Probability Distributions

Data Science Campus

Sample Space and Events

- The list of all possible outcomes of a random experiment is called **sample space** and it's usually denoted with letter S
- A subset E of the sample space is called **event**

For Example

Toss a coin

Sample space $S = \{H, T\}$

Possible events $\emptyset \quad \{H\} \quad \{T\} \quad \{H, T\}$

For Example

Roll a die

Sample space $S = \{1, 2, 3, 4, 5, 6\}$

Possible events $\{1\}$ $\{1, 2\}$ $\{1, 3, 5\}$ $\{1, 2, 3, 4, 5, 6\}$

Probability

Given a sample space S of a random experiment, a probability over S is a function p that takes an event as input and provides a number as output such that

1. $0 \leq p(E) \leq 1$ for any possible event E
2. $p(S) = 1$
3. $p(E \cup F) = p(E) + p(F)$ for any pair of disjoint events

For Example

Toss a *fair* coin

Sample space $S = \{H, T\}$

Probability $p(H) = p(T) = 0.5$

For Example

Roll a fair die

Sample space $S = \{1, 2, 3, 4, 5, 6\}$

Probability $p(i) = 1/6$ for $i = 1, \ldots, 6$

For Example

Toss a *biased* coin

Sample space $S = \{H, T\}$

Probability $p(H) = 0.3 \quad p(T) = 0.7$

For Example

Roll a biased die

Sample space $S = \{1, 2, 3, 4, 5, 6\}$

Probability $p(1) = 0.3$ $\quad p(i) = 0.14$ for $i = 2, \ldots, 6$

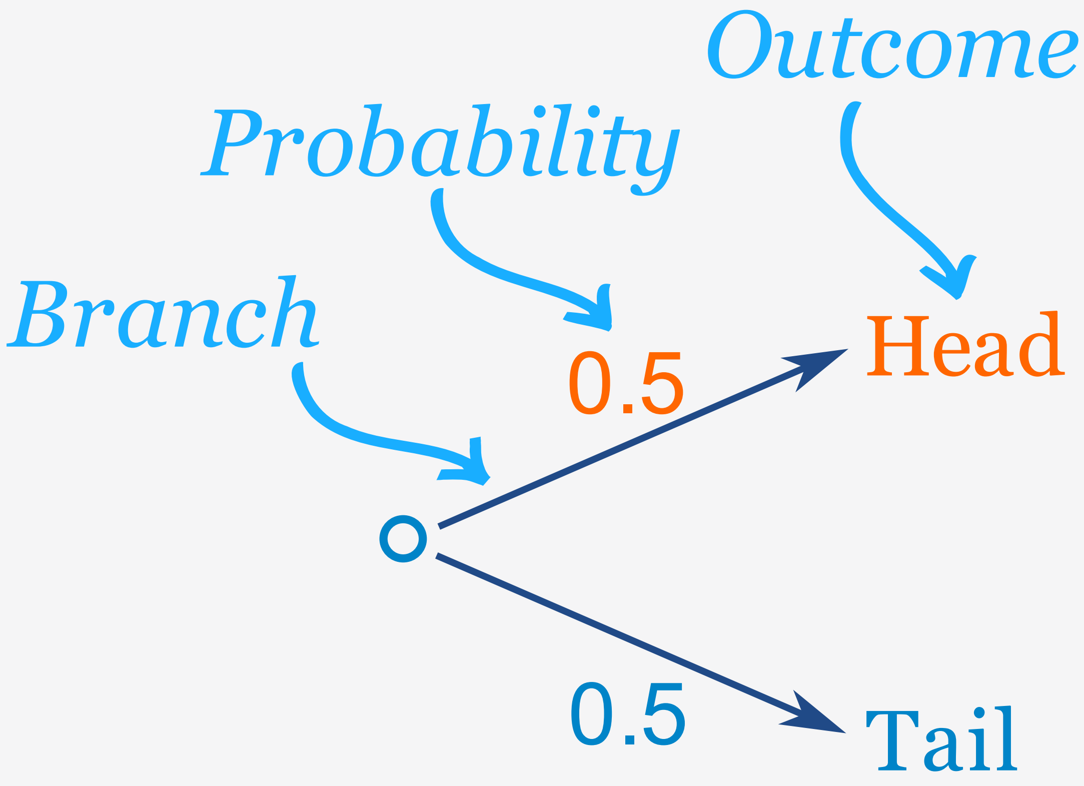
Random Variable

- A **random variable**, denoted by X, is a measurable quantity which can take any value on range of values. Its value is the result of a random experiment (or observation). Actual measured values are represented by x.

For Example

- Number of heads in n coin tosses
- Sum of results after rolling two dices
- Maximum result of rolling two dices

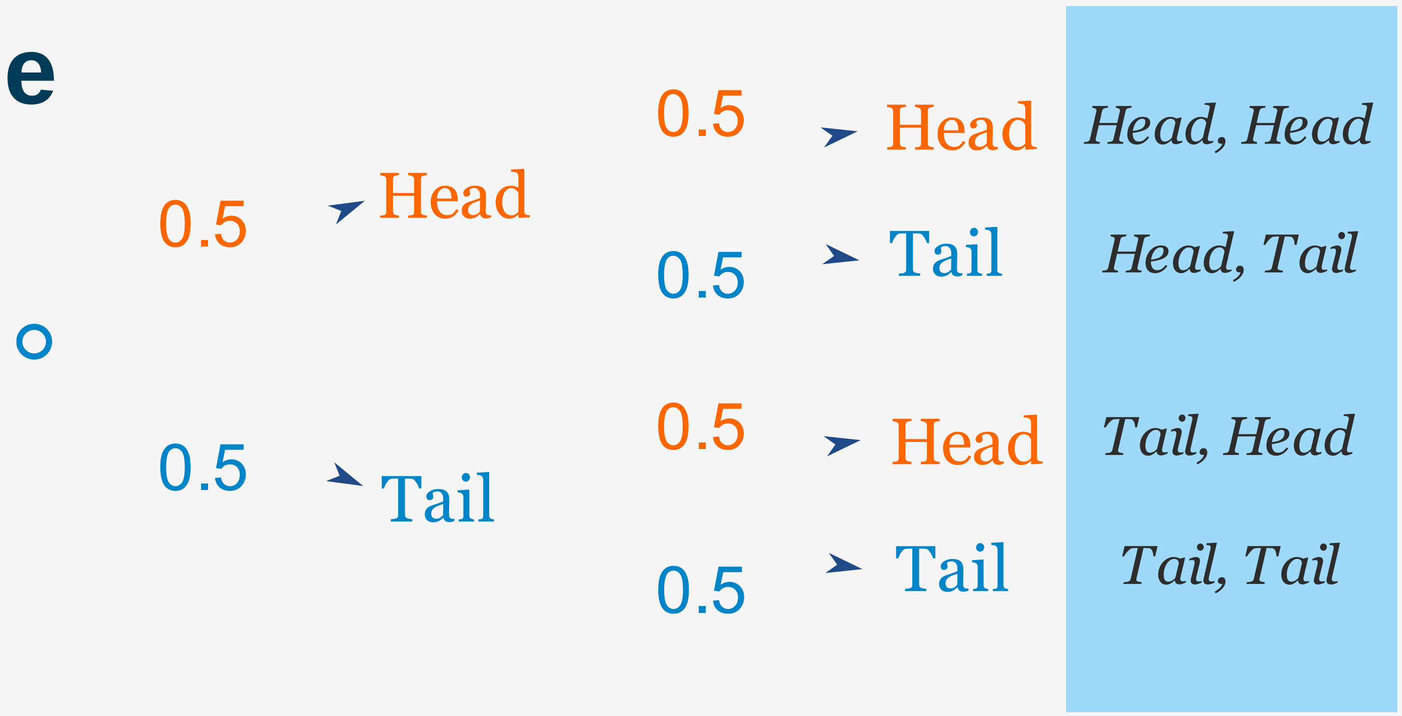
For Example



X = Number of heads when tossing 1 coin

x	0	1
$P(X = x)$	0.5	0.5

For Example

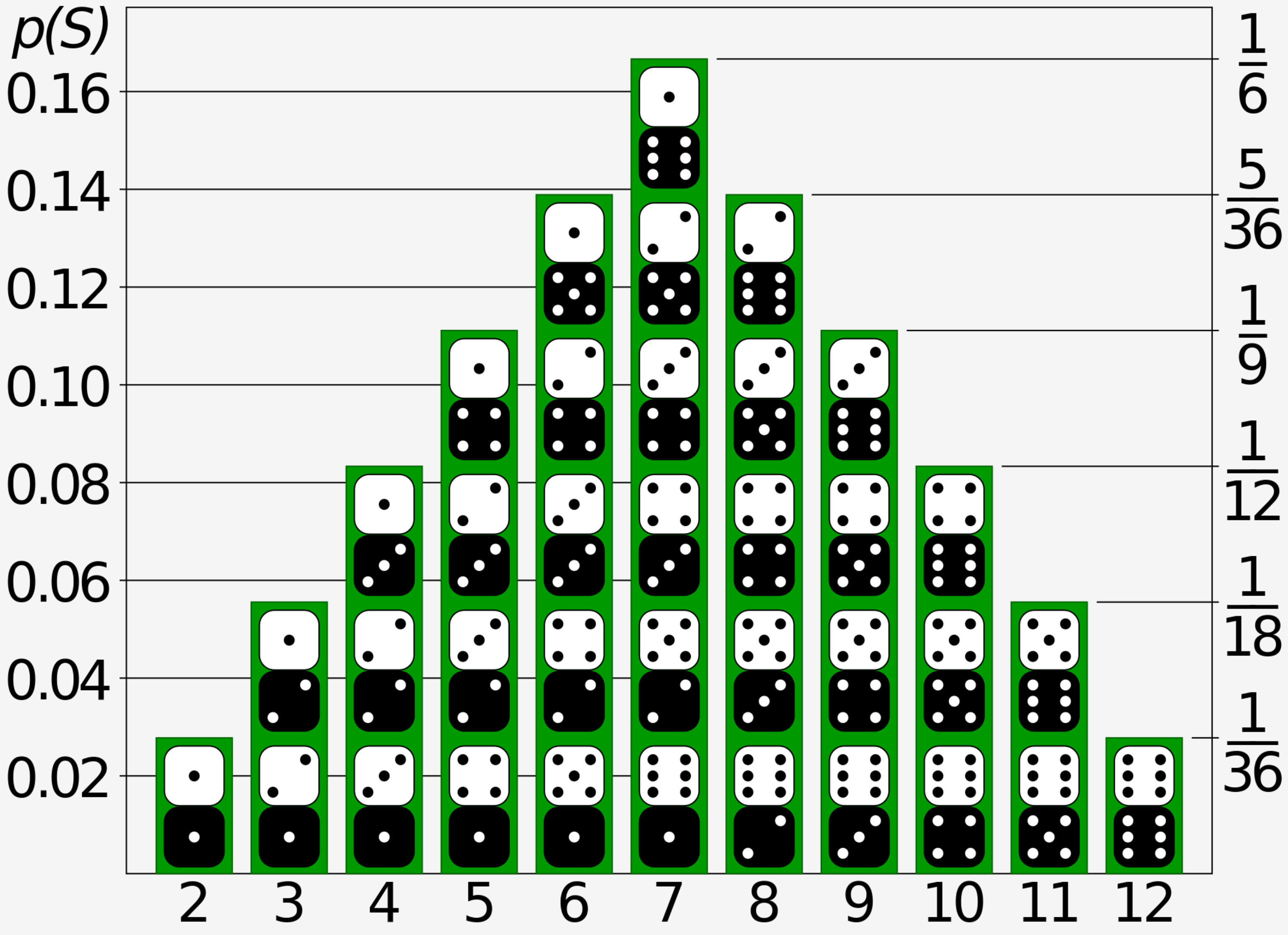


X = Number of heads when tossing 2 coins

x	0	1	2
$P(X = x)$	0.25	0.5	0.25

For Example

X = Sum of results of two dices



Bernoulli Distribution

Binary outcome

Sample space $S = \{0, 1\}$

Probability $p(1) = p \qquad p(0) = 1 - p$

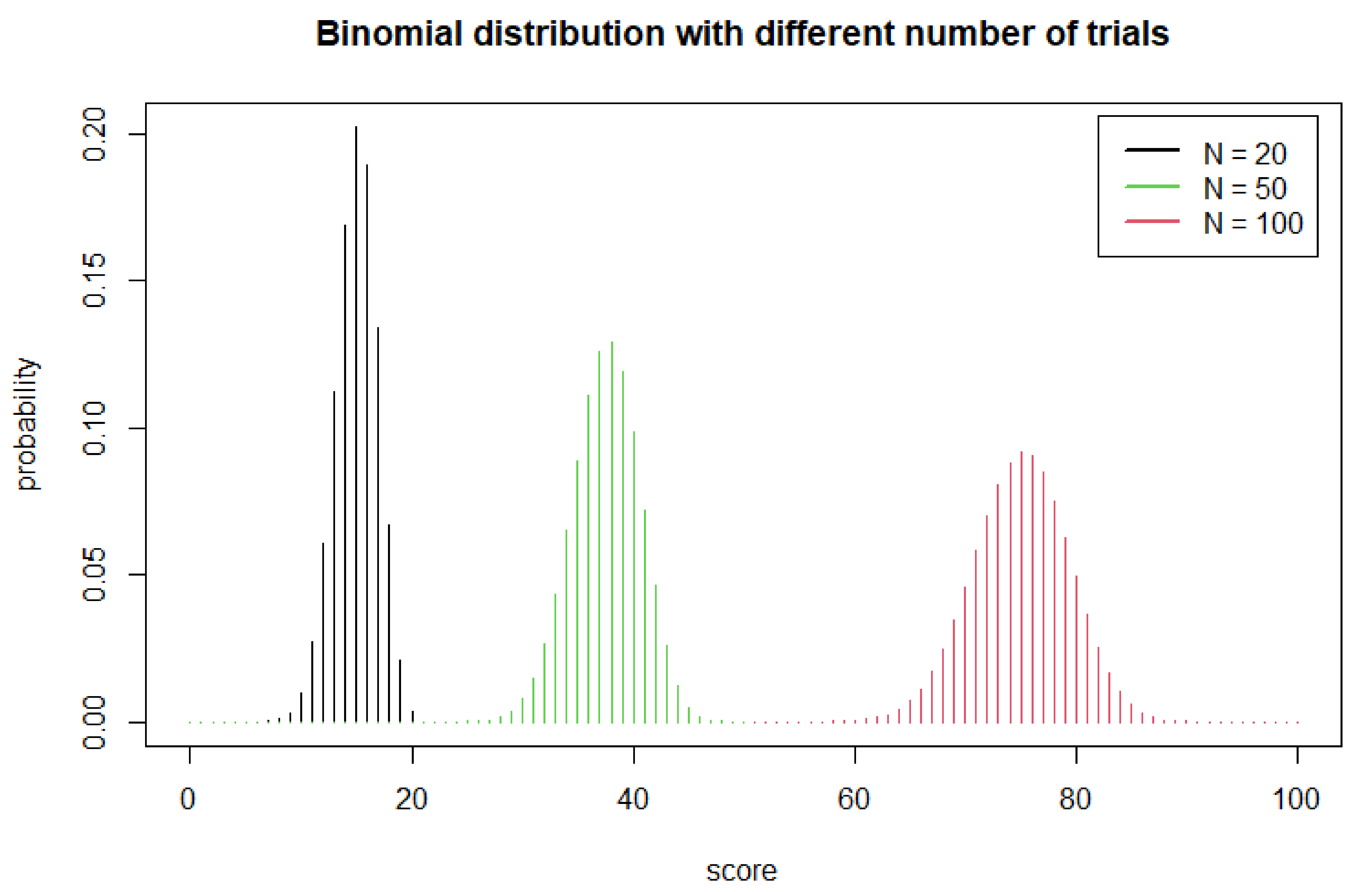
Binomial Distribution

n independent trials

with probability p of success at each trail

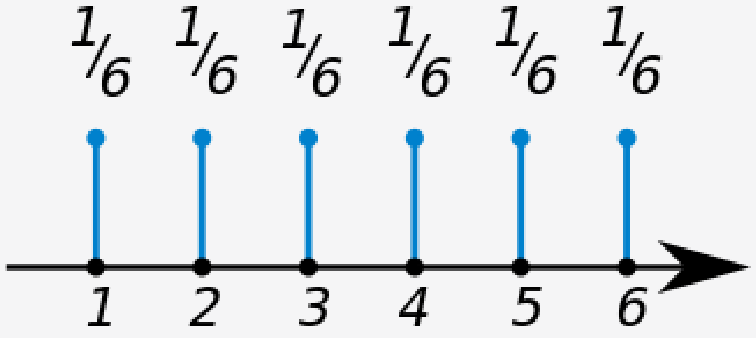
X = Number of successes in n trials

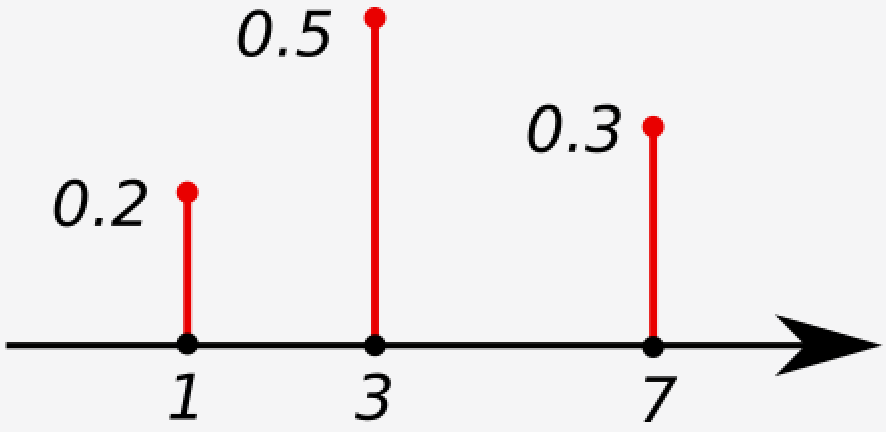
Probability distribution $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$


Binomial distribution with different number of trials
N = 20
N = 50
N = 100
probability
0.00
0.05
0.10
0.15
0.20
score
0
20
40
60
80
100

Discrete Random Variables

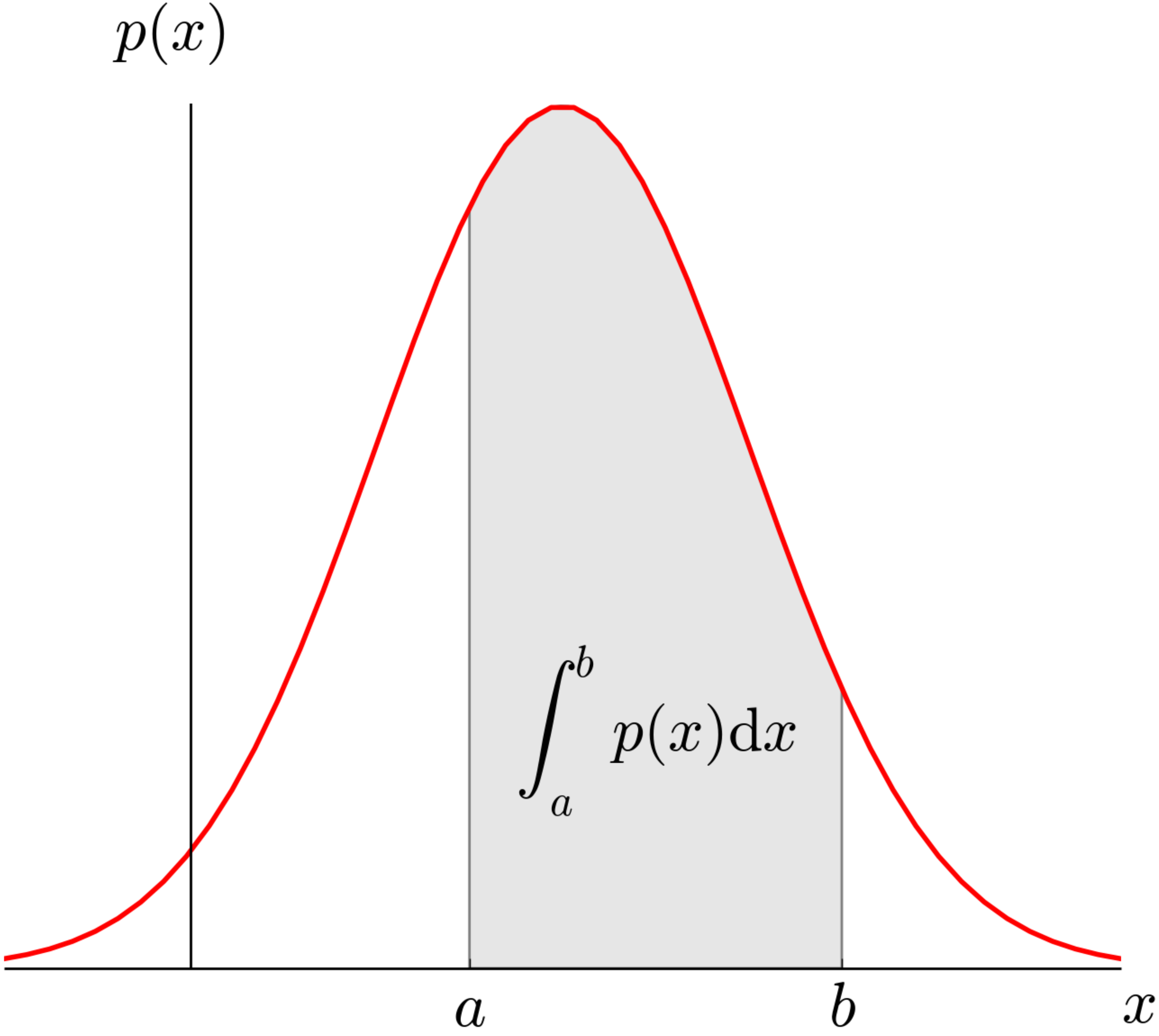
- So far, we talked about **discrete** random variables, variables that can take on only a countable number of values
- The distribution of a discrete random variable is also called probability mass function





Continuous Random Variables

- Continuous random variables can take values in a (possibly unbounded) interval of numbers
- The distribution of a continuous random variable is also called probability density function

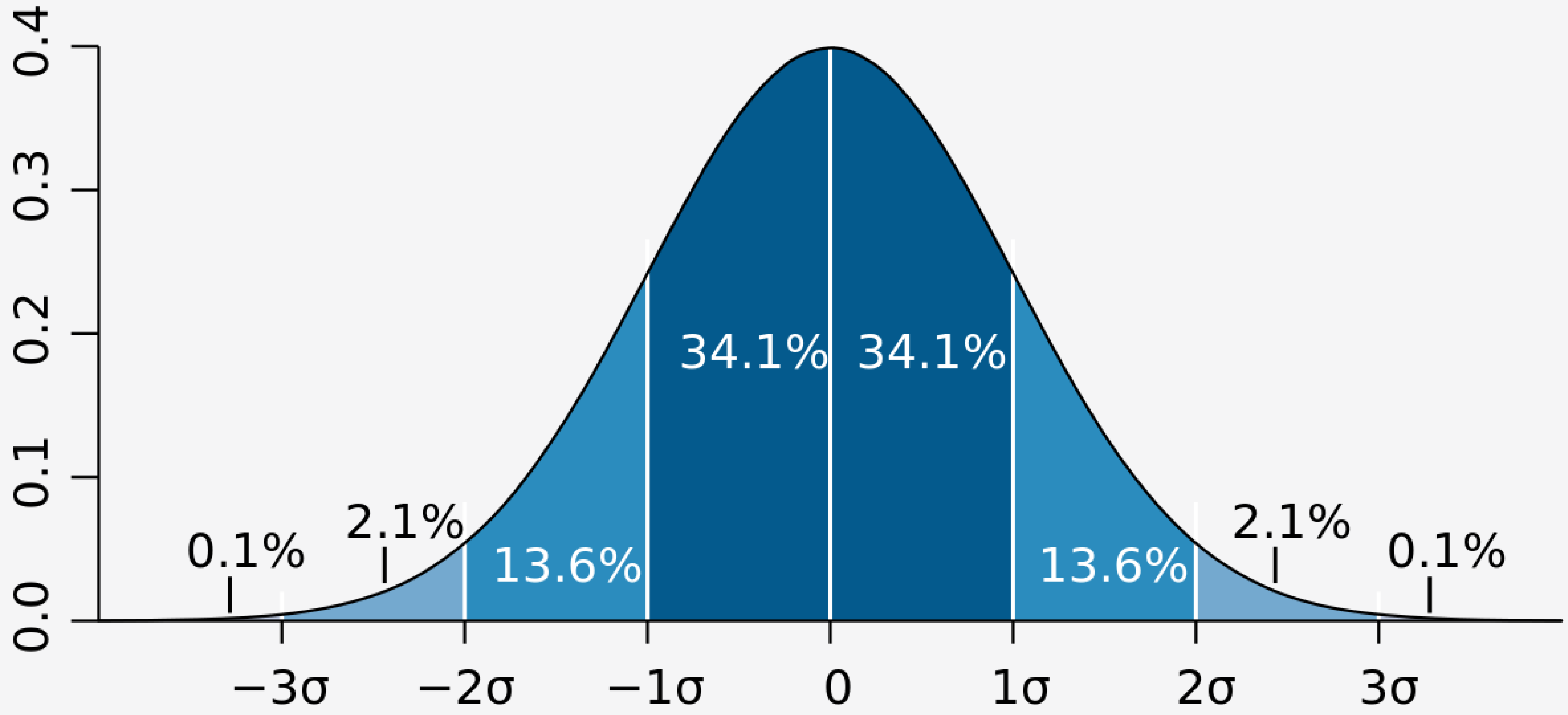

$p(x)$
$\int_a^b p(x)\mathrm{d}x$
a
b
x

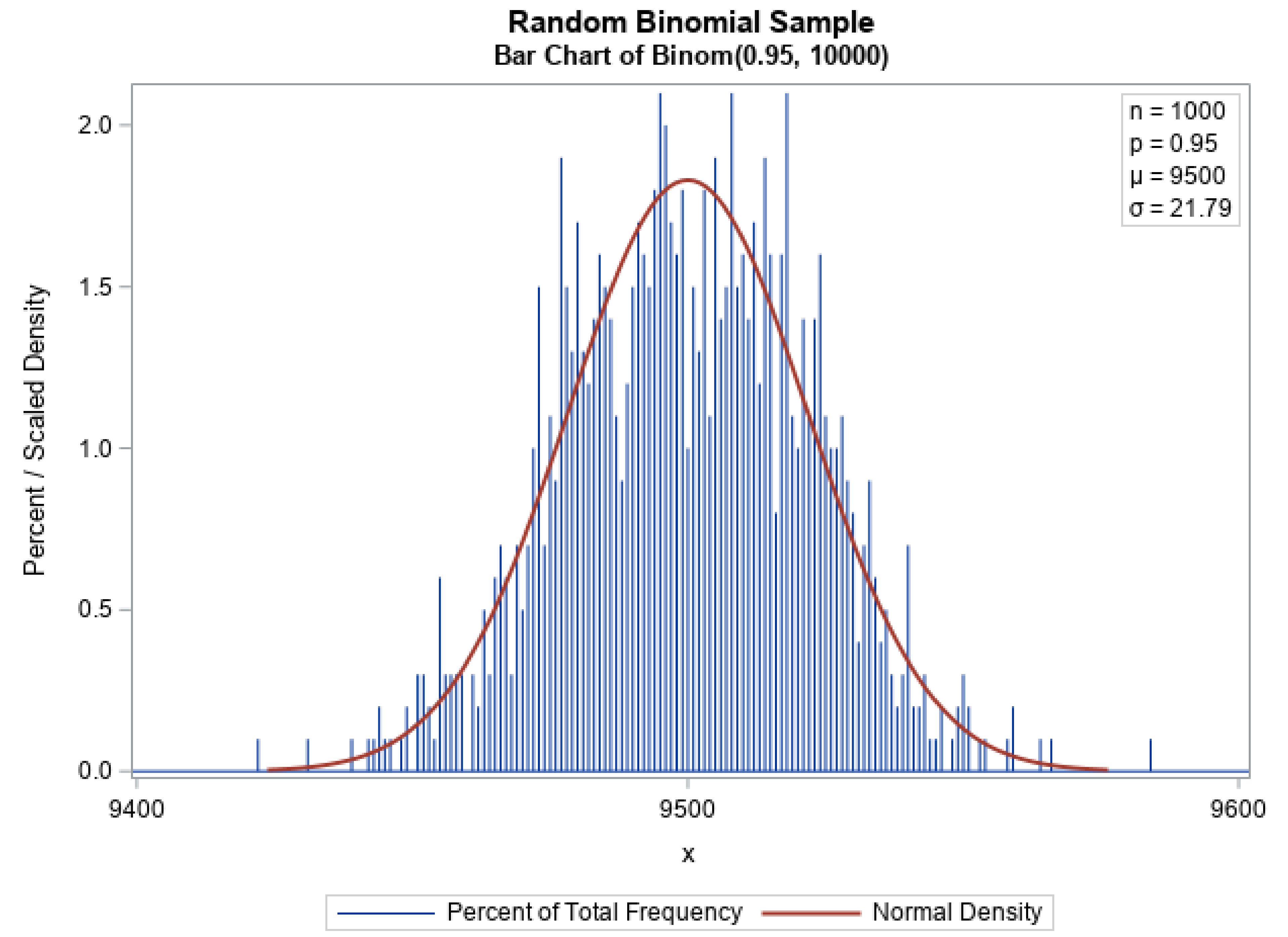
Normal Distribution

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

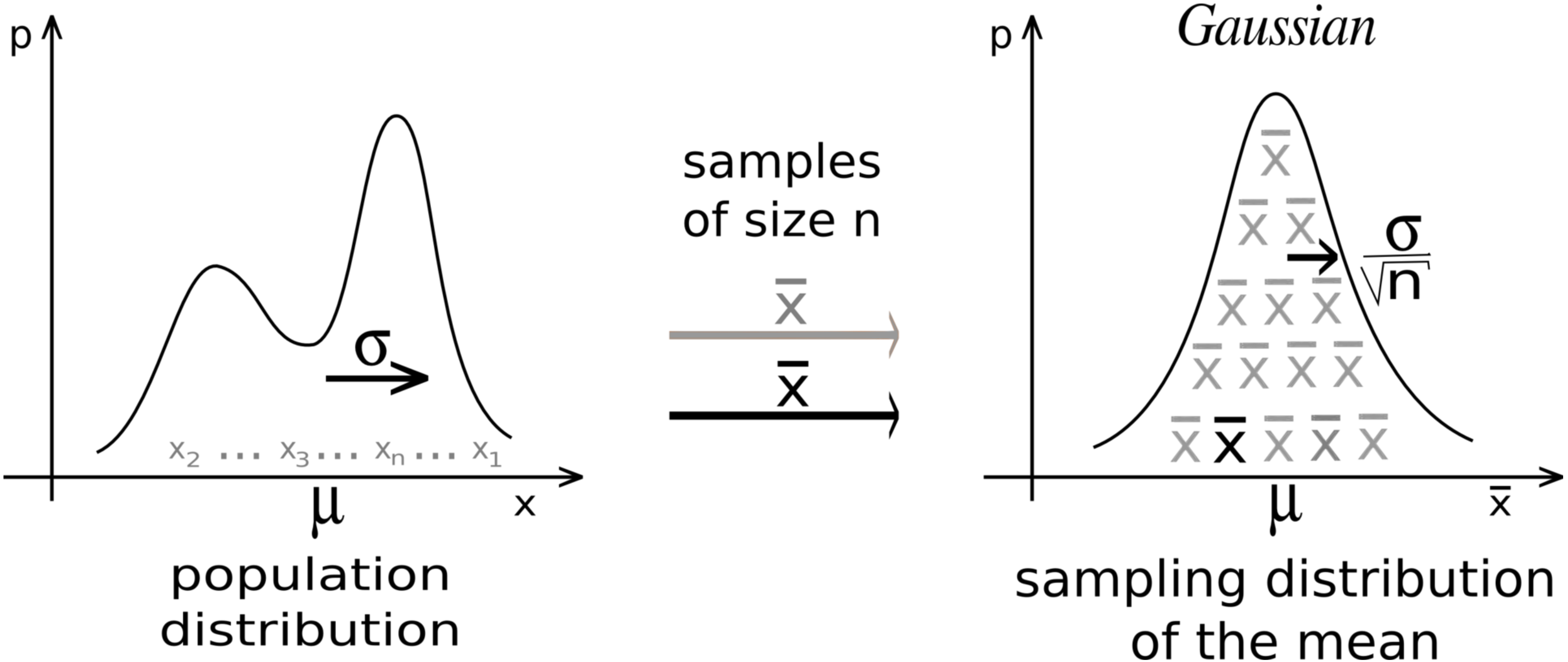
$$\mu = \text{ Mean}$$
$$\sigma = \text{ Standard Deviation}$$
$$\pi \approx 3.14159\cdots$$
$$e \approx 2.71828\cdots$$

Normal Distribution




Random Binomial Sample
Bar Chart of Binom(0.95, 10000)
n = 1000
p = 0.95
μ = 9500
σ = 21.79
Percent / Scaled Density
2.0
1.5
1.0
0.5
0.0
9400
9500
9600
x
Percent of Total Frequency
Normal Density

Why Normal Distribution is Important?



Independent study

- R read chapter 1 of statistics in R course (learning hub)
- Python read chapter 3 of statistics in Python course (learning hub)

Additional references

- Statistical distributions with practical examples https://datasciencedojo.com/blog/types-of-statistical-distributions-in-ml/
- If you want to dig into mathematical details a free (full!) course in probability https://online.stat.psu.edu/statprogram/stat414

Additional references

- R probability distributions https://rstudio.github.io/r-manuals/r-intro/Probability-distributions.html
- Python stats models distributions https://www.statsmodels.org/stable/distributions.html