Data 8, Lab 8

The Central Limit Theorem, Sample Means, and Correlation

Hubert Luo

Spring 2020

10 April 2020

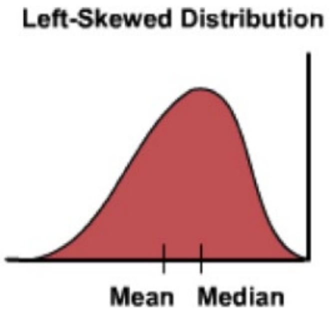
Berkeley
UNIVERSITY OF CALIFORNIA

Agenda

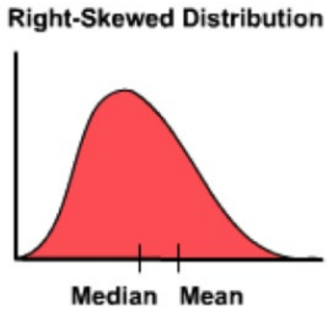
1. Skewness
2. Variability
3. Chebyshev's Bounds
4. Standard Units
5. Normal Distribution
6. Central Limit Theorem
7. Distribution of Sample Means
8. Correlation

Skewness

- Left skew
  - Long left tail
  - Mean < Median



- Right skew
  - Long right tail
  - Mean > Median



Variability

- Variance: How spread out is the data?
- Standard Deviation: Square root of the variance
  - Same unit as the data
  - The larger the SD, the more spread out the data is

Berkeley
UNIVERSITY OF CALIFORNIA

Chebyshev's Bounds

- **Regardless of the distribution**, the proportion of values in the range "average ± z 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 (88.888…%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)
average ± 5 SDs	at least 1 - 1/25 (96%)

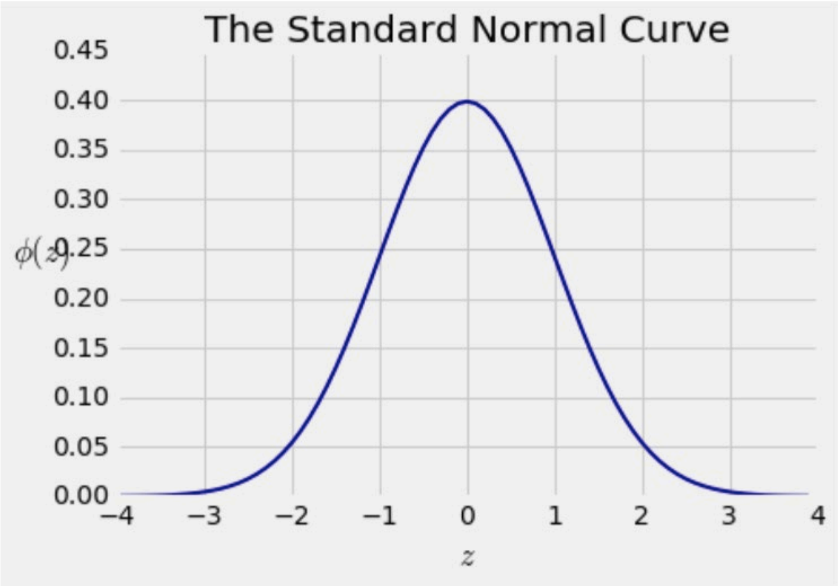
Berkeley
UNIVERSITY OF CALIFORNIA

Standard Units

- Standard Unit: Number of SD's above or below average
- Allows us to easily compare different distributions and units
- Z = (value-average)/SD
- Average of standard units is always 0
- SD of standard units is always 1

Normal Distribution

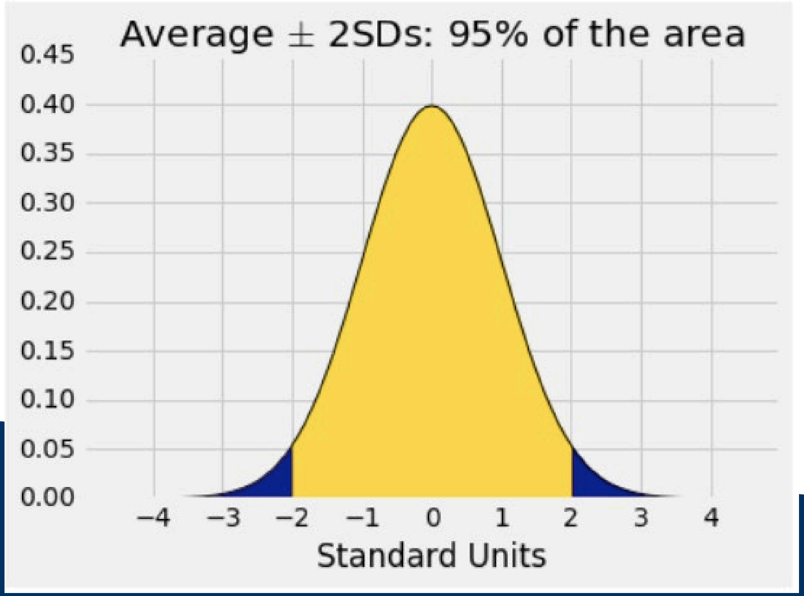
- An extremely common distribution in statistics, shaped like a bell curve
- Most of the data is within a few SD's of the mean



Berkeley
UNIVERSITY OF CALIFORNIA

Normal Distribution (cont'd)

Range	All Distributions	Normal
average ± 1 SDs	at least 0%	68%
average ± 2 SDs	at least 75%	95%
average ± 3 SDs	at least 88.9%	99.7%



Berkeley
UNIVERSITY OF CALIFORNIA

Central Limit Theorem

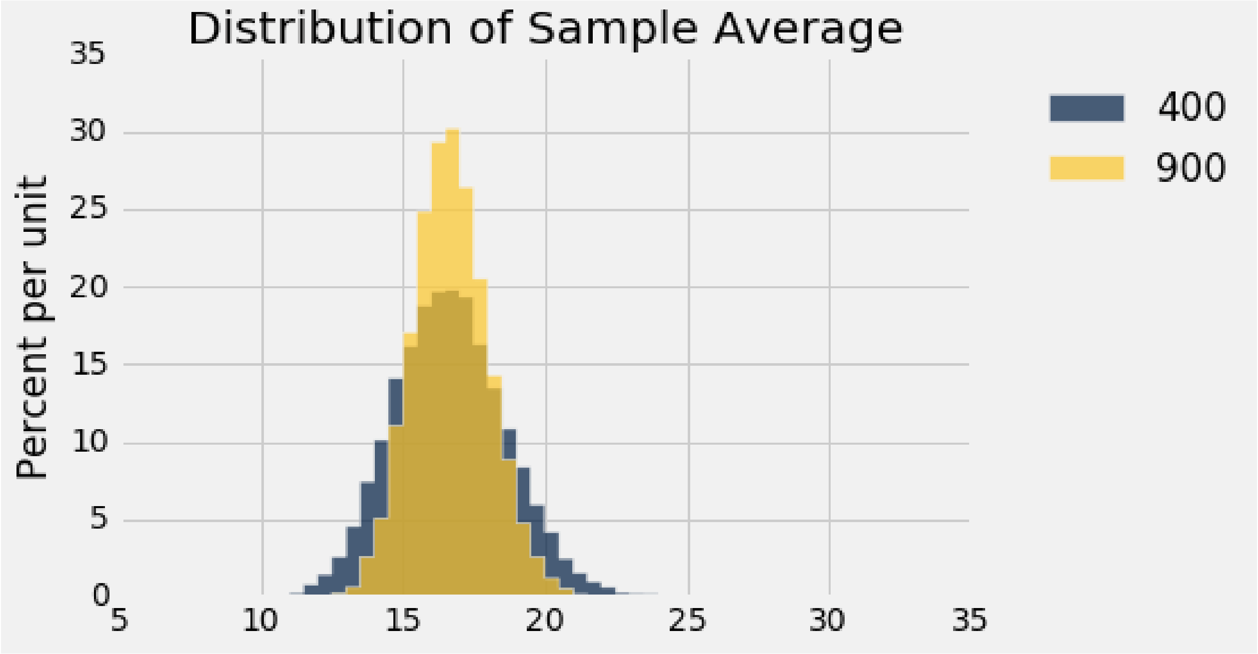
- If the sample is large and drawn at random with replacement
- Regardless of the distribution of the population, the distribution of the sample sum or average is **roughly normal**
- Distribution of the sample sum/average:
  - Many possible random samples of the same size
  - Distribution is based on the sum/average of different samples

Distribution of Sample Mean

- As the sample size increases, the sample mean is more likely to be closer to the population mean
- As a result, the distribution of sample means will have lower SD – a "narrower bell shape" when the sample size increases

$$SD\ of\ Sample\ Means = \frac{Population\ SD}{\sqrt{Sample\ Size}}$$

Distribution of Sample Mean



Berkeley
UNIVERSITY OF CALIFORNIA

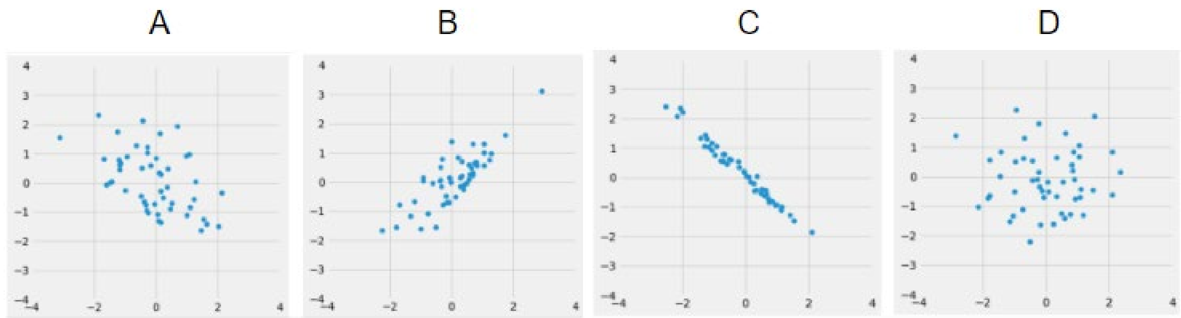
Correlation

- Measure the strength of the linear relationship between two variables
- Correlation must be between -1 and 1 (inclusive)
- When r is positive, there is a **positive linear association** between the two variables
- When r is negative, there is a **negative linear association** between the two variables
- Correlation of x and y is the same as correlation of y and x

Calculating Correlation

- Calculated as r = average of element-wise product of two variables in standard units
- Algorithm:
    1. Given two numeric arrays `x` and `y` of the same length
    2. Convert both `x` and `y` into standard units `x_su` and `y_su`
    3. Calculate array of the product of the arrays in standard units
       `xy_product = x_su*y_su`
        1. The ith element in this product array is the product of the ith element in the `x_su` array and the ith element in the `y_su` array
    4. Correlation is the mean of this array `np.mean(xy_product)`

Correlation Example: Worksheet Q6



A. Small negative correlation: Weak negative linear association since the points are not clustered tightly around a line
B. Positive correlation: Positive linear association since the points are clustered somewhat tightly around a line
C. Strong negative correlation: Strong negative linear association since the points are clustered tightly around a line
D. No correlation: No visible trend since the points are just a blob

Announcements

- Project 2 Checkpoint 2 is due today 4/10
  - Entire project is due next Friday 4/17, bonus point for early submission by 4/16
  - If you're working with a partner make sure you add their email onto Okpy so you both receive credit
- HW10 due on 4/16. Bonus point for early submission by 4/15
- Lab 8 extension by one day to Saturday 4/11 at 11:59PM for this week only
- This semester's lecture on privacy has been cancelled, but you can watch last year's version here. The material covered in that lecture will be on the final.