DATA SCIENCE

Statistics for Data Science

Probability Distributions



UNC GREENSBORO

FOUNDATIONS

A variable can be counted or measured

Discrete values are countable. Continuous values can fall anywhere within an interval.

Discrete: count	Continuous: measure
Number of customers	Customer waiting time
Number of purchases	Customer height
Number of errors	Temperature
Number of clicks	Response time

Discrete means countable, not small

Example

Website visits in one day

$X \in \{0, 1, 2, 3, \ldots, 1000, \ldots\}$

There may be infinitely many possible counts, but they are countable.

A value such as 2.37 visits has no meaning in this context.

Example: Ω = 5 coin tosses = {<HHHHH>, <HHHHT>, <HHHTH>, <HHHTH>…}

We may just care about how many tails? Thus,

$X(<HHHHH>) = 0$

$X(<HHHTH>) = 1$

$X(<TTTHT>) = 4$

$X(<HTTTT>) = 4$

X only has 6 possible values: 0, 1, 2, 3, 4, 5

What is the probability that we end up with $k = 4$ tails?

$\mathbf{P}(X = k) := \mathbf{P}(\{\omega : X(\omega) = k\})$ where $\omega \in \Omega$

X is a *discrete random variable* if it takes only a countable number of values.

DISCRETE PROBABILITY

A distribution pairs values with probabilities

X = number of purchases made by a customer

x	P(X = x)
0	0.40
1	0.30
2	0.20
3	0.07
4	0.03

Properties of a PMF

$$P(X = x) \geq 0$$

Every probability is nonnegative.

$$\Sigma_x P(X = x) = 1$$

All possible outcomes together have probability 1.

The PMF answers exact-value questions

For discrete random variables, we use a Probability Mass Function (PMF). A PMF describes the probability distribution of a discrete random variable.

x	P(X = x)
0	0.40
1	0.30
2	0.20
3	0.07
4	0.03

Probability mass function

For a given *discrete* random variable X, *probability mass function* (*pmf*), *fx:* $\mathbb{R} \to [0, 1]$, is defined by:

$$f_X(x) = \mathrm{P}(X = x)$$

p(2) = 0.20

P(X ≥ 2) =
P(X=2)+P(X=3)+P(X=4)

= 0.20 + 0.07 + 0.03 = 0.30

An interval or grouped event adds the relevant point probabilities.

DISCRETE PROBABILITY

A PMF places probability on individual values



The PMF gives the probability that the random variable takes a specific value.

BINOMIAL DISTRIBUTION

Binomial questions count successes across fixed trials

A simple data science example

A fraud model reviews 10 known fraudulent transactions.

For this example, assume each transaction is independently flagged with probability 0.20.

X = number of transactions flagged

Question: What is the chance that exactly 3 fraudulent transactions are flagged?

BINOMIAL DISTRIBUTION

A binomial model needs four conditions

Use it when

- The number of trials n is fixed
- Each trial has two outcomes (success or failure)
- Trials are independent
- Success probability p stays constant

Pause if

- The number of trials can change
- More than two outcomes matter
- p changes from trial to trial
- Trials depend on earlier outcomes

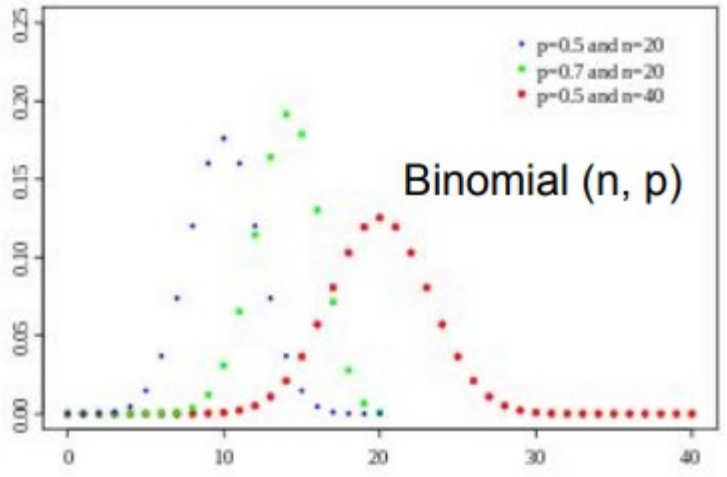
X = number of successes in n trials

X ~ Binomial(n, p)

n number of trials

p probability of success on each trial

X ∈ {0, 1, 2, …, n}



BINOMIAL DISTRIBUTION

The binomial PMF combines arrangements and likelihood

Probability of exactly k successes

$$P(X = k) = C(n, k)\, p^k\, (1 - p)^{n-k}$$

$$C(n, k) = n! / [k!(n - k)!]$$

C(n, k) counts the arrangements of k successes.

p^k models the successes; $(1 - p)^{n-k}$ models the failures.

BINOMIAL DISTRIBUTION

Probability of Exactly 3 successes

A simple data science example

A fraud model reviews 10 known fraudulent transactions.

For this example, assume each transaction is independently flagged with probability 0.20.

X = number of transactions flagged

For X ~ Binomial(10, 0.2)

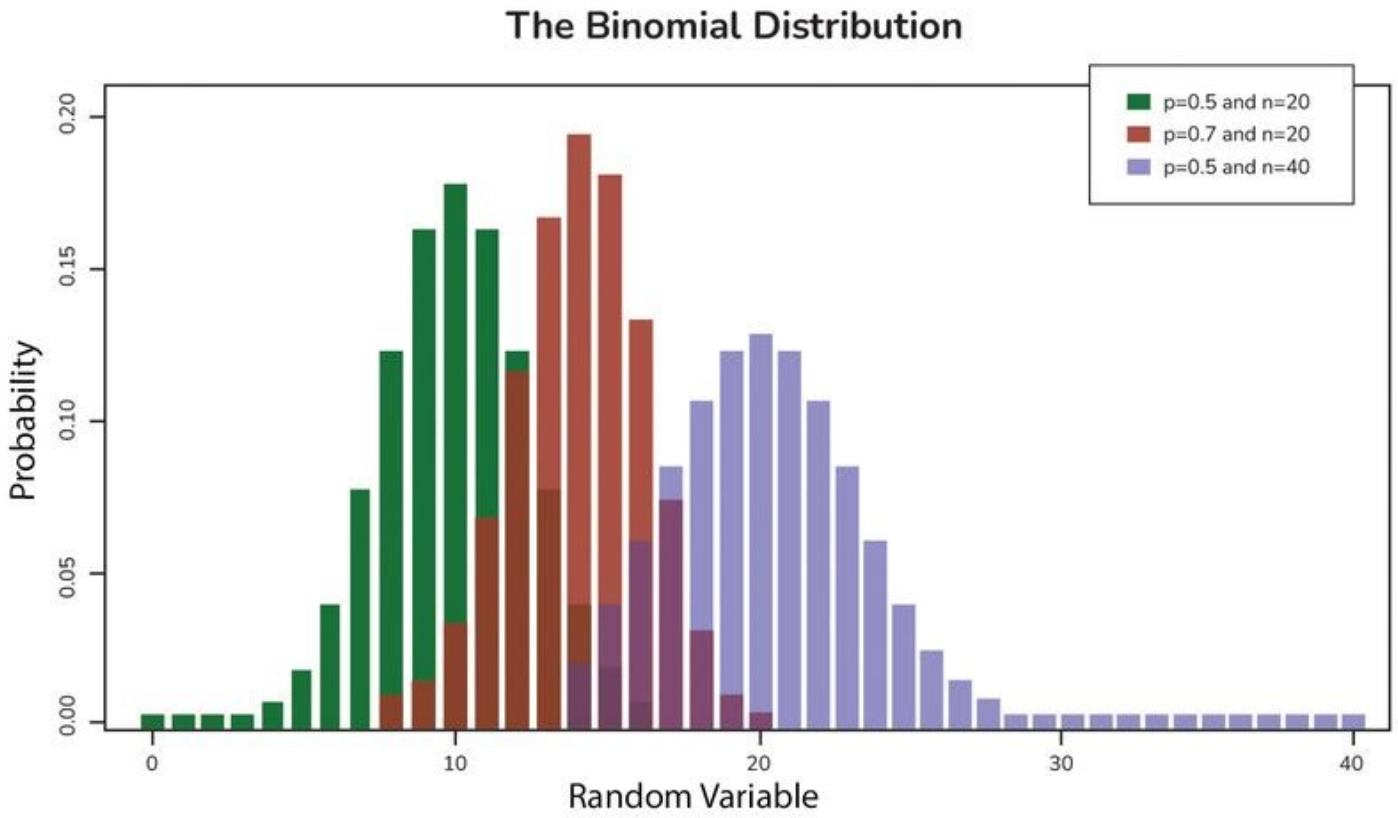
$$P(X = 3) = C(10, 3)(0.2)^3(0.8)^7$$

$$= 120 \times 0.008 \times 0.2097152$$

$$\approx 0.2013$$

About a 20.1% chance of exactly three successes.

Changing p shifts the binomial distribution



Discuss

- What changes as p rises?
- Where does probability concentrate?
- Is each shape symmetric?

CUMULATIVE PROBABILITY

The CDF answers ‘at most’ questions

PMF

What is the probability of exactly k?

$$F(x) = P(X = x)$$

CDF

What is the probability of at most k?

$$F(x) = P(X \leq x)$$

A discrete CDF accumulates point probabilities up to x.

For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

$$P(X \leq x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}$$

CUMULATIVE PROBABILITY

Probability of At most 3 successes

A simple data science example

A fraud model reviews 10 known fraudulent transactions.

For this example, assume each transaction is independently flagged with probability 0.20.

X = number of transactions flagged

For X ~ Binomial(10, 0.2)

$P(X \leq 3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)$

$\approx 0.1074 + 0.2684 + 0.3020 + 0.2013$

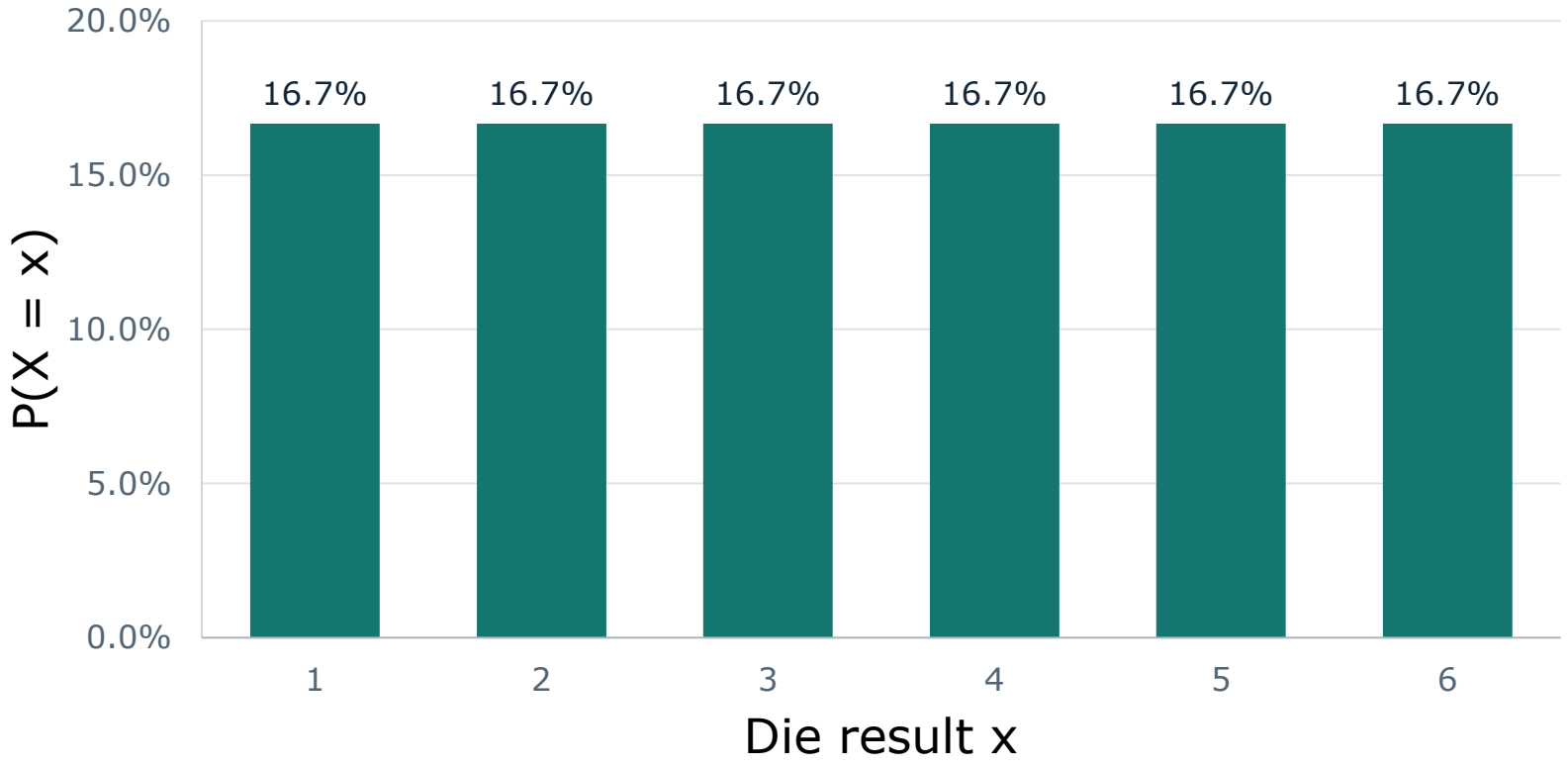
≈ 0.879

About an 87.9% chance of three or fewer successes.

DISCRETE UNIFORM

A discrete uniform model gives every value equal probability

Fair six-sided die



X ∈ {1, 2, 3, 4, 5, 6}

P(X = x) = 1/6

for each die result

DISCRETE UNIFORM

Die probabilities are counts divided by six

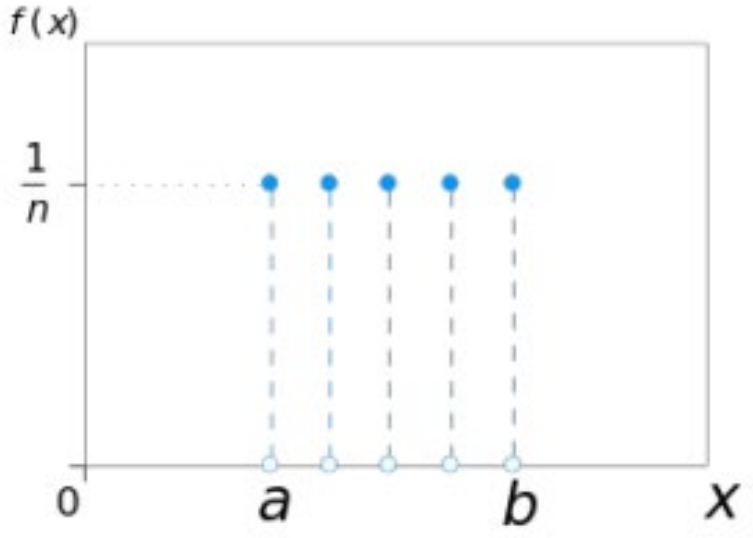
For a fair die

P(X = 4) = 1/6

P(X ≤ 4) = 4/6 = 2/3

P(X > 4) = 2/6 = 1/3

P(X is even) = 3/6 = 1/2



Count the outcomes in the event, then divide by the 6 equally likely outcomes.

GEOMETRIC DISTRIBUTION

The geometric model counts trials until the first success

Binomial

How many successes in n trials?

Geometric

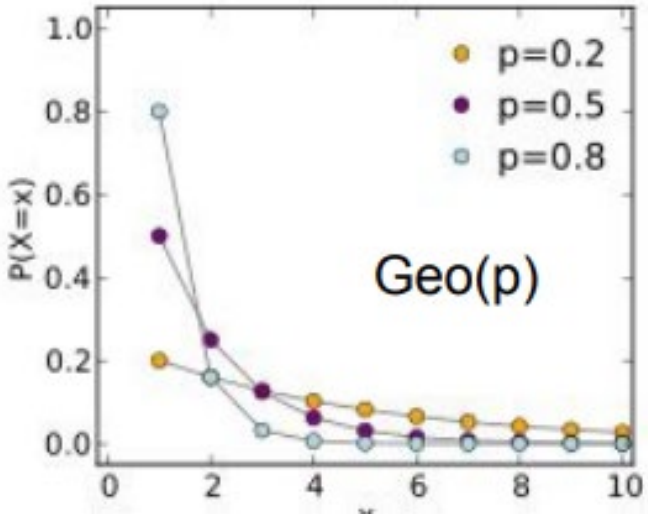
How many trials until the first success?

X = trial of the first success

$$P(X = k) = (1 - p)^{k-1}p$$

First k − 1 trials fail; trial k succeeds.

If p = 0.2: $P(X = 4) = (0.8)^3(0.2) = 0.1024$



Put Things Together

Probability Mass Function (PMF)

- Binomial(n, p)
  - *Example: number of heads after n coin flips (p, probability of heads)*
- Bernoulli(p) = Binomial(1, p)
  - *Example: one trial of success or failure*
- Discrete Uniform(a, b)
  - *Example: one trial of coin flips/dice roll*
- Geometric(p)
  - *Example: coin flips until first head*

A continuous value has probability over intervals

- Example: Number of heads in 3 coin tosses

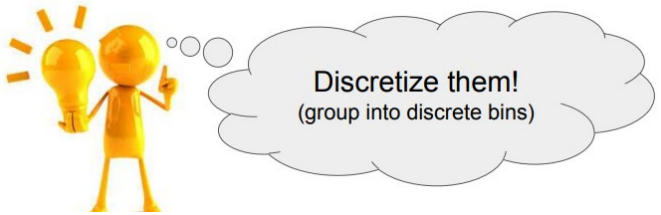
$$P(X=0),\ P(X=1),\ P(X=2),\ P(X=3) \qquad \sum P(X=x)=1$$

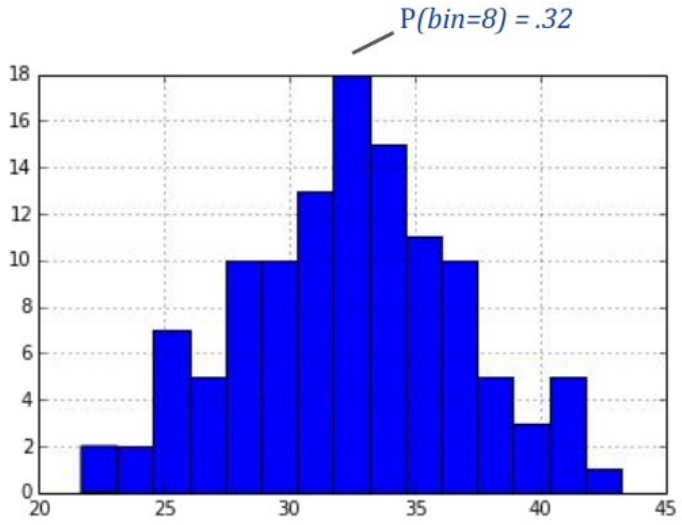
Each specific value has a probability.

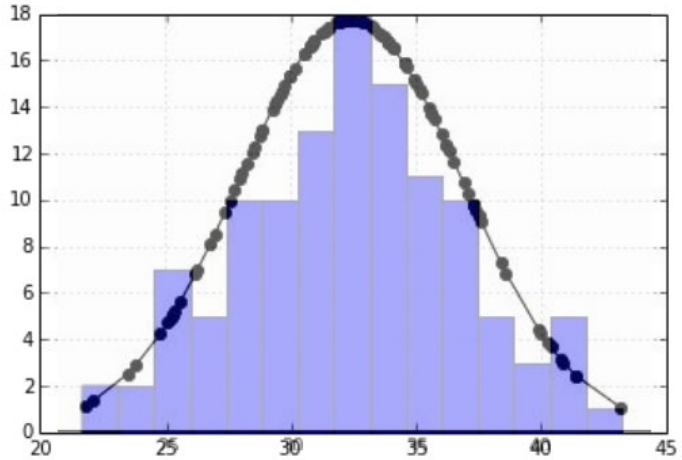
- But for continuous random variables, we learned:

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

How to model?







PROBABILITY DENSITY

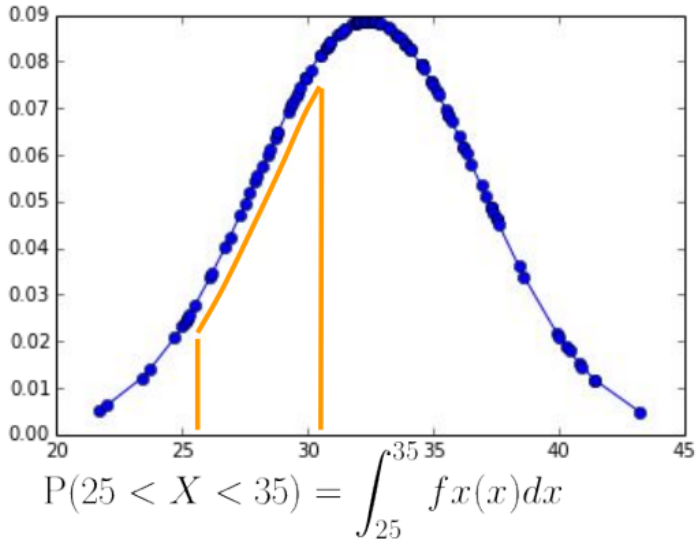
A PDF describes density; area gives probability

A PDF describes how probability is distributed over possible values of a continuous variable.

Probability density function

$$f(x)$$

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$



The area under the curve between a and b is the probability.

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

Total area under a valid PDF equals 1.

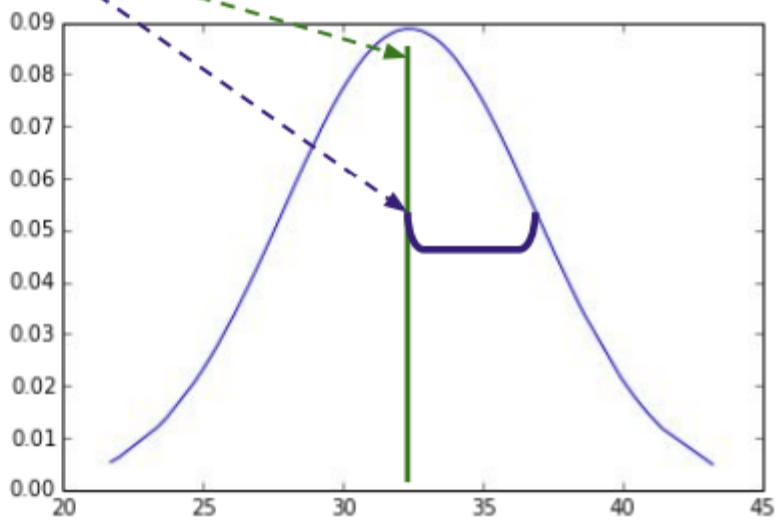
- $f(x)$ does not yield a probability
- $\int_a^b f(x)\,dx$ does
- x may be anything ($\mathbb{R}$)
- thus, may be $f(x) > 1$

The normal model describes symmetric variation

A bell-shaped, symmetric, unimodal distribution centered at its mean.

Common *pdfs*: Normal(μ, σ^2)

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



μ: mean (or "center")
= expectation

σ^2: variance,
σ: standard deviation

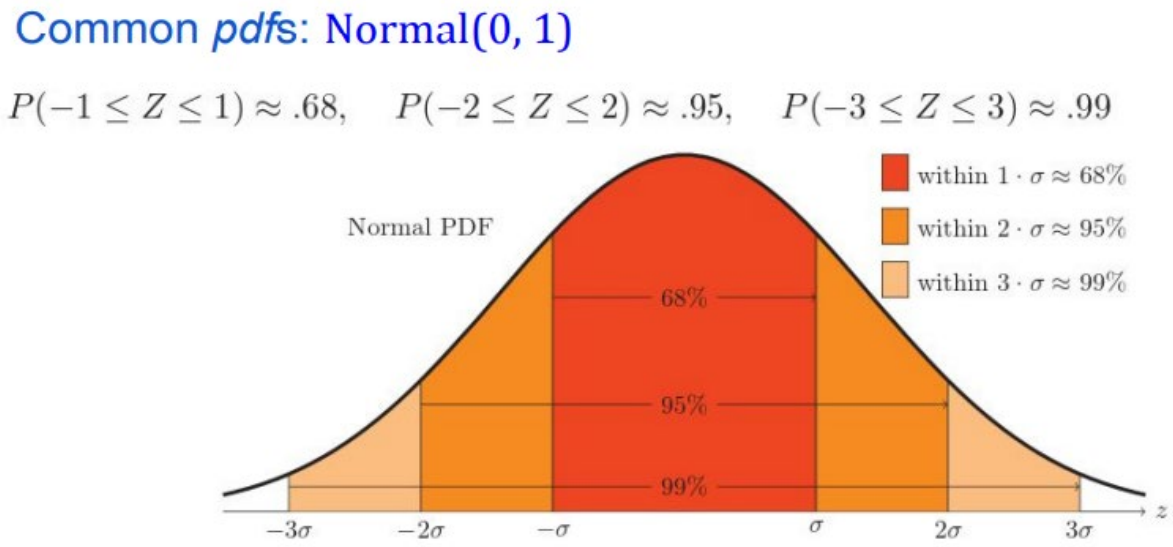
Mean = median = mode

NORMAL DISTRIBUTION

The empirical rule summarizes normal-model spread

Range around μ	Approx. share
μ ± 1σ	68%
μ ± 2σ	95%
μ ± 3σ	99.7%

68–95–99.7 rule



NORMAL DISTRIBUTION

A score 3 standard deviations below the mean is unusual

Exam scores are approximately normal

$\mu = 75$ $\sigma = 10$

68%	65 to 85
95%	55 to 95
99.7%	45 to 105

A score of 45 is 3σ below μ. It is very unusual under this normal model.

Use the rule only when the distribution is approximately normal.

Z-SCORES

A raw score needs context

Both students scored 85.

Student A

$\mu = 75 \quad \sigma = 10$

Student B

$\mu = 70 \quad \sigma = 5$

How far is each score from its own distribution?

A z-score measures distance from the mean in standard deviations.

Z-SCORES

A z-score expresses distance in standard deviations

$$z = (x - \mu) / \sigma$$

$z > 0$ **above the mean**

$z < 0$ **below the mean**

$z = 0$ **at the mean**

Student A

$$zA = (85 - 75) / 10 = 1$$

One standard deviation above the mean.

Student B

$$zB = (85 - 70) / 5 = 3$$

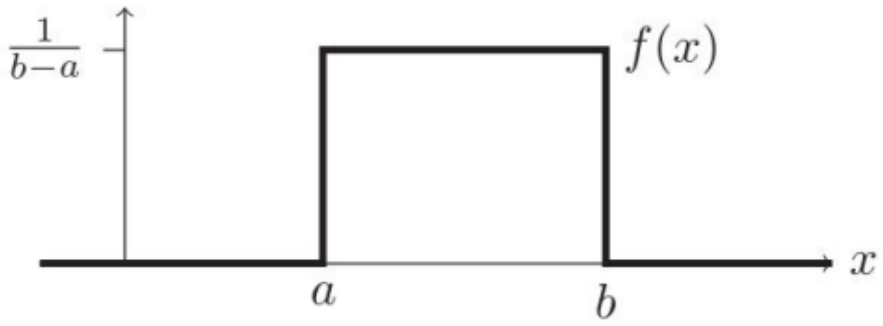
Three standard deviation above the mean.

Z-scores make standardized comparisons possible.

A continuous uniform model has constant density

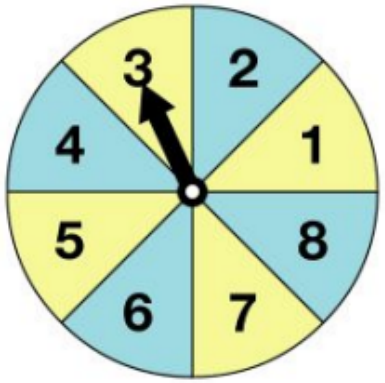
Common *pdf*s: Uniform(a, b)

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



X ~ Uniform(a, b), examples:

- spinner in a game
- random number generator
- analog to digital rounding error



Bus arrival example

The bus arrives at a random time between 10:00 and 10:20.

Waiting time X ~ Uniform(0, 20) minutes

CONTINUOUS UNIFORM

A 5-minute wait covers one quarter of the interval

$X \sim \text{Uniform}(0, 20)$

$P(X < 5)$ = interval length / total length

$= 5 / 20 = 0.25$

25% chance of waiting less than 5 minutes.

For a continuous uniform model, equal-length intervals have equal probability.

The exponential model describes waiting times

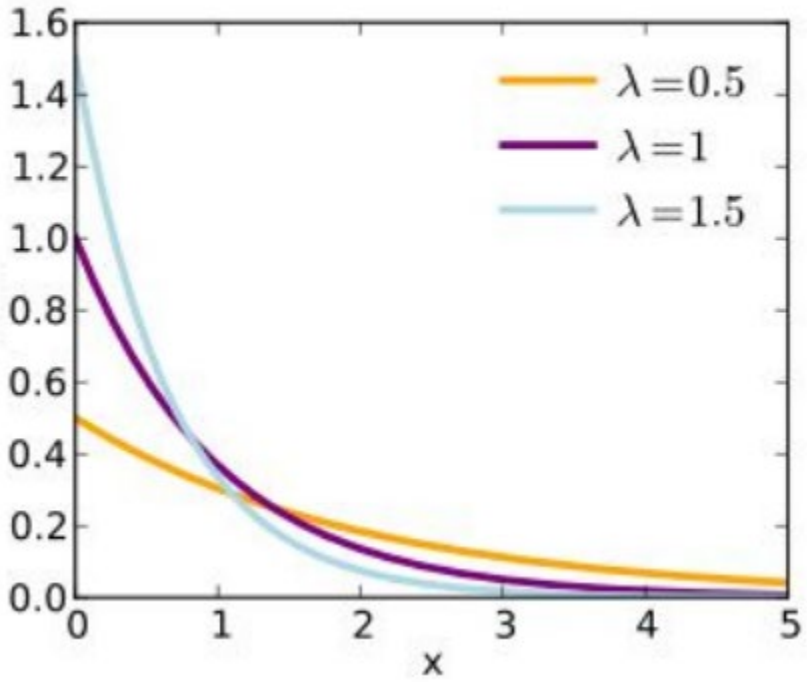
Common *pdf*s: Exponential(λ)

$$f_X(x) = \lambda e^{-\lambda x}, x > 0$$

λ: rate or inverse scale

β: scale ($\lambda = \frac{1}{\beta}$)

Credit: Wikipedia



For example, if buses arrive at a bus stop every 15 minutes on average, the time you wait for the next bus can be modelled using an exponential distribution.

EXPONENTIAL DISTRIBUTION

The exponential CDF also gives the chance of waiting longer

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}$$

$$P(X > x) = e^{-\lambda x}$$

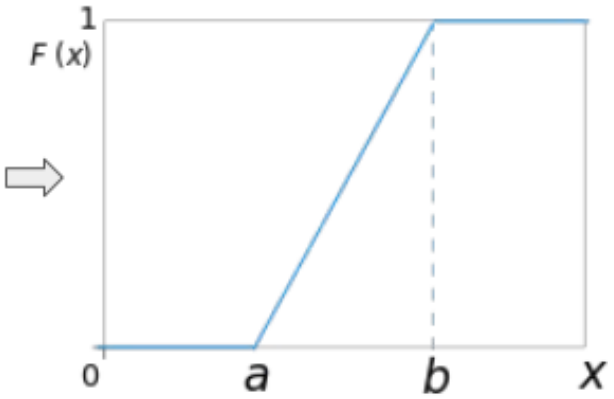
The first formula accumulates probability up to x.

The second gives the probability that the wait exceeds x.

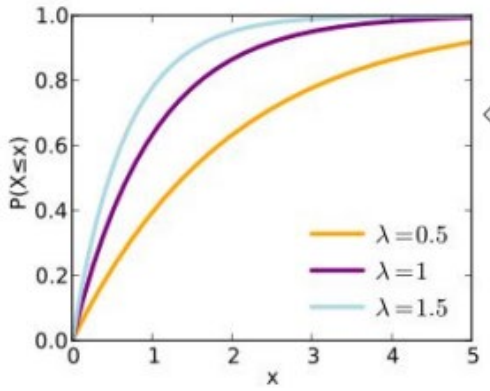
Example question: What is the chance of waiting more than 5 minutes?

CDF for continuous distribution

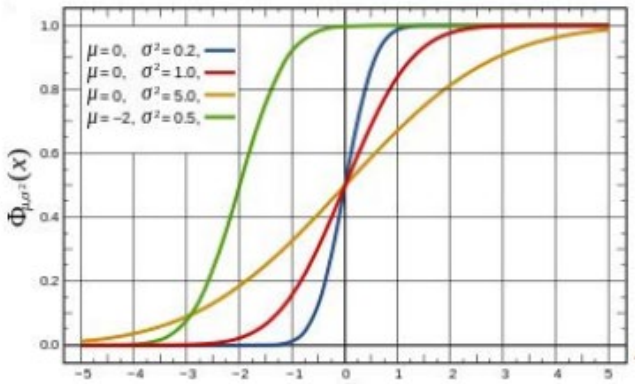
Uniform ⇨



⇦ Exponential



Normal ⇨



For a given random variable X, the *cumulative distribution function* (CDF), *Fx:* $\mathbb{R} \rightarrow [0, 1]$, is defined by:

$$F_X(x) = \mathrm{P}(X \leq x)$$

fx **:**
probability density function (pdf)

$$f_X(x) \geq 0, \text{for all } x \in X,$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1, \quad \text{and}$$

$$\mathrm{P}(a < X < b) = \int_a^b f_X(x)dx$$

INTEGRATED PRACTICE

Practice 1: choose a distribution for each scenario

Choose from Binomial, Geometric, Discrete Uniform, Normal, or Exponential.

- A. Result of rolling a fair six-sided die
- B. Waiting time until the next customer arrives
- C. Number of heads in 20 independent coin flips
- D. Height of randomly selected adults
- E. Number of attempts until the first successful login

INTEGRATED PRACTICE

Practice 2

A data scientist records the number of system failures per day:

What type of distribution is this?

What is $P(X \leq 1)$?

What is $P(X \geq 2)$?

Failures (X)	Probability
0	0.50
1	0.30
2	0.15
3	0.05

If the probability density function is given as: $f(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$

Find P (0.2≤ X ≤ 0.6)

INTEGRATED PRACTICE

Practice 3: standardize an exam score

1. The exam scores in a statistics class follow a normal distribution with: Mean $\mu=70$, Standard deviation $\sigma=10$. A student scored 85.

- Compute the z-score.
- Is the score above or below average?

2. In a sample of size 7 with values 6, 8, 10, 12, 14, 16, 18 calculate the Z-score for x = 16.