The Amazing Simpson's Paradox

Simpson's Paradox is another example of how confounding variables can complicate the interpretation of statistical data. It is a classic case of how data can be sliced and diced.

Example #20 — The following data comes from an article in the Florida Law Review that studied the association between race and whether defendants convicted of homicide received the death penalty, based on 674 homicide cases between 1976 and 1987

	Death Penalty			
Defendant's Race	**YES**	**NO**	**TOTAL**	**DP %**
White	53	430	483	11.0%
Black	15	176	191	7.9%

- A higher percentage of white defendants (11.0%) received the death penalty than black defendants (7.9%)

- But when we analyze the data separately for white and black **victims**, the direction of the apparent association is reversed. For both white and black victims, a higher percentage of blacks received the death penalty than whites. What is going on here?

		Death Penalty			
Victim's Race	**Defendant's Race**	**YES**	**NO**	**TOTAL**	**DP %**
White	**White**	53	414	467	11.3%
	Black	11	37	48	22.9%
Black	**White**	0	16	16	0%
	Black	4	139	143	2.8%
TOTAL	**White**	53	430	483	11.0%
	Black	15	176	191	7.9%

This reversal of apparent relationship between two variables when looking at aggregated vs. disaggregated data is called Simpson's Paradox. It is caused by the confounding variable Victim's Race –

which is creating a spurious relationship between Defendant's Race and the Death Penalty.

- 467/483 = 96.7% of victims of white defendants were white
- 143/191 = 74.9% of victims of black defendants were black
- 64/515 = 12.4% of defendants received the death penalty when the victim was white vs. 4/159 = 2.5% when the victim was black

So, what is the effect of Defendant's Race on the Death Penalty — **with all other things (e.g. Victim's Race) being equal**? Conclusion: Black defendants do show a higher likelihood of getting the death penalty than white defendants.

So, why did the aggregate data show that white defendants had a higher likelihood of getting the death penalty? (Please pardon my direct language here.) Because whites tended to murder whites and blacks tended to murder blacks, and the death penalty was more likely when whites were murdered.

Example #21 — Other real-world examples of Simpson's Paradox

- At UC Berkeley in 1973, the overall graduate school acceptance rate for men was higher than for women (44% vs. 35%). But most departments showed a higher acceptance rate for women. The confounding variable was that women tended to apply to more competitive departments with lower acceptance rates.

- Early in the pandemic, Italy's overall COVID-19 case fatality rate was higher than China's. However, when analyzed by age group (e.g. 20-30, 30-40, 40-50, …), China actually had higher fatality rates in every category. Italy's higher aggregate death rate was driven by its much older population demographic.

- For kidney stones, Treatment B had higher overall success rate than Treatment A. But when split by stone size, Treatment A was more effective for both small & large stones. Why? Doctors gave Treatment A to the most severe cases (large stones).

<u>Example #22</u> — Statistics Can Be Deceptive

- <u>Assuming causality</u> just because two things are correlated; Also, misinterpretations of data due to confounding variables, Simpson's Paradox and tricky conditional probabilities

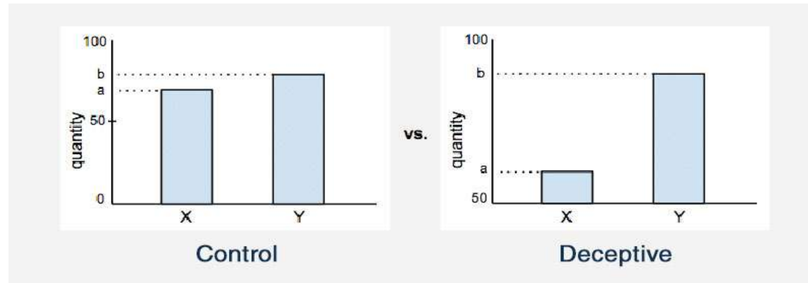
- <u>Sometimes statistics presented in a misleading way</u> — In 2007, an ad in the UK claims that 80% of dentists recommend Colgate. But most dentists recommended other toothpastes as well, and Colgate was usually further down the list. The ad was banned.

- <u>Showing cumulative data to hide trends in annual data</u>
  In a 2013 presentation, Apple CEO Tim Cook was criticized for showing cumulative sales of iPhones, perhaps to hide slipping annual sales

- <u>Misleading averages</u> — e.g. salaries of accounting department
  $43K , $45K , $46K , $49K , $51K , $52K , $55K, $124K , $138K
  Mean = $67K (misleads), Median = $51K (more representative)

- <u>Loaded questions can greatly slant the responses</u>

- <u>Truncating an axis</u> on a graph (starting at a value other than zero) can exaggerate the size of small changes



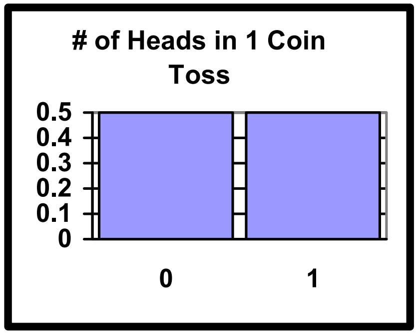
- <u>Some statistics are just made up</u> — In 2011, Reebok fined $25 million for false claims that their EasyTone shoes can tone a person's butt up to 28% more. Number completely made up.

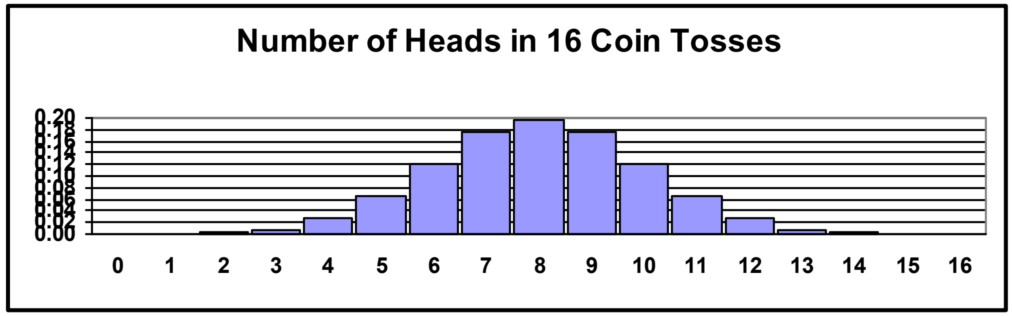
- <u>Using percentages to sensationalize a difference</u> — A medical study reports that a certain exposure doubles a person's risk for a very rare cancer (relative risk). But the absolute risk increases from only 0.01% (1 in 10,000) to 0.02% (2 in 10,000).

- <u>Survey bias</u> — e.g. 1936 Literary Digest presidential poll of 10 million voters drew from subscribers, telephone directories, car registrations — i.e. biased toward more affluent Republicans. Prediction: Landon wins 57% of vote ----- Actual: FDR won 61%

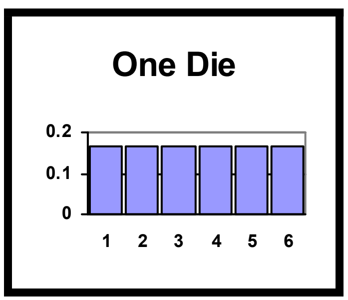
Why are so many natural phenomena in the real world "normally distributed" as a bell-shaped curve?

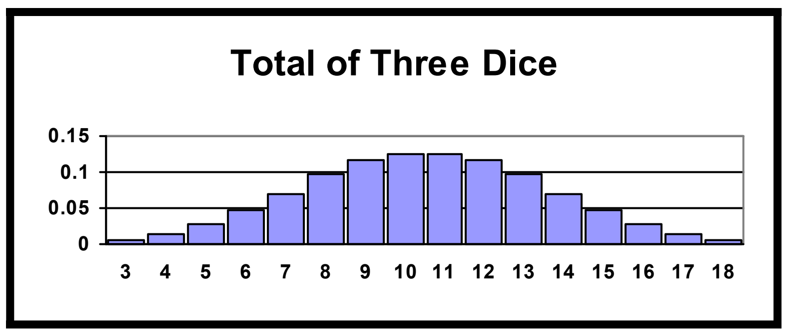
The number of *Heads* becomes more "normally distributed" as a bell-shaped curve as the number of coins tossed is increased





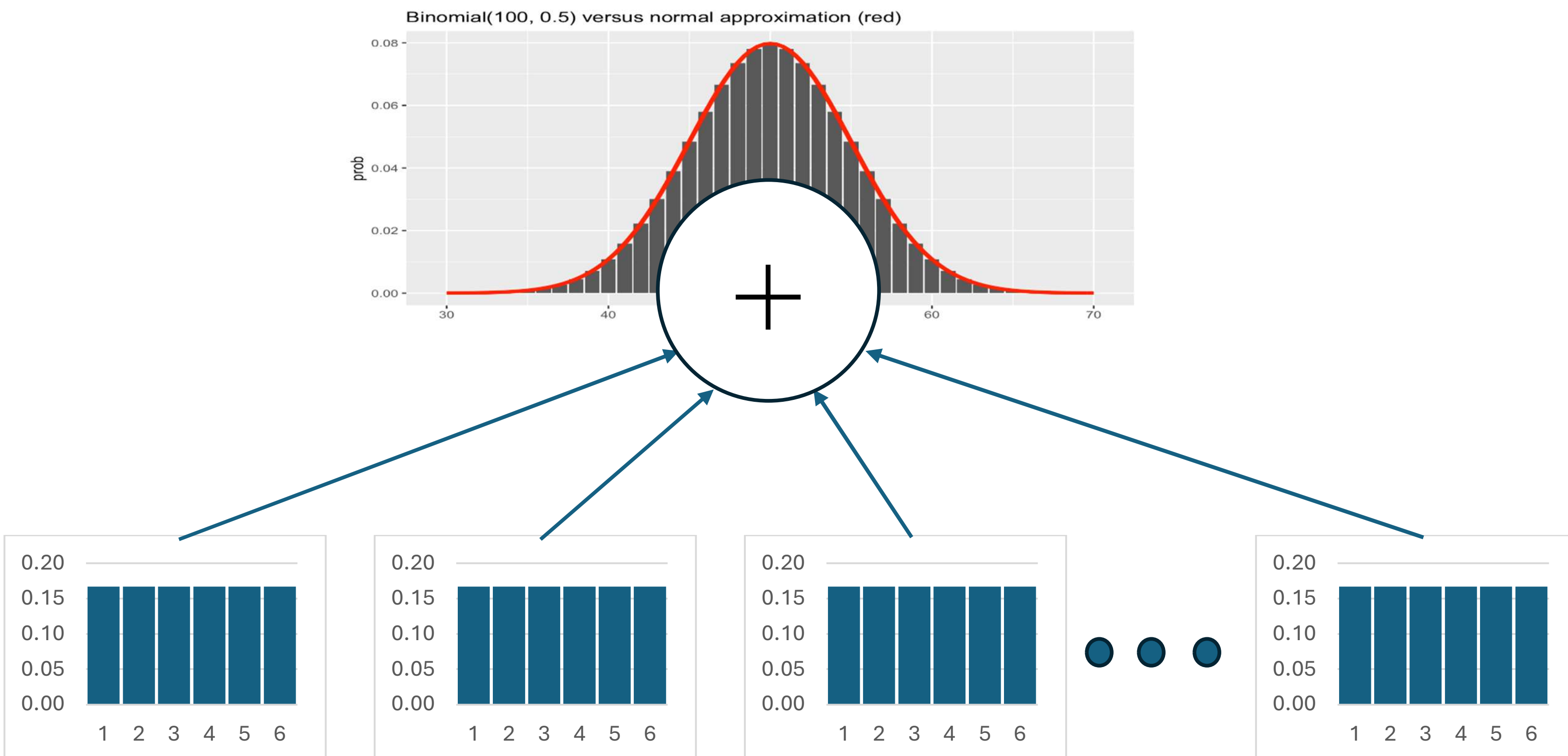
The total outcome of tossing dice becomes more "normally distributed" as bell-shaped curve as number of dice is increased





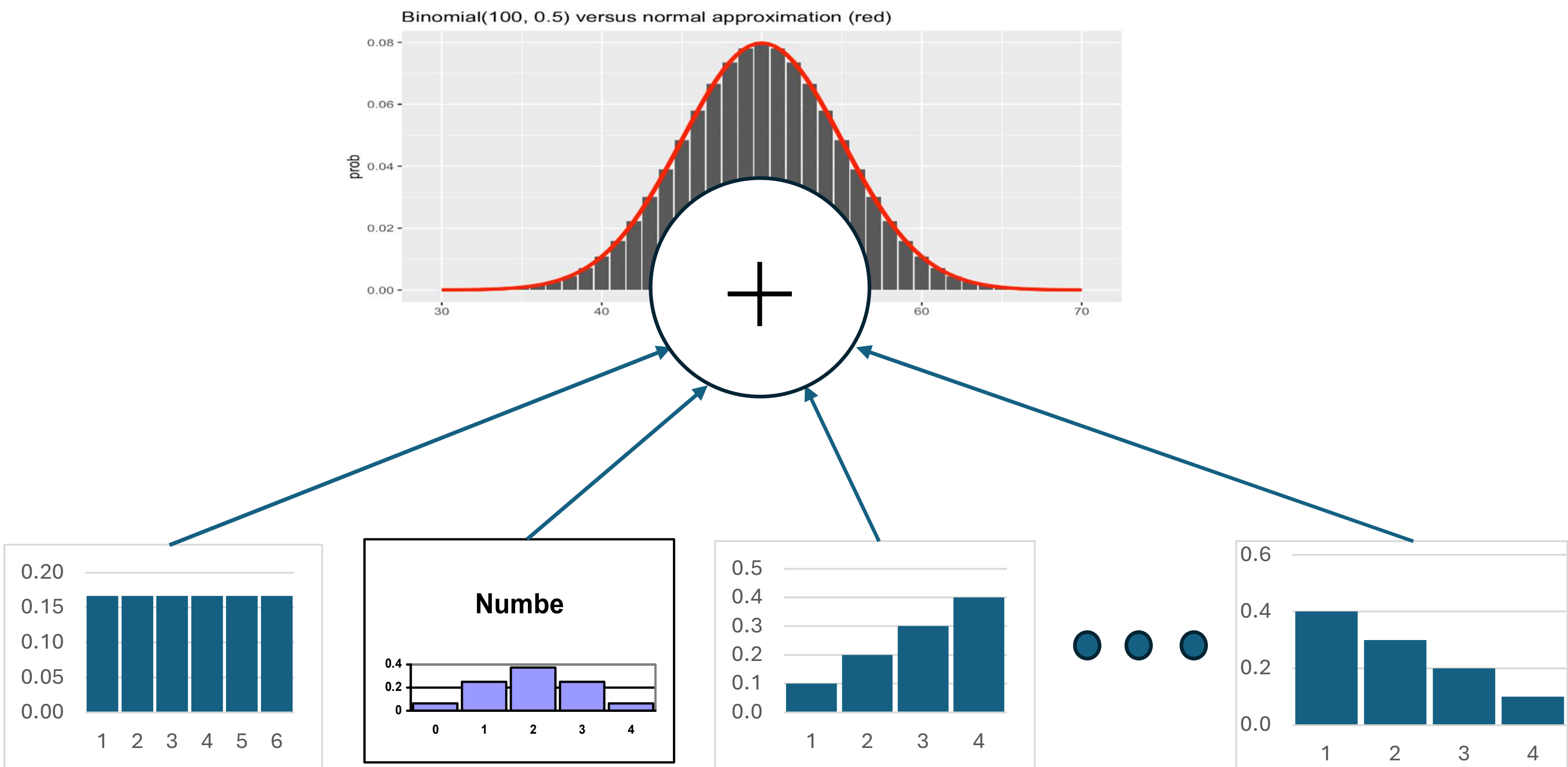
Why is a bell-shaped normal distribution so common in real world?

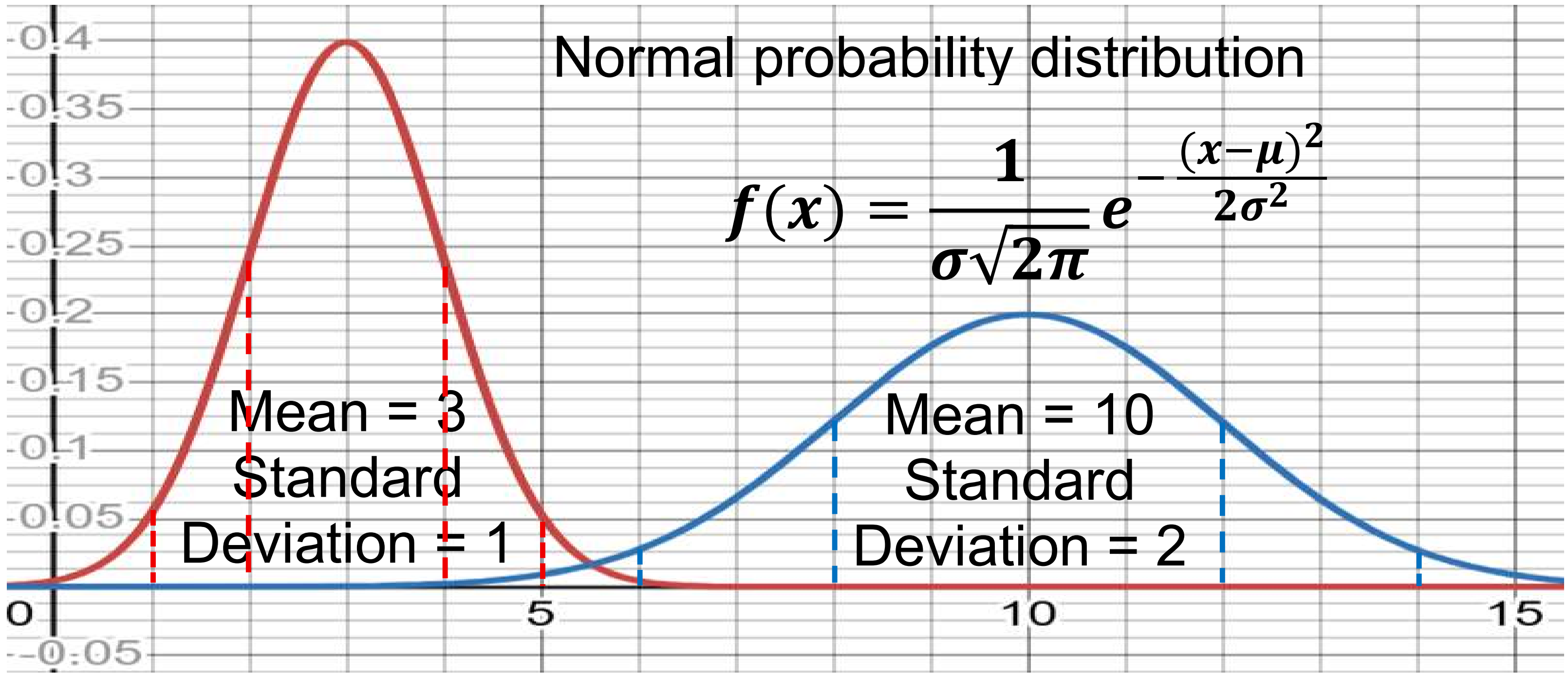
- The sum of many independent random numbers from ***any*** probability distribution tends toward a bell-shaped normal distribution, provided that the sample size is sufficiently large
- This is known as the Central Limit Theorem



The Central Limit Theorem is central to statistics

- The Central Limit Theorem is generally true even if the original probability distributions are all different
- The “normal distribution” is so-named because it is the “normal” or typical probability distribution for real-world phenomena



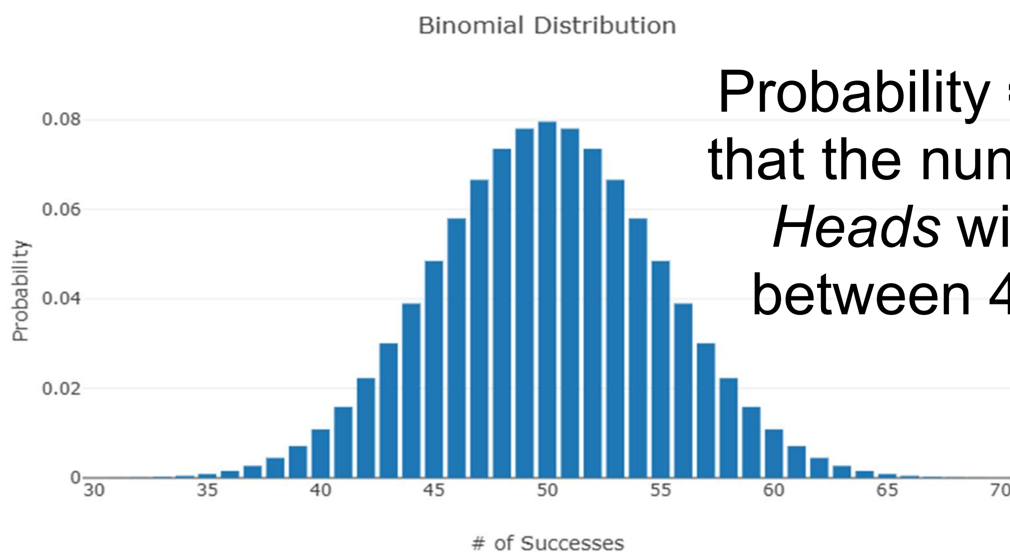


- Unlike a discrete probability distribution, the Normal distribution is a <u>continuous probability distribution</u> and does <u>not</u> give the probability for any particular value. Rather, the <u>area under the curve</u> gives the probability for an interval of values — e.g. the probability that a man is between 67-73 inches tall.
- About 68% of normal distribution lies within ±1 std. dev. of mean
- About 95% of normal distribution lies within ±2 std. dev. of mean

Example #23 — Normal Distributions in the Real World

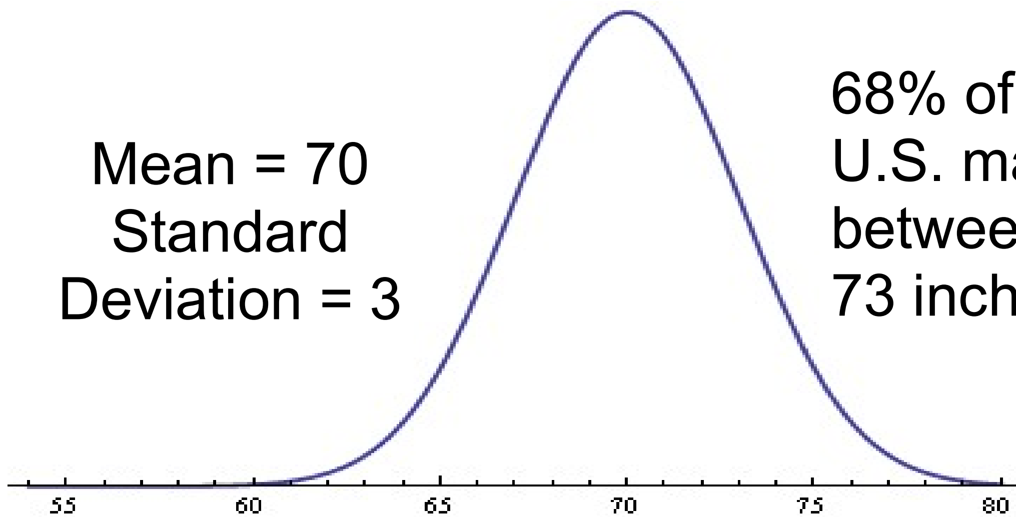
Since real-world quantities (e.g. height, blood pressure, IQ score, monthly temperatures) are often the cumulative effect of many independent random factors, the Central Limit Theorem is the main reason for the prevalence of the normal distribution in nature

- Number of *Heads* in 100 coin tosses is almost exactly normally distributed with mean = 50 and standard deviation = 5



Probability ≈ 68% that the number of *Heads* will be between 45-55

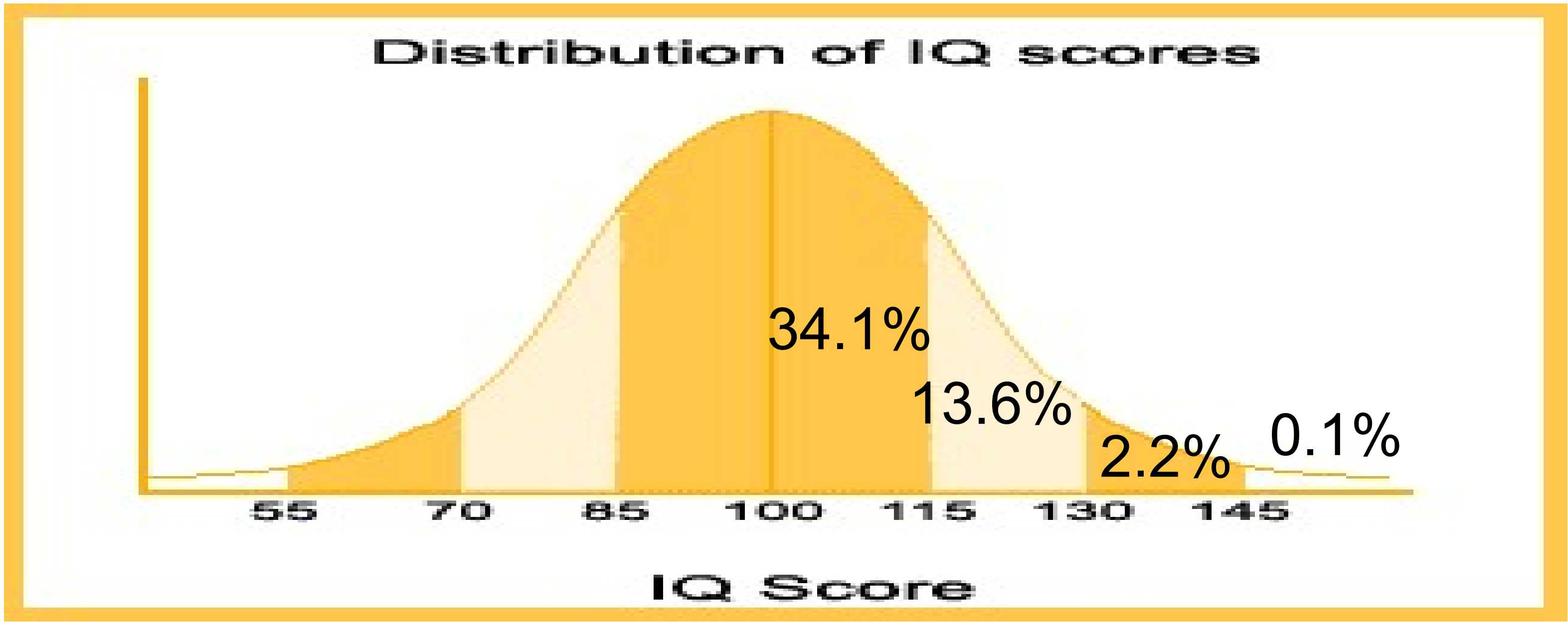
- Heights of adult U.S. males are roughly normally distributed with mean = 70 inches and standard deviation = 3 inches.
  A person's height is the combined effect of many genetic, environmental, diet, exercise and other independent factors.



- Heights of adult U.S. females are roughly normally distributed with mean = 65.5 inches and standard deviation = 2.5 inches (68% of adult females are between 63 and 68 inches tall)

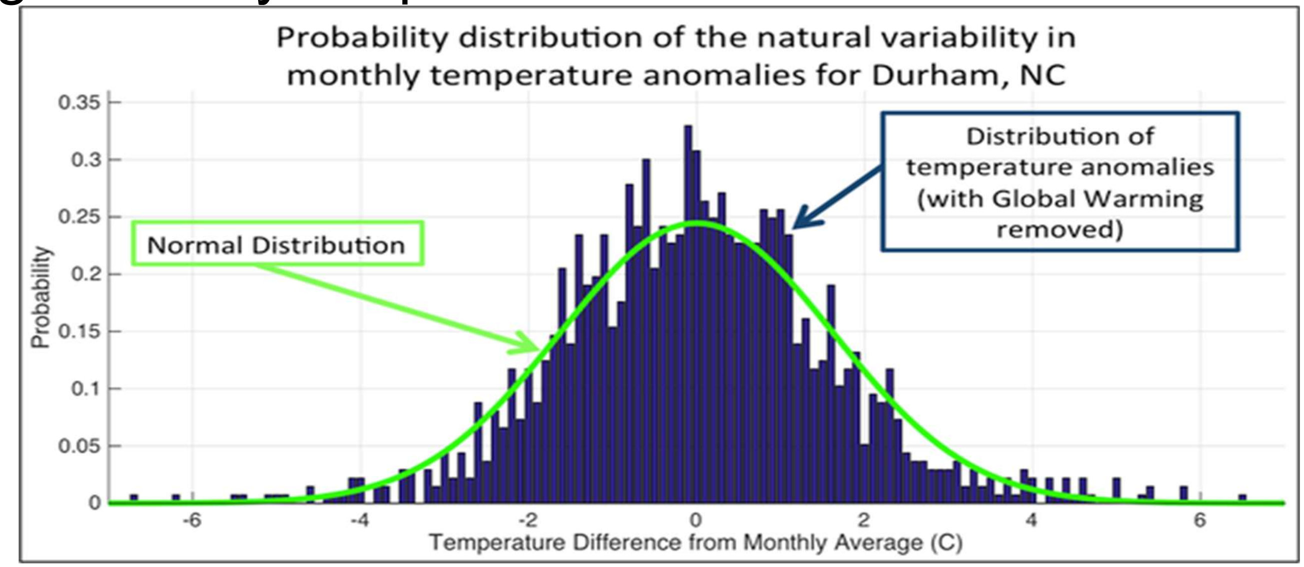
- Birth weights of full-term babies are roughly normally distributed with mean = 7.5 pounds and standard deviation = 1.1 pounds

- IQ scores are approximately normally distributed with a mean = 100 and a standard deviation = 15



- SAT scores roughly follow a normal distribution within a range of 200-800. In fact, the SATs were originally designed to have a mean of 500 and a standard deviation of 100.

- Average monthly temperatures



- Other human measurements are roughly normally distributed — e.g. weight, chest size, shoe size, body temperature

- Most measurement errors assumed to be normally distributed

- According to portfolio theory, investment returns of a diversified stock portfolio follow a normal distribution