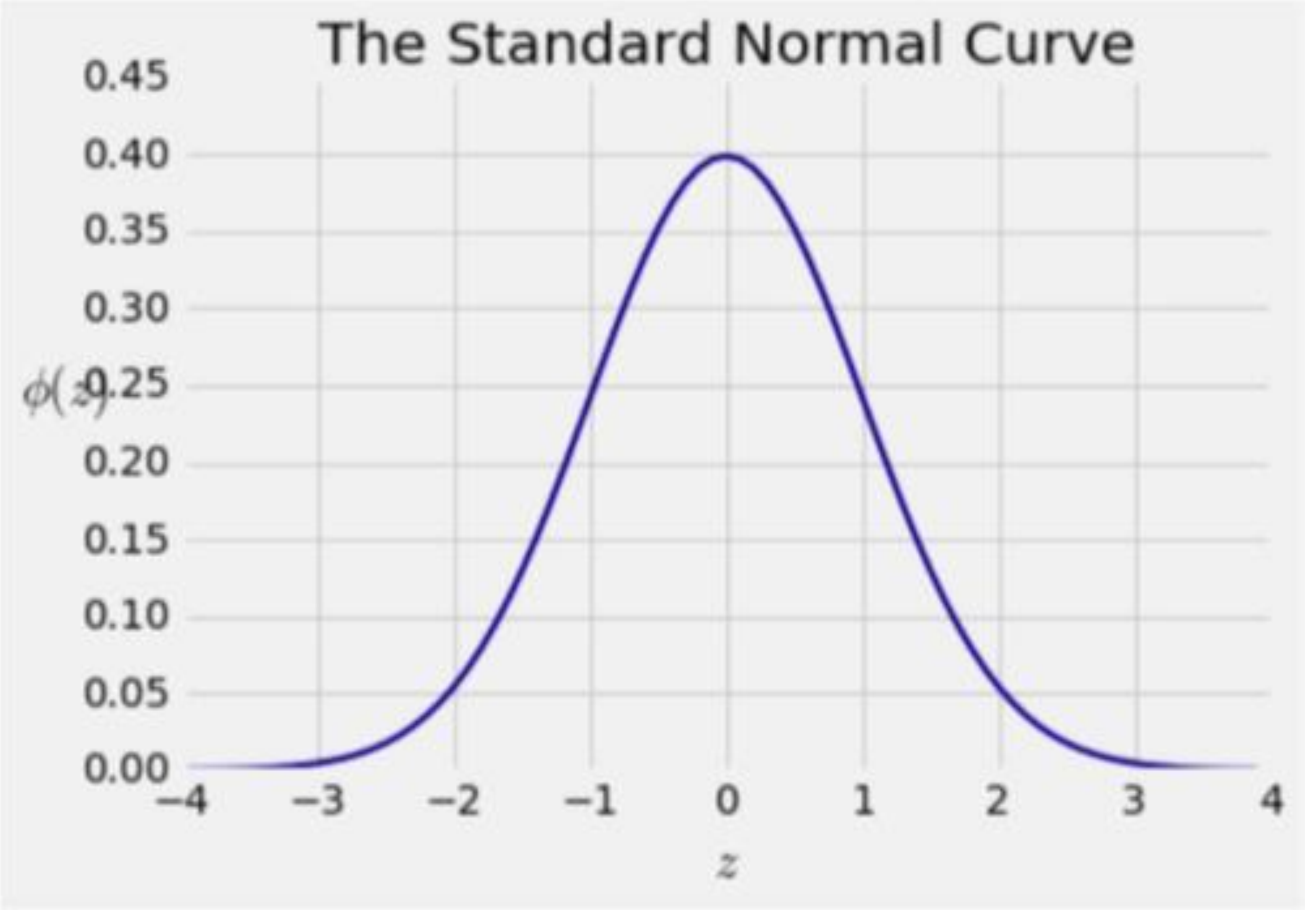
	Mean	Standard Deviation
FGPct	0.464	0.053
Points	994	414
Assists	220	170
Steals	68.2	31.5

Let's explore this in Jupyter!

Question: Relative to his peers, which statistic is most and least impressive?

The normal distribution

The standard normal curve



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

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

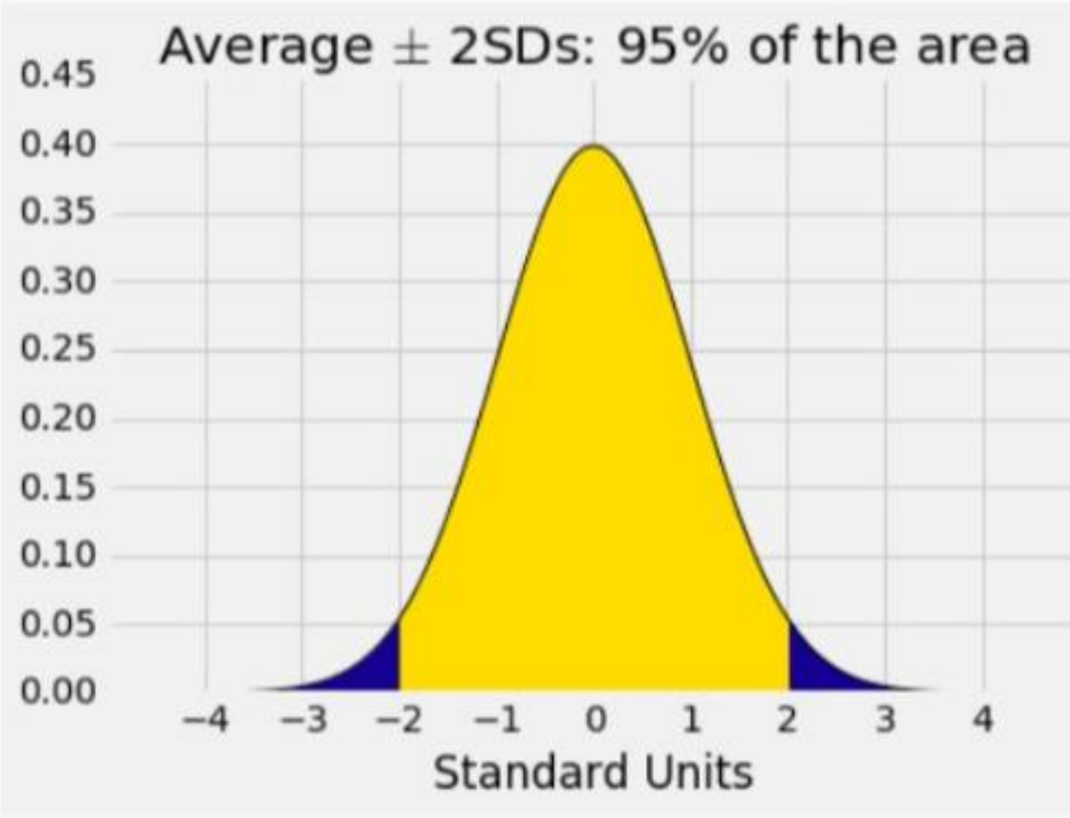
Bounds and normal approximations

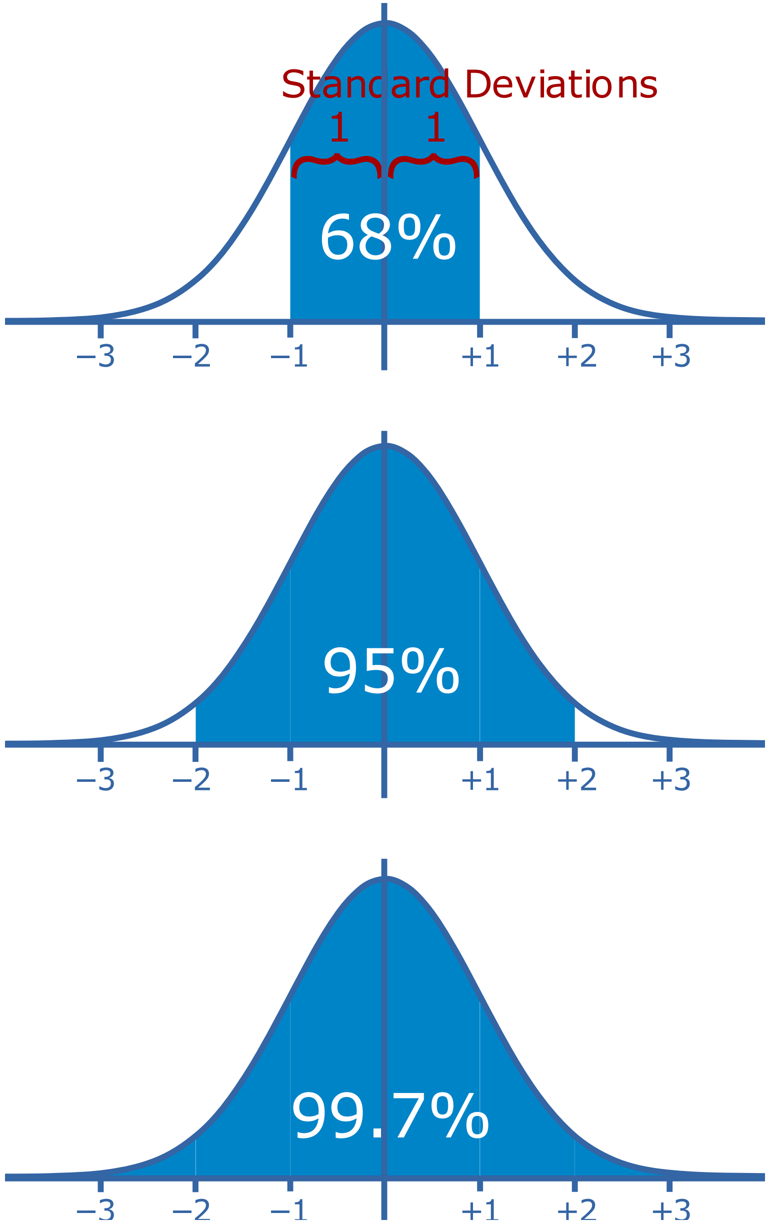
Chebyshev's Inequality: No matter what the shape of the distribution, the bulk of the data are in the range average ± a few SDs"

If a histogram is bell-shaped, then almost all of the data are in the range "average ± 3 SDs"

Percent in Range	All Distributions	Normal Distribution
Average ± 1 SDs	at least 0%	About 68%
Average ± 2 SDs	at least 75%	About 95%
Average ± 3 SDs	at least 88.88%	About 99.73%

The “Central” Area





Let's explore this in Jupyter!

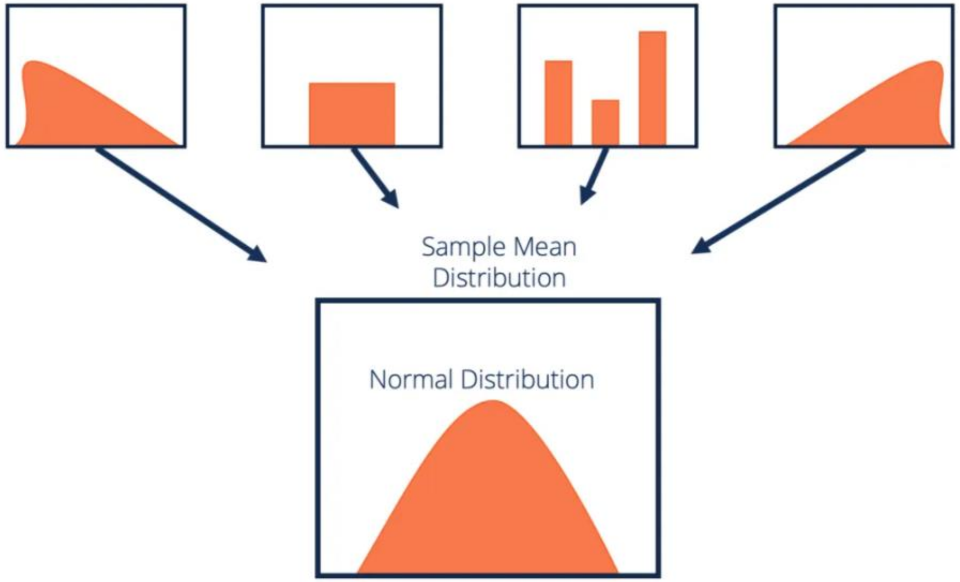
The Central Limit Theorem

The 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 normal



Let's explore this in Jupyter!