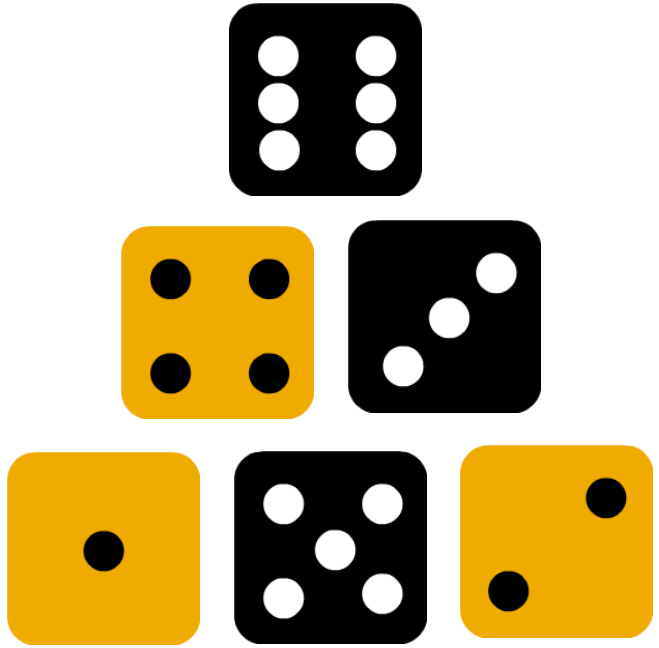
Week 5: Probability Distributions

Unit 1: Properties of Distributions

Introduction

- A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.



http://statisticsbyjim.com/basics/probability-distributions/

https://en.wikipedia.org/wiki/Probability_distribution

Types of probability distribution

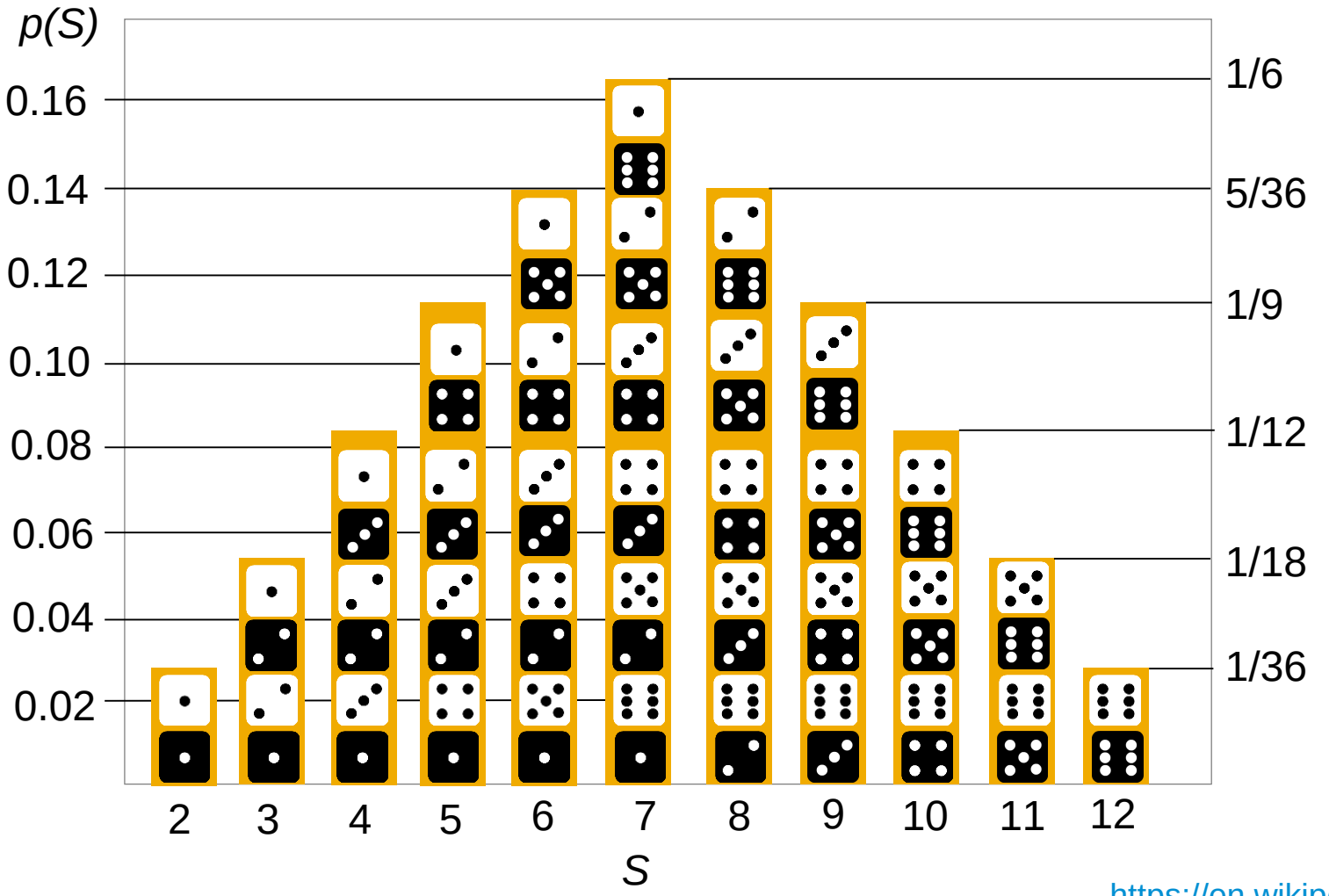
1. Discrete probability distribution

- The set of possible outcomes is discrete

2. Continuous probability distribution

- The set of possible outcomes can take on values in a continuous range

Discrete probability functions



The probability mass function (pmf) of counts from two dice

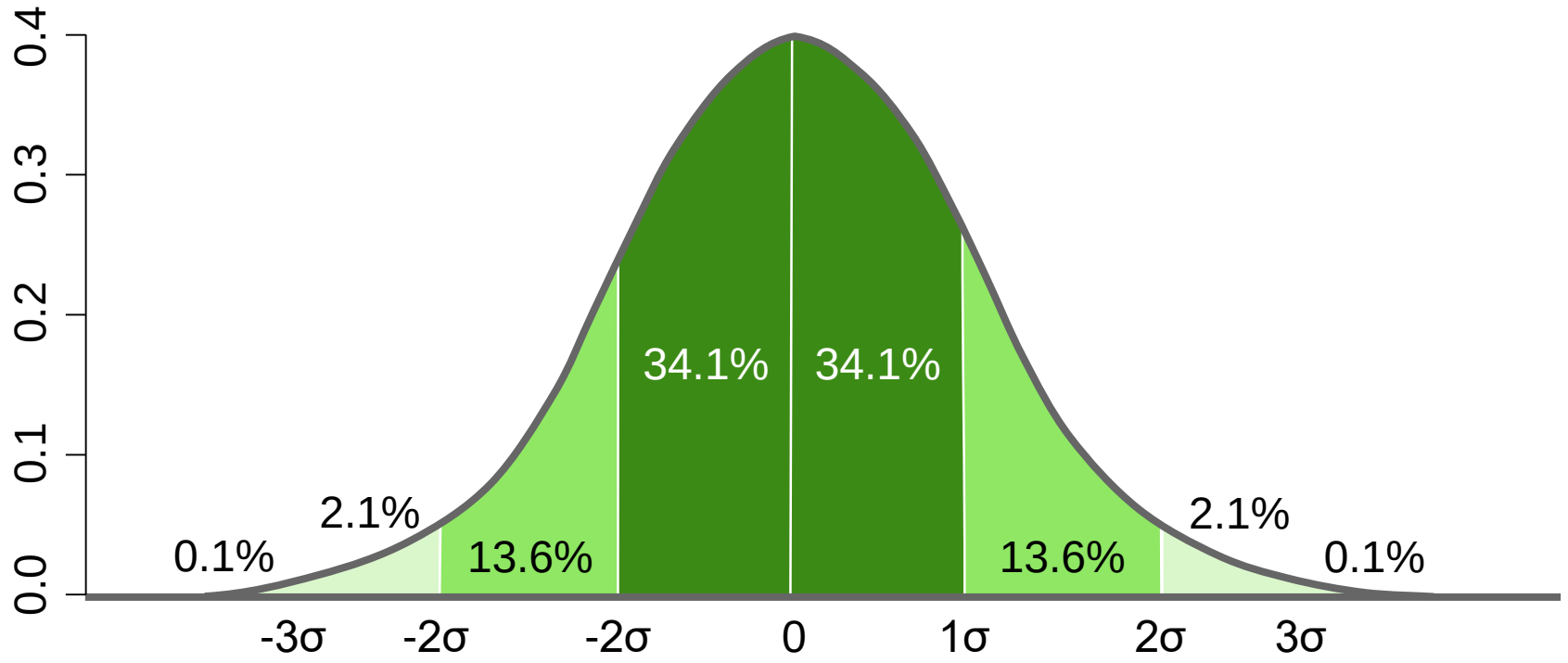
https://en.wikipedia.org/wiki/Probability_distribution

Discrete probability example

Number of Heads	Probability
0	0.25
1	0.50
2	0.25

Flip a coin two times

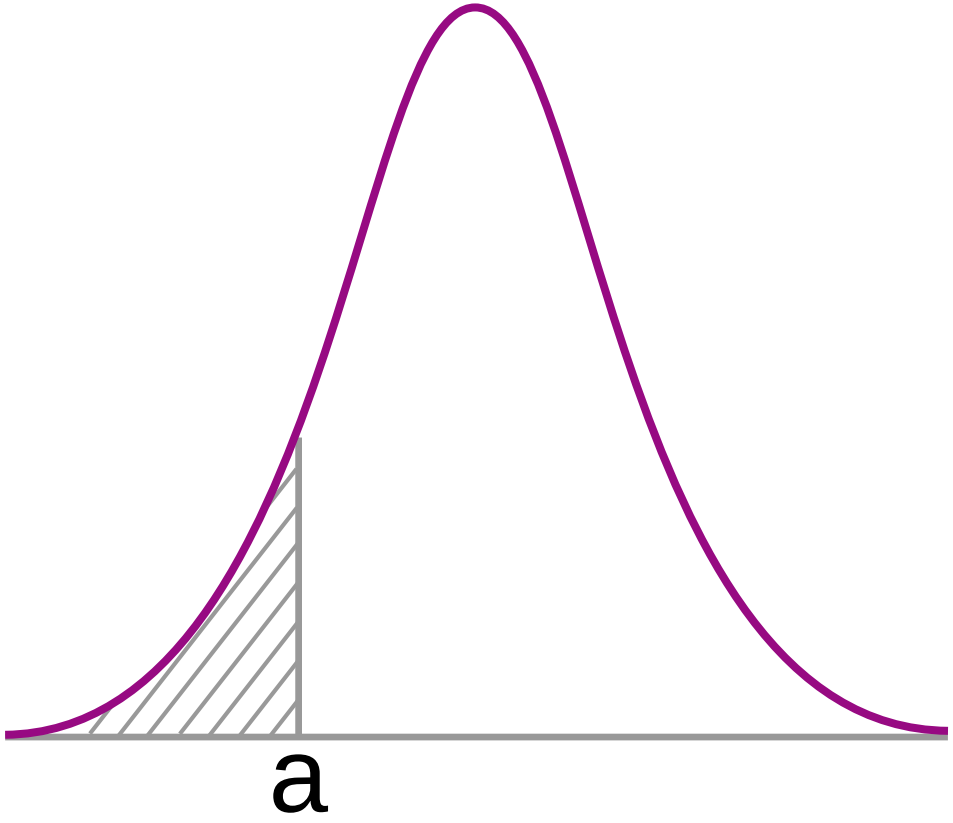
Continuous probability functions



The probability density function (pdf) of the normal distribution

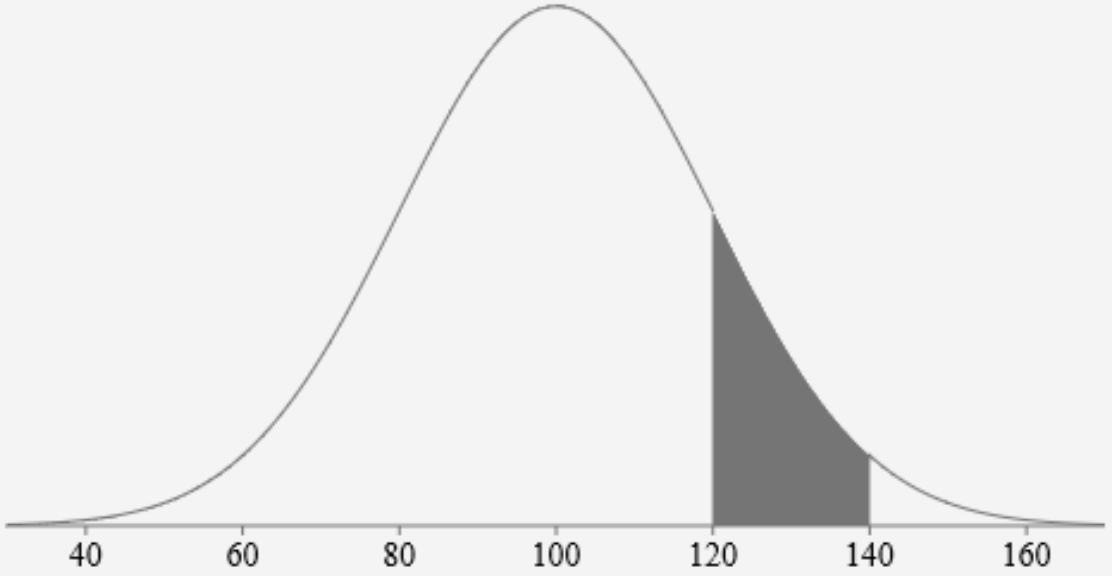
https://en.wikipedia.org/wiki/Probability_distribution

Continuous probability example 1



Refer to https://stattrek.com/probability-distributions/discrete-continuous.aspx for more information.

Continuous probability example 2



Specify Parameters:

Mean 100
SD 20

- Above 1.96
- Below 1.96
- Between 120 and 140
- Outside -1.96 and 1.96

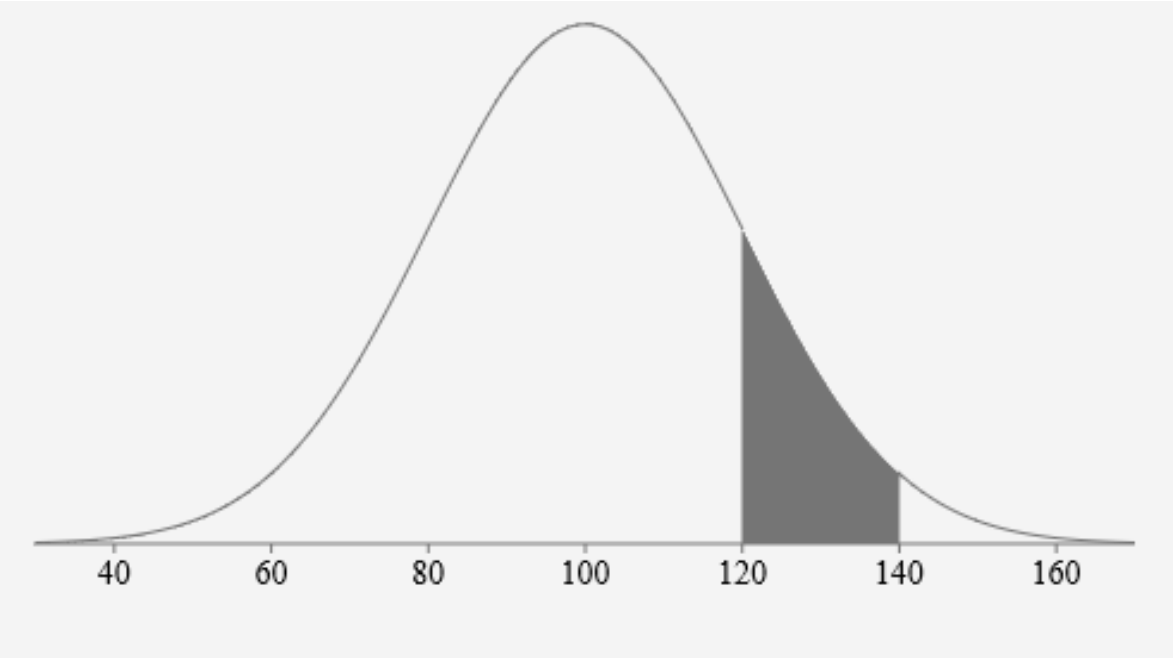
Results:
Area (probability) = 0.1359
Recalculate

http://onlinestatbook.com/2/calculators/normal_dist.html to draw a normal distribution

Continuous probability vs discrete probability distribution

Number of Heads	Probability
0	0.25
1	0.50
2	0.25

Discrete



Continuous

Summary

- A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.
- A discrete random variable can take only a finite number of different values like 0,1,2,3,4, etc., whereas a continuous random variable is a variable that can take an infinite number of possible values.
- Discrete probability functions are also known as “**probability mass functions**” and can assume a discrete number of values.
- Continuous probability functions are also known as “**probability density functions**” and the probabilities are measured over ranges of values rather than single points.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

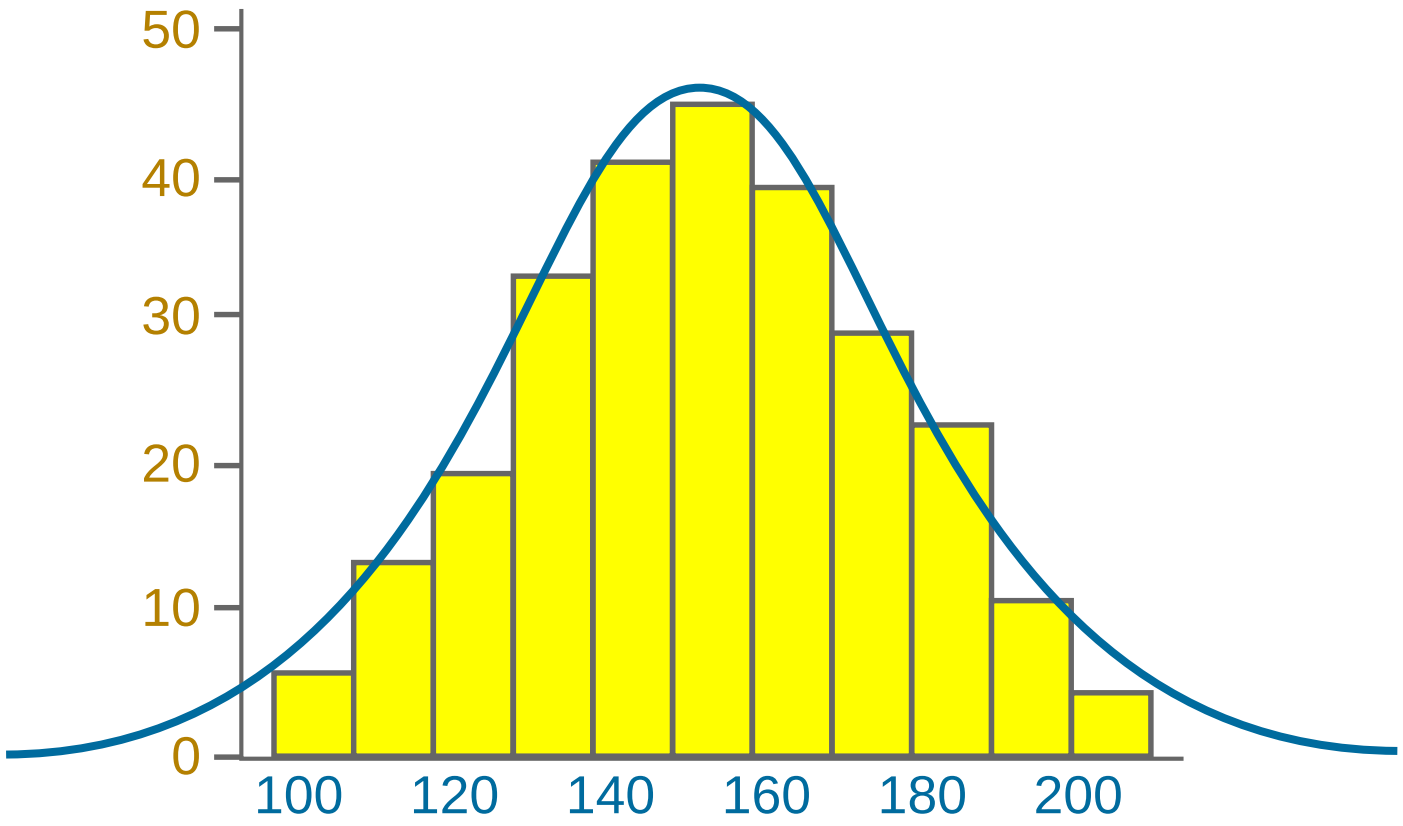
www.sap.com/contactsap

THE BEST RUN SAP®

Week 5: Probability Distributions

Unit 2: The Normal Distribution

Introduction



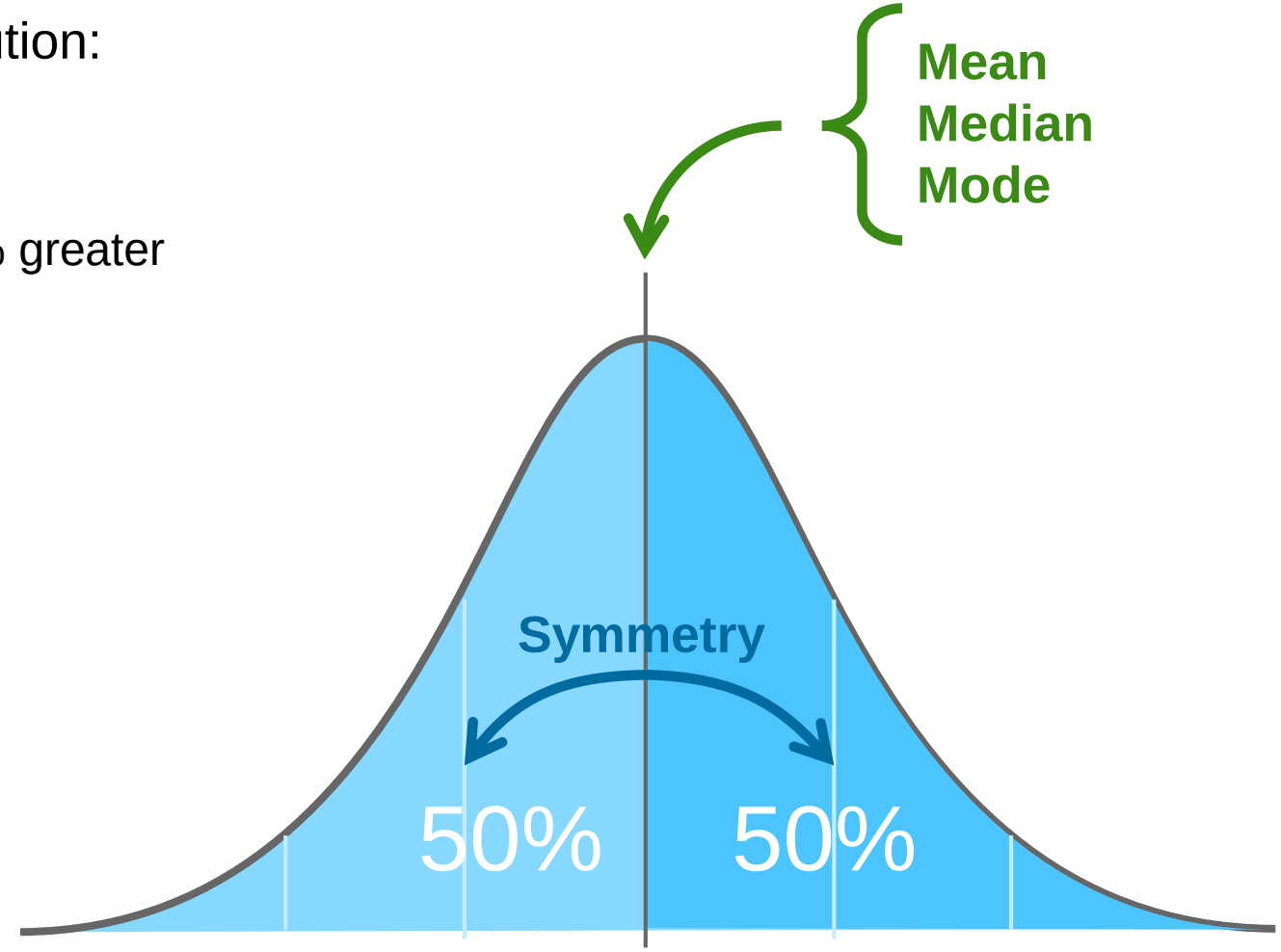
The Normal Distribution

https://www.mathsisfun.com/data/standard-normal-distribution.html
https://en.wikipedia.org/wiki/Normal_distribution

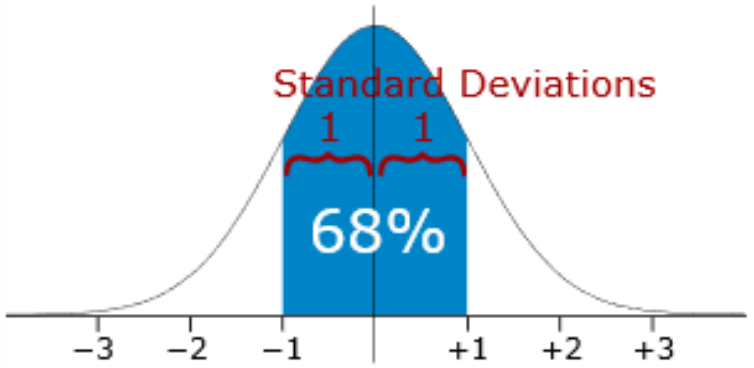
Characteristics

The characteristics of the normal distribution:

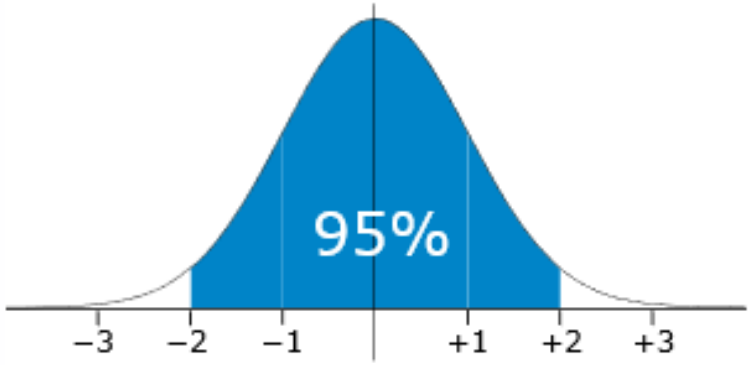
- mean = median = mode
- symmetry about the centre
- 50% of values less than the mean and 50% greater than the mean



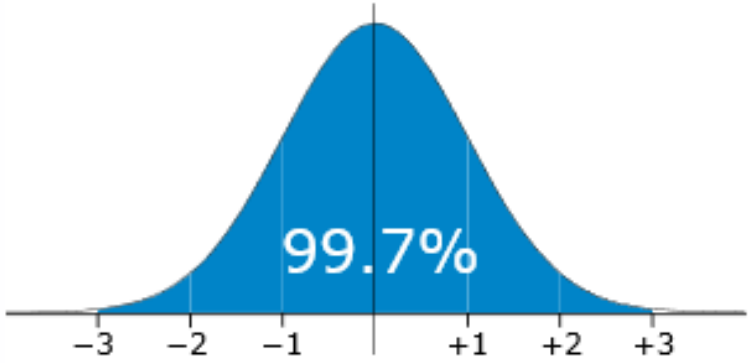
Standard deviation



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



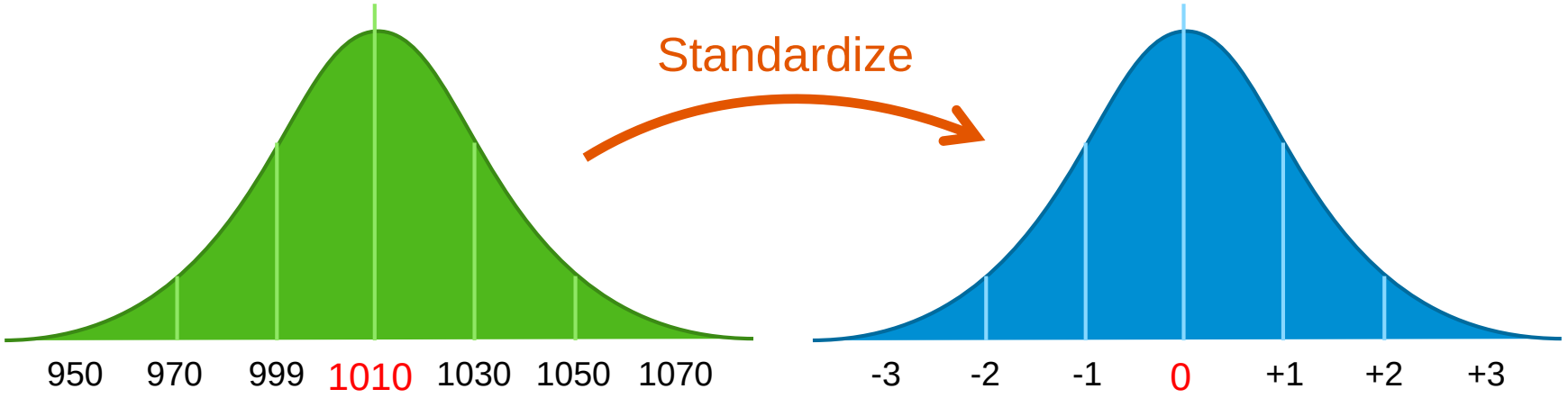
95% of values are within 2 standard deviations of the mean



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

For a standard deviation calculator, see:
https://www.mathsisfun.com/data/standard-deviation-calculator.html

Standard normal distribution



A Normal Distribution

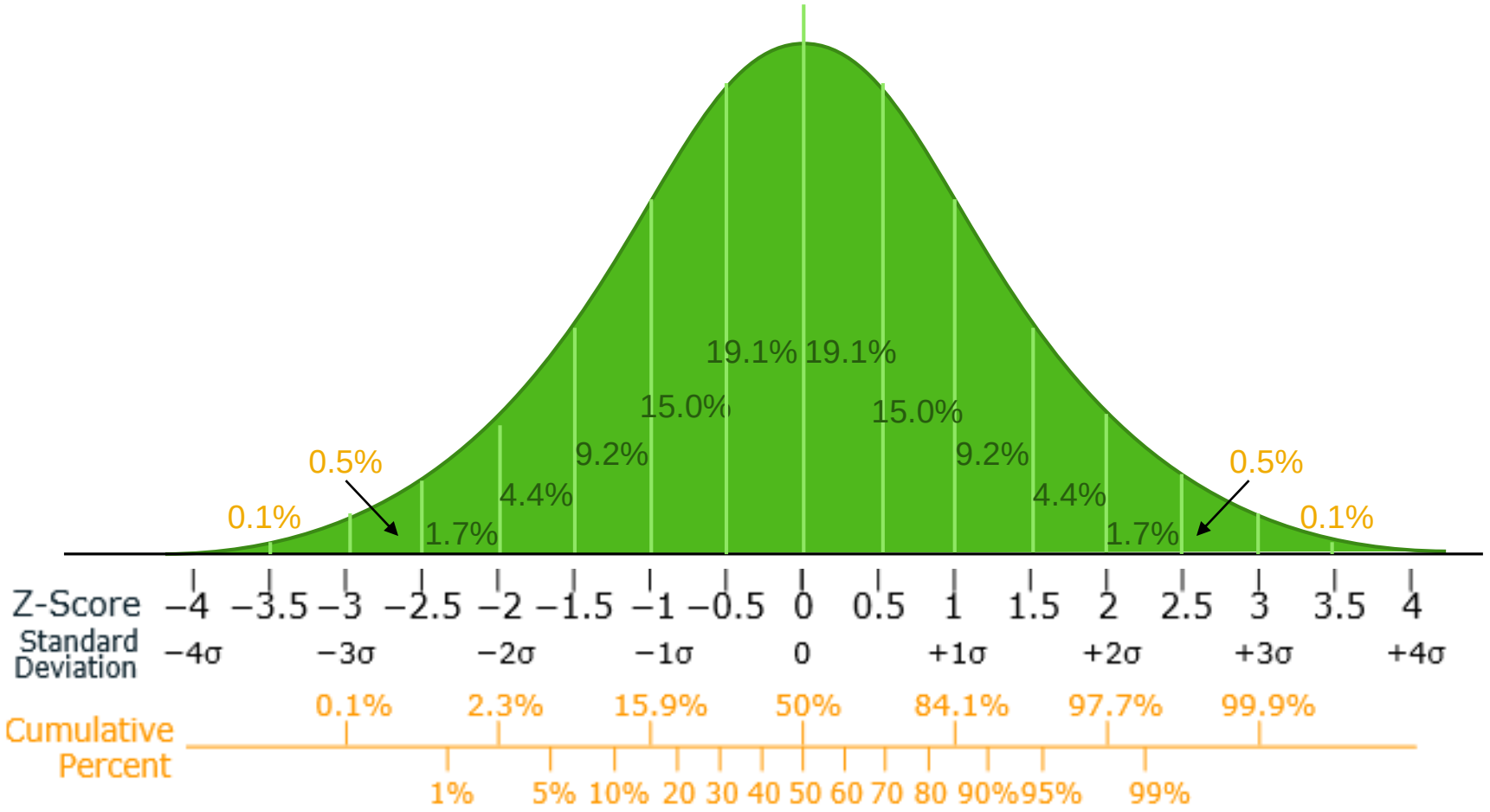
The Standard Normal Distribution

The formula for the z-score: $$z = \frac{x - \mu}{\sigma}$$

z is the "z-score" (standard score)
x is the value to be standardized
μ is the mean
σ is the standard deviation

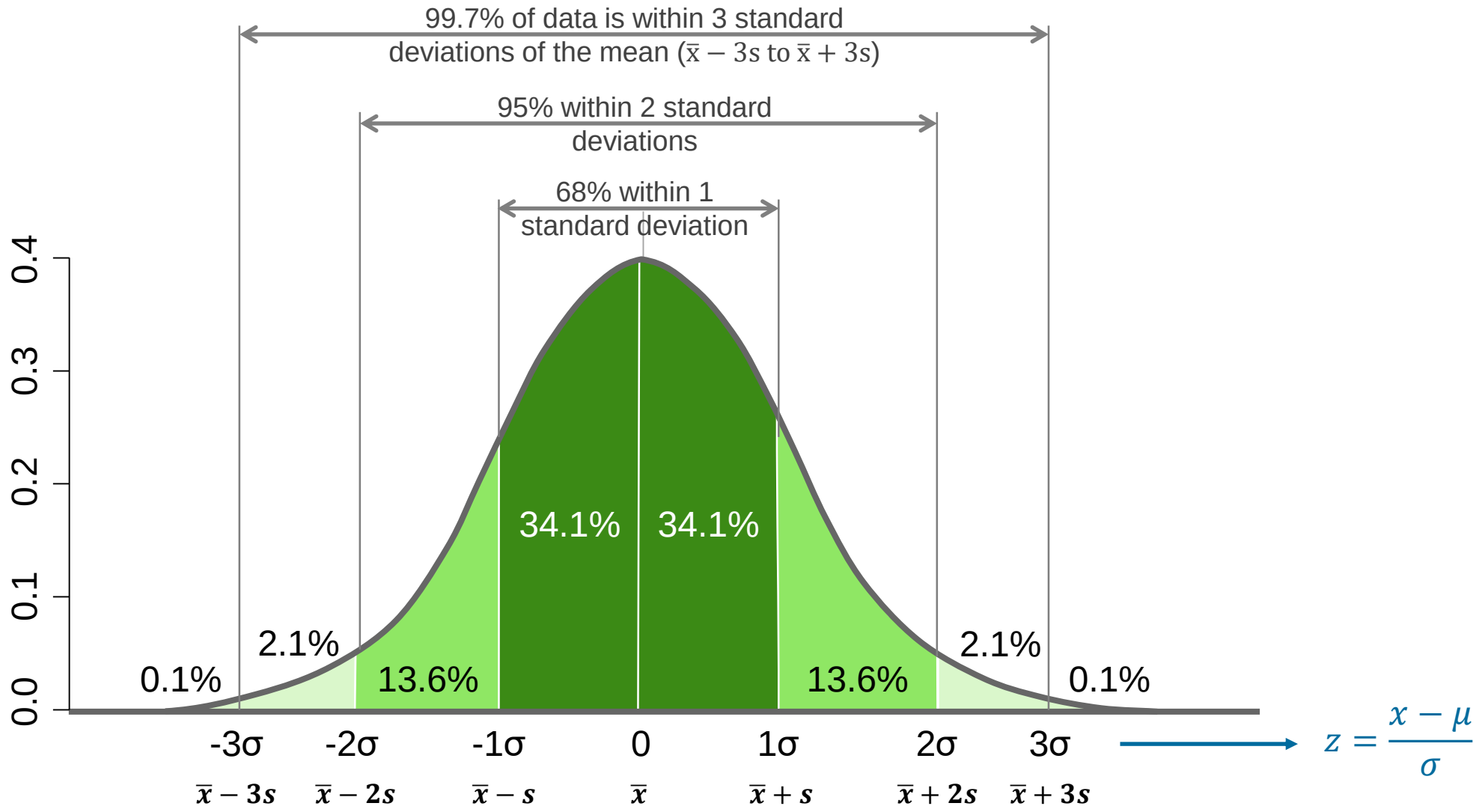
For an interactive standard normal distribution calculator, see:
https://www.mathsisfun.com/data/standard-normal-distribution-table.html

Standard normal distribution example



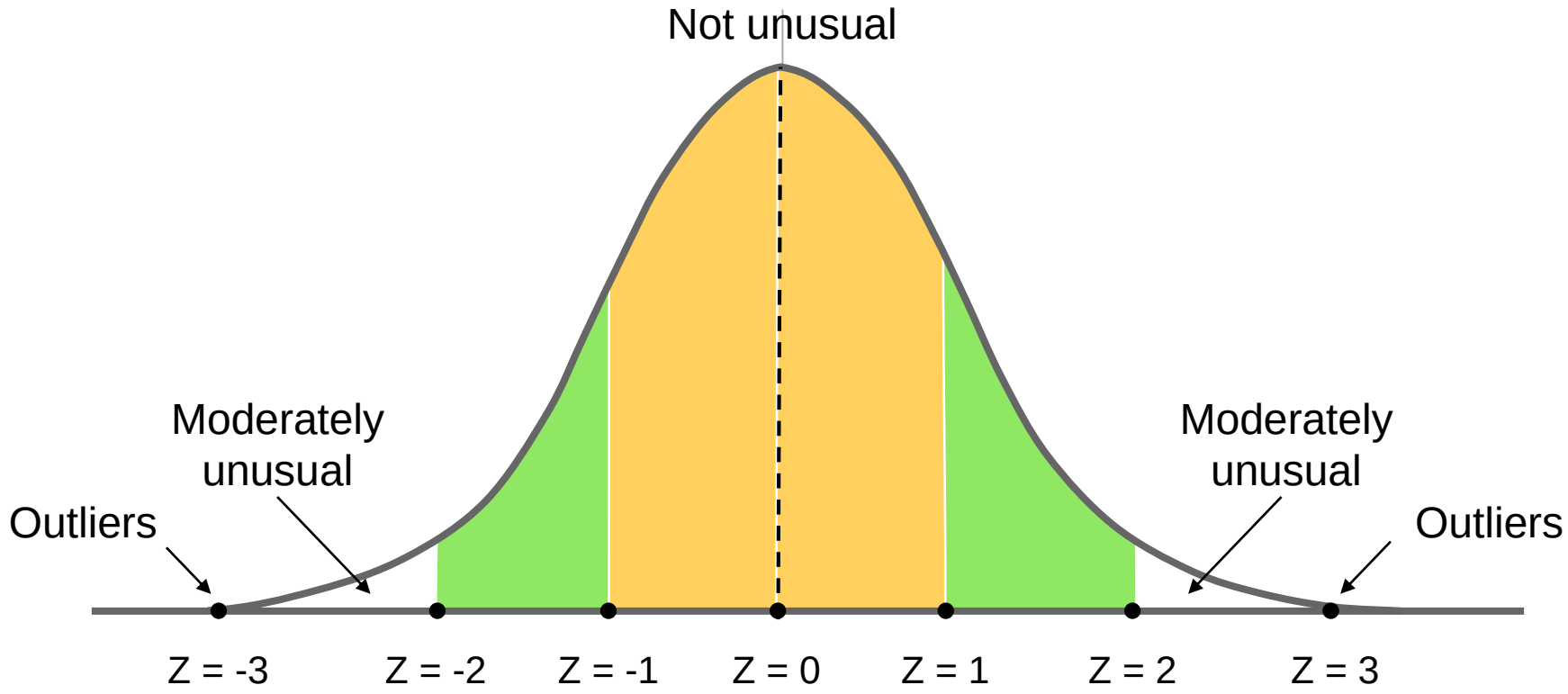
Standard Normal Distribution

The empirical rule

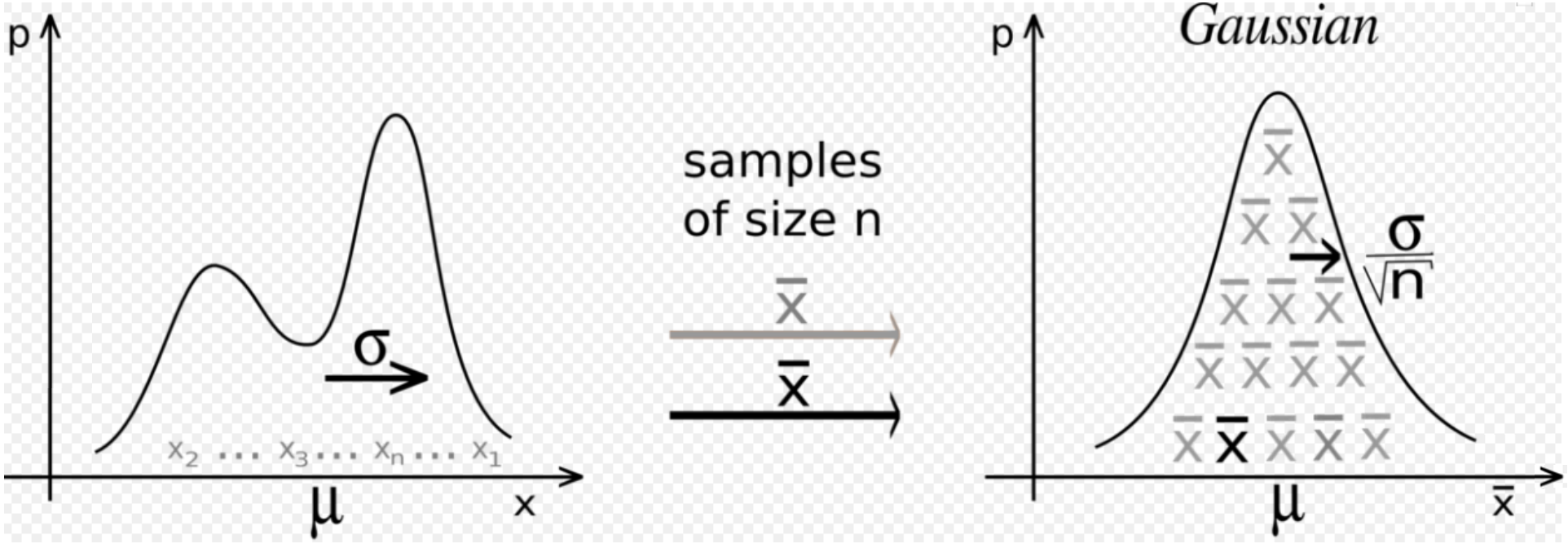


Rules of thumb for detecting outliers

Possible Outliers	***Outliers***
$\|z\| > 2$	$\|z\| > 3$



Central limit theorem



population distribution

sampling distribution of the mean

https://en.wikipedia.org/wiki/Central_limit_theorem

https://machinelearningmastery.com/a-gentle-introduction-to-the-central-limit-theorem-for-machine-learning/

Summary

- The **normal** distribution is a very commonly encountered continuous probability distribution.
- The characteristics of the normal distribution are:
  - mean = median = mode
  - symmetry about the centre
  - 50% of values less than the mean and 50% greater than the mean
- When we calculate the standard deviation, we find that generally:
  - 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
- The **empirical rule** states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean.
- The **central limit theorem** (CLT) establishes that when independent random variables are added, their properly normalized sum tends towards a normal distribution even if the original variables themselves are not normally distributed.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

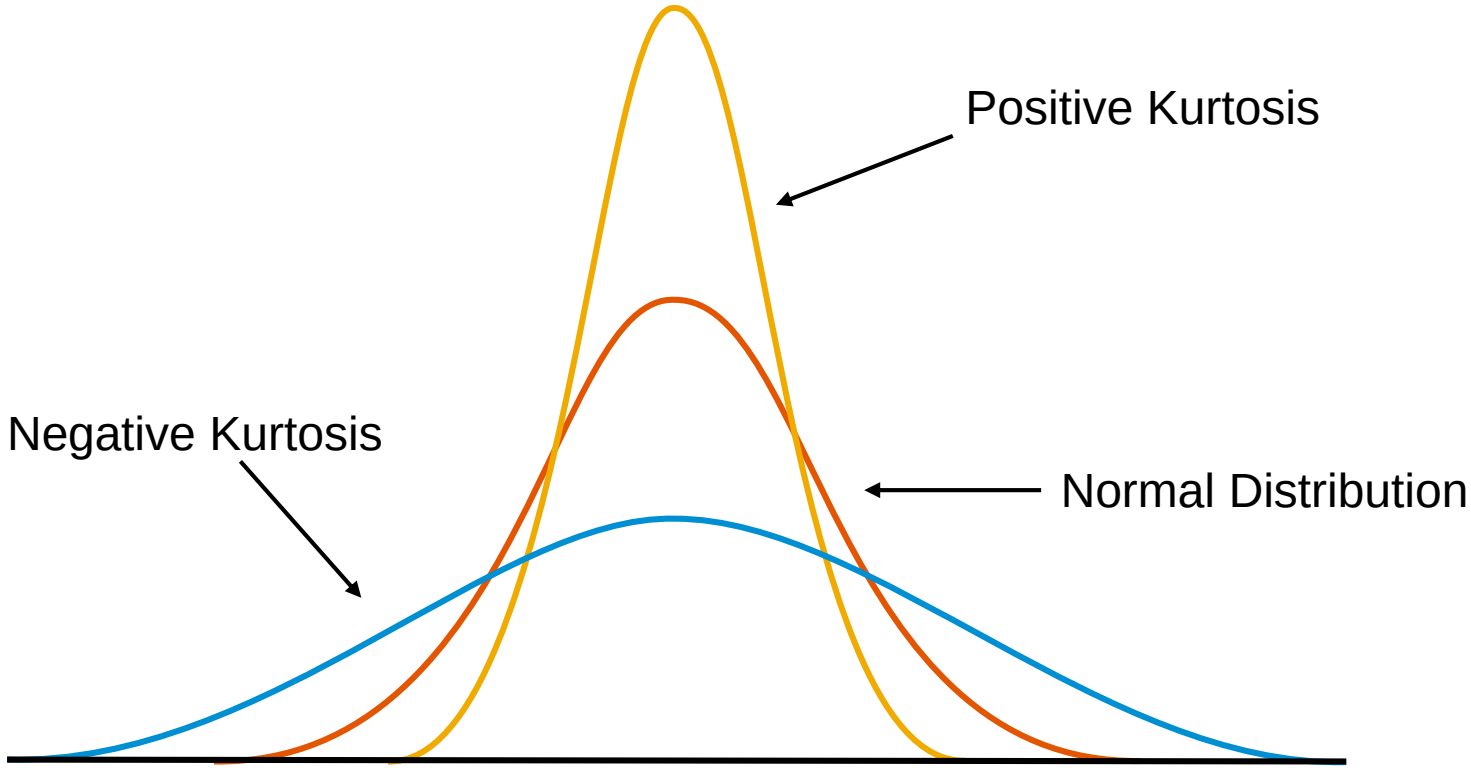
www.sap.com/contactsap

THE BEST RUN SAP®

Week 5: Probability Distributions

Unit 3: Kurtosis and Skewness

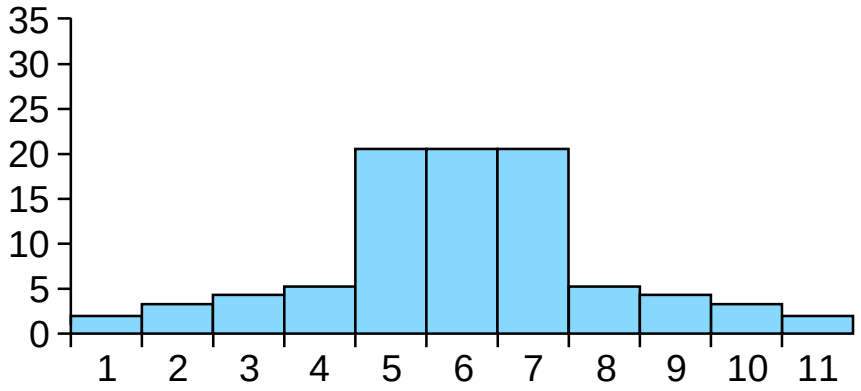
Introduction to kurtosis



https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/

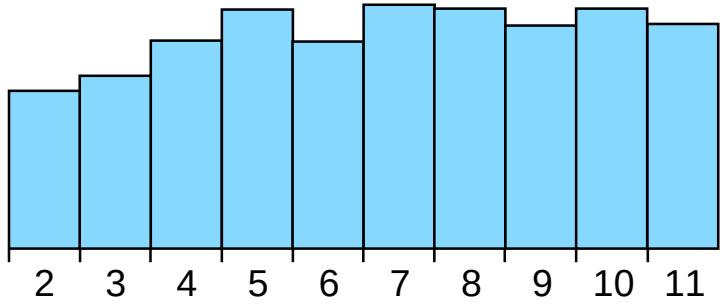
Kurtosis

- Data sets with high, positive kurtosis tend to have heavy tails, or outliers.



- This distribution has positive kurtosis (heavier tails compared to the normal distribution)

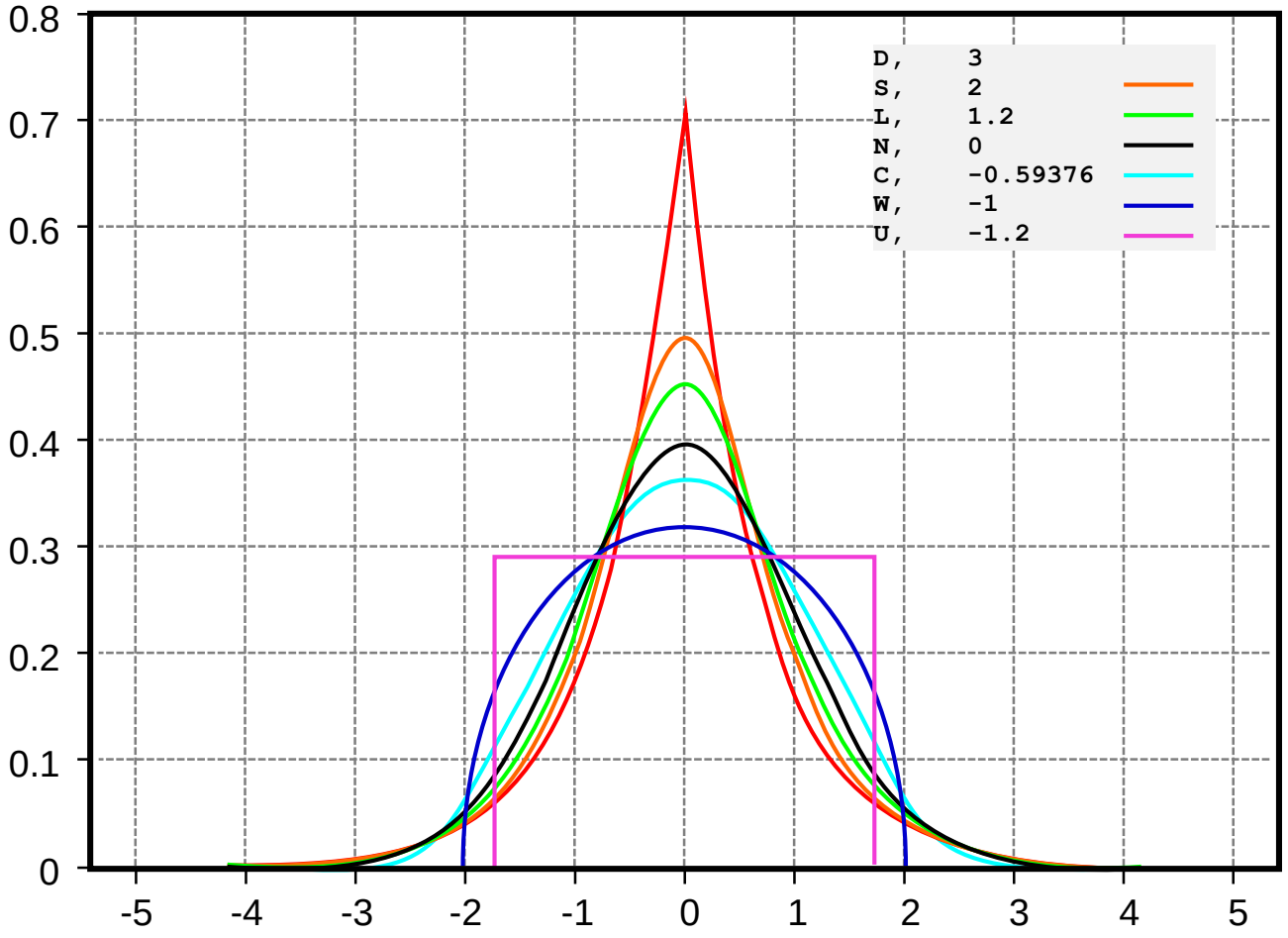
- Data sets with low kurtosis tend to have light tails, or lack of outliers.



- This distribution has low kurtosis (no tails)

https://en.wikipedia.org/wiki/Kurtosis
https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics

Excess kurtosis



Key:

Red, kurt 3, Laplace (D)ouble exponential distribution;

Orange, kurt 2, hyperbolic (S)ecant distribution;

Green, kurt 1.2, (L)ogistic distribution;

Black, kurt 0, (N)ormal distribution;

Cyan, kurt −0.593762…, raised (C)osine distribution;

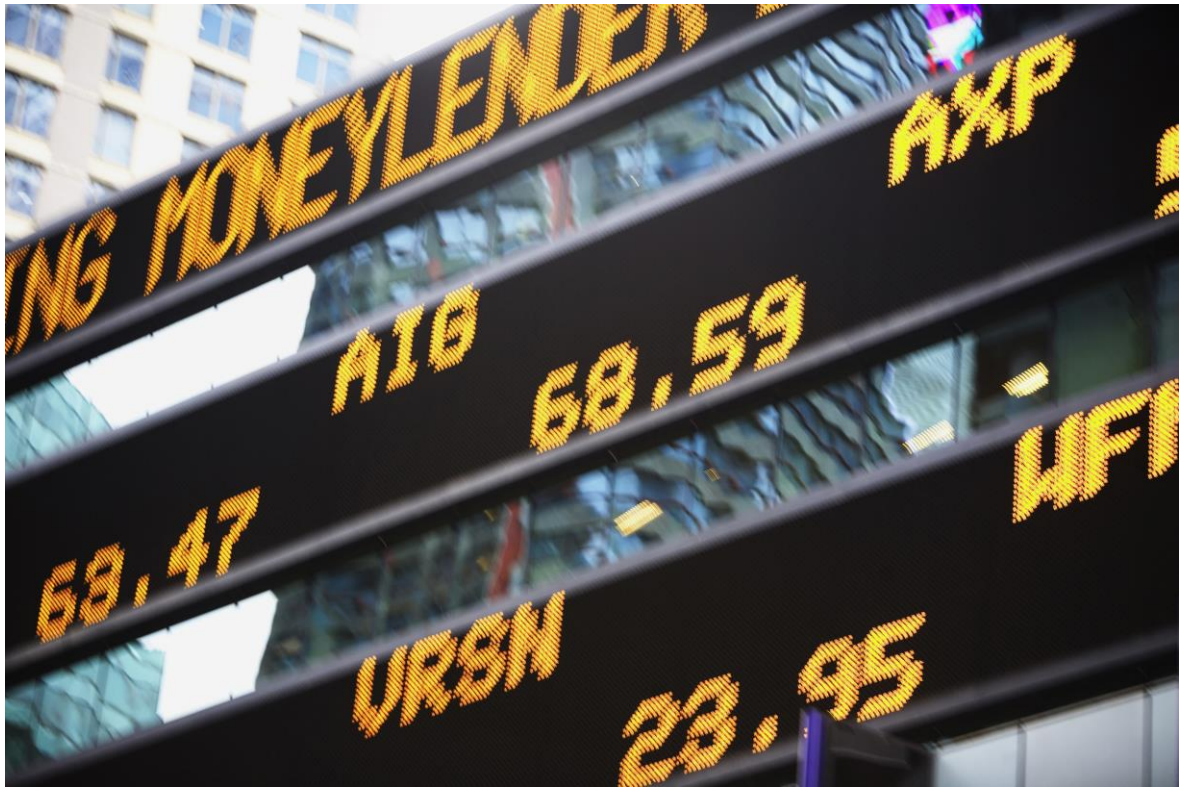
Blue, kurt −1, (W)igner semicircle distribution;

Magenta, kurt −1.2, (U)niform distribution.

https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/

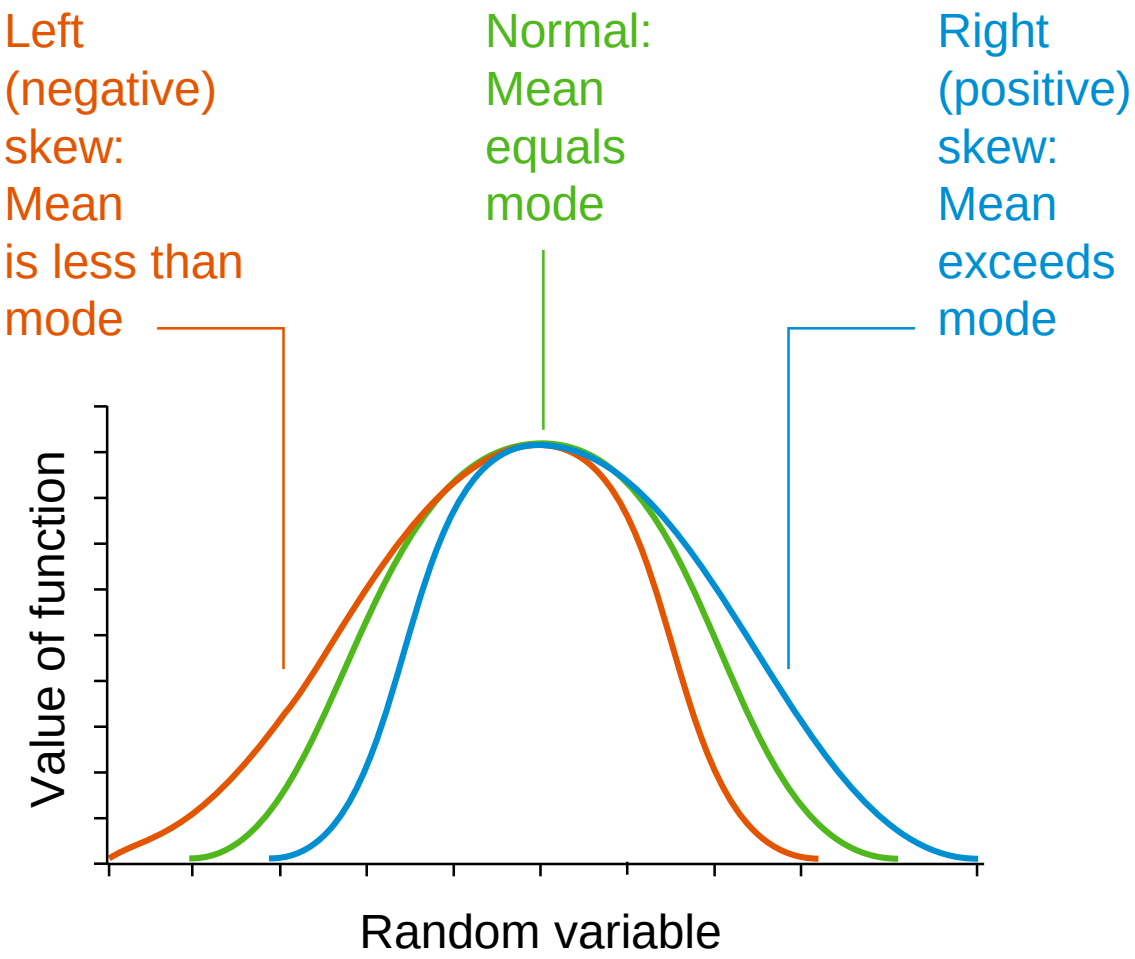
Kurtosis in financial markets

- Real estate (with a kurtosis of 8.75) and high yield US bonds (8.63) are high risk investments.
- Investment grade US bonds (1.06) and small cap US stocks (1.08) would be considered safer investments.



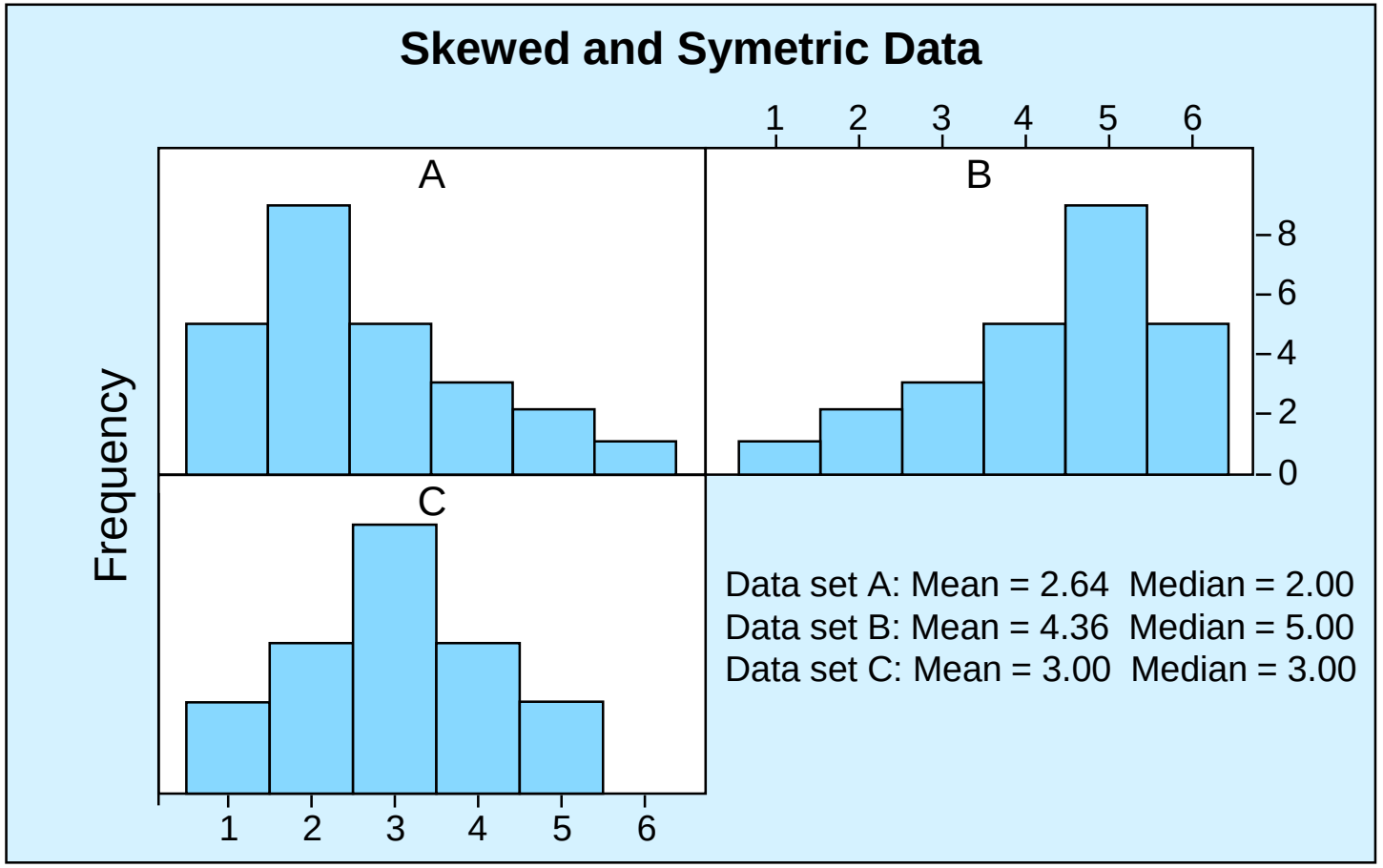
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/

Introduction to skewness



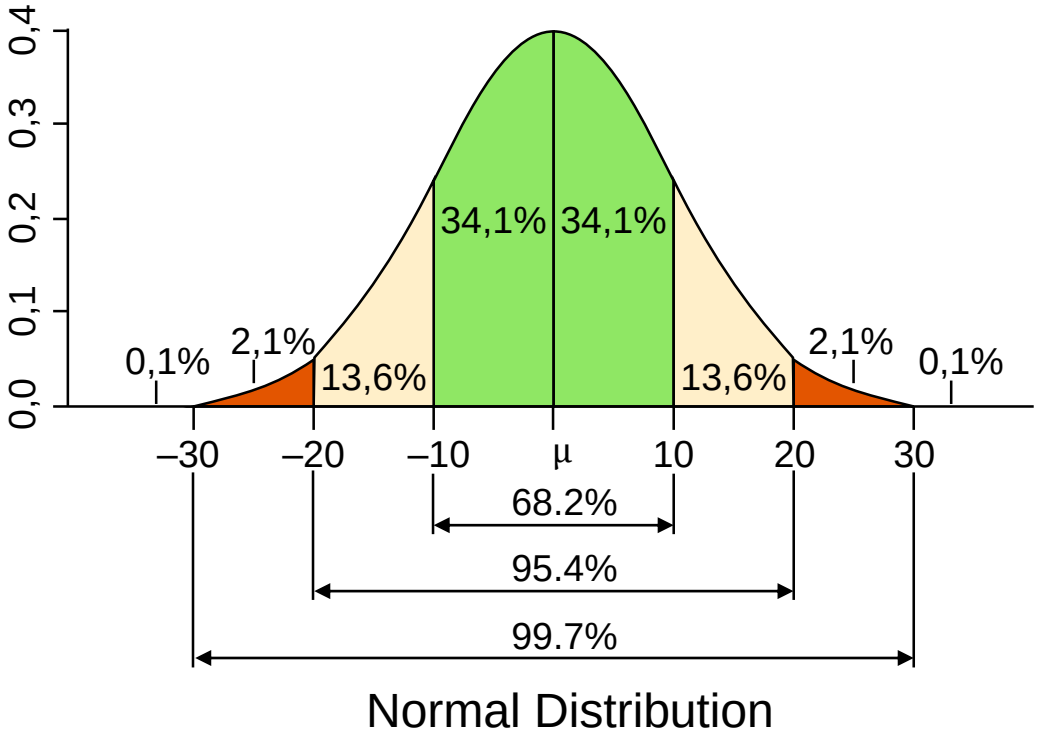
https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
https://whatis.techtarget.com/definition/skewness

Mean and median

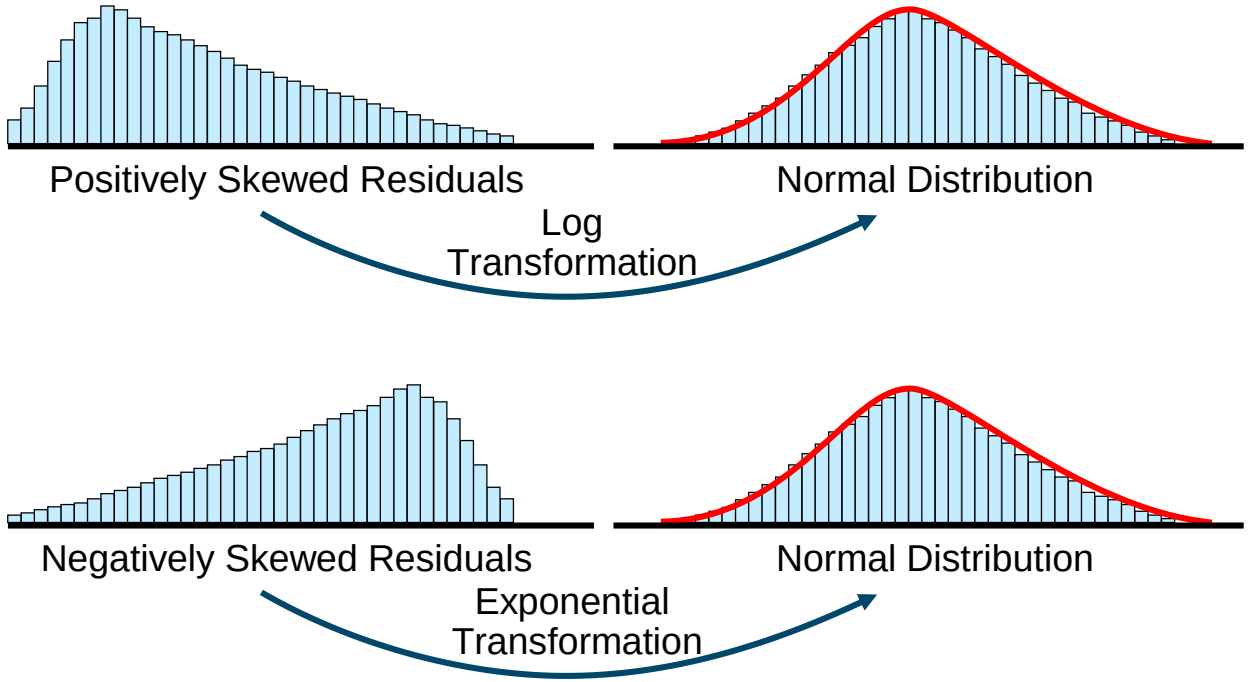


von Hippel, Paul T. (2005). "Mean, Median, and Skew: Correcting a Textbook Rule". Journal of Statistics Education. ***13*** *(2).*
https://en.wikipedia.org/wiki/Skewness

Why is skew important?



Normal Distribution



Data Transformations

https://www.sheffield.ac.uk/polopoly_fs/1.579181!/file/stcp-marshallsamuels-NormalityS.pdf
https://www.quora.com/How-does-skewness-impact-regression-model
https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
https://www.linkedin.com/pulse/question-does-skewness-variable-impact-predictive-data-mosaddar for more information

Summary

- Kurtosis is a measure of the "tailedness" of the probability distribution:
  - Data sets with high kurtosis tend to have heavy tails, or outliers (“**leptokurtic”**).
  - Data sets with low kurtosis tend to have light tails, or lack of outliers (“**platykurtic”**).
  - Distributions with zero excess kurtosis are called “**mesokurtic”** (normal distribution family).
- Skewness is a measure of the asymmetry of a probability distribution.
  - A distribution is symmetric if it looks the same to the left and right of the center point.
  - If most of the data is on the left side of the histogram but a few larger values are on the right, the data is said to be *skewed to the right (positive skew).*
  - If most of the data is on the right, with a few smaller values showing up on the left side of the histogram, the data is *skewed to the left (negative skew).*
  - If the distribution is symmetric, then the mean is equal to the median and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean = median = mode.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

www.sap.com/contactsap

THE BEST RUN SAP®

Week 5: Probability Distributions

Unit 4: Using the Normal Distribution to Calculate Probability

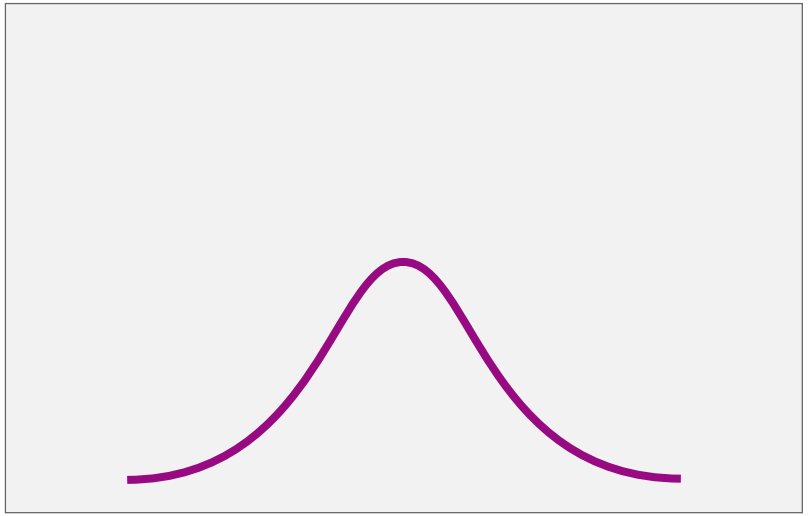
Normal distribution recap

Normal Curve

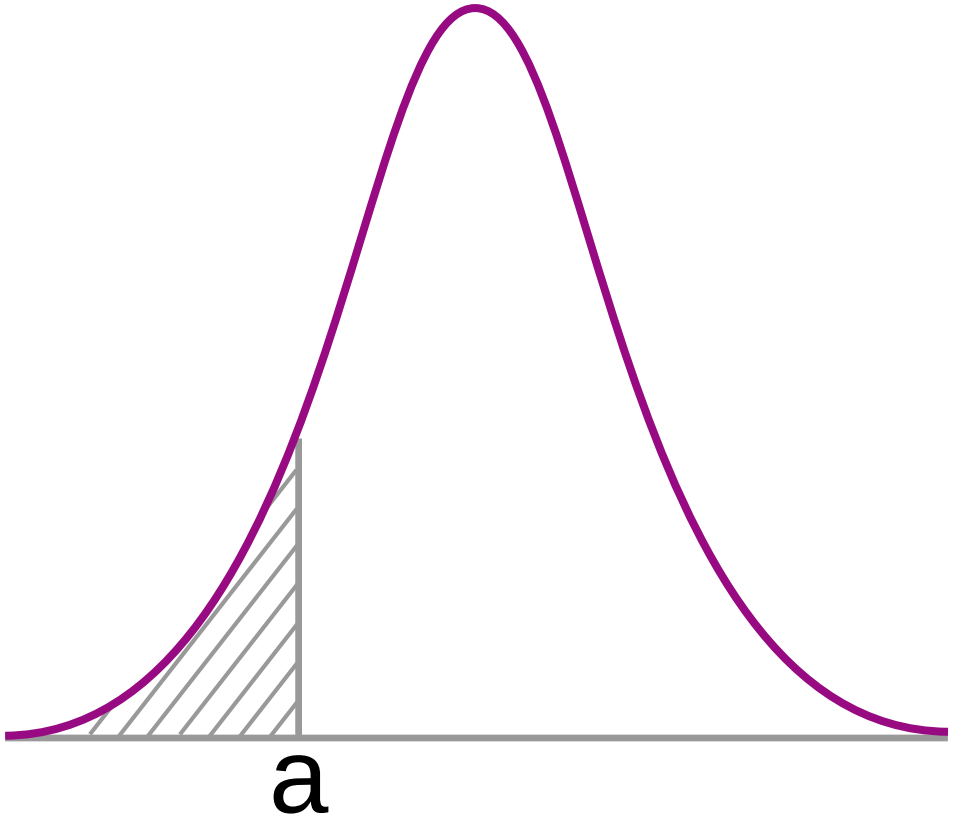
Smaller Standard Deviation

Normal Curve

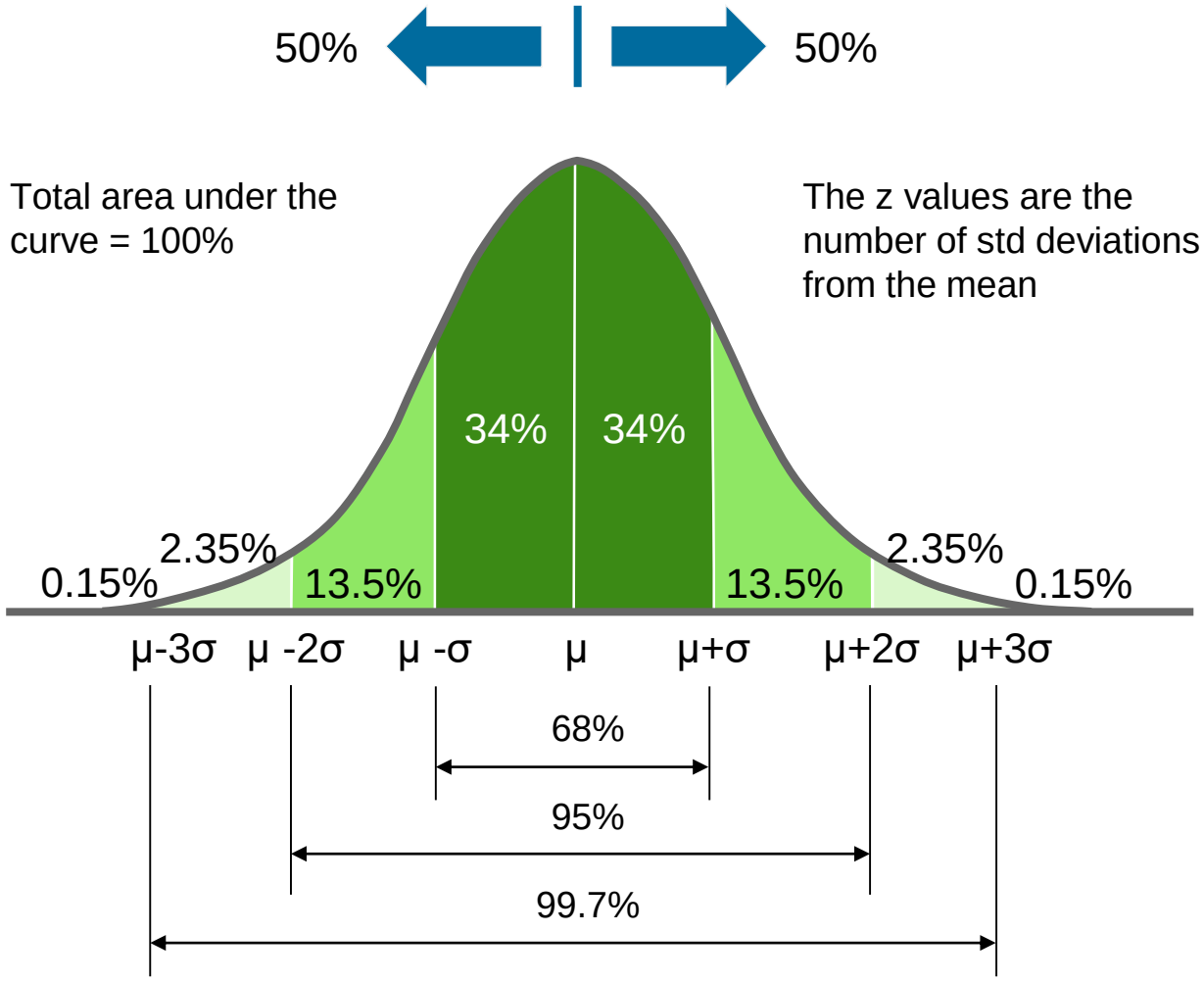
Larger Standard Deviation



Probability and the normal distribution recap



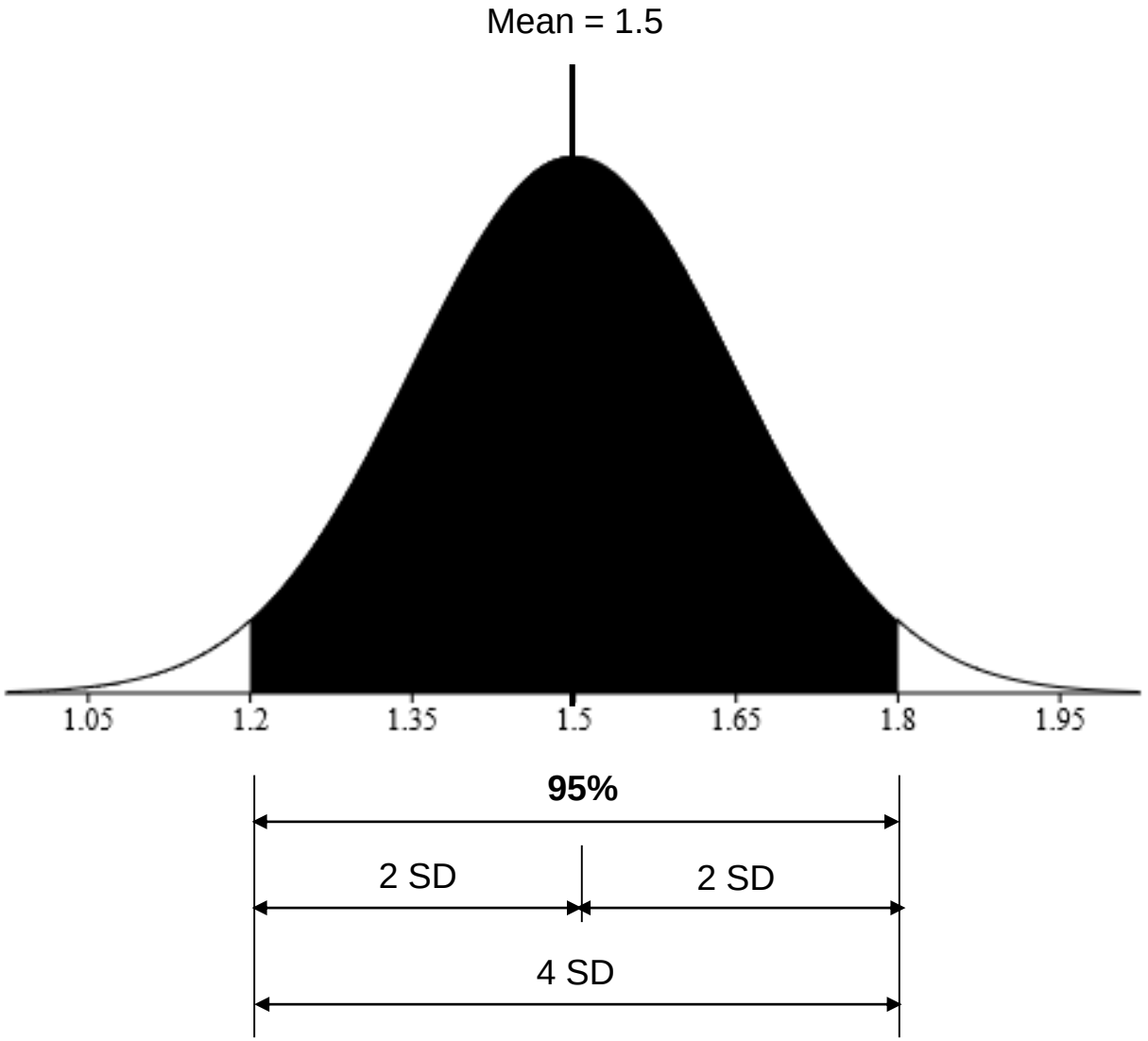
Empirical rule recap



Empirical rule example

Question

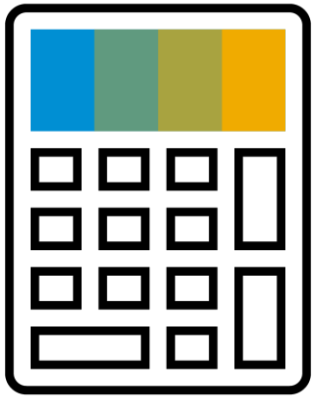
- 95% of students at school are between 1.2m and 1.8m tall.
- Assuming this data is normally distributed, calculate the mean and standard deviation.



http://davidmlane.com/hyperstat/z_table.html

Find probabilities

- How can you use this theory in practice?
- To find the probability associated with a normal random variable, use a graphing calculator, an online normal distribution calculator, or a normal distribution table.
- There are lots of normal distribution calculators available online.



- Here are some examples for you:

https://www.mathportal.org/calculators/statistics-calculator/normal-distribution-calculator.php

https://stattrek.com/online-calculator/normal.aspx

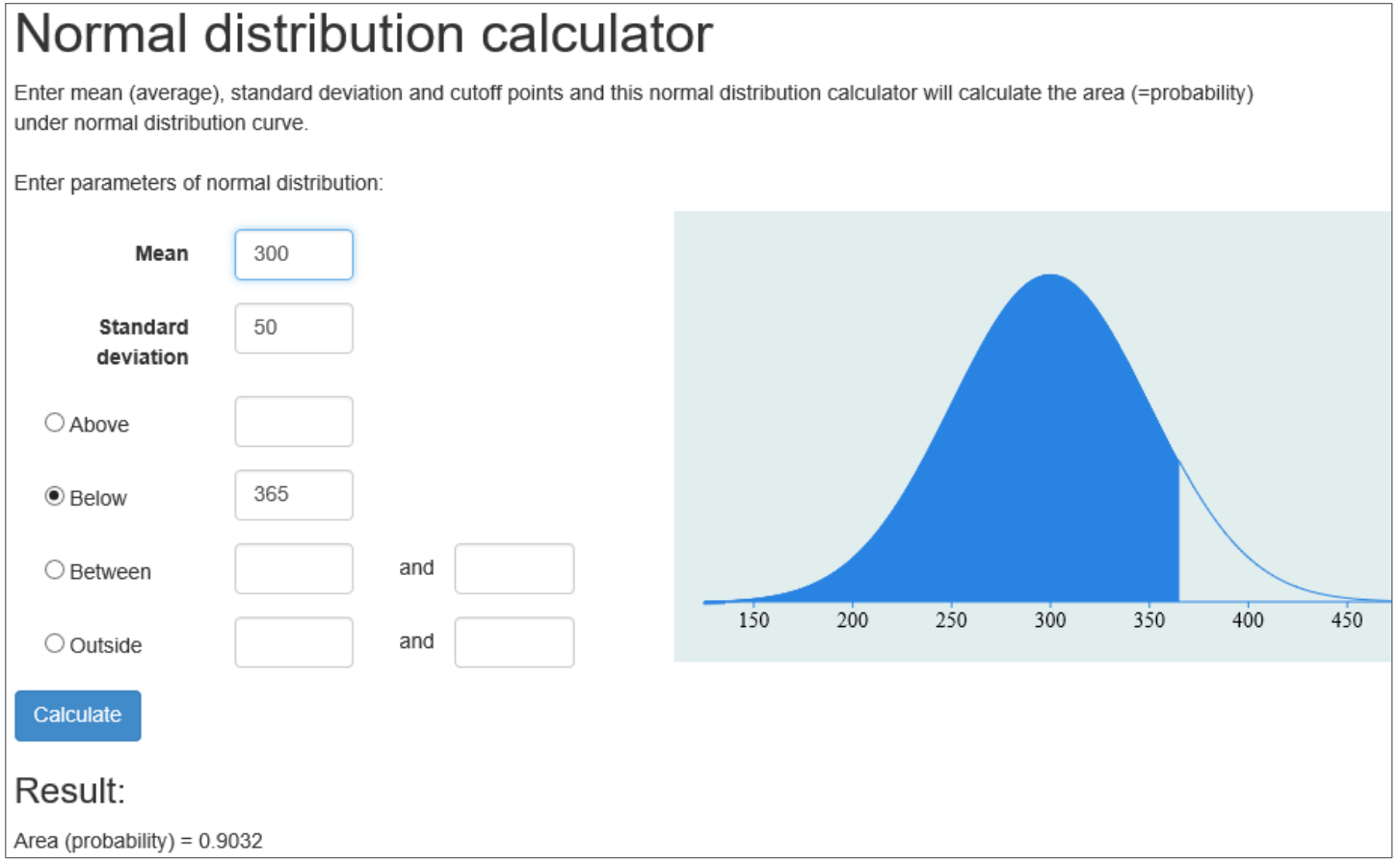
https://www.hackmath.net/en/calculator/normal-distribution

http://davidmlane.com/hyperstat/z_table.html

Example 1

Question

- On average, a light bulb lasts 300 days with a standard deviation of 50 days.
- Assuming that bulb life is normally distributed, what is the probability that the light bulb will last at most 365 days?



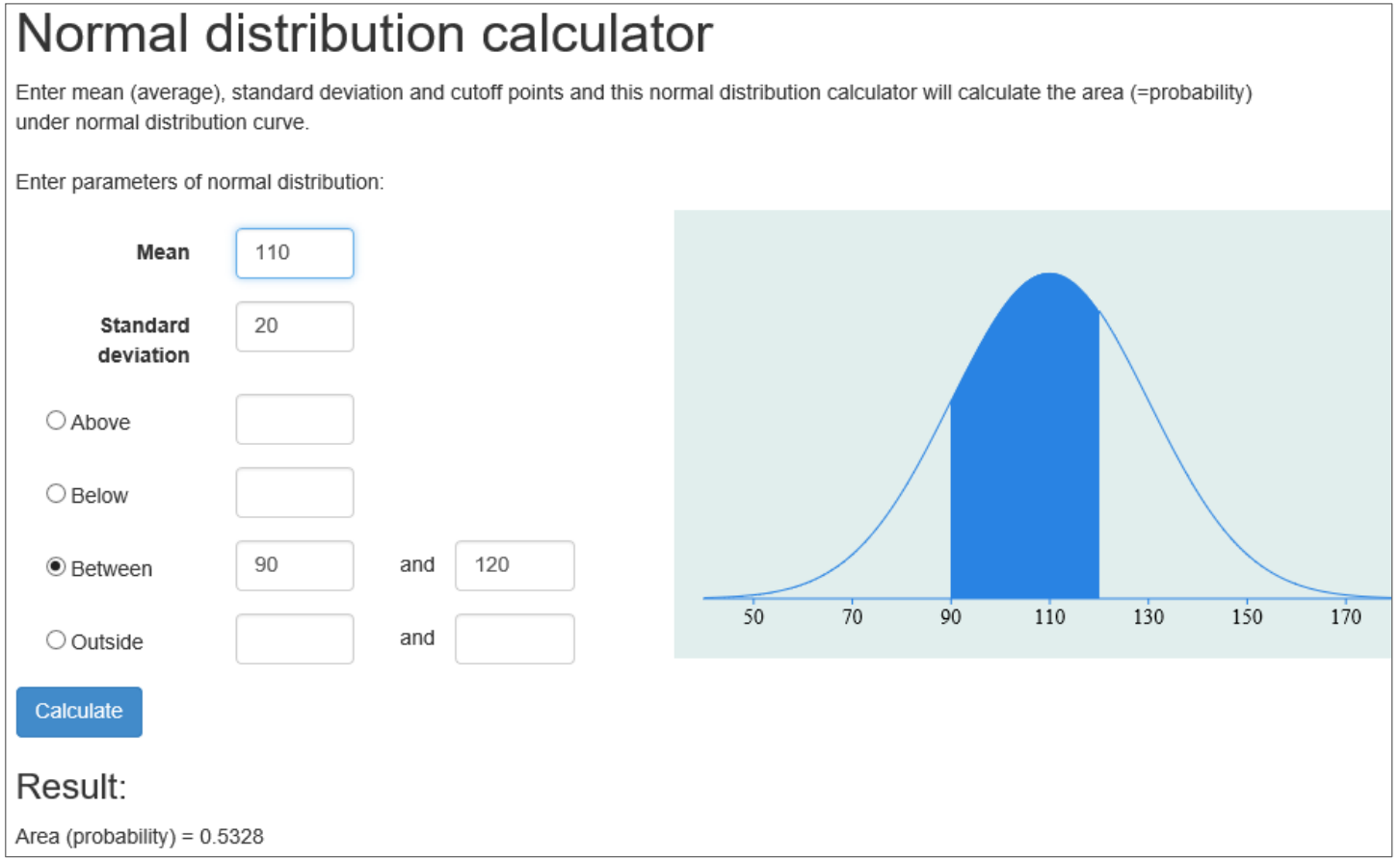
https://www.hackmath.net/en/calculator/normal-distribution

https://www.hackmath.net/en/calculator/normal-distribution?mean=300&sd=50&above=&area=below&below=365&ll=&ul=&outsideLL=&outsideUL=&draw=Calculate

Example 2

Question

- Scores on an IQ test are normally distributed.
- If the test has a mean of 110 and a standard deviation of 20, what is the probability that a person who takes the test will score between 90 and 120?

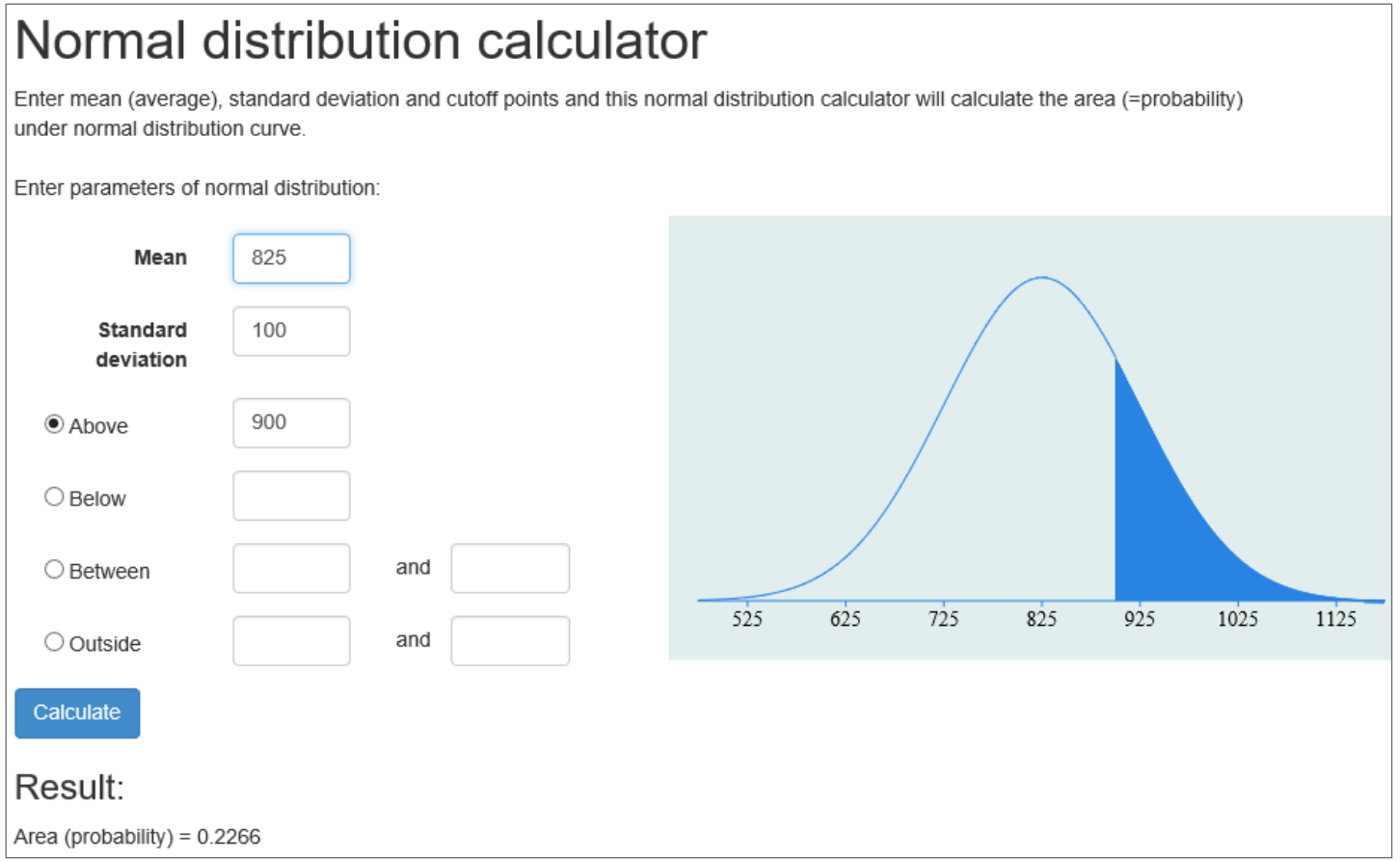


https://www.hackmath.net/en/calculator/normal-distribution?mean=110&sd=20&above=&below=&area=between&ll=90&ul=120&outsideLL=&outsideUL=&draw=Calculate

Example 3

Question

- A student achieved a score of 900 in an exam.
- The mean test score was 825 with a standard deviation of 100.
- Assuming that test scores are normally distributed, what proportion of students achieved a higher score than 900?



https://www.hackmath.net/en/calculator/normal-distribution?mean=825&sd=100&area=above&above=900&below=&ll=&ul=&outsideLL=&outsideUL=&draw=Calculate

Summary

- The **normal distribution** refers to a family of continuous probability distributions.
- The area under the normal distribution curve can be used to calculate probabilities for a normally distributed random variable.
- There are lots of normal distribution calculators available online. Given the mean and standard deviation, the calculator can be used to calculate the area under the normal curve (the probability):
  - less than a value
  - greater than a value
  - between values
  - outside two values

https://stattrek.com/probability-distributions/normal.aspx

https://www.mathsisfun.com/data/standard-normal-distribution.html

https://statistics.laerd.com/statistical-guides/normal-distribution-calculations.php

Thank you.

Contact information:

open@sap.com

Follow all of SAP

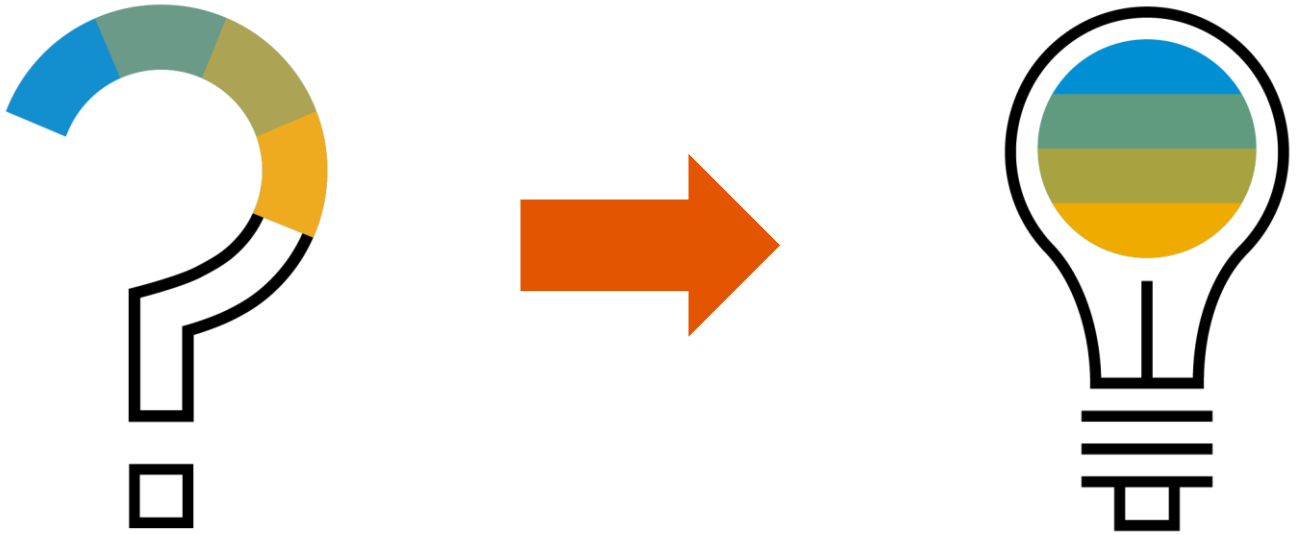
www.sap.com/contactsap

THE BEST RUN SAP®

Week 5: Probability Distributions

Unit 5: Hypothesis Testing

Introduction

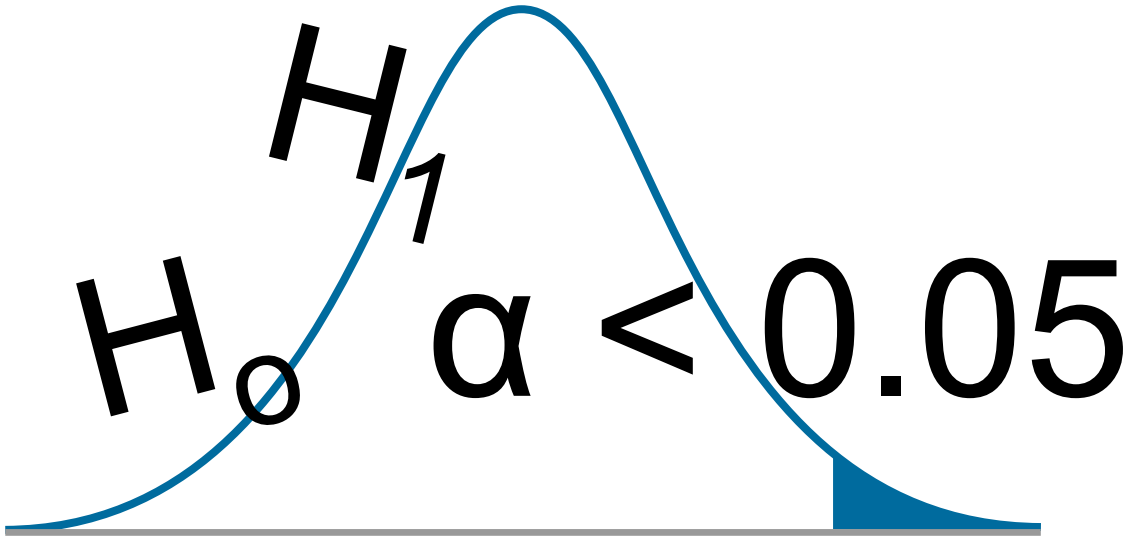


Null and alternative hypotheses

Determine whether a coin was fairly balanced:

- A null hypothesis H_0 might be that half the flips would result in Heads and half in Tails.
- The alternative hypothesis H_a might be that the number of Heads and Tails would be very different.

$$H_o: P = 0.5$$
$$H_1: P \neq 0.5$$



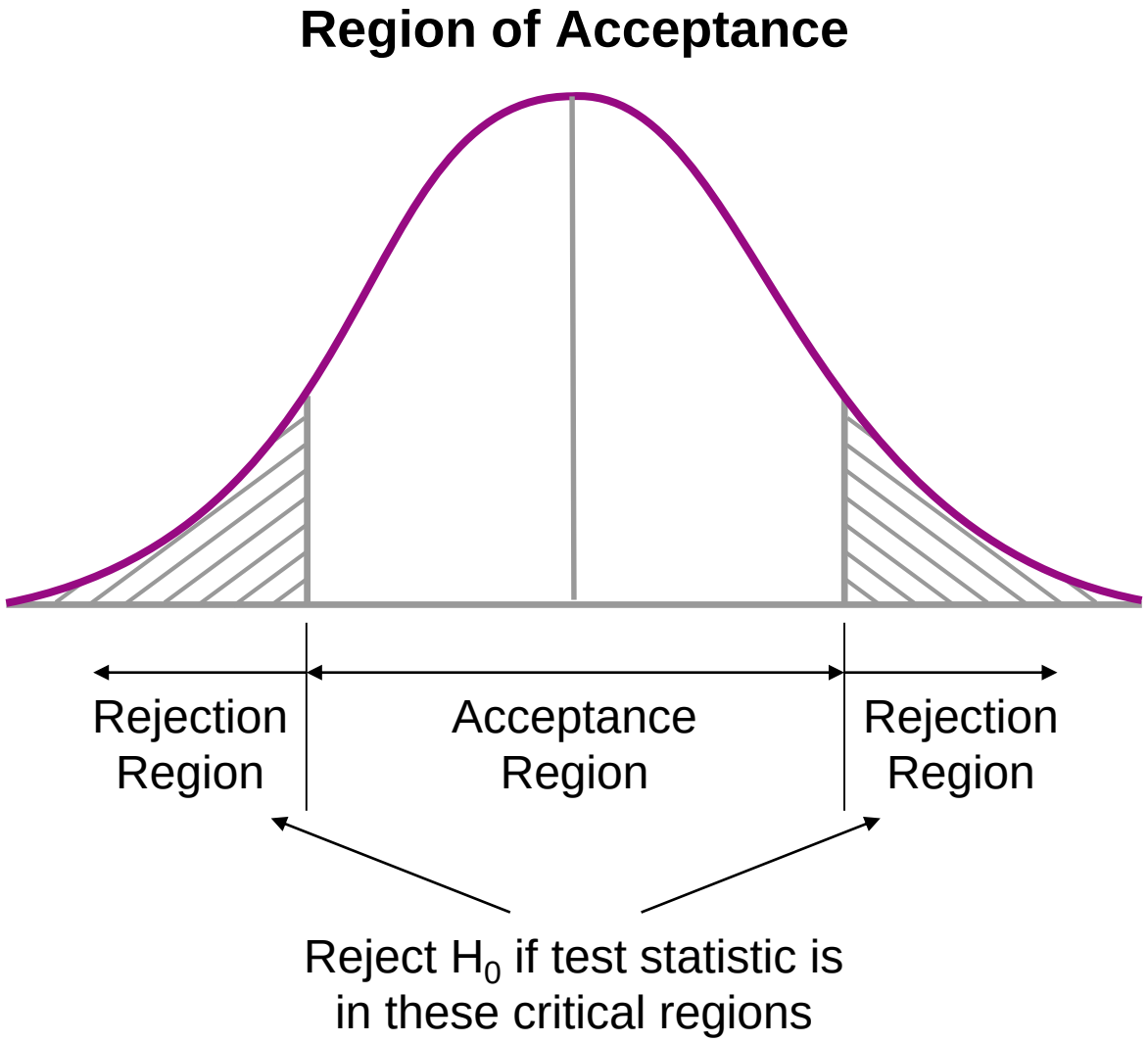
https://stattrek.com/hypothesis-test/hypothesis-testing.aspx

Testing

Hypothesis Testing

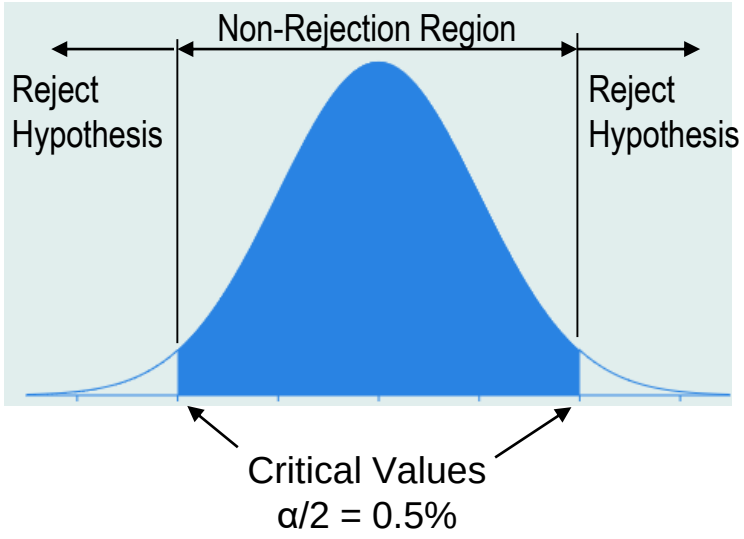
1. State the hypotheses
2. Formulate an analysis plan
3. Analyze sample data
4. Interpret results

Decision rules

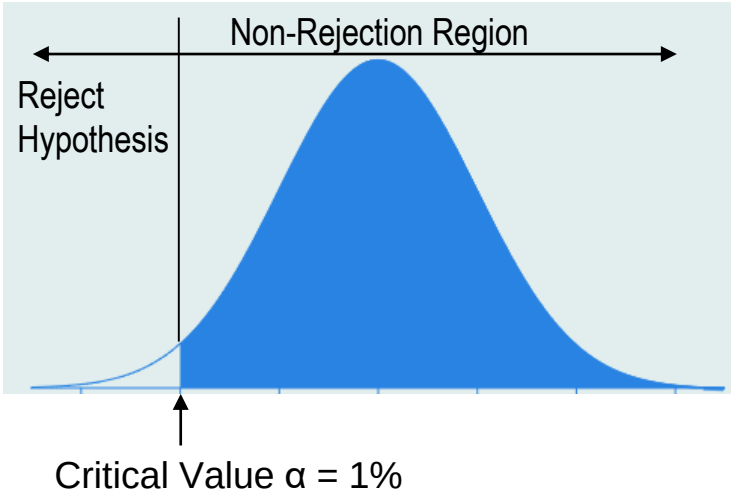


One-tailed and two-tailed tests

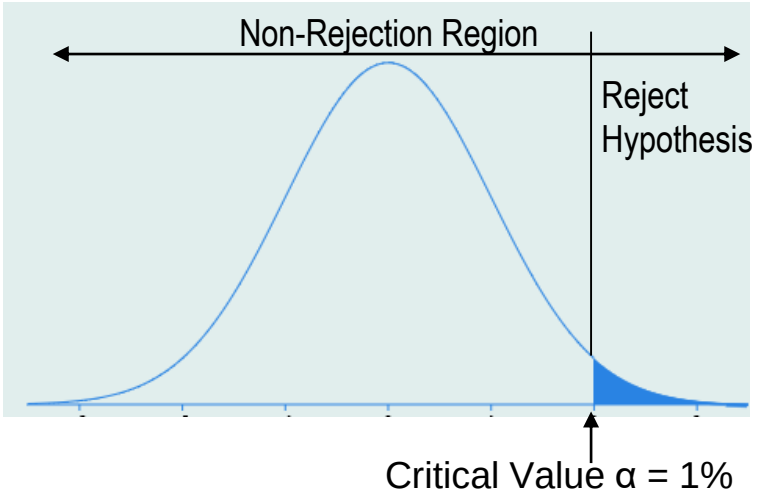
Two-Tailed Test



Left-Sided (One-Tailed) Test



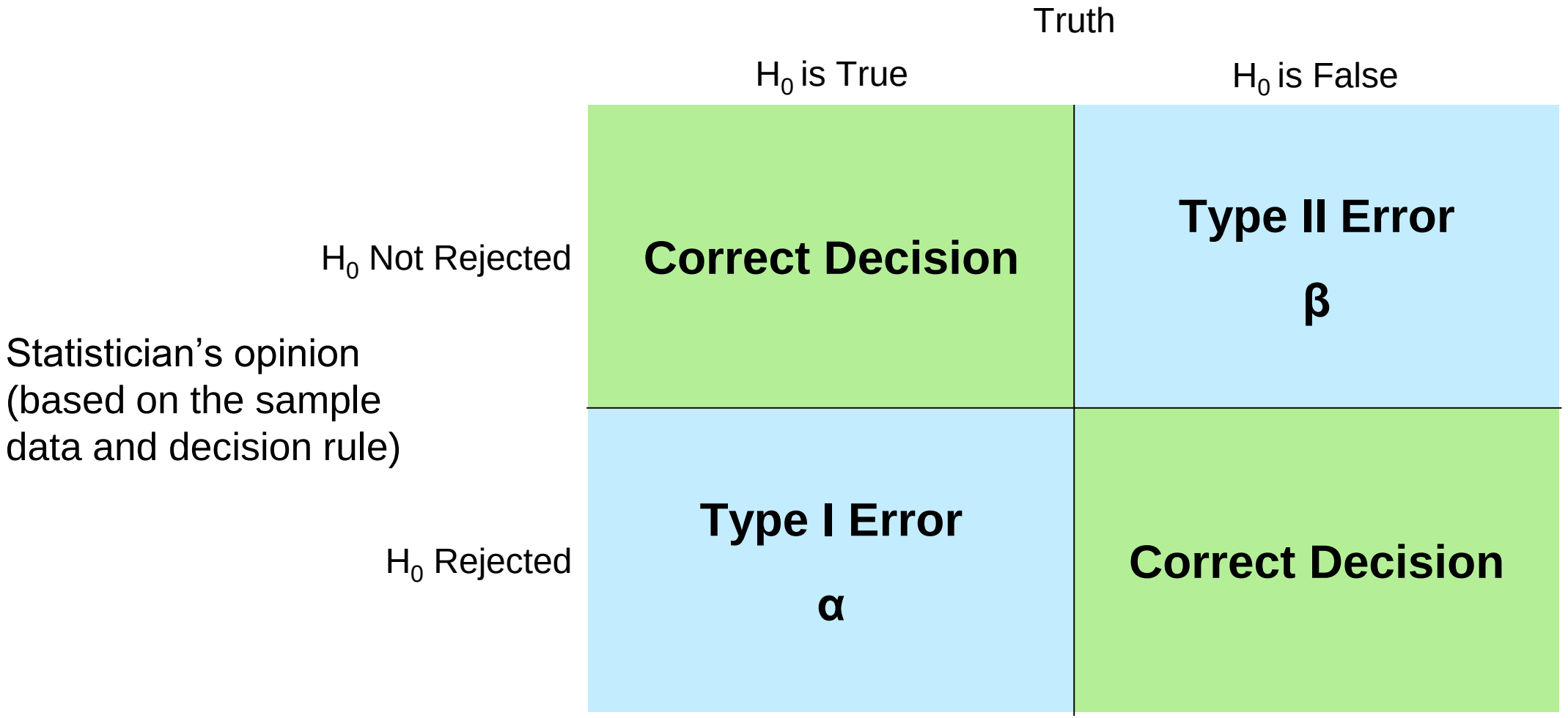
Right-Sided (One-Tailed) Test



http://www.stat.yale.edu/Courses/1997-98/101/sigtest.htm

https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests-significance-levels-alpha-and-p-values-in-statistics

Decision errors



https://en.wikipedia.org/wiki/Power_(statistics)

Summary

- "**Hypothesis testing**" refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.
- There are two types of statistical hypotheses:
  1. **Null hypothesis** (H_0) is usually the hypothesis that the sample observations result purely from chance.
  2. **Alternative hypothesis** (H_1 or H_a) is the hypothesis that the sample observations are influenced by some non-random cause.
- An analysis plan includes decision rules for rejecting the null hypothesis. Statisticians describe these decision rules in two ways – with reference to a P-value or with reference to a region of acceptance.
- Two types of errors can result from a hypothesis test.
  - A Type I error occurs when the researcher rejects a null hypothesis when it is true.
  - A Type II error occurs when the researcher fails to reject a null hypothesis that is false.

Thank you.

Contact information:

open@sap.com

Follow all of SAP

www.sap.com/contactsap

THE BEST RUN SAP®