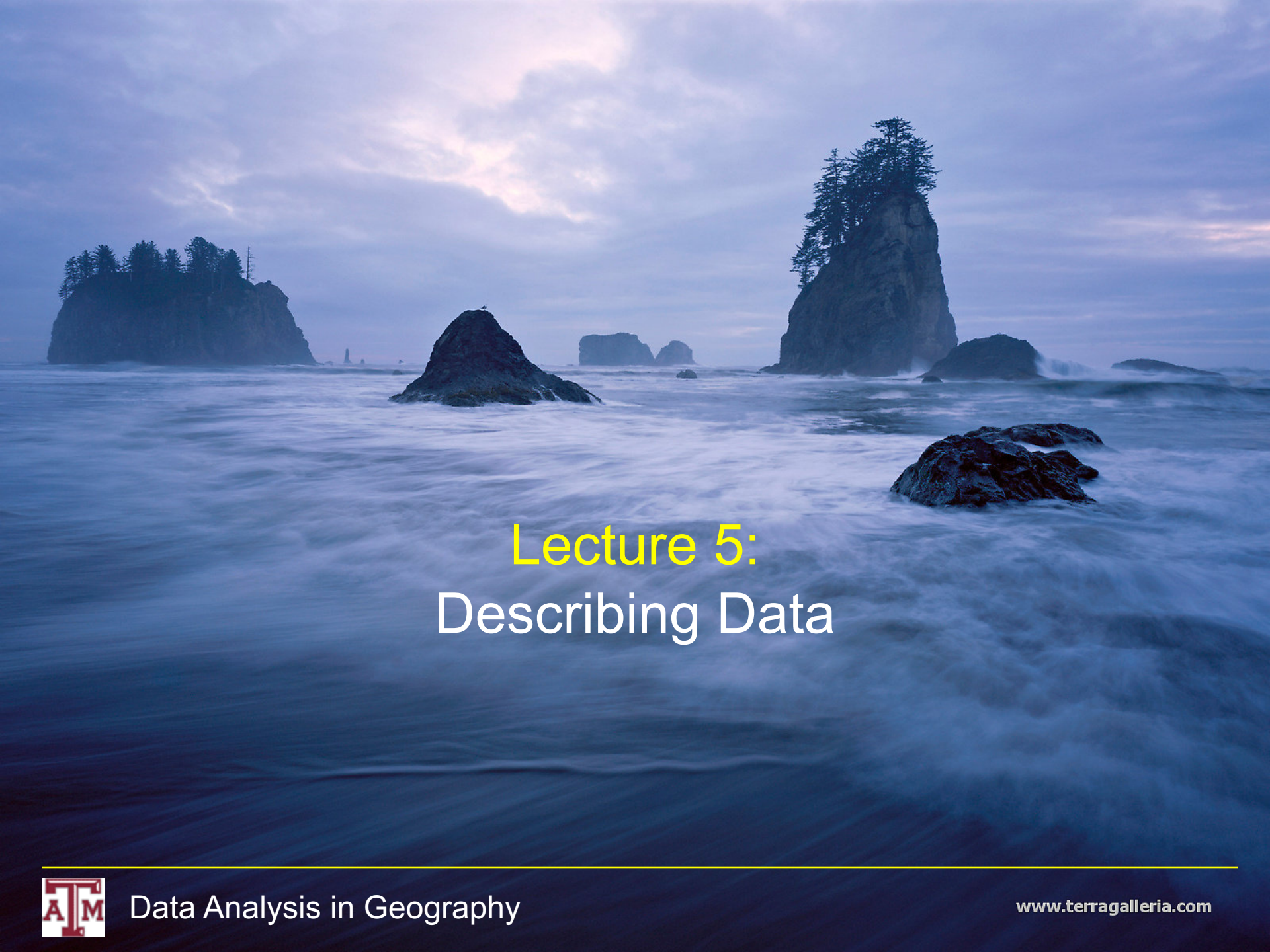


Lecture 5:
Describing Data

Announcements

1. If you missed class last Thursday, please come see me after class today

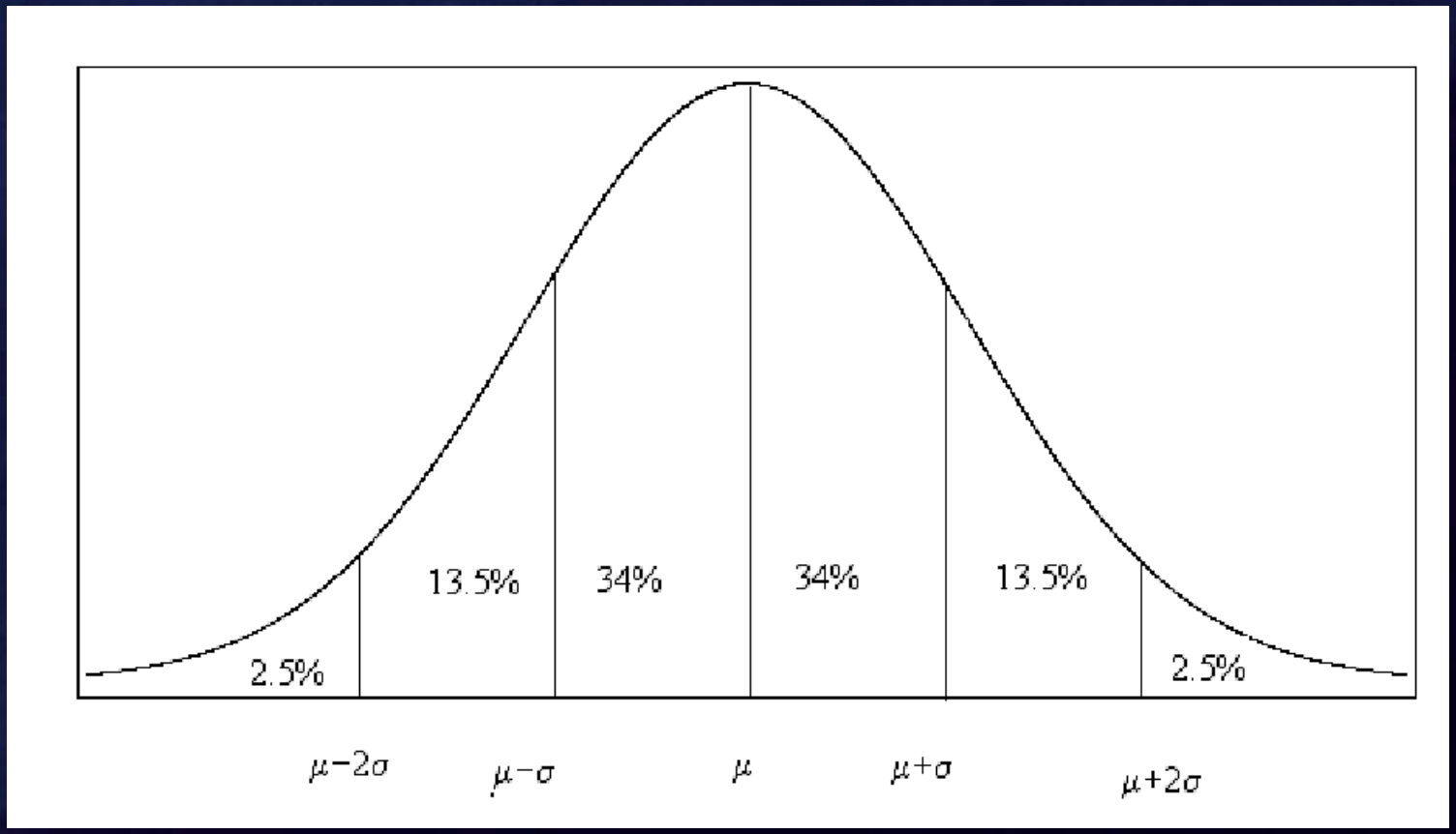
2. Weather allowing, we will be going back outside on Tuesday to collect some more samples (for HW#2)

Reminders:

1. HW#0 is graded and on eCampus

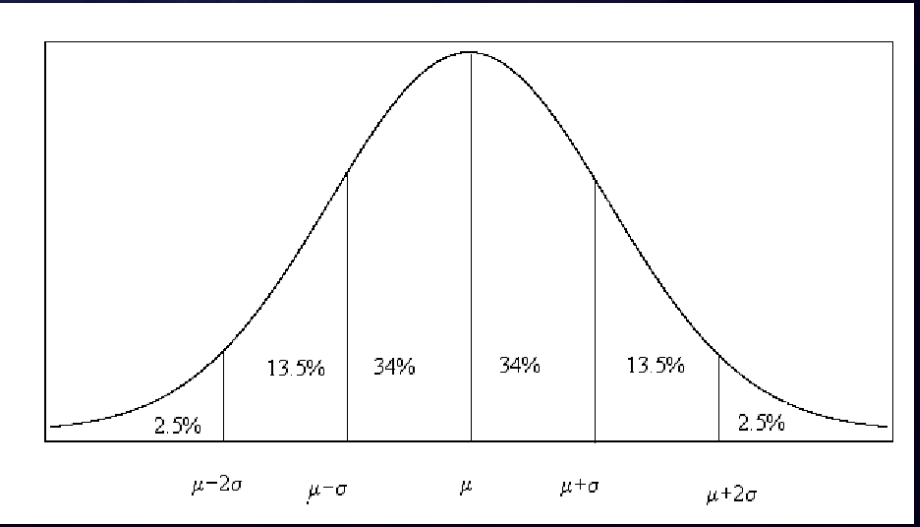
2. Measuring tapes are due to me today

Probability Distribution Function



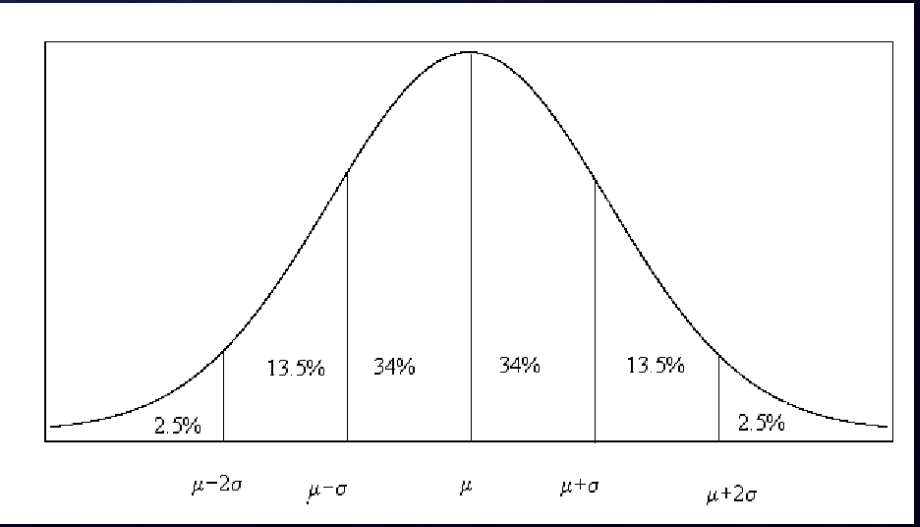
Normal Distribution

- The shape of the normal curve is often illustrated as a bell-shaped curve.
- The highest point on the normal curve is at the mean of the distribution.
- The normal curve is symmetric.
- The standard deviation determines the width of the curve.



Normal Distribution

- The total area under the curve the same as any other probability distribution is 1.
- Probabilities for a random variable are given by areas under the curve



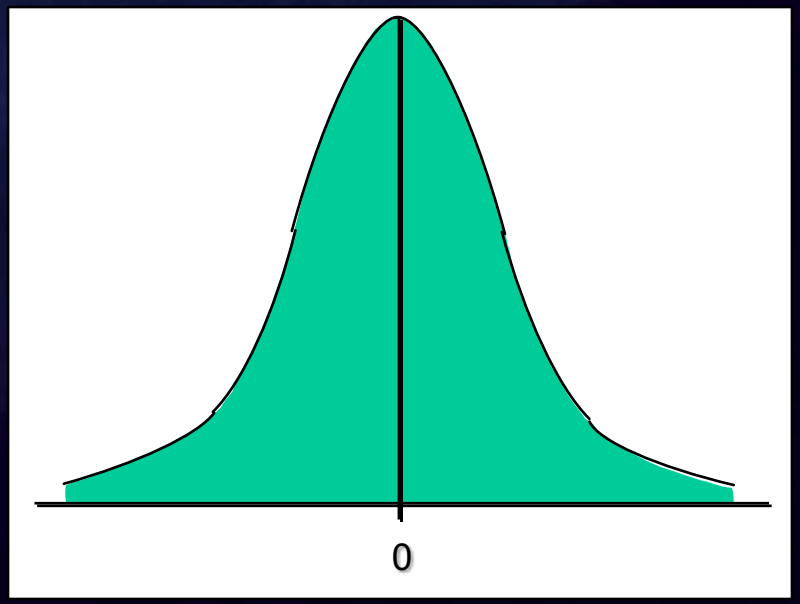
Central Tendency

- Represents the center of the distribution
- Mean: Sum of all samples divided by the number of samples

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Point on the distribution where the sum of all deviations is equal to zero

$$0 = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x})$$



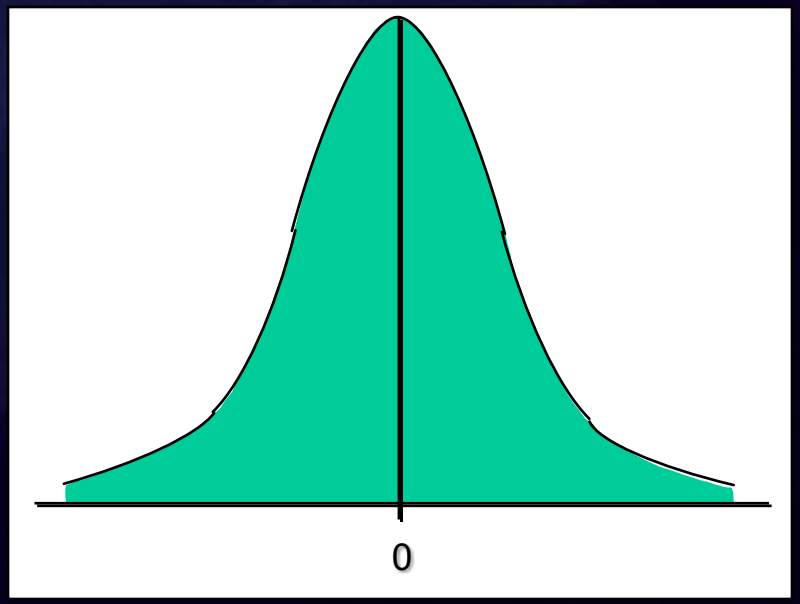
Central Tendency

- Mode: Value that occurs most frequently- the peak of the distribution

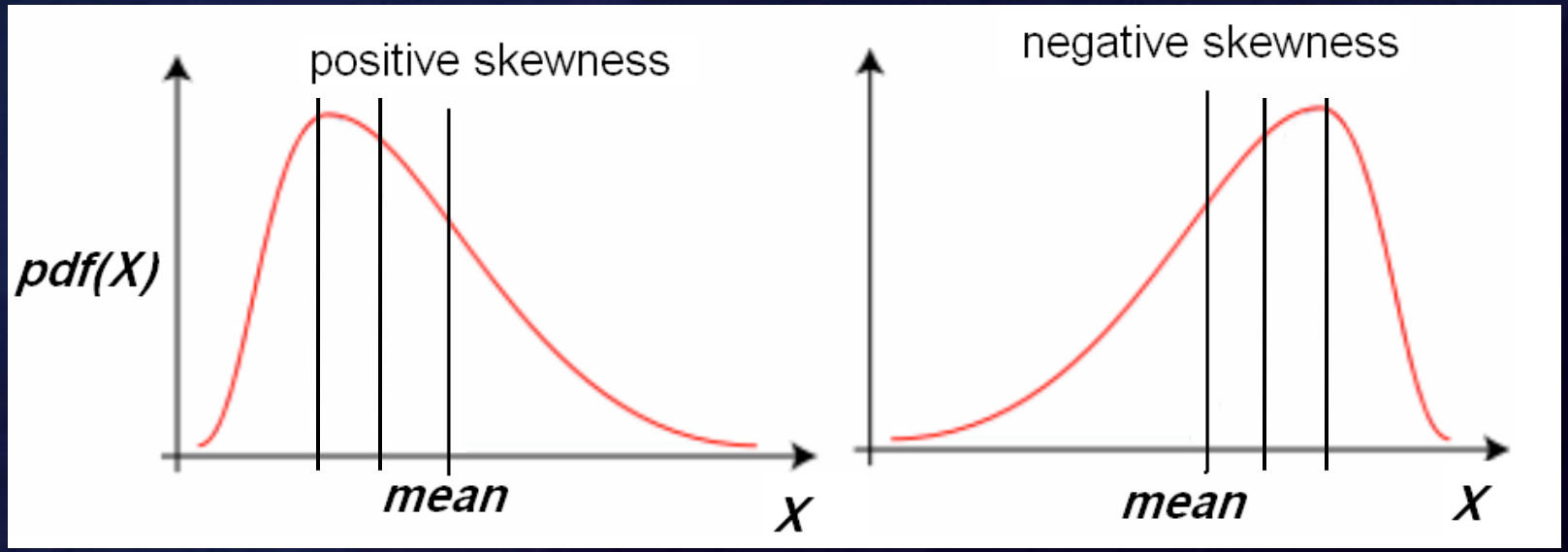
- Median: is the middle value from a set of observations- 50% are larger and 50% are smaller

- In a normal distribution:

mean = median = mode



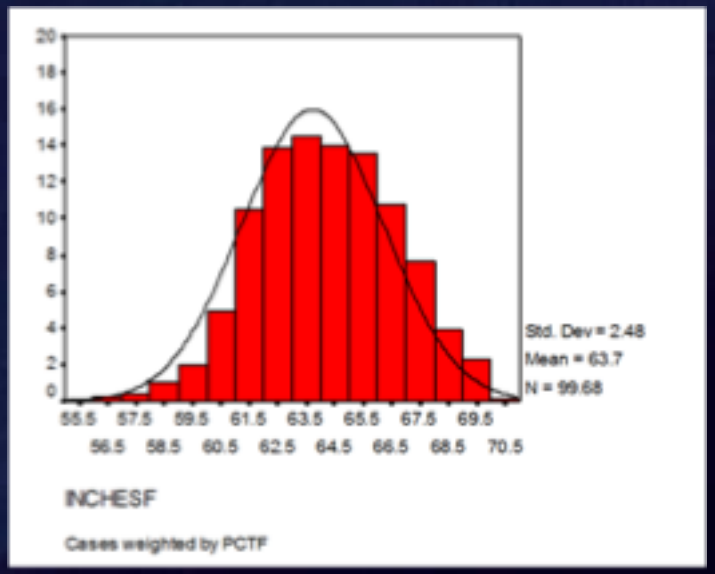
Skewed Distributions

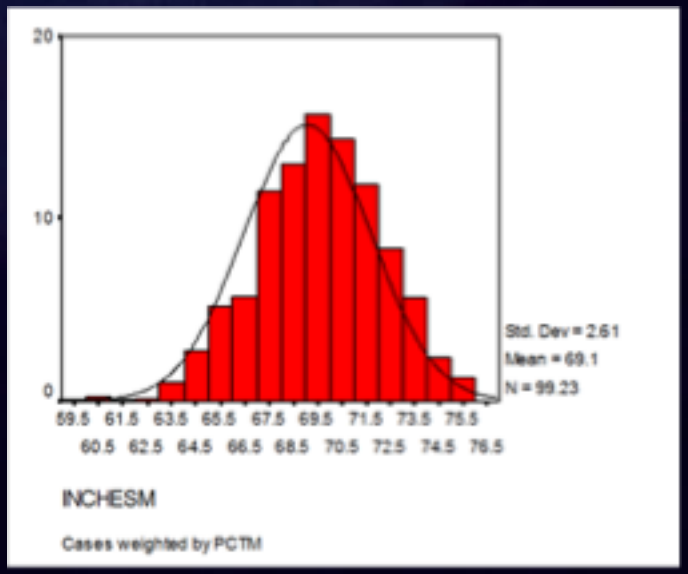


- Mean, median and mode are no longer the same in a skewed distribution
- Mean is strongly influenced by very large or very small values and is moved furthest from the mode

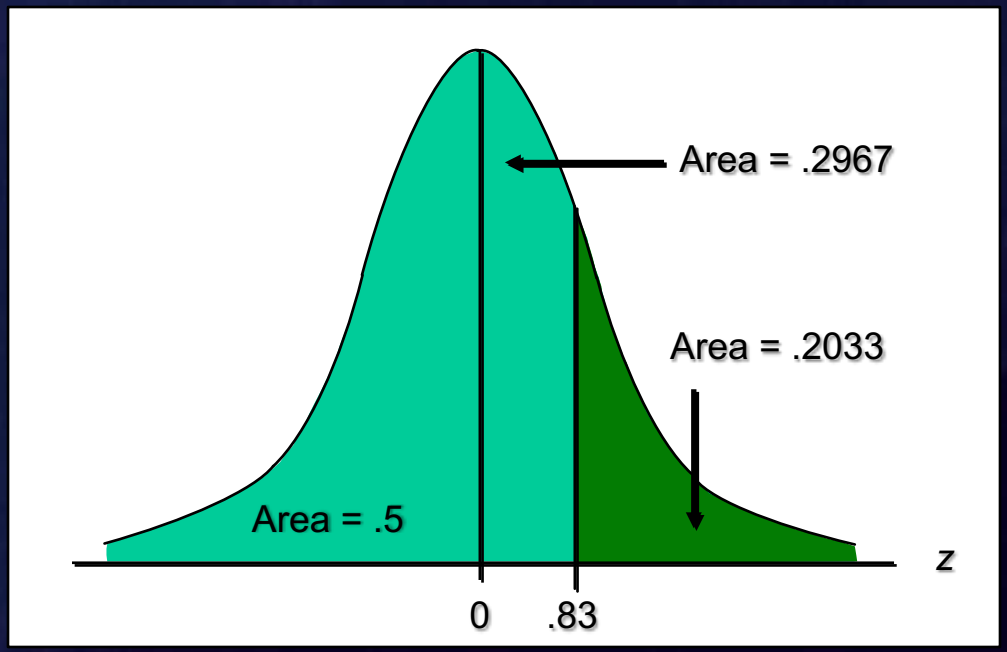
Adult Heights

- Female (left) and Male (right) adult heights are well approximated by normal distributions



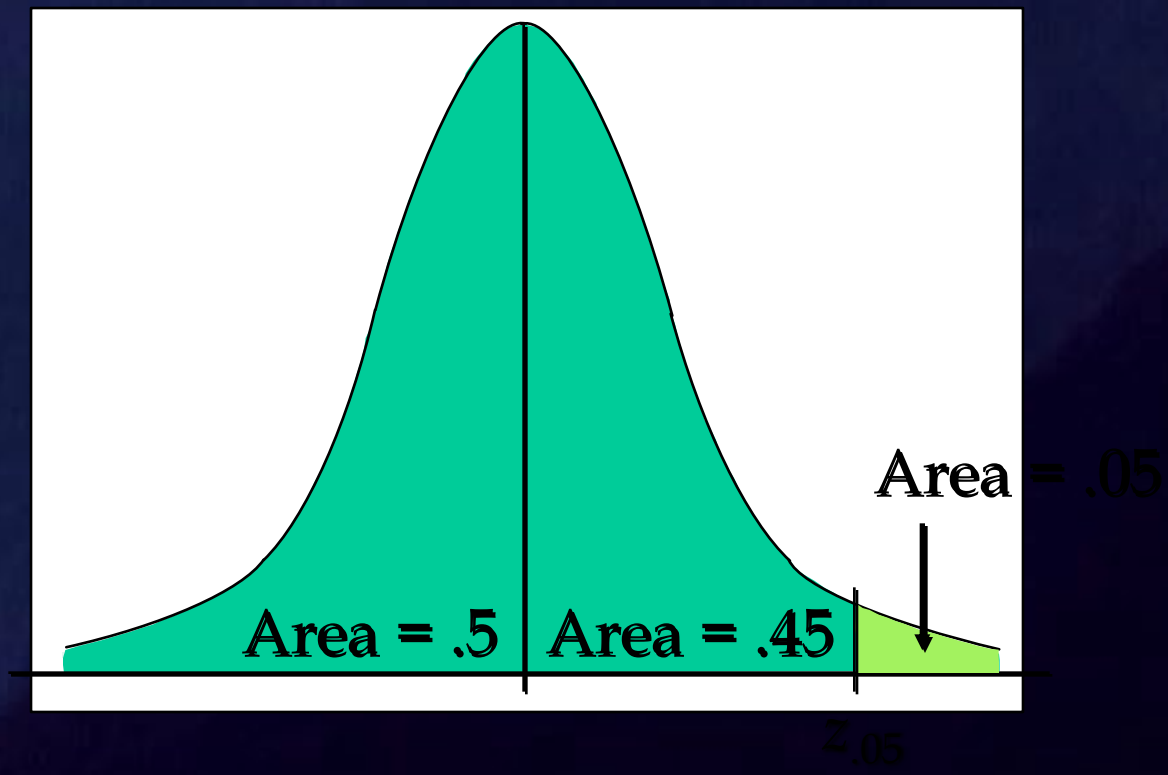


Probability



The Standard Normal table shows an area of .2967 for the region between the $z = 0$ line and the $z = .83$ line above. The shaded tail area is .5 - .2967 = .2033.

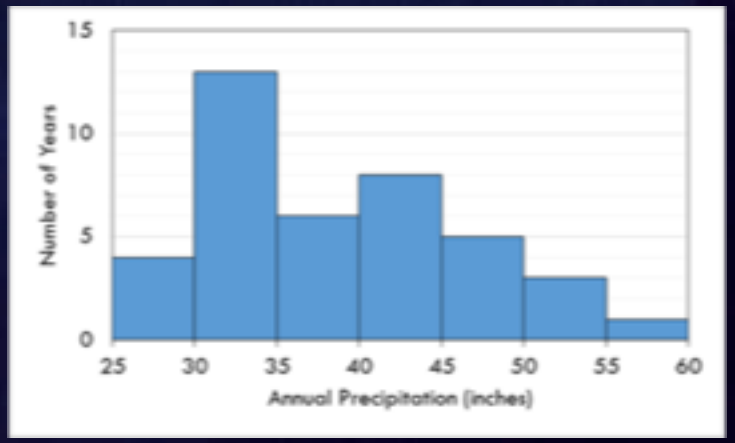
Probability



If I take a sample, what is the probability that I will get one of the largest values?

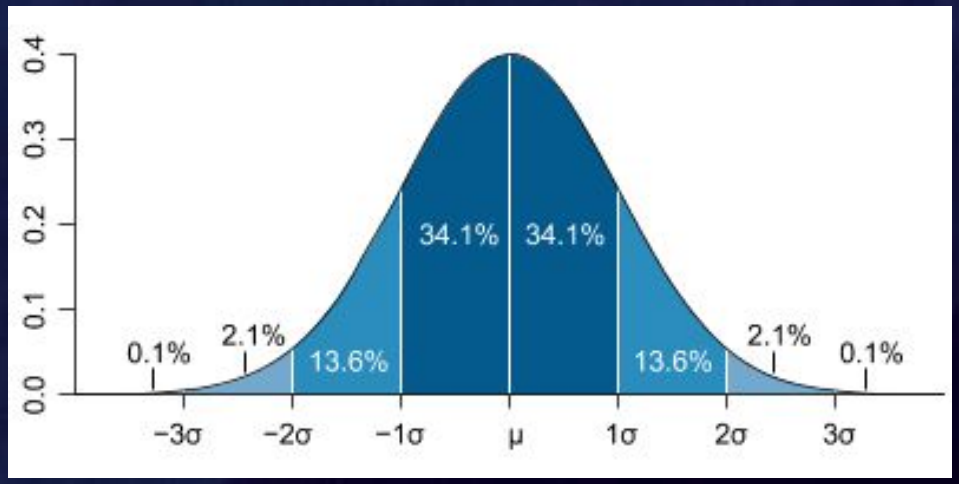
Example: Annual Precipitation in Washington, DC over 40 years

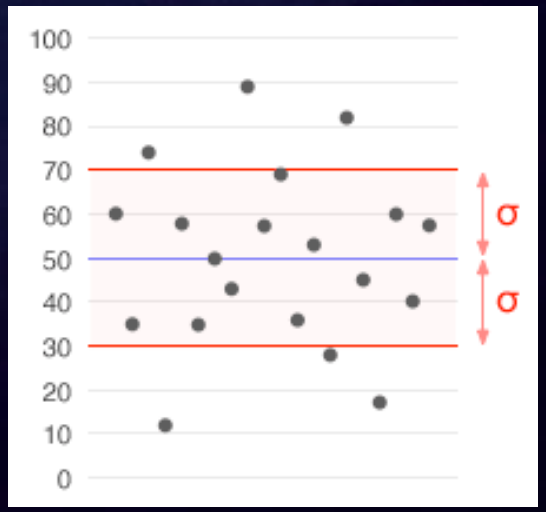
Ordered Sample			
26”	35”	39”	45”
26”	35”	40”	46”
28”	35”	40”	47”
29”	36”	41”	47”
32”	36”	41”	48”
32”	36”	41”	50”
33”	36”	41”	51”
33”	38”	41”	51”
34”	39”	43”	54”
35”	39”	43”	57”



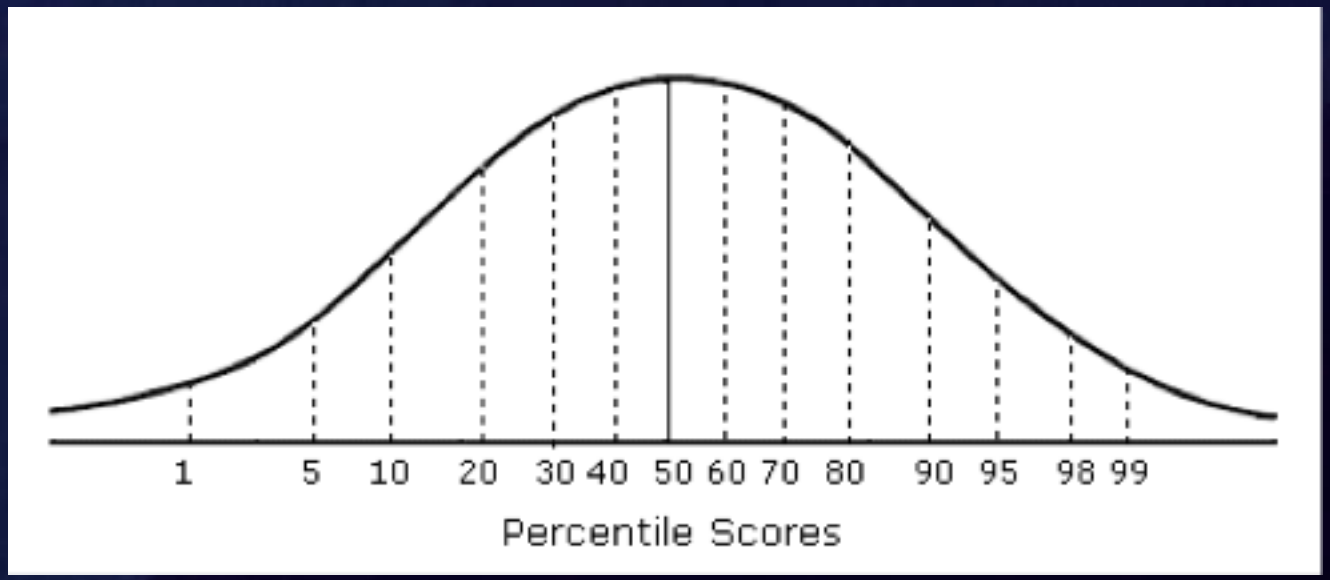
Statistic	Value
Mean	39.5”
Median	39.0”
Mode	41.0”
Standard Deviation	7.4”

Normal Distribution



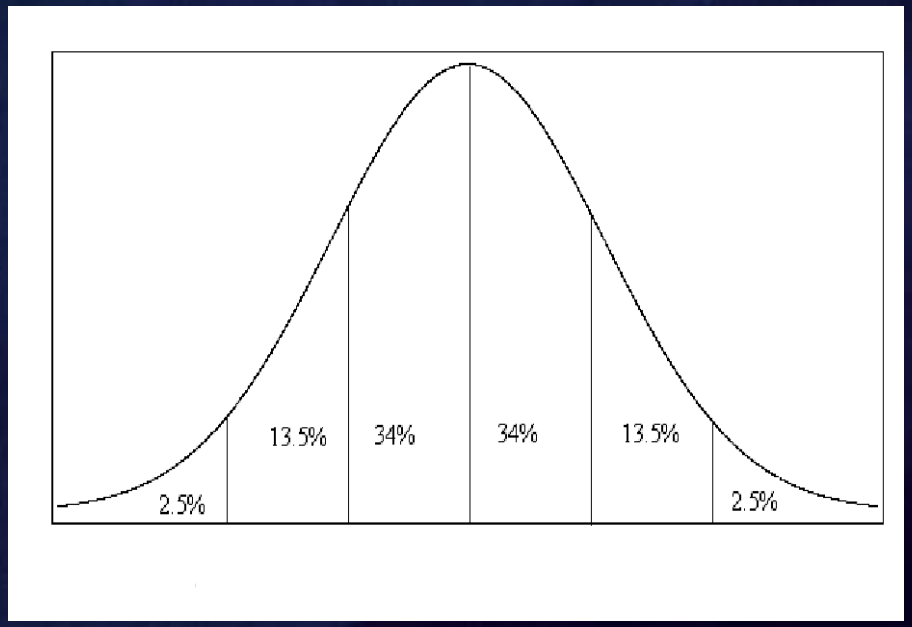


Percentiles



Percentiles, including the median, are calculated based on the area under the curve : where is the 50th percentile?

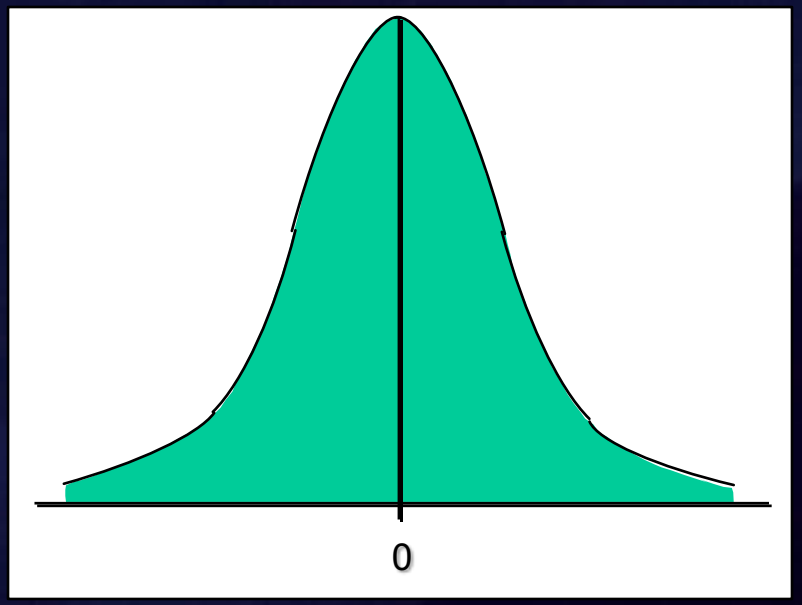
Cumulative Distribution Function

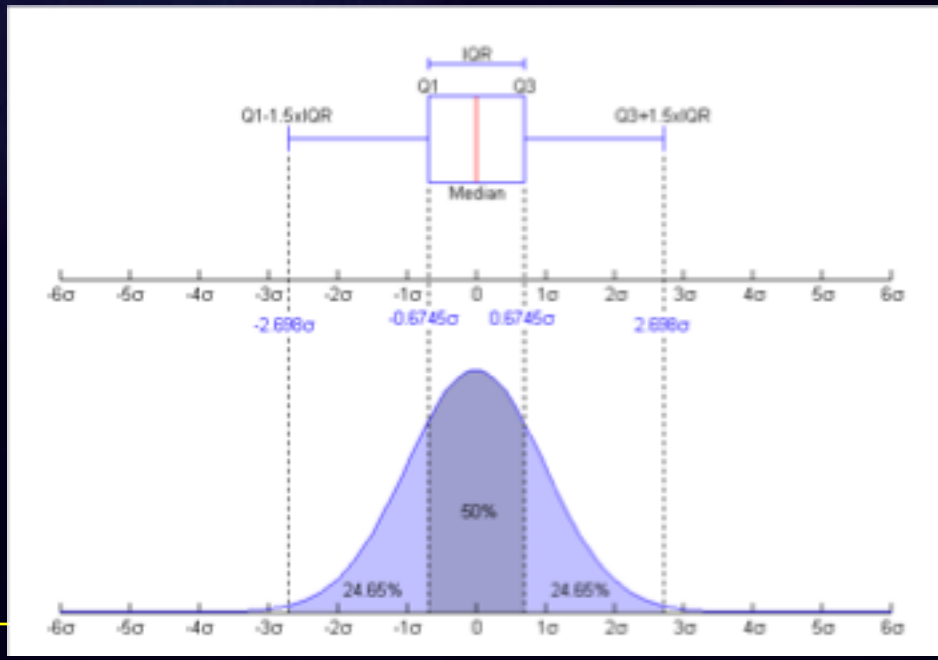


Cumulative Distribution: aggregate frequencies from value to value and present the cumulative frequency at each value

Dispersion

- Amount of spread/variability in a sample
- Range: Maximum value – Minimum value
  - 100% of all samples fall within this range
- Interquartile range: difference between the 25th and 75th percentile
  - middle half of the data
  - 50% of all samples fall within this range





Dispersion

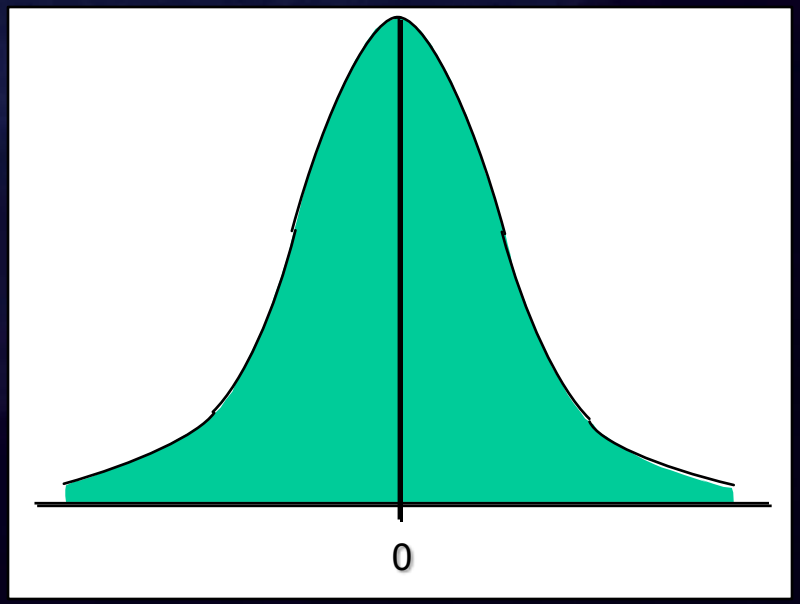
- Standard Deviation: Average deviation of the sample values from the mean

- Remember: Sum of deviations is equal to 0

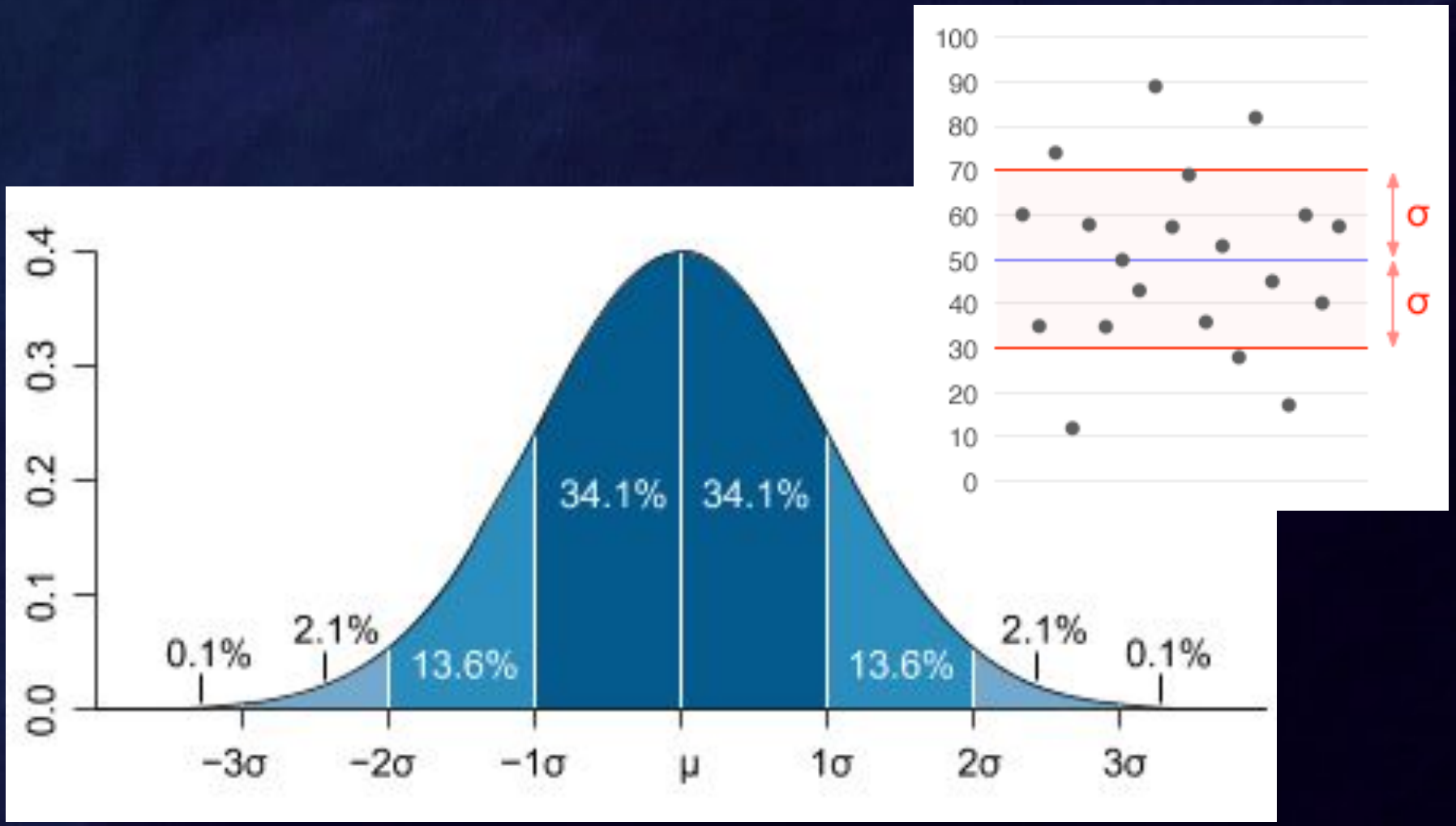
$$0 = (x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_n - \bar{x})$$

- Average of the square of the deviations

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

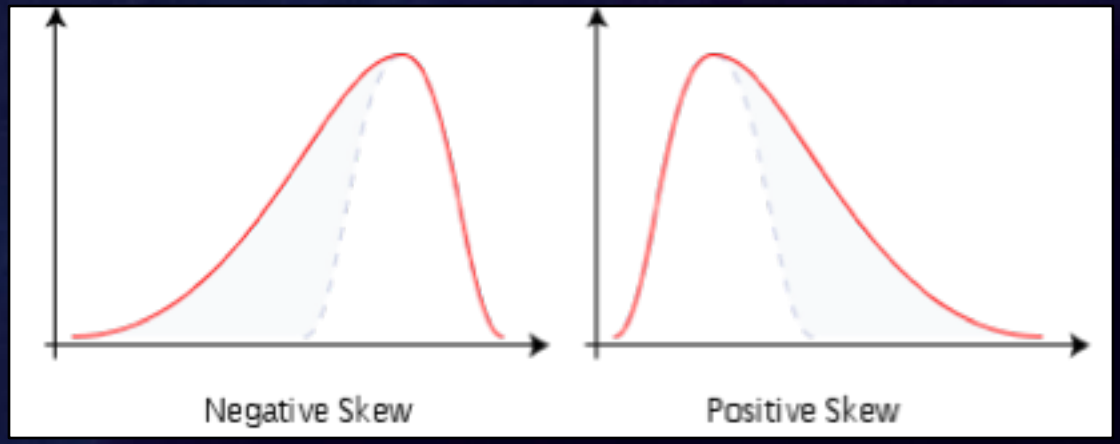


Normal Distribution



Shape

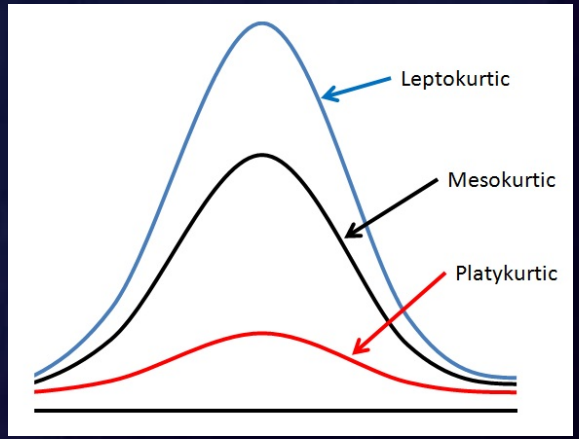
- Skewness: Asymmetry of a distribution
  - Is the standard deviation made up more from large or small values
- Negative skewness: left tail (of smaller values) is longer and the mass of the distribution is concentrated on the right side
- Positive skewness: right tail (of larger values) is longer and the mass of the distribution is concentrated on the left side



$$Skewness = \frac{\sum (x_i - \bar{x})^3}{n\sigma^3}$$

Shape

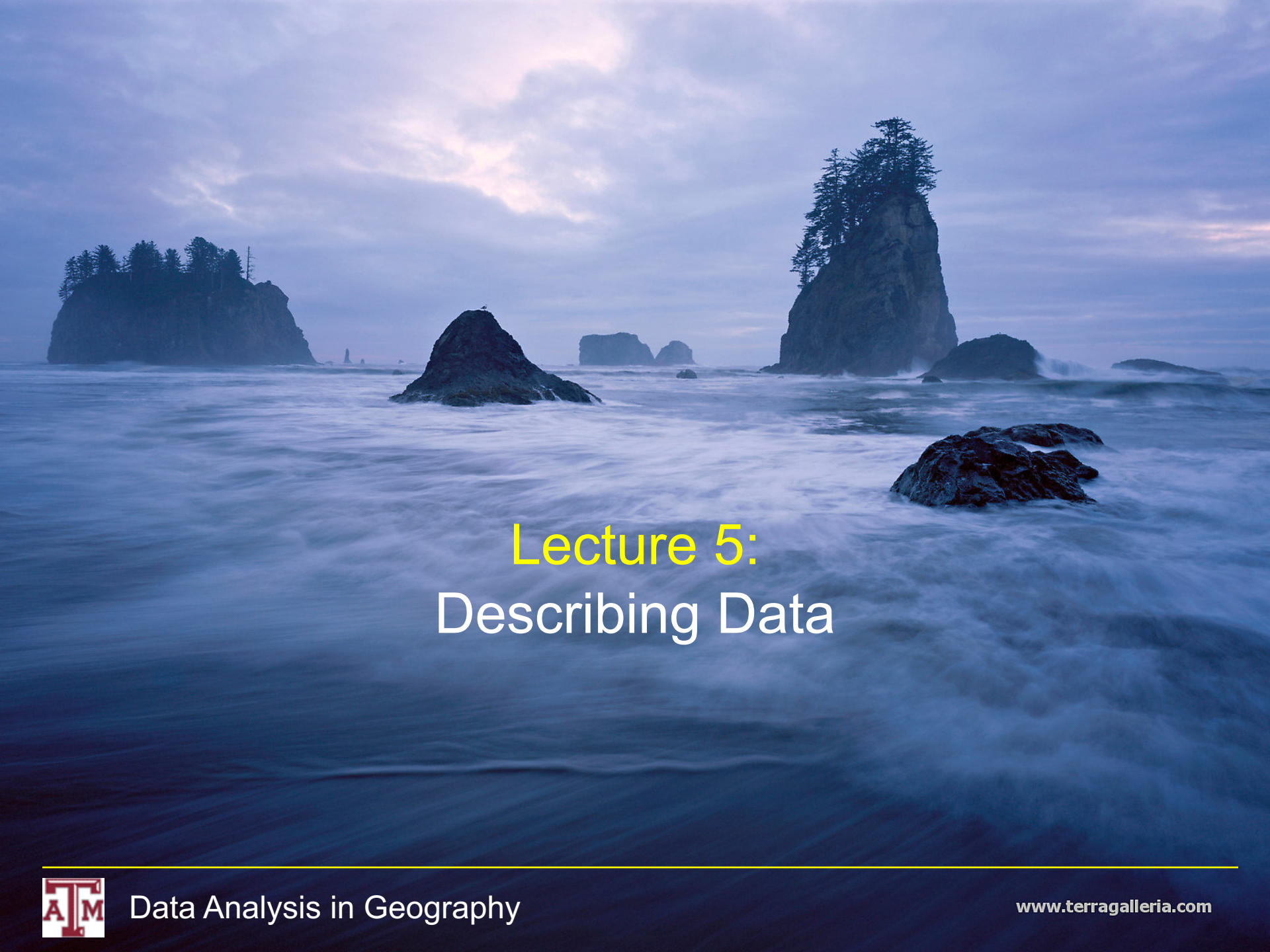
- Kurtosis: Measure of peakedness of the distribution
  - Is the distribution peaked and narrow or flat and wide
- Platykurtic: flat distribution without concentration at the mean- evenly dispersed
- Leptokurtic: peaked distribution with little spread- unevenly dispersed



$$Kurtosis = \frac{\sum (x_i - \bar{x})^4}{n\sigma^4}$$

Normal distribution has a kurtosis of 3

Let's Open up Rstudio…



Lecture 5:
Describing Data