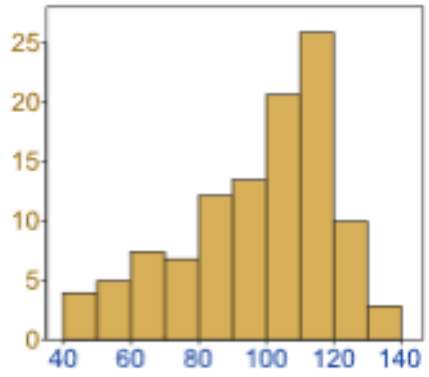
CSE303

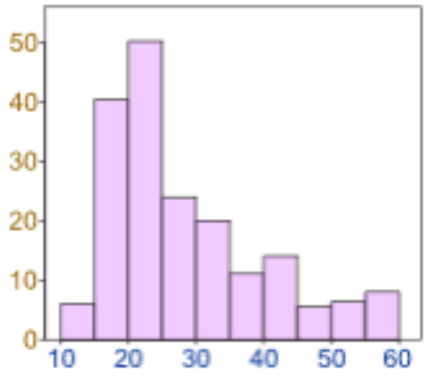
Lecture 4: Different Data Distributions

DATA DISTRIBUTION

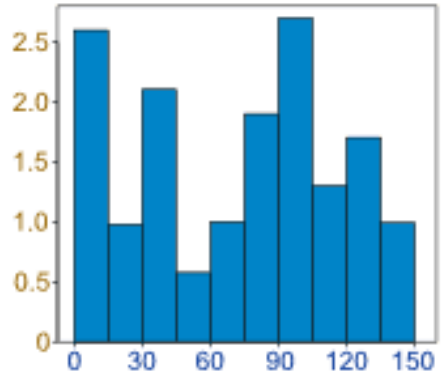
It can be spread out more on the left

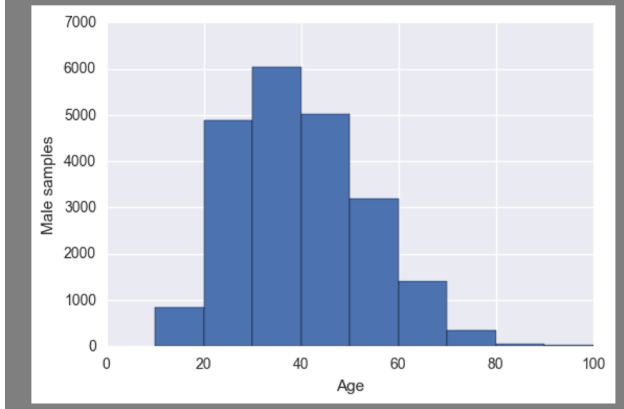


Or more on the right



Or it can be all jumbled up





NORMAL DISTRIBUTION

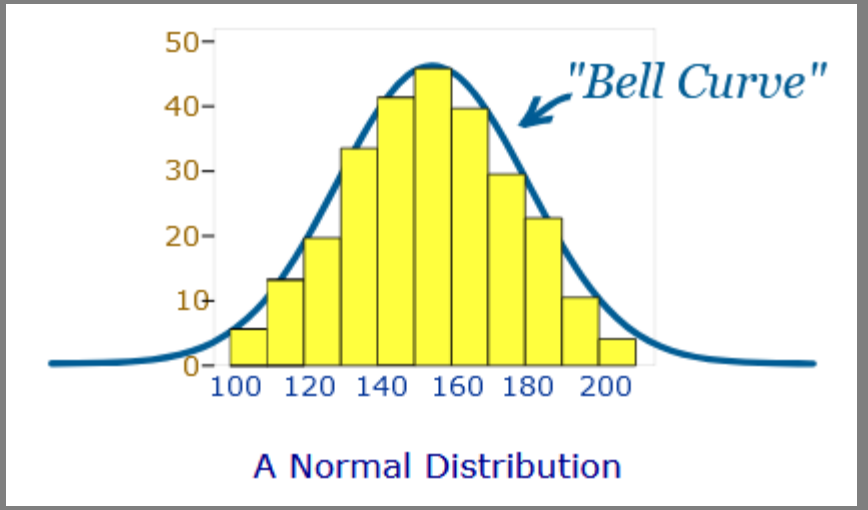
- In statistics, a **normal distribution** or **Gaussian distribution** is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

NORMAL DISTRIBUTION

- In probability theory, the **normal** (or **Gaussian** or **Gauss** or **Laplace-Gauss**) **distribution** is a very common continuous probability distribution
- The probability density of the normal distribution is

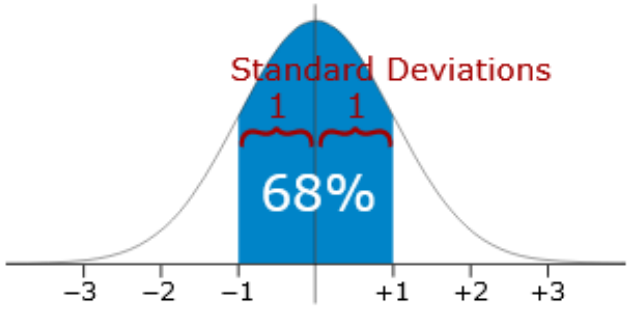
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



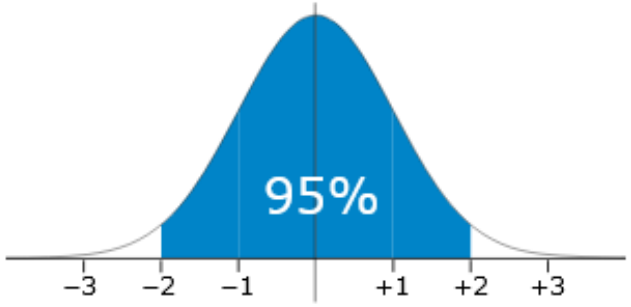
The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

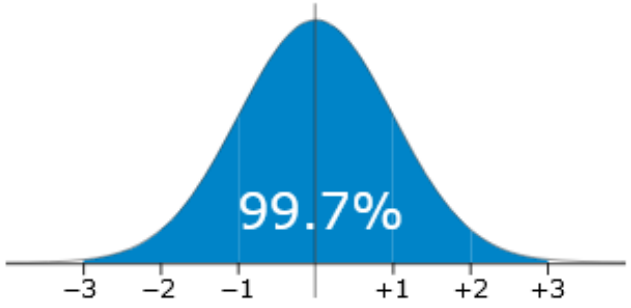
PROPERTIES OF NORMAL DISTRIBUTION



68% of values are within
1 standard deviation of the mean



95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the mean

EXAMPLE 1

- **95% of students at school are between 1.1m and 1.7m tall.** Assuming this data is **normally distributed** can you calculate the mean and standard deviation?

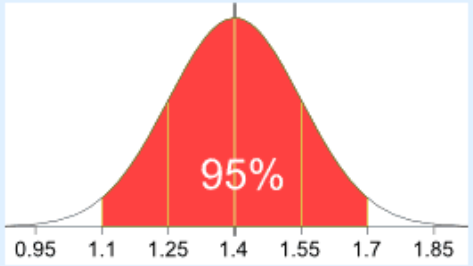
The mean is halfway between 1.1m and 1.7m:

Mean = (1.1m + 1.7m) / 2 = **1.4m**

95% is 2 standard deviations either side of the mean (a total of 4 standard deviations) so:

1 standard deviation = (1.7m-1.1m) / 4
= 0.6m / 4
= **0.15m**

And this is the result:



STANDARD SCORE OR “Z-SCORE”

- The number of **standard deviations from the mean** is also called the "Standard Score", "sigma" or "z-score“
- **Example 2: In that same school one of your friends is 1.85m tall. Find out his z-score.**
- **z-score (for one sample) = $(x – \mu) / \sigma = 1.85 – 1.4 / 0.15 = 3.0$**

WHY DO WE NEED Z-SCORE?

- Example 4: **Professor Willoughby is marking a test.** Here are the students results (out of 60 points):

  20, 15, 26, 32, 18, 28, 35, 14, 26, 22, 17

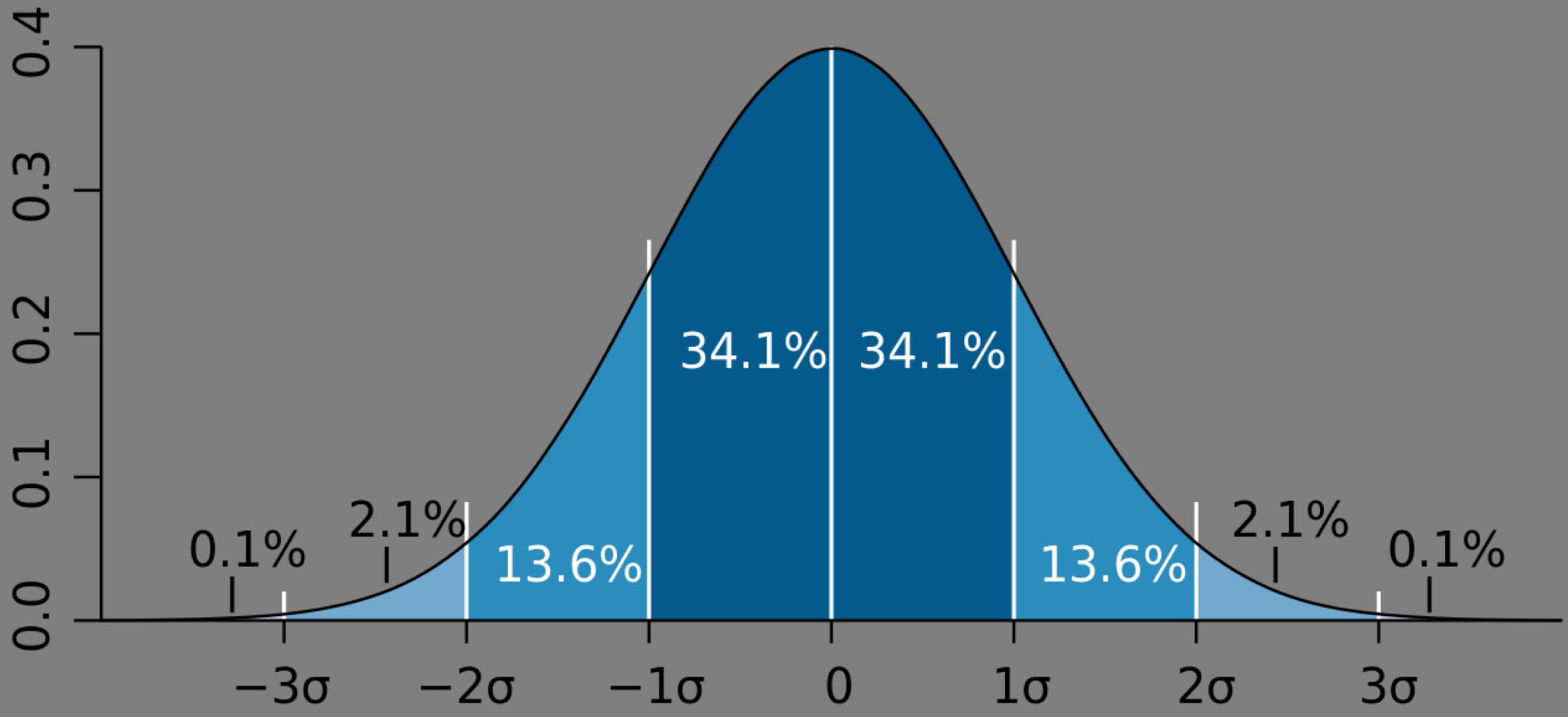
  Most students didn't even get 30 out of 60, and most will fail.

- Professor decides to Standardize all the scores and only fail people 1 standard deviation below the mean.
- The **Mean is 23**, and the **Standard Deviation is 6.6**, and these are the Standard Scores:

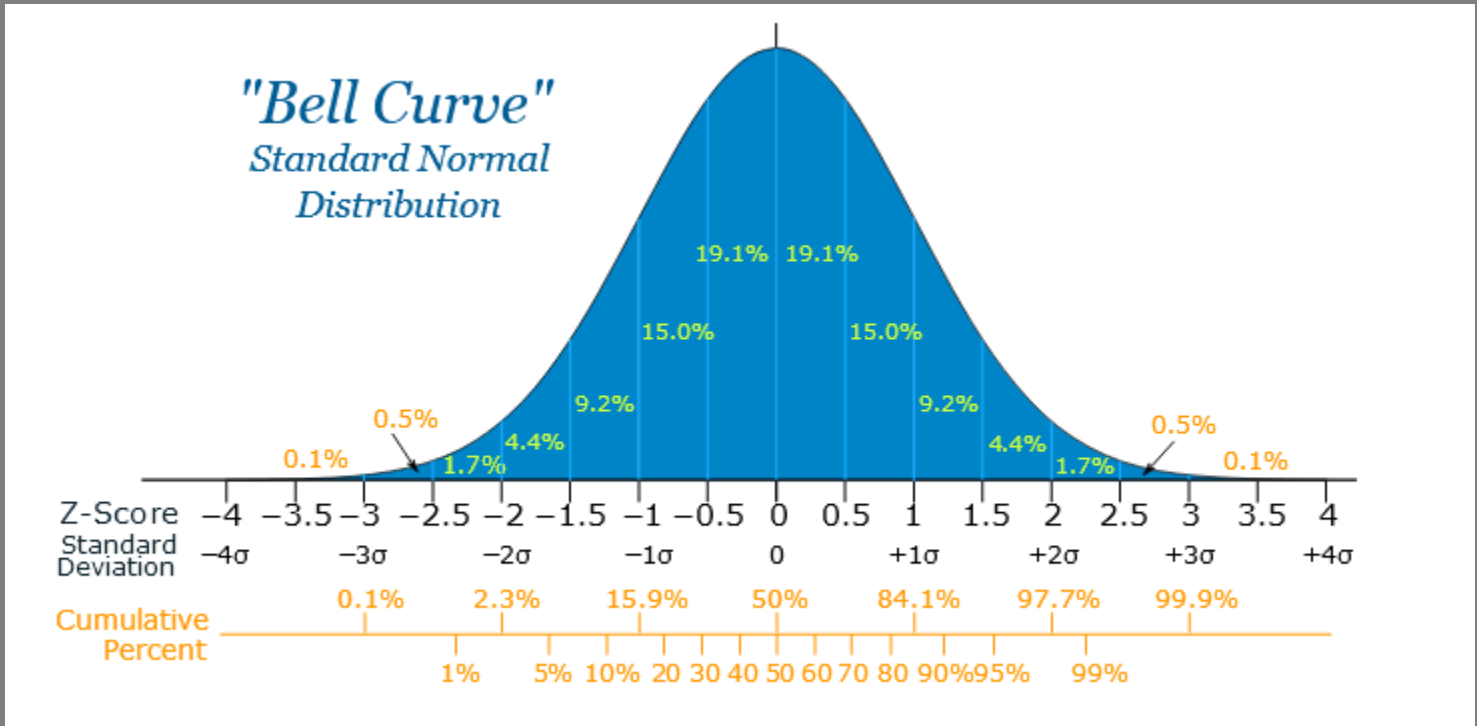
  -0.45, -1.21, 0.45, 1.36, -0.76, 0.76, 1.82, -1.36, 0.45, -0.15, -0.91

- Now only 2 students will fail (the ones who scored 15 and 14 on the test)
- **Much fairer!**

STANDARD NORMAL DISTRIBUTION

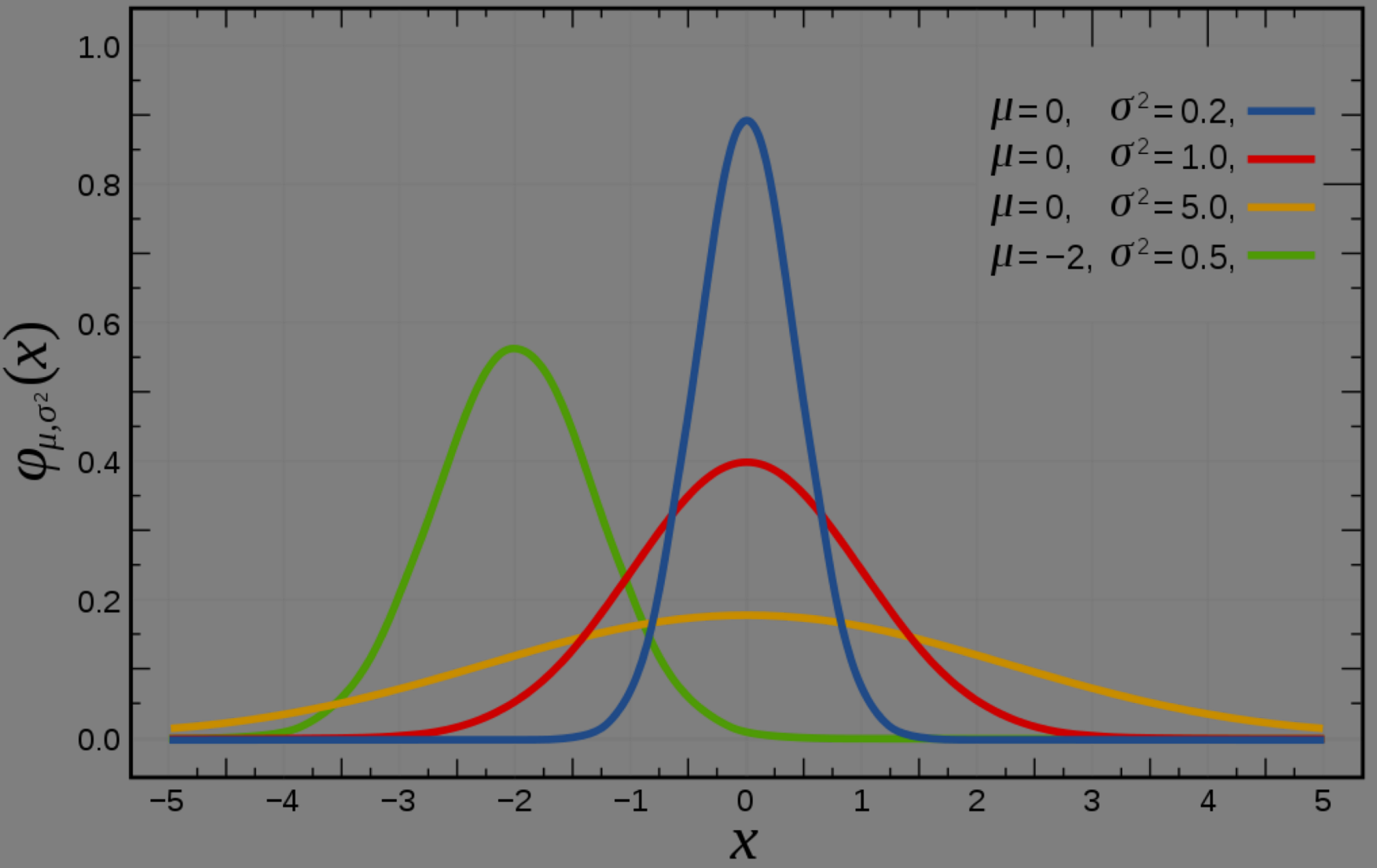


ANOTHER EXAMPLE



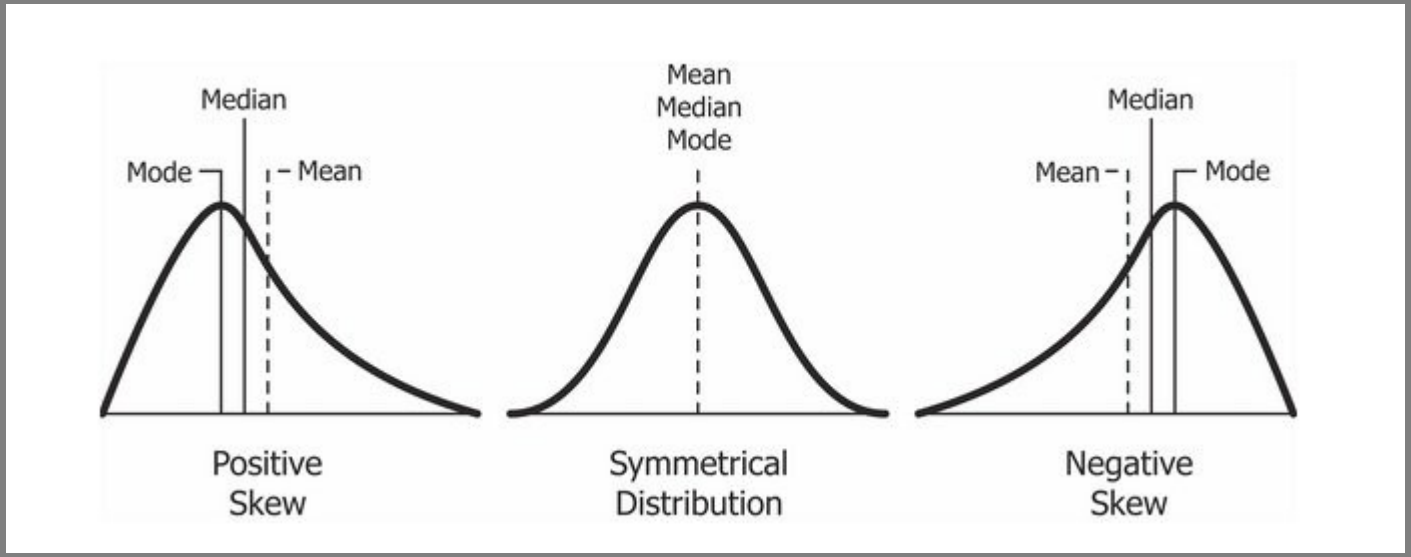
- Your score in a recent test was **0.5 standard deviations** above the average, how many people scored **lower** than you did?

NORMAL DISTRIBUTIONS



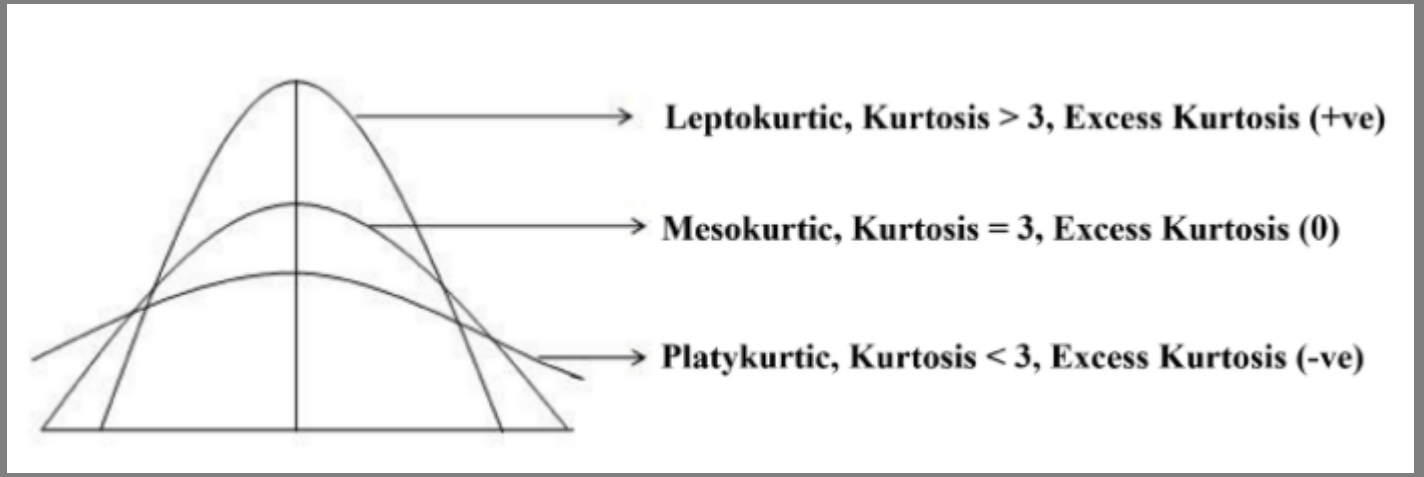
SKEWNESS

- It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution.
- It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.



KURTOSIS

- Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution.



FORMULA FOR SKEWNESS AND KURTOSIS

$$\text{Skewness} = \frac{\sum \left(x - \bar{x}\right)^3}{(n-1) \cdot S^3}$$

$$\text{Kurtosis} = \frac{\sum \left(x - \bar{x}\right)^4}{(n-1) \cdot S^4}$$

BINOMIAL DISTRIBUTION

- A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.
- The binomial is a type of distribution that has two possible outcomes (the prefix "bi" means two, or twice). For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.
- Binomial Distribution Function: $\mathbf{b(x; n, P)} = nCx * p^x * (1 - p)^{n-x}$
- Mean = n * P
- Variance = n * P * (1-P)

PRACTICE PROBLEMS

- A coin is tossed 10 times. What is the probability of getting exactly 6 heads?

- 60% of people who purchase sports cars are men. If 10 sports car owners are randomly selected, find the probability that exactly 7 are men.

POISSON DISTRIBUTION

- A Poisson distribution is a tool that helps to predict the probability of certain events from happening when you know how often the event has occurred. It gives us the probability of a given number of events happening in a fixed interval of time.
- Poisson Distribution Function: $\mathbf{P(x; \mu) = (e^{-\mu} * \mu^{x}) / x!}$

PRACTICE PROBLEMS

- The average number of major storms in your city is 2 per year. What is the probability that exactly 3 storms will hit your city next year?

CORRELATION ANALYSIS (NOMINAL DATA)

- **X^2 (chi-square) test**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

CHI-SQUARE CALCULATION: AN EXAMPLE

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

PRACTICE PROBLEM

- Let's say you want to know if gender has anything to do with political party preference. You poll 440 voters in a simple random sample to find out which political party they prefer. The results of the survey are shown in the table below:

	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	440

CHI-SQUARE TABLE

Critical values of the Chi-square distribution with d degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, Table D.1

THANK YOU