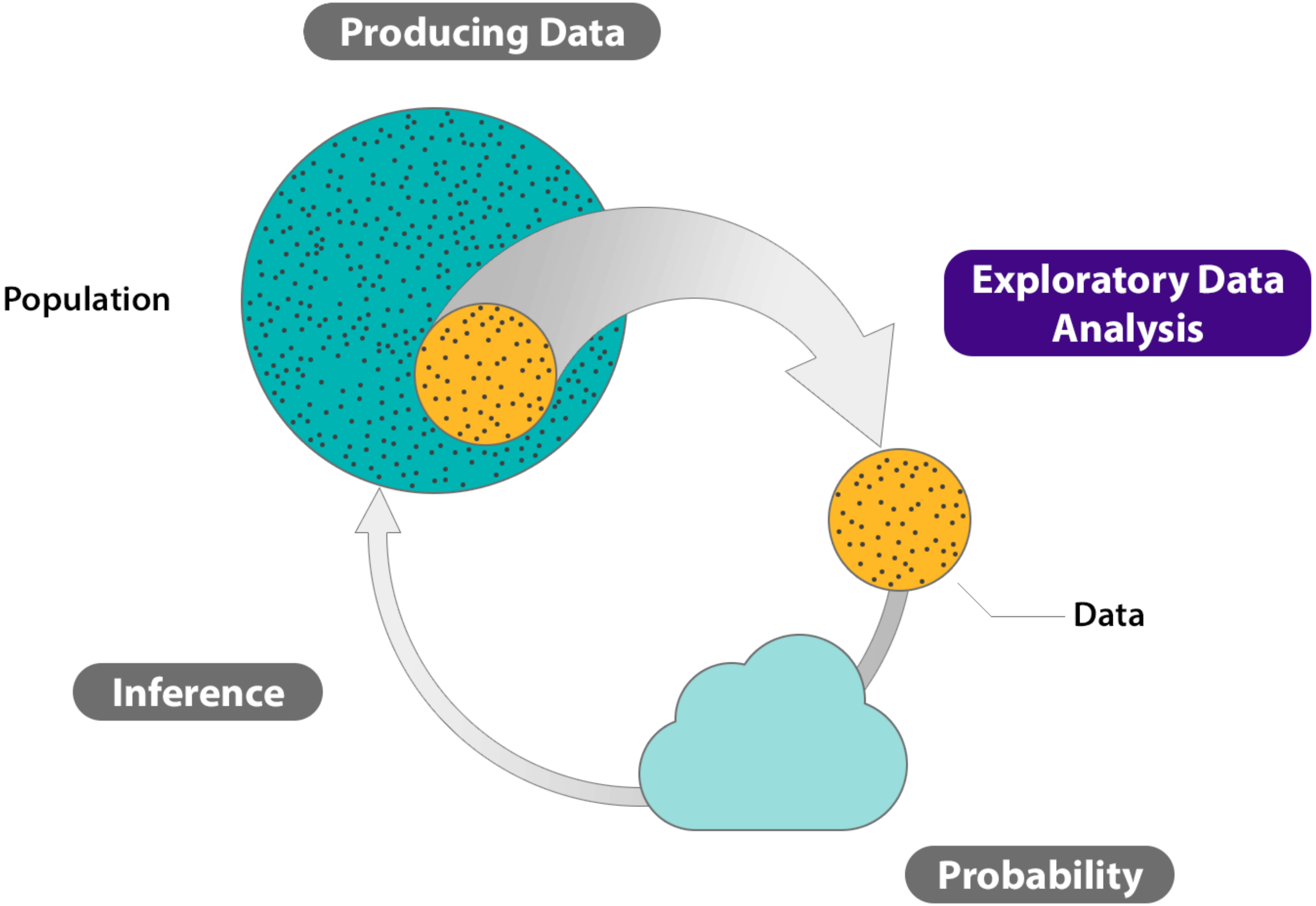
Biostatistics
Week III

Ege Ülgen, MD, PhD

20 October 2022

ACIBADEM MEHMET ALİ AYDINLAR ÜNİVERSİTESİ

The Big Picture



Unit 1: exploratory data analysis [Internet]. [cited 2021 Sep 27]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/

Exploratory Data Analysis (EDA)

- Examining Distributions — exploring data one variable at a time
- **Examining Relationships — exploring data two variables at a time**

Contingency table/Cross tabulation/Crosstab

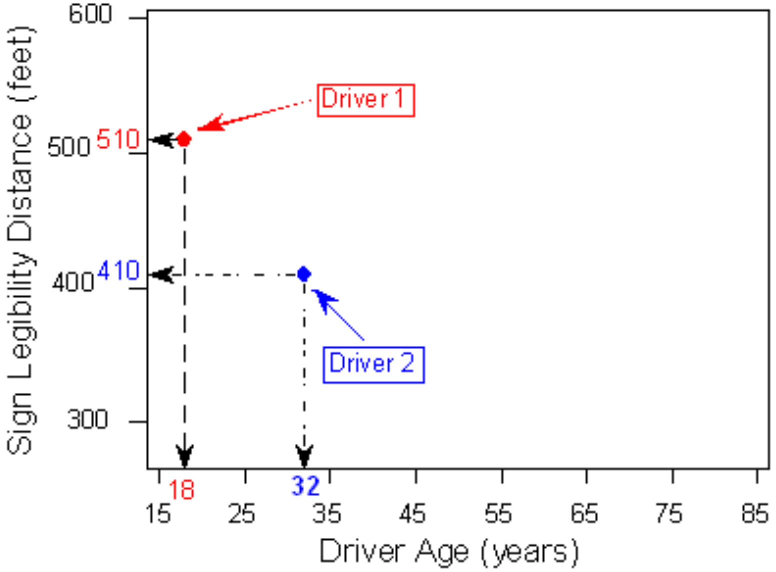
- Tables in which two categorical variables are investigated together

	Male	Female
No education	4	10
Primary school	3	5
High school	2	8
Bachelor’s degree	7	9

https://bolt.mph.ufl.edu/6050-6052/unit-1/case-q-q/scatterplots/

Scatter Plots

	Age (X)	Distance (Y)
Driver 1	18	510
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

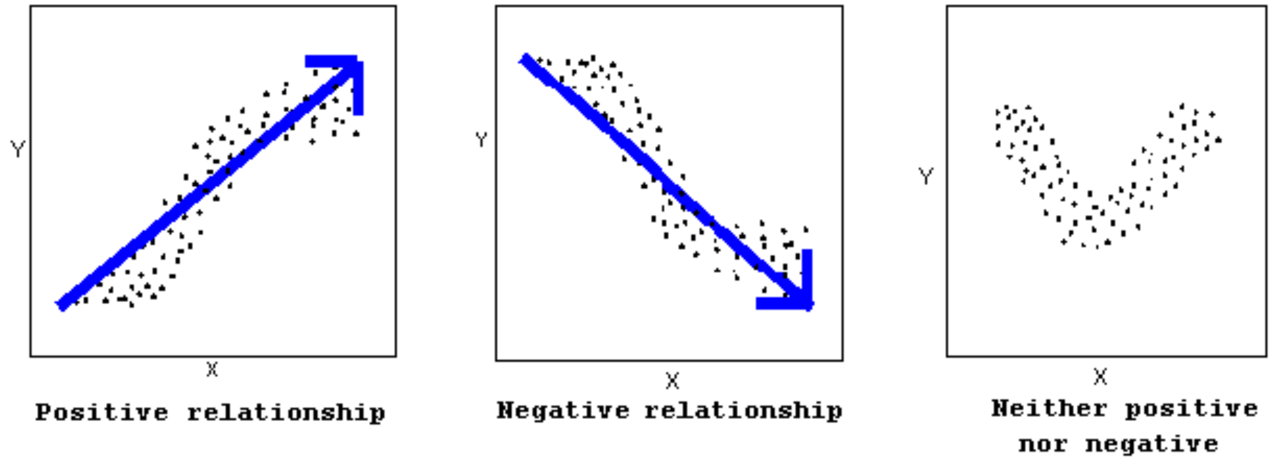


X – Explanatory
Y – Response

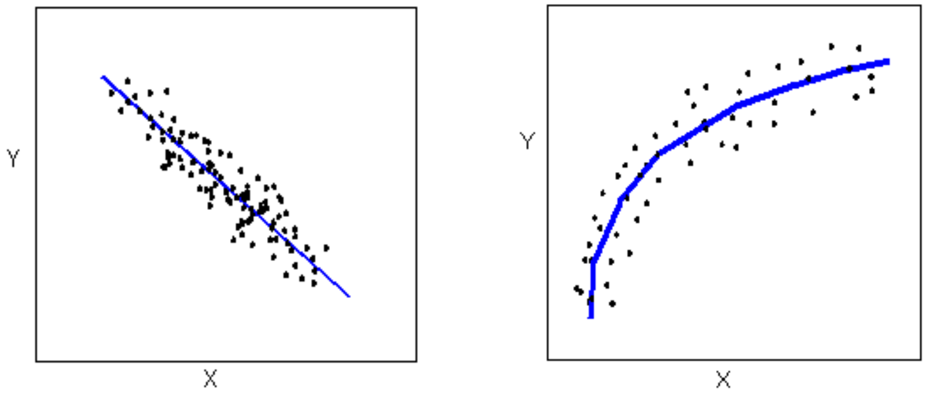
https://bolt.mph.ufl.edu/6050-6052/unit-1/case-q-q/scatterplots/

Interpreting Scatter Plots

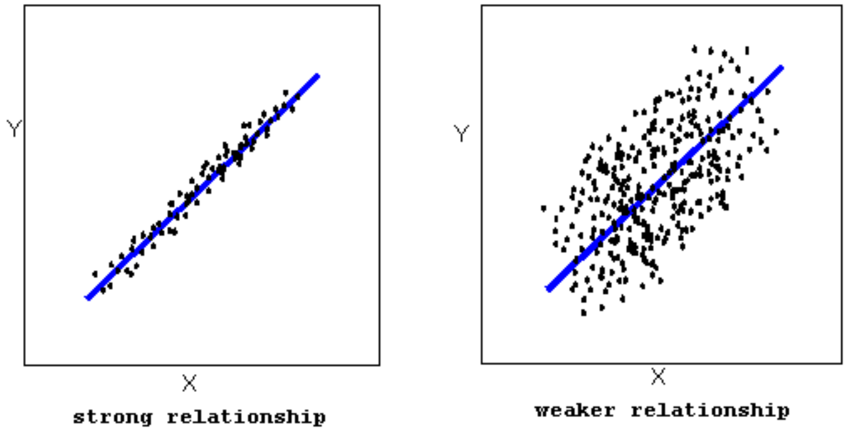
Direction



Form



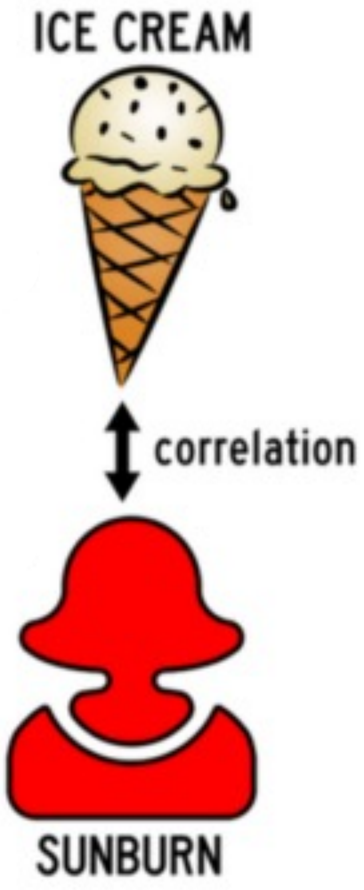
Strength



Correlation

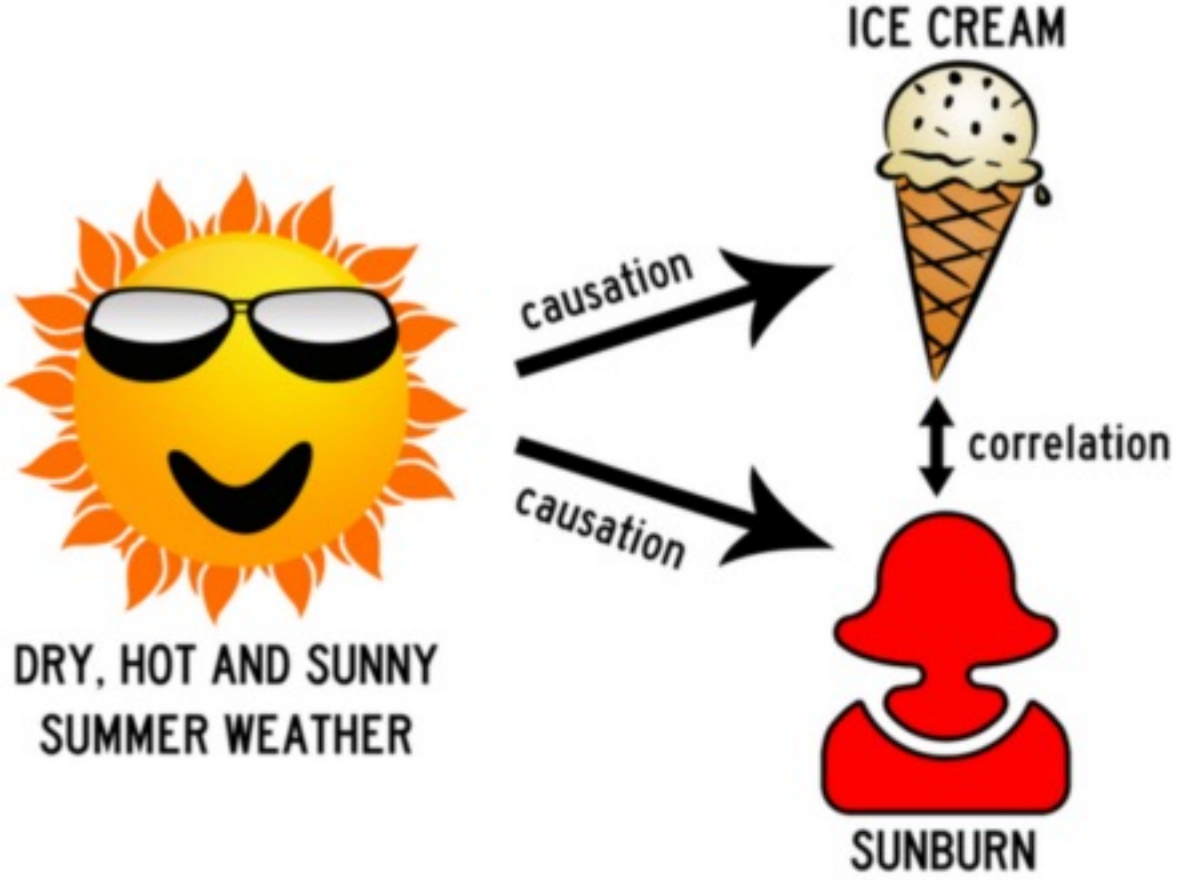
- Correlation is a bivariate analysis that measures **the strength of association** between two variables and **the direction** of the relationships
- In terms of the strength of relationship, the value of the correlation coefficient varies **between +1 and -1**
- **Correlation does not mean causation**

Correlation does not mean causation



Figueroa A. Correlation is not causation [Internet]. Medium. 2019 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

Correlation does not mean causation



Figueroa A. Correlation is not causation [Internet]. Medium. 2019 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

Correlation Coefficient

- A statistic that measures the relationship between two variables

- Pearson’s r
  - Measures **linear** relationship
  - Both variables have to be normally distributed
- Spearman’s ρ
  - Measures **monotonic** relationship
  - Based on rank – non-parametric

Pearson Correlation Coefficient

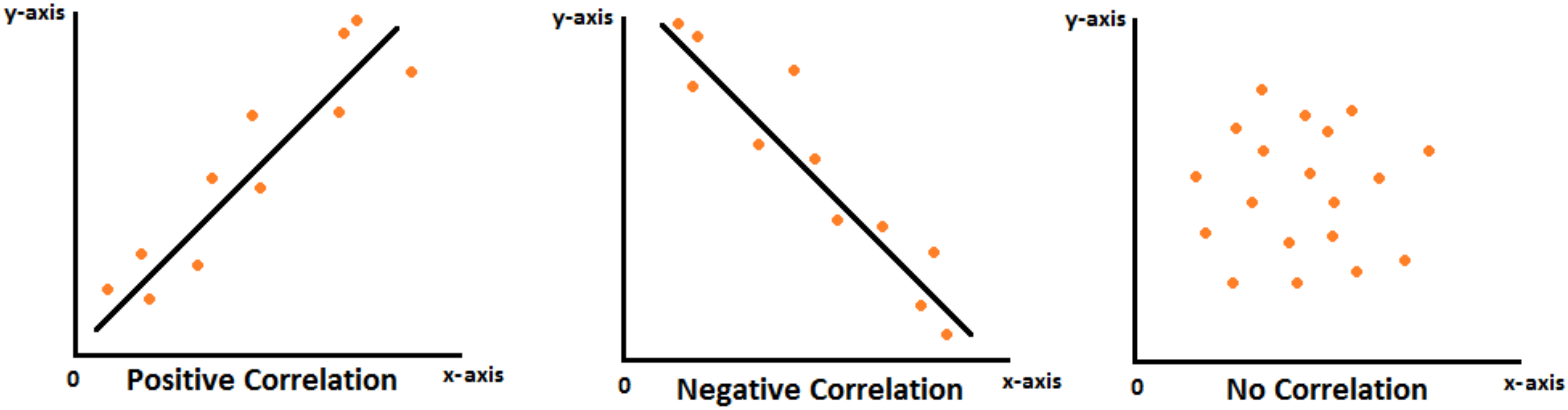
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- A measure of the **linear** correlation between two variables X and Y
- takes values between -1 and 1
- unitless
- $r_{X,Y} = r_{Y,X}$
- $r_{X,Y} = 0$ means **no linear relationship**

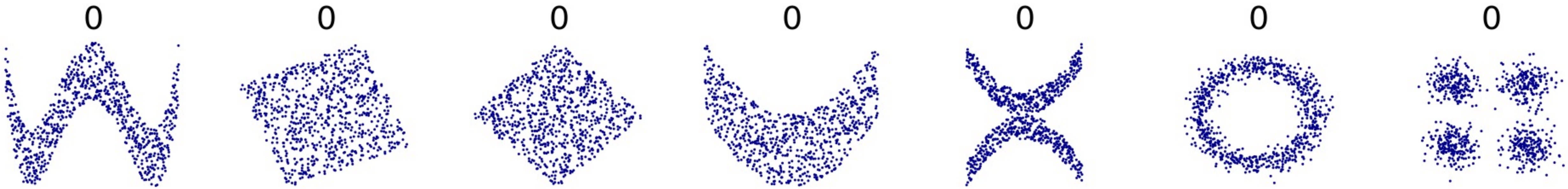
Pearson Correlation Coefficient

Cohen's (1988) conventions to interpret effect size:

- $|r| = 0.10 - 0.29$: Weak
- $|r| = 0.30 - 0.49$: Moderate
- $|r| \geq 0.50$: Strong

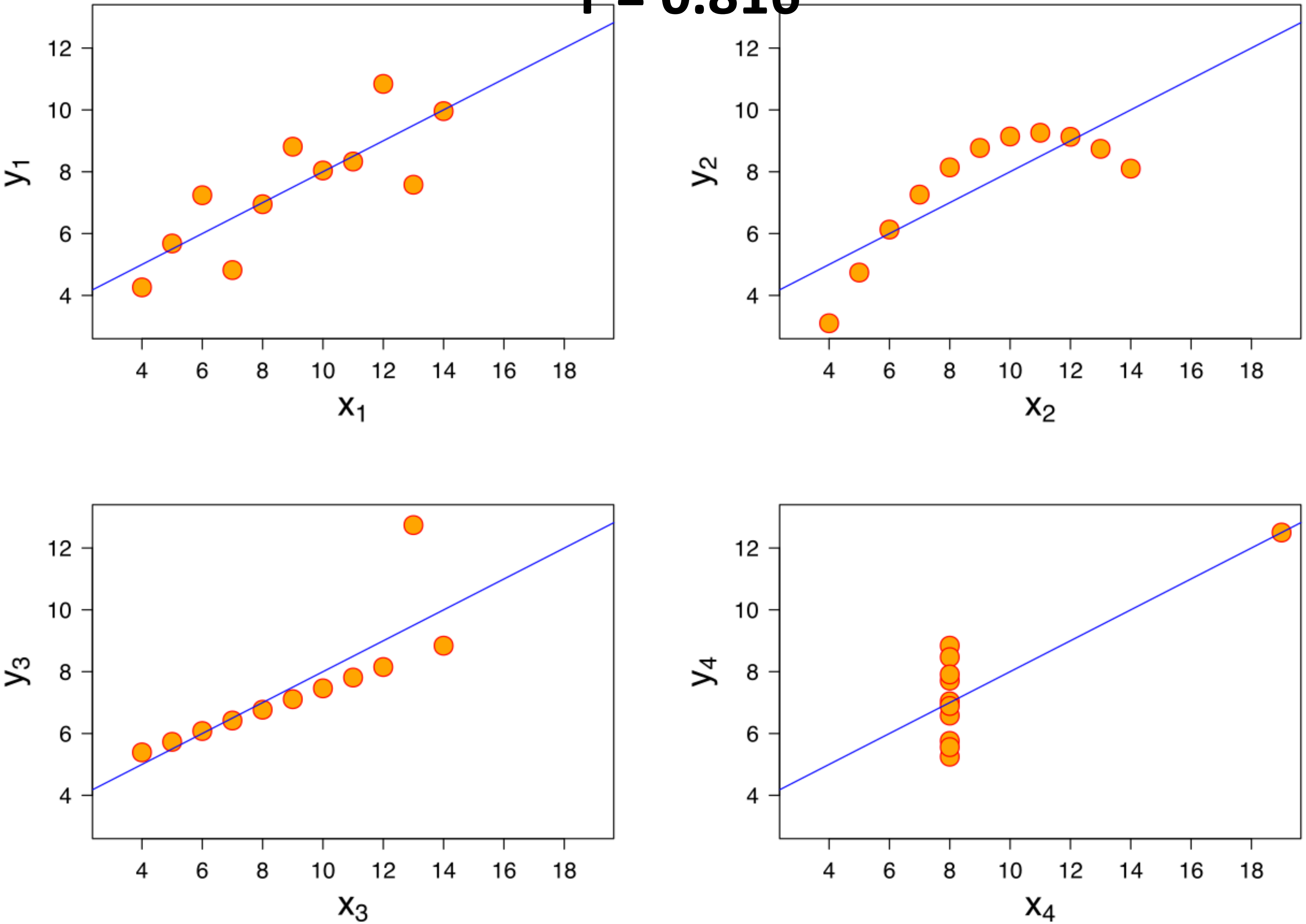


Suresh S. From correlation to causation [Internet]. Medium. 2020 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/from-correlation-to-causation-49f566eea954



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

r = 0.816

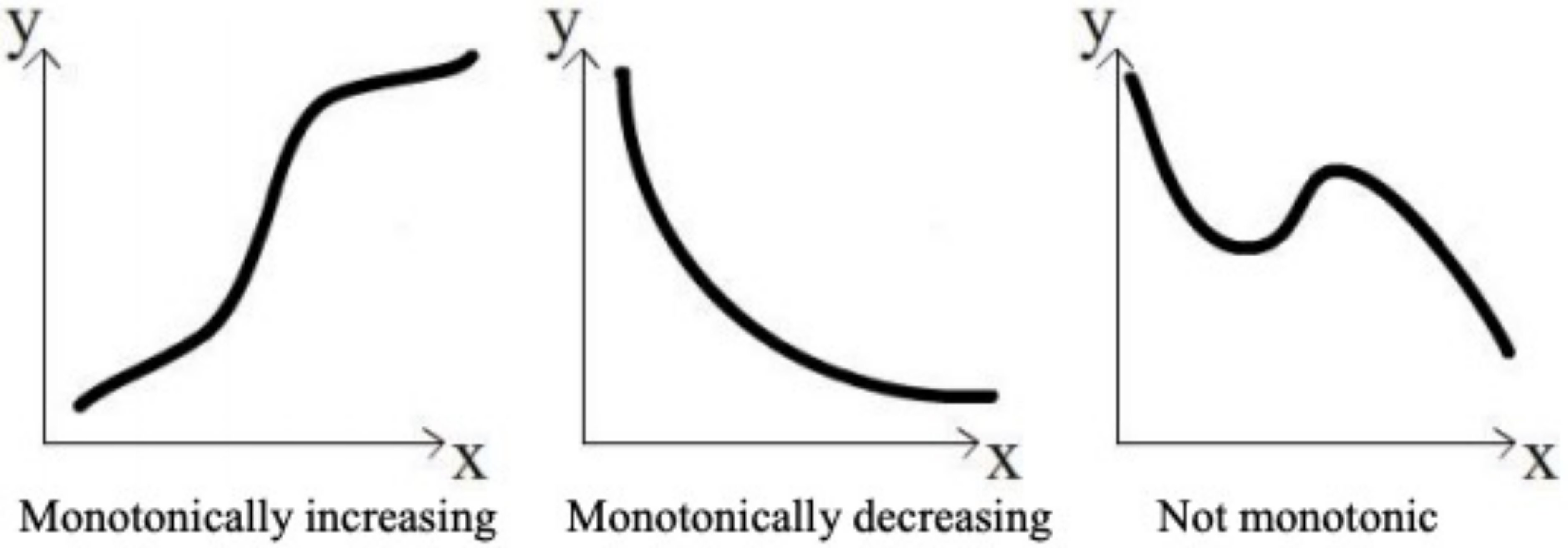


https://en.wikipedia.org/wiki/Correlation_and_dependence

Spearman Rank Correlation

- It assesses how well the relationship between two variables can be described **using a monotonic function**
- It **does not carry any assumptions about the distribution** of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal

Spearman Rank Correlation



Spearman Rank Correlation

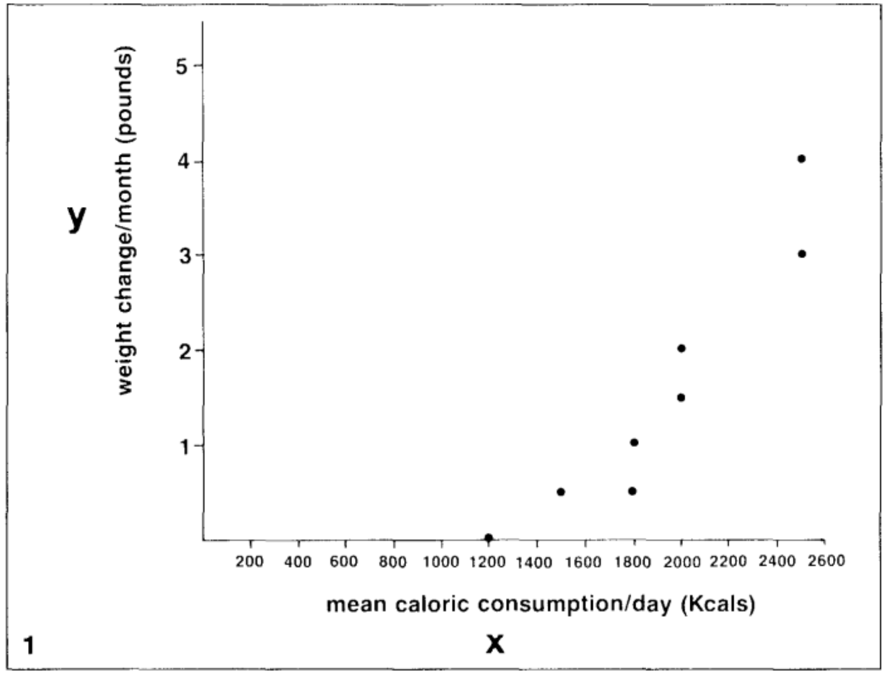
$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

- d_i := the difference between the ranks of corresponding variables (i.e., $d = X_i - Y_i$)
- n := number of observations

TABLE 1. *Sample data: Caloric consumption versus weight change*

Patient	(X) Mean Caloric Consumption/Day	(Y) Weight Change/ Month
1	1,200	0.0
2	1,500	0.5
3	1,800	0.5
4	2,000	1.5
5	2,500	4.0
6	1,800	1.0
7	2,500	3.0
8	2,000	2.0

FIGURE 1. *Scatter diagram for sample data given in Table 1 (caloric consumption vs weight change).*



There is a strong positive relationship between mean caloric consumption/day and weight change/month

r = 0.94 or
ρ = 0.97

Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 6, Correlation and regression. Ann Emerg Med. 1990 Dec;19(12):1462–8.

Brief Summary

- The relationship between two continuous variables can be visualized using scatter plots
- The relationship between two variables can be assessed using correlation
  - Pearson
  - Spearman

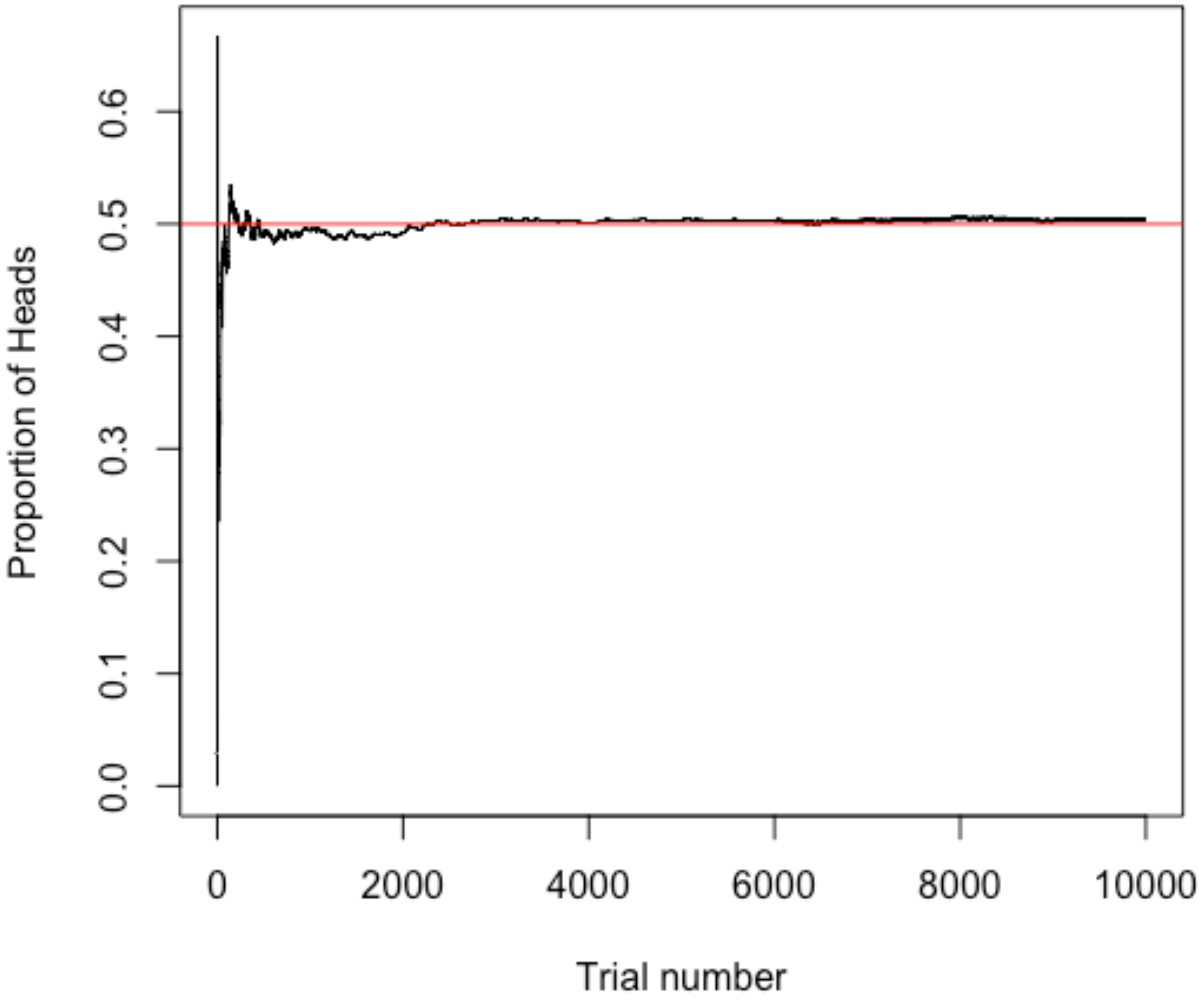
Probability

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n}$$

- *P(A)*: probability of event A
- *n(A)*: frequency of event A out of n trials
- *n*: number of trials

Olasılık

"Heads, you get a quadruple bypass.
Tails, you take a baby aspirin."



Probability - Definitions

- **Experiment:** a process that produces an outcome/outcomes
  Sample Space (Ω): the set of all possible outcomes from an experiment
- **Event:** any set of outcomes of an experiment

Probability - Definitions

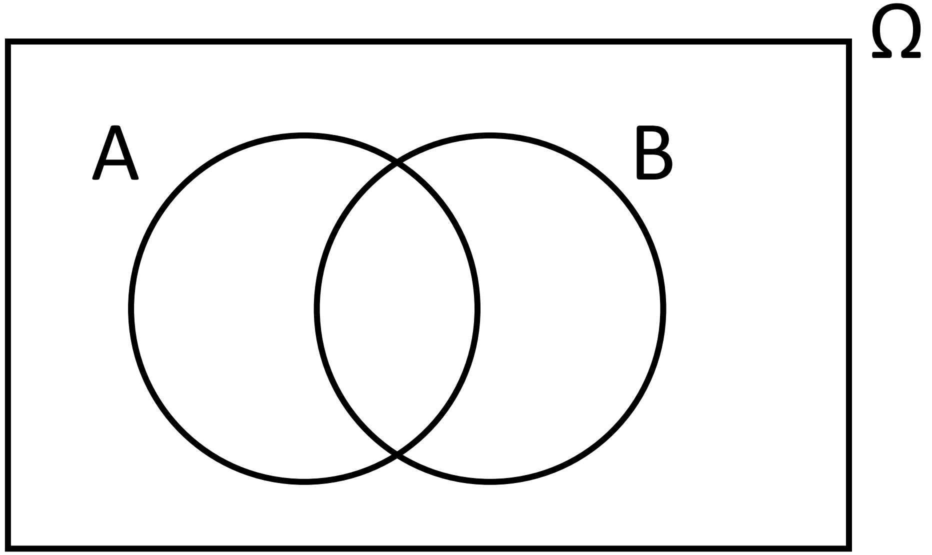
- **Experiment:** flipping a coin and rolling a die at the same time
- **Sample Space:**

  Ω = {(H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6),
  (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6),}
- **Event:**

  A: {rolling an even number} P(A) = 6 / 12

  B: {getting heads and an odd number} P(B) = 3 / 12

Probability - Properties



- $P(\Omega) = 1$

- $0 \leq P(A) \leq 1$
- $P(A^c) = 1 - P(A)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

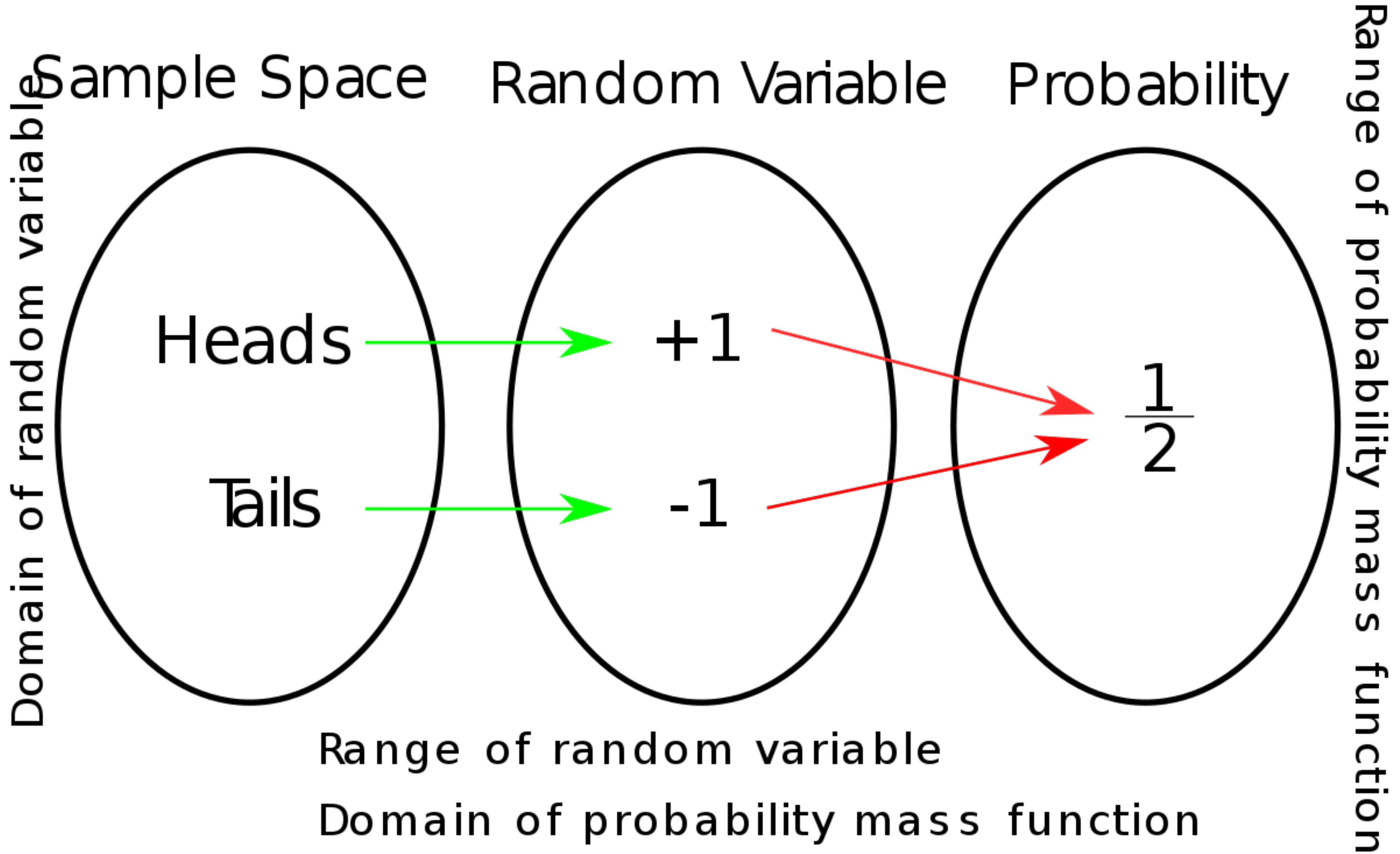
- If $A \cap B$ is an empty set (i.e., if A and B do not occur at the same time), A and B are called disjoint (mutually-exclusive)

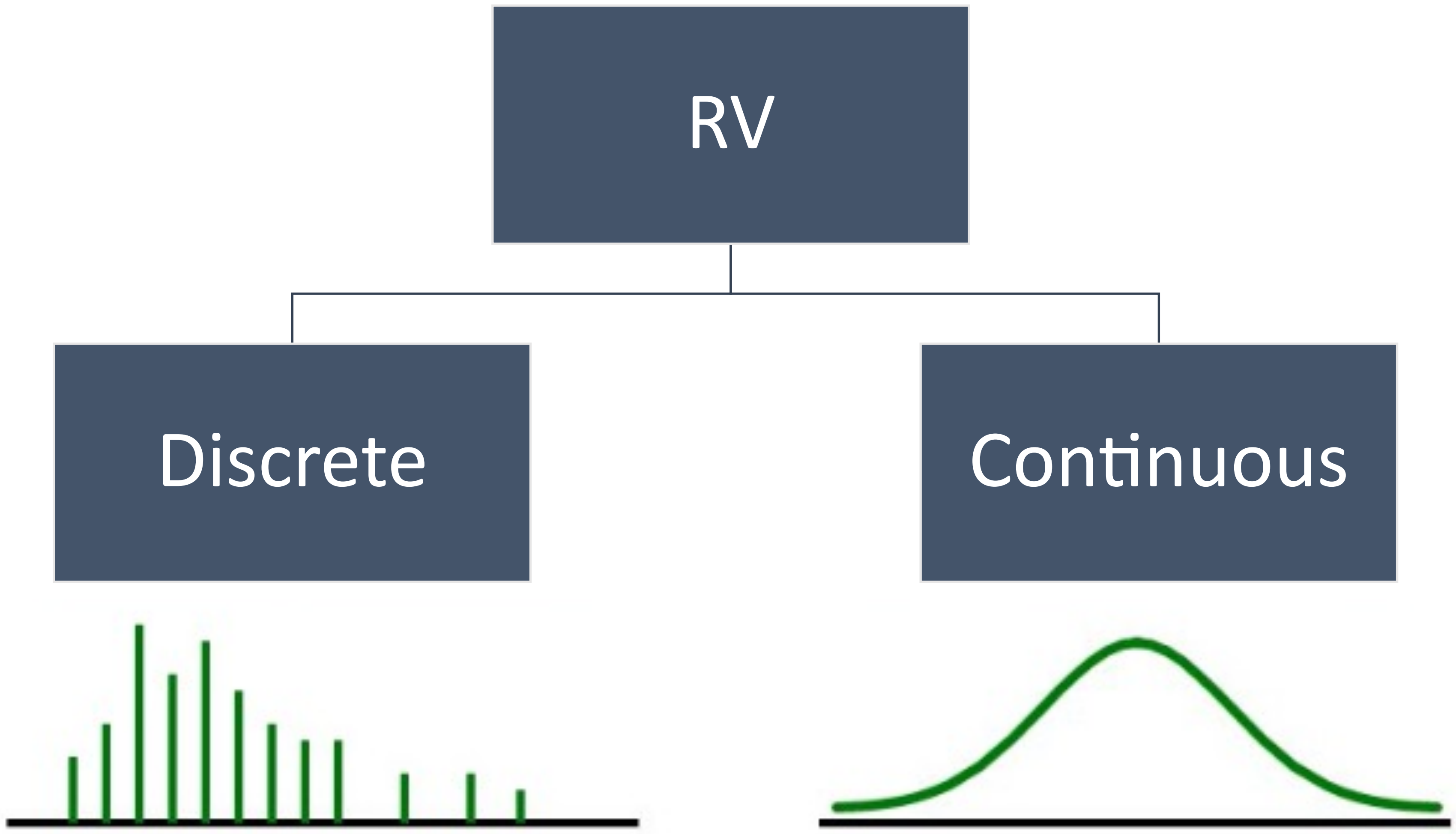
Random Variable

- A random variable (RV) is a variable whose possible values are **numerical outcomes of a random phenomenon**

- There are two types of random variables:
  - ***Discrete*** – flipping a coin, rolling a die, number of pancreatic cancer cases in a year ...
  - **Continuous** – systolic blood pressures of hypertensive patients, progression-free survival time of glioblastoma patients, expression level of a certain gene …

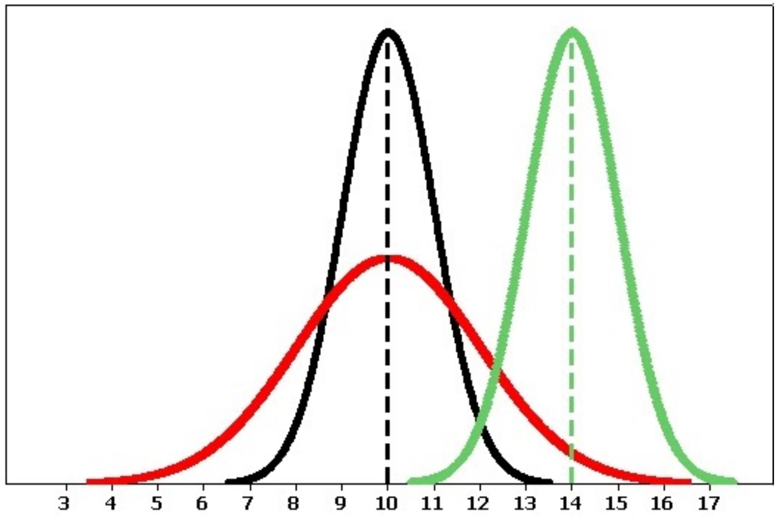
https://en.wikipedia.org/wiki/Random_variable




RV
Discrete
Continuous

Normal Distribution

- The distributions of many variables follow a “normal distribution”
- The **bell-shape** indicates that values closer to the mean are more likely, and it becomes increasingly unlikely to take values far from the mean in either direction

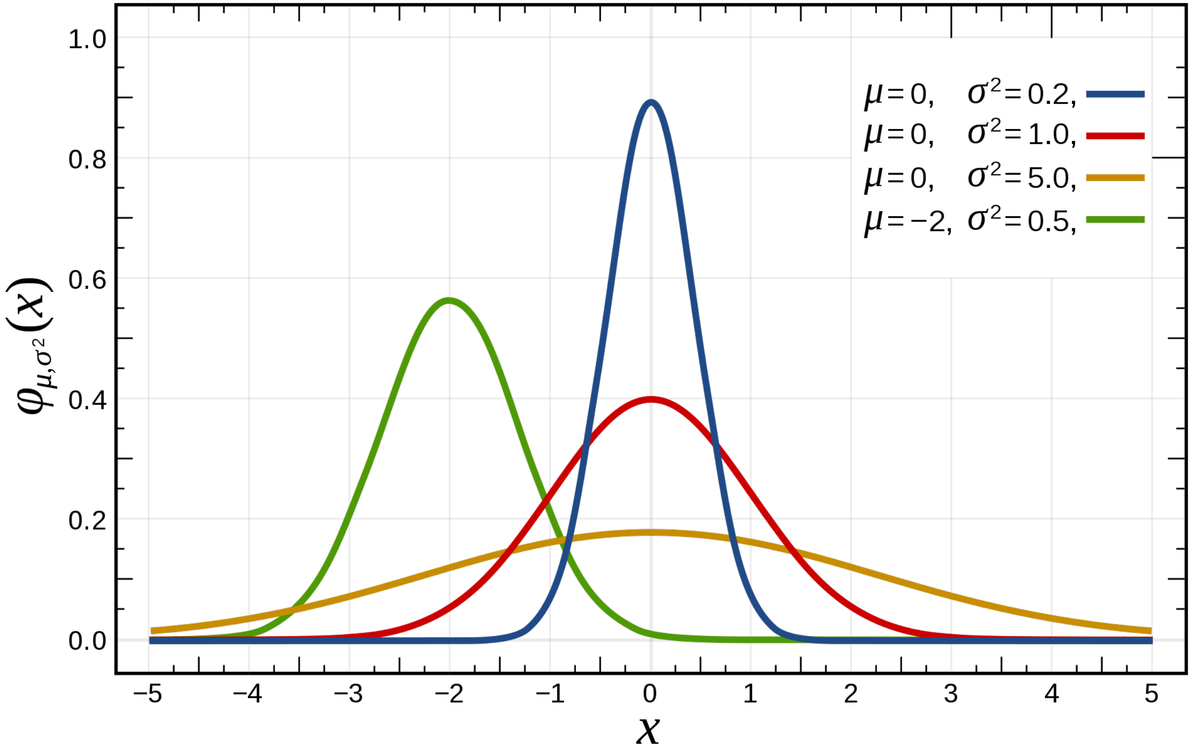


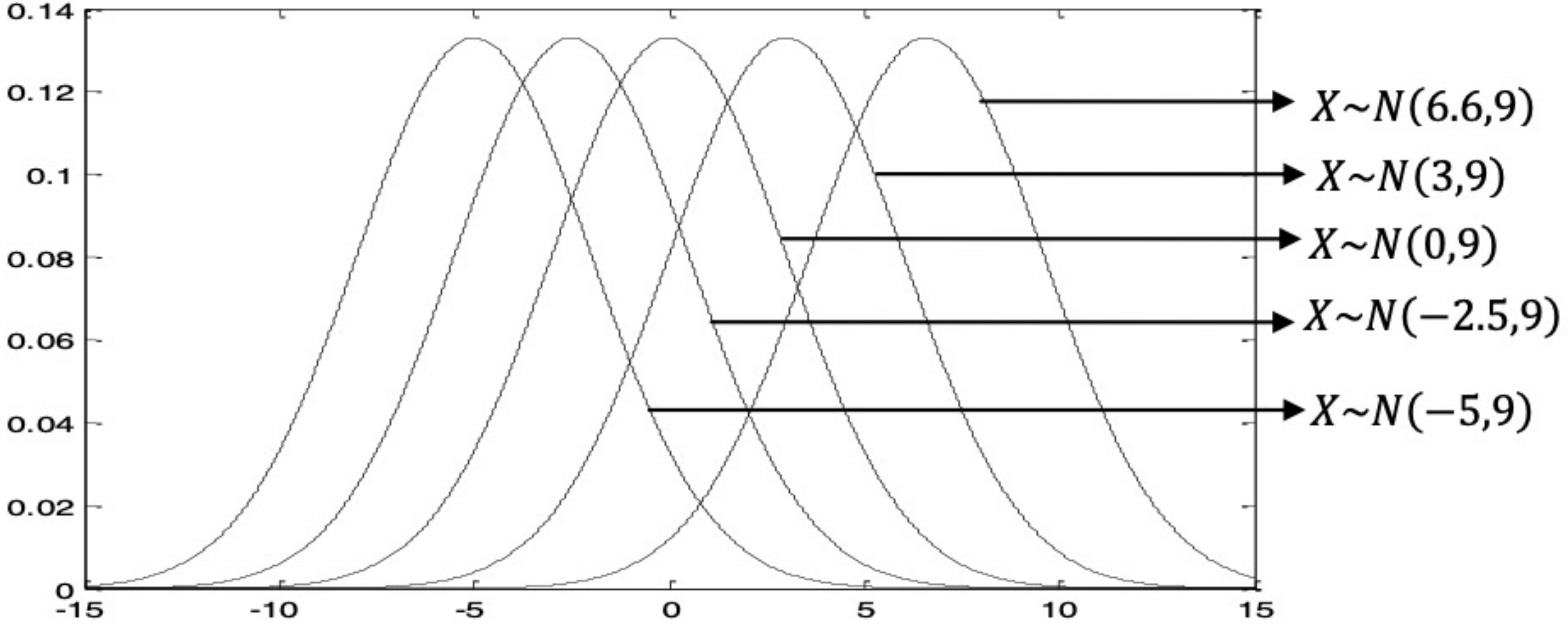
Normal Distribution

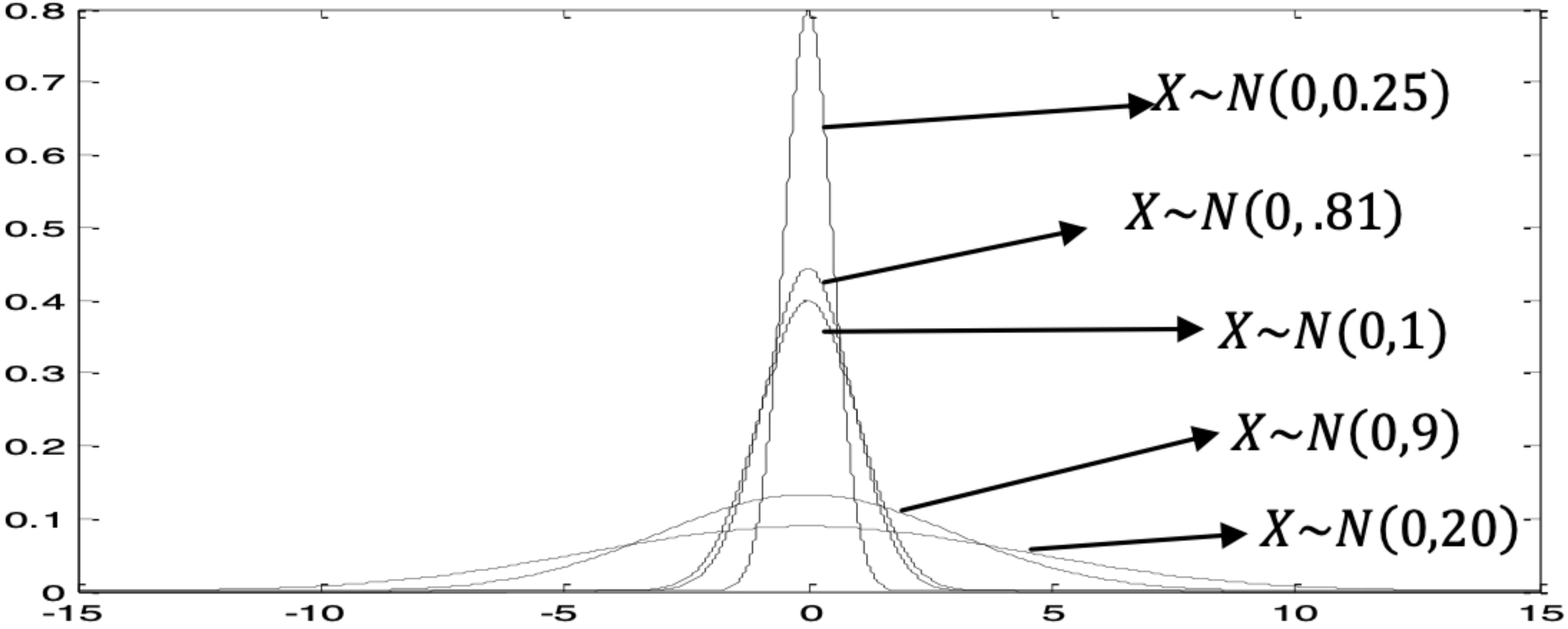
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma^2 > 0$$

- Mean = Median = Mode= μ
- Variance = σ^2

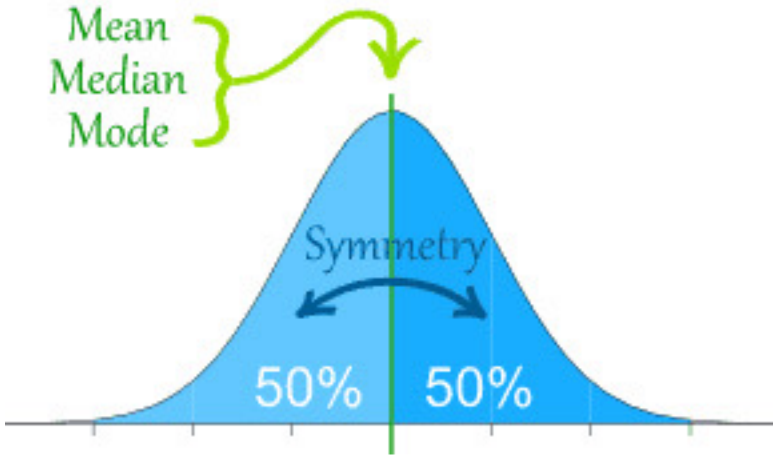
$X \sim N(\mu, \sigma^2)$

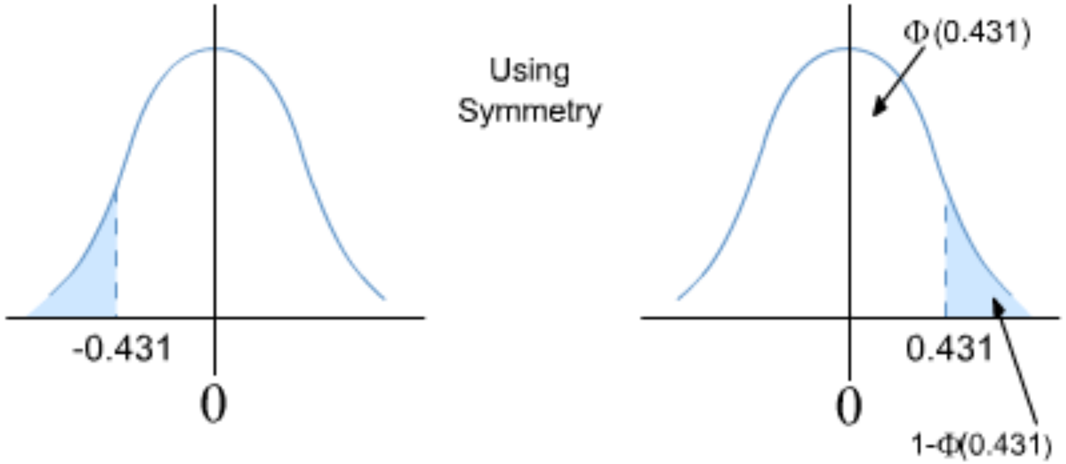



0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-15
-10
-5
0
5
10
15
$X \sim N(6.6,9)$
$X \sim N(3,9)$
$X \sim N(0,9)$
$X \sim N(-2.5,9)$
$X \sim N(-5,9)$



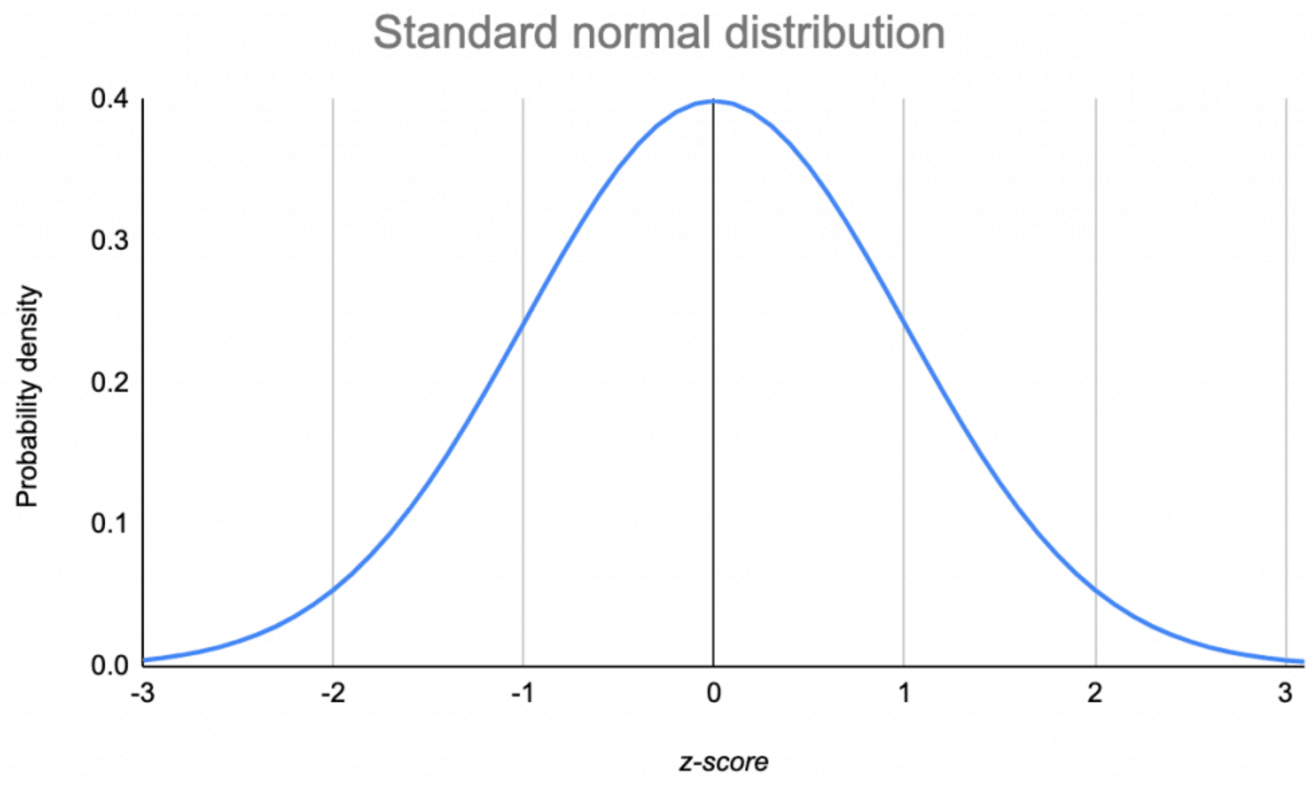
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-15
-10
-5
0
5
10
15
X~N(0,0.25)
X~N(0,.81)
X~N(0,1)
X~N(0,9)
X~N(0,20)


Mean
Median
Mode
Symmetry
50%
50%


-0.431
0
Using
Symmetry
Φ(0.431)
0.431
0
1-Φ(0.431)

Standard Normal Distribution

- Normal distribution for which $\mu = 0$ and $\sigma^2 = 1$
- Usually denoted with Z



https://byjus.com/maths/z-score-table/

STANDARD NORMAL PROBABILITIES

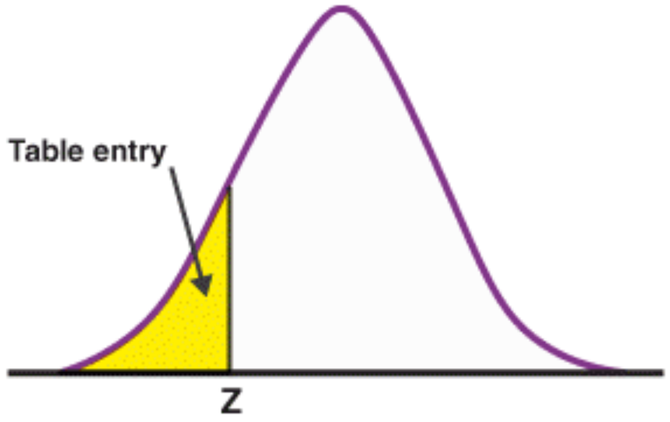


Table entry for z is the area under the standard normal curve to theleft of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143

Standard Normal Probabilities

https://byjus.com/maths/z-score-table/

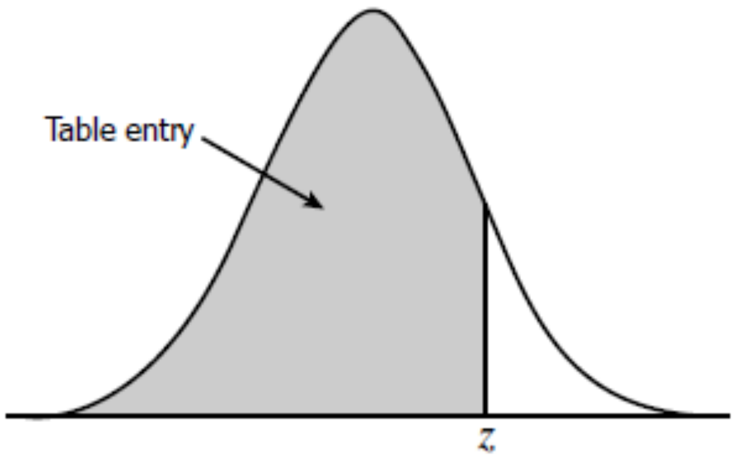


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Standardization

$$X \sim N(\mu, \sigma^2) \implies \boldsymbol{Z} = \frac{\boldsymbol{X} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \sim N(0,1)$$

Normal distribution [Internet]. [cited 2021 Oct 4]. Available from: https://www.mathsisfun.com/data/standard-normal-distribution.html

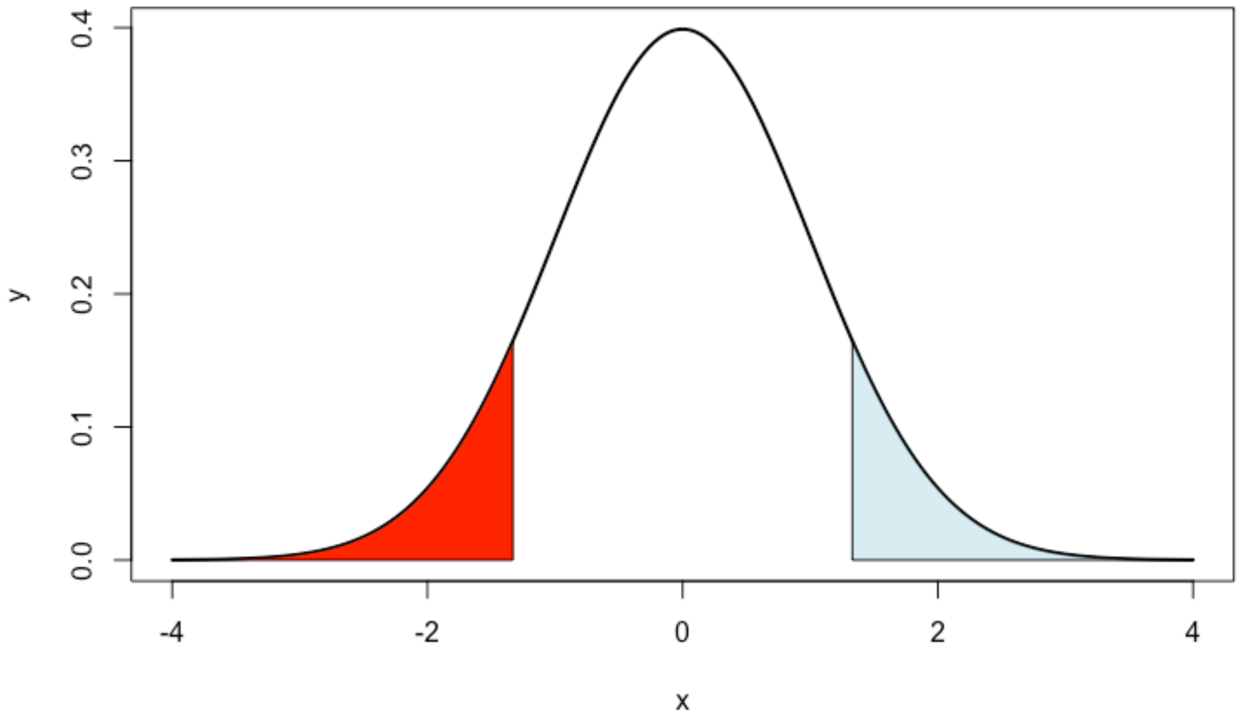
Normal Distribution - Example

- In a hospital, the systolic blood pressures of patients follow a normal distribution with mean = 15, variance = 9 $X \sim N(15,9)$
- For a randomly selected patient, what is the probability that their SBP is:
  a) Smaller than 11?
  b) Larger than 12?
  c) Between 9 and 16?

$X \sim N(15,9)$

a) < 11

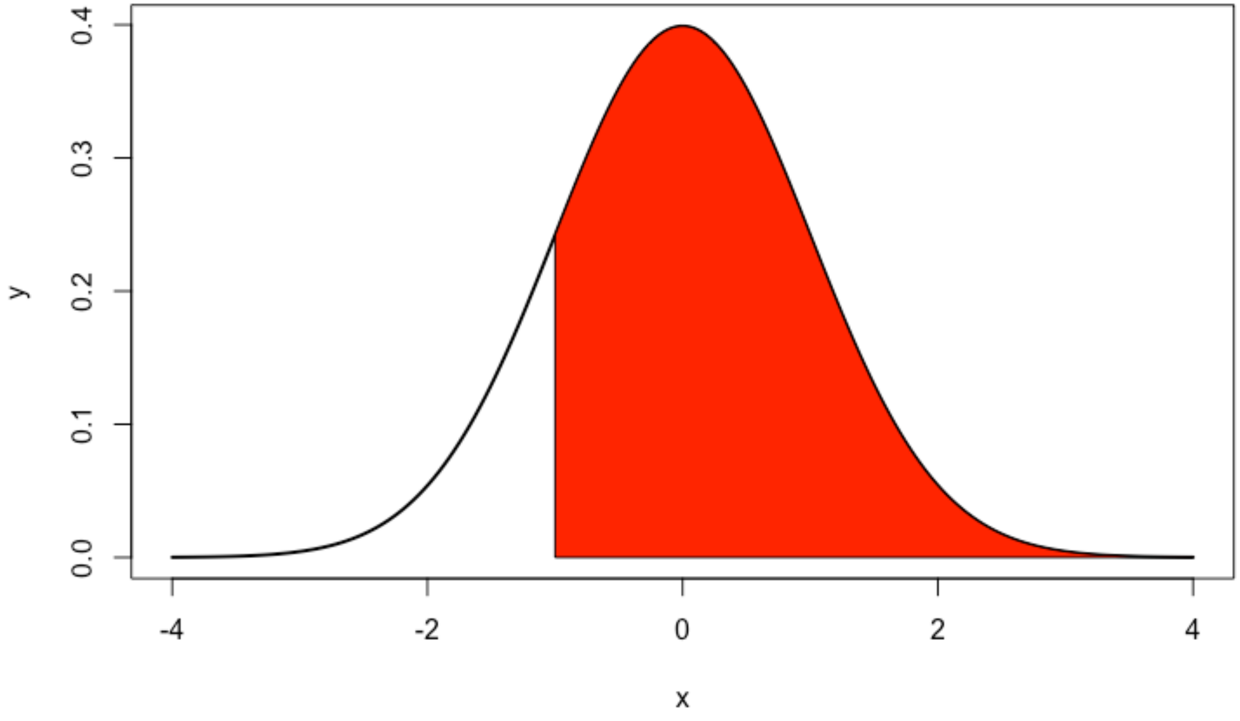
$$P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{11-15}{3}\right) = P(Z \leq -1.33) = P(Z \geq 1.33) = 0.0918$$



$X \sim N(15,9)$

b) > 12

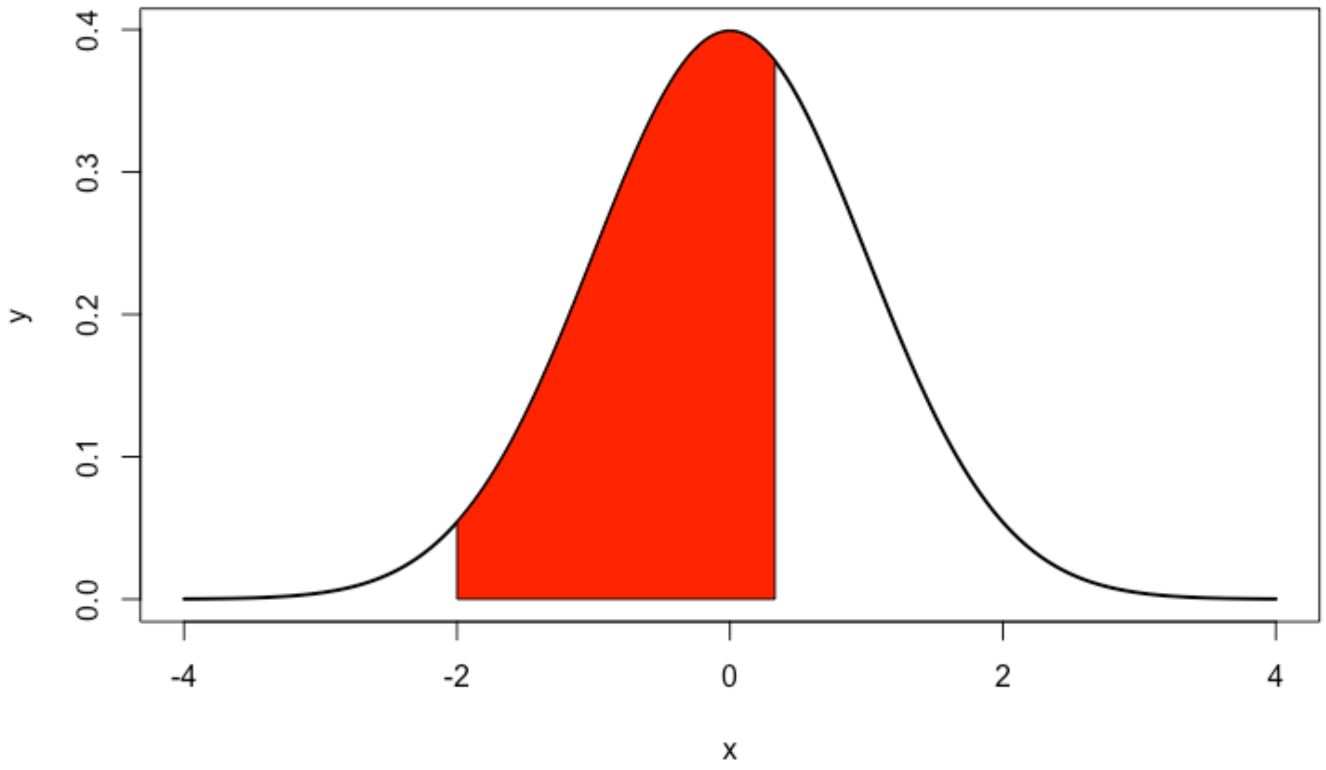
$$P(X > x) = P\left(Z > \frac{x-\mu}{\sigma}\right) = P\left(Z > \frac{12-15}{3}\right) = P(Z > -1) = 0.8413$$



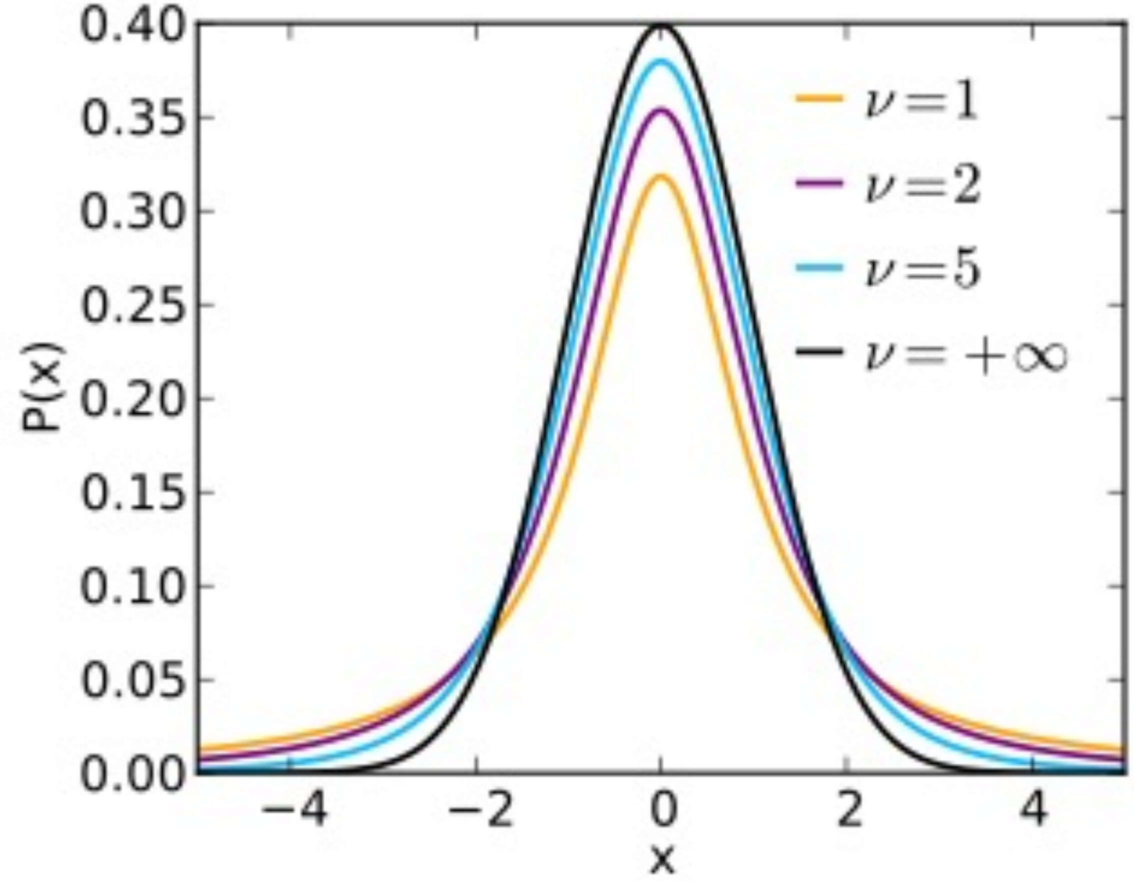
$X \sim N(15{,}9)$

c) Between 9 and 16

$$P(9 < X < 16) = P\left(\frac{9-15}{3} < Z < \frac{16-15}{3}\right) = P(-2 < Z < 0.33) = P(Z < 0.33) - P(Z \leq -2) = 0.6065$$



(Student's) t Distribution



https://en.wikipedia.org/wiki/Student%27s_t-distribution