Lecture 2: Review of basic maths

Population genetic PCB4553/6685

Part 1: Probability

- Evolution is fundamentally a random process.
    - Mendelian transmission is random
    - Mutation is random
    - Mating and reproduction is random
    - Allele frequency changes in finite population is a random process
    - Many results in popgen rely on probabilistic processes in their core arguments

Random variables

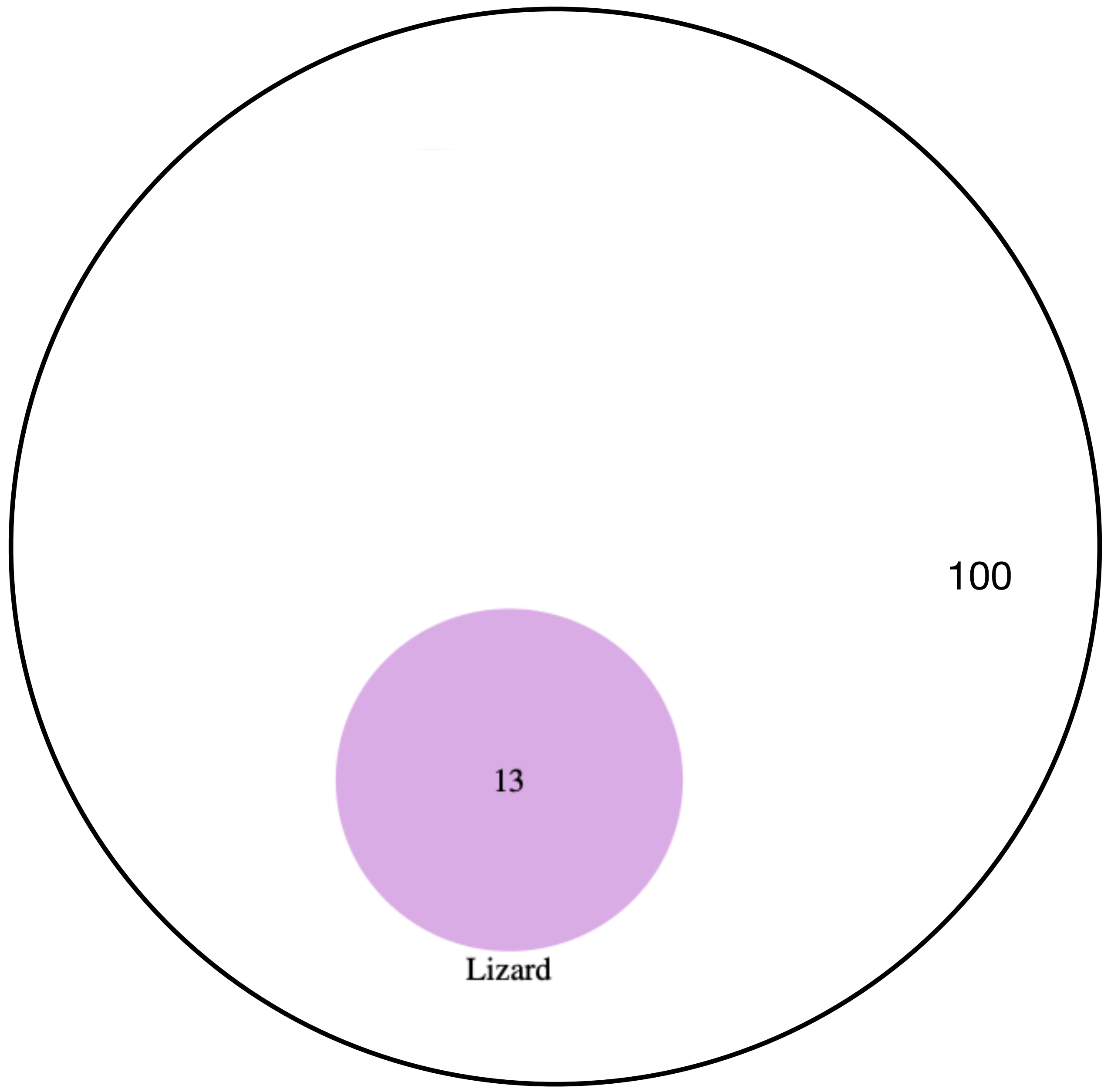
- A random variable X, roughly, is a variable that takes on values drawn randomly from a sample space, with the likelihood of each outcome specified by some function
  - Example: a six-sided die
    - Sample space: set of all possible outcomes, e.g. $\{1, 2, 3, 4, 5, 6\}$
    - Probability of each number occurring is 1/6

Types of random variables

- Discrete random variables take on a countable number of values
  - Probability *mass* function
  - $\mathbb{P}[X = x_i] = p_i$ "the probability that X equals x_i is p_i"
- Continuous random variables, which can take on values on a continuum, e.g. real number line
  - Examples: height of a person, time until the next incoming call, etc.
  - Probability *density* function
  - $\mathbb{P}[a \leq X \leq b] = \int_a^b p(x)\,dx$ "the probability that X is interval $[a, b]$ equals the area under the curve $p(x)$ from a to b"

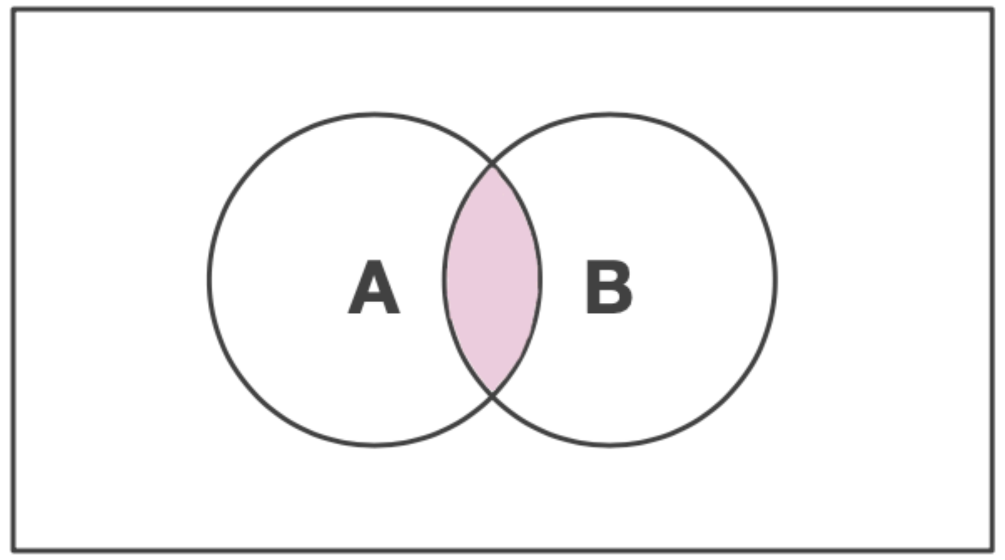
Events

- An event is a subset of the sample space
- Example 1: “a 6-sided die lands on an even number” = $\{2,4,6\} \subset \{1,2,3,4,5,6\}$
- Example 2: “a person’s height is between 160 and 165cm” = [160, 165] $\subset \mathbb{R}^+$

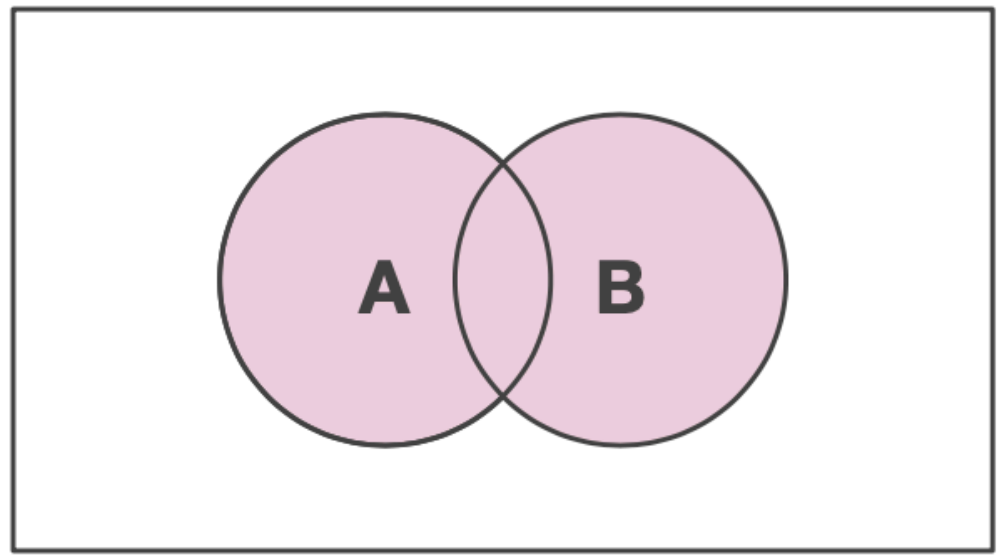


Algebra of events

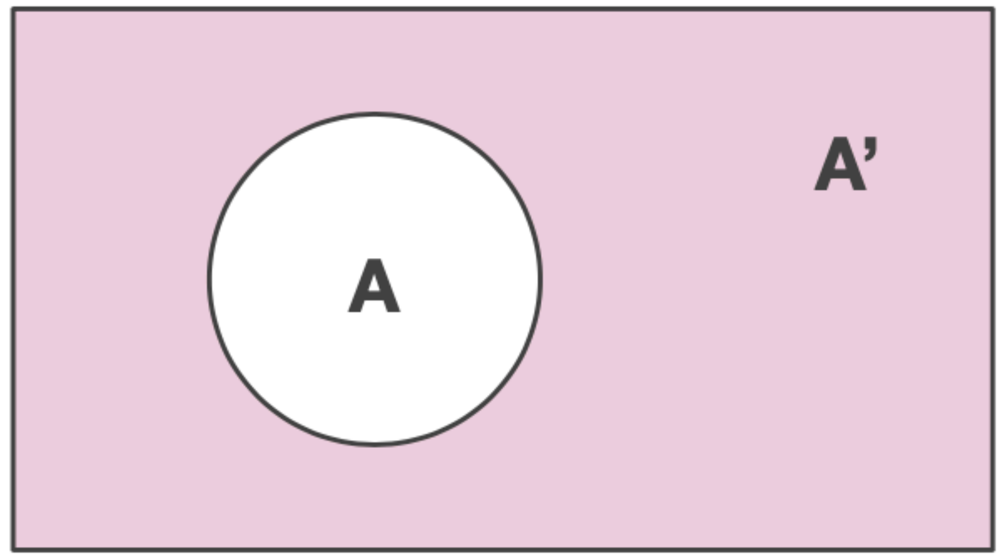
- $P(A) = 1 - P(A')$
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \varnothing$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, if if $A \cap B \neq \varnothing$



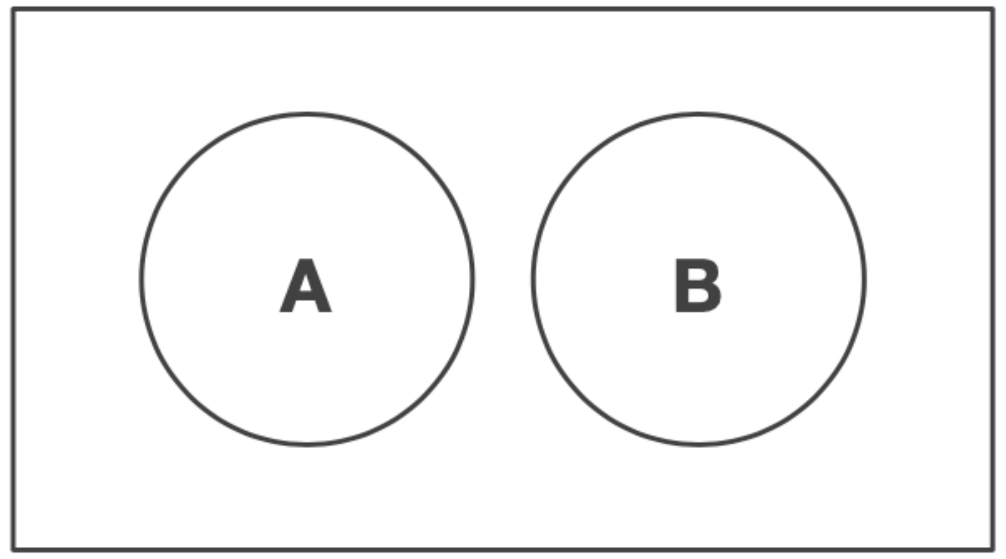
intersection: A ∩ B



union: A ∪ B



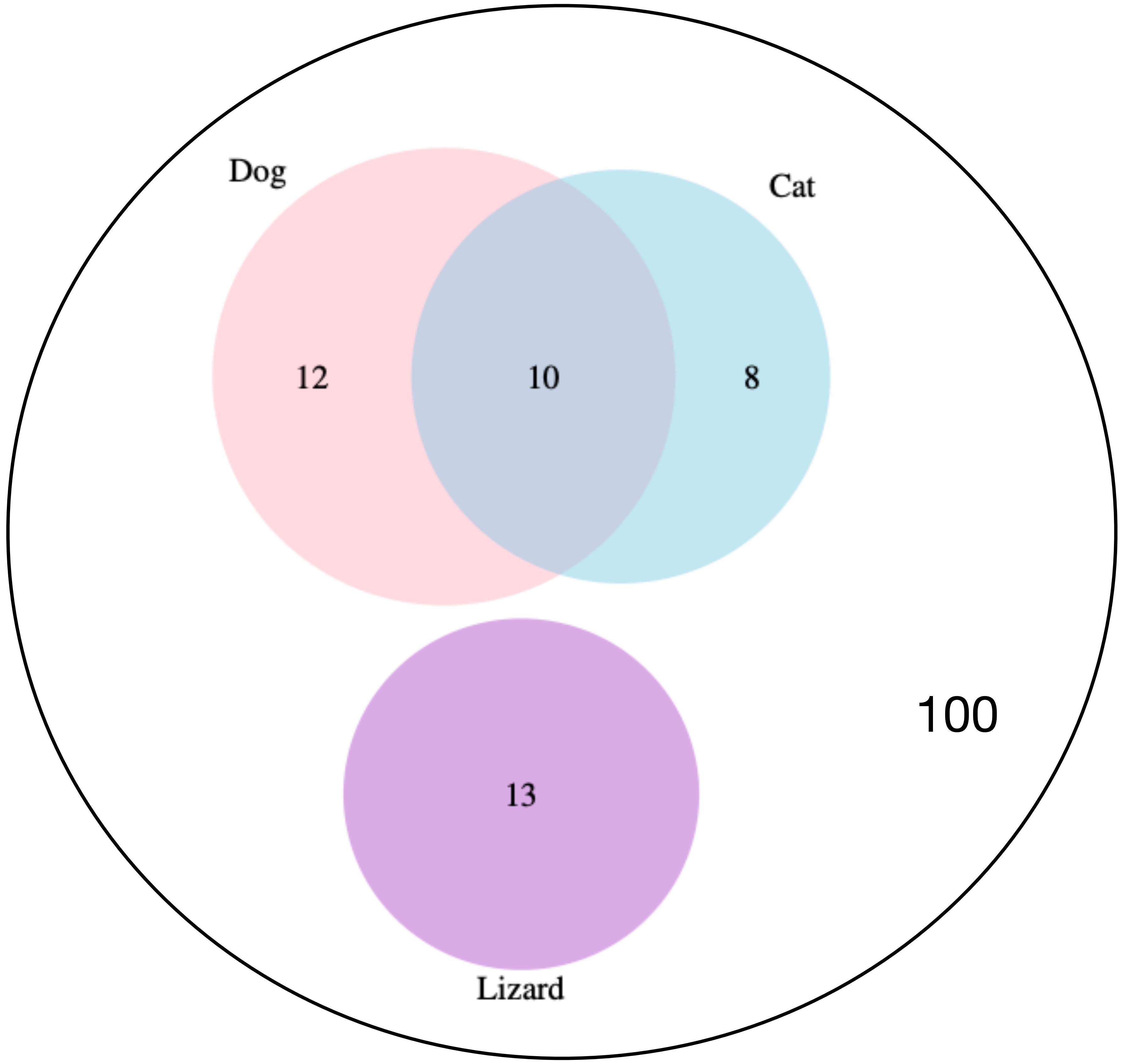
complement of A: A’



mutually exclusive (disjoint)

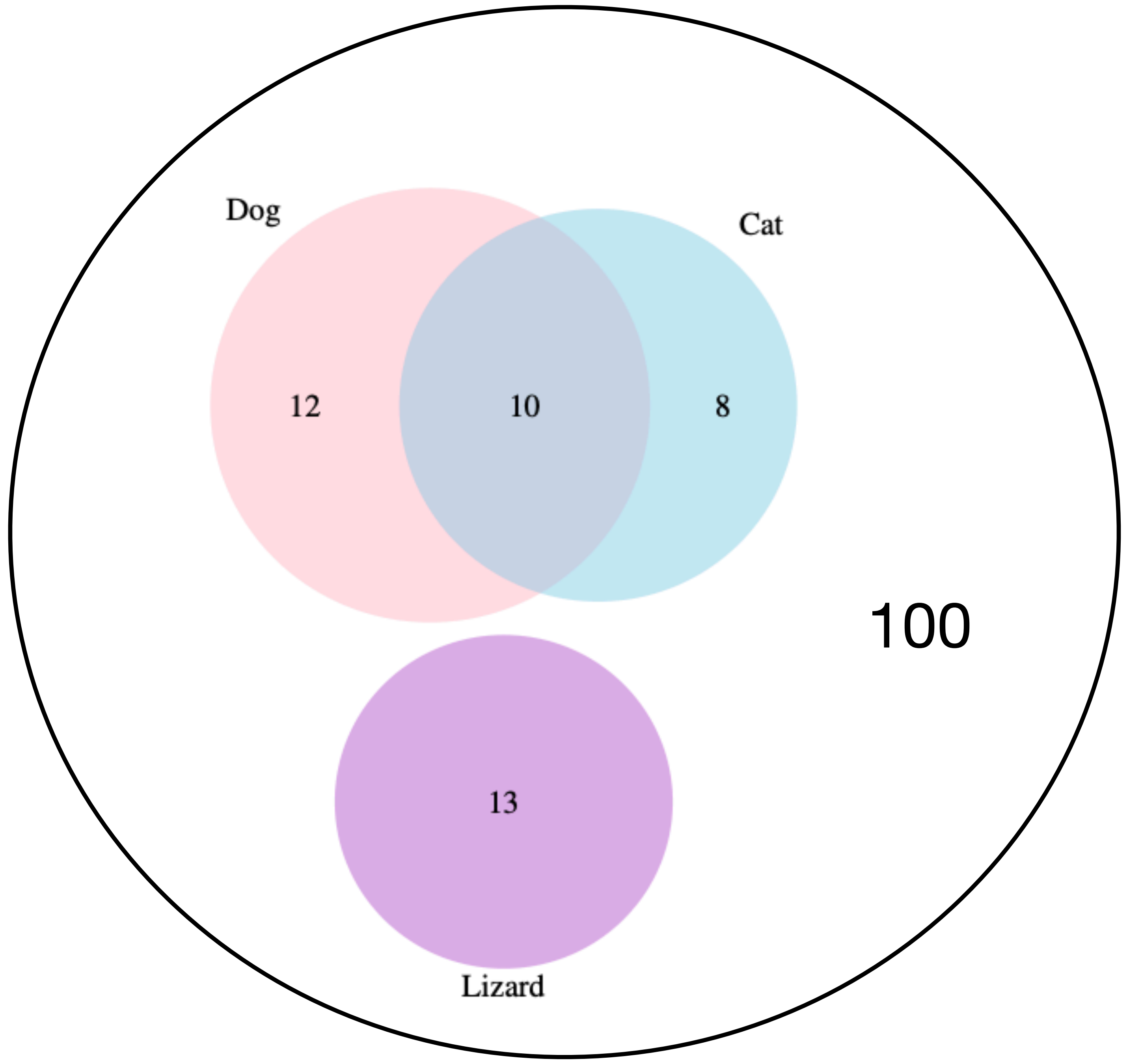
Exercise

- P(Dog) = ?
- P(Dog or not dog) = ?
- P(Dog or Lizard) = ?
- P(Dog or cat) = ?

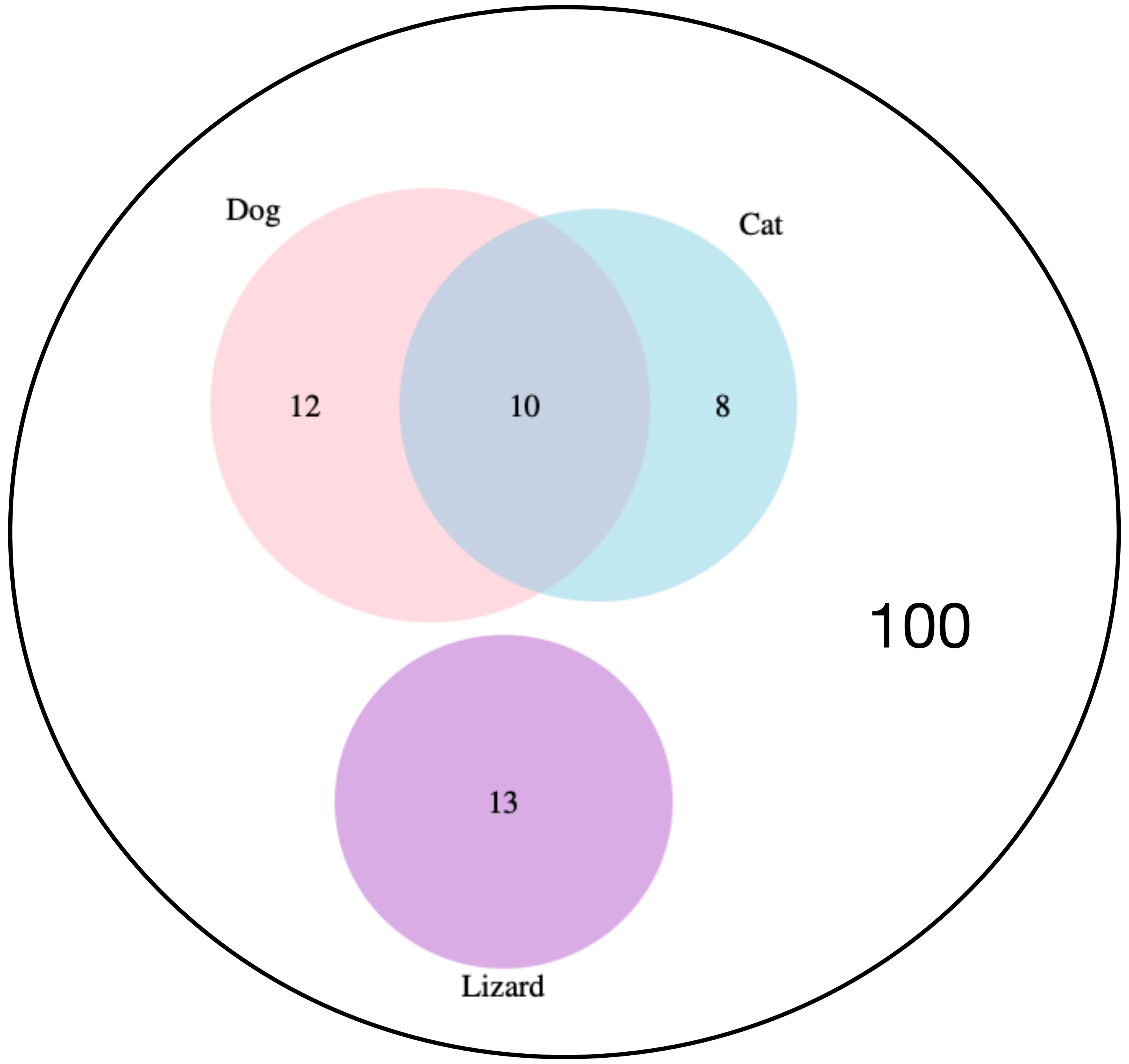


Conditional probability

- conditional probability: the probability of an event conditional on some other particular event happening.
  - $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
  - $P(\text{cat} \mid \text{dog}) = \dfrac{P(\text{cat} \cap \text{dog})}{P(\text{dog})}$
- Joint probability
  - $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$

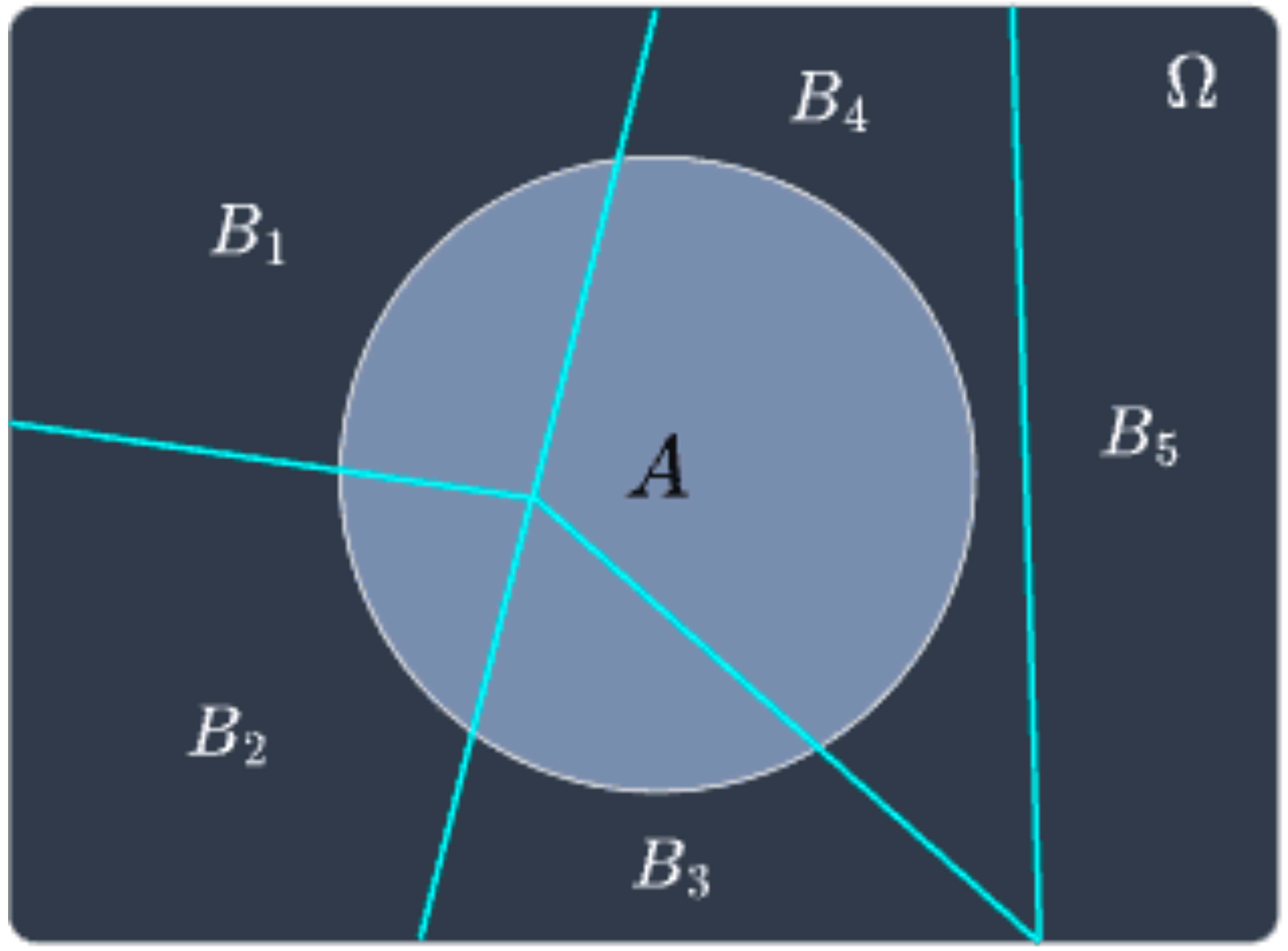


- $P(\text{cat}) = P(\text{cat} \cap \text{dog}) + P(\text{cat} \cap \text{not dog})$
- $P(\text{cat}) = P(\text{cat} \mid \text{dog})P(\text{dog}) + P(\text{cat} \mid \text{dog})P(\text{not dog})$



Law of total probability

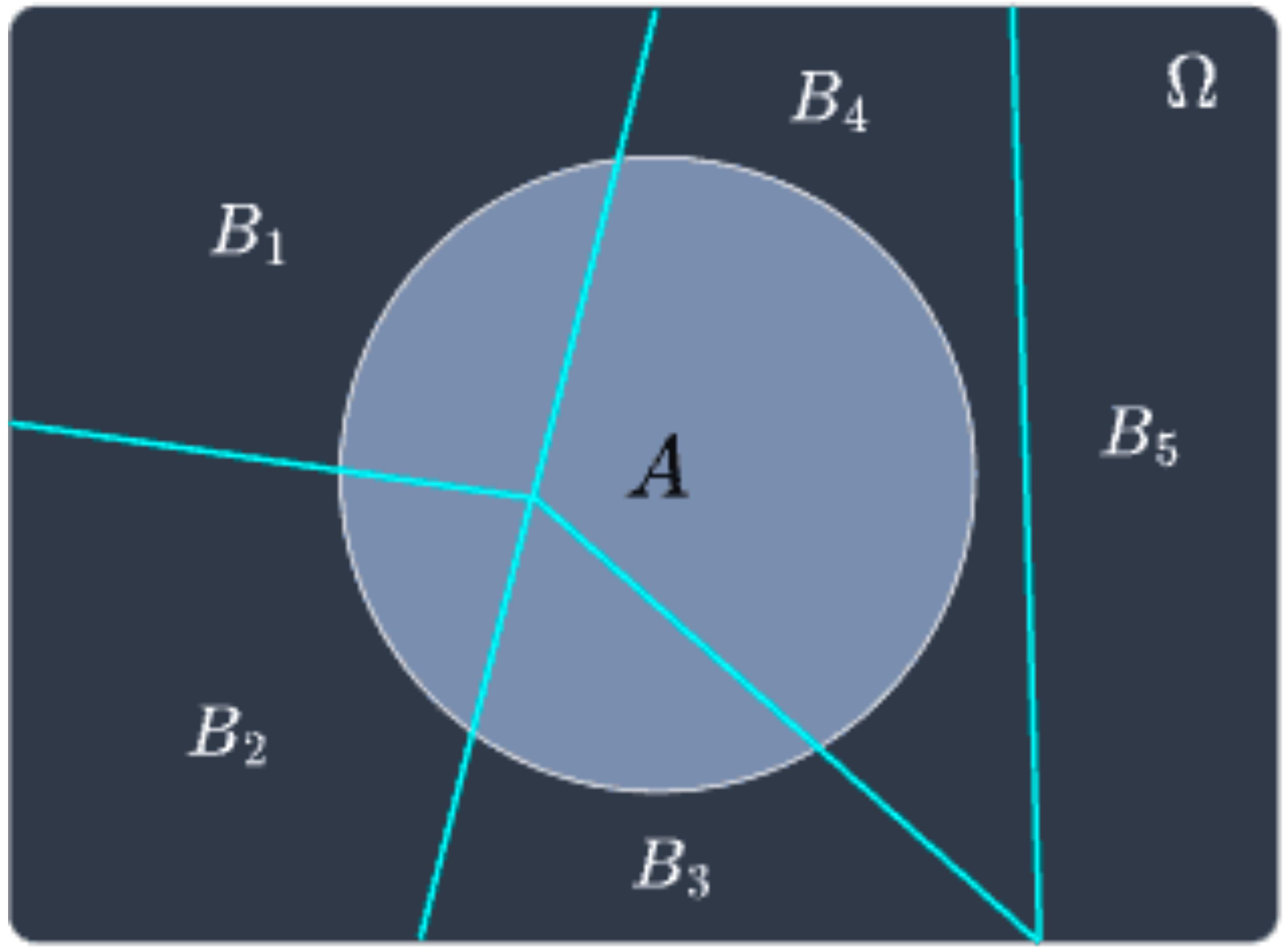
- The total probability of an event A can be obtained by summing over all of the L mutually exclusive ways that A can happen
- $B_1, B_2, \cdots, B_L$ are disjoint events, so that their union is the whole sample space
- $P(A) = \sum_{i=1}^{L} P(A \mid B_i)P(B_i) = \sum_{i=1}^{L} P(A \cap B_i)$
- $\sum_{i=1}^{L}$ is the sigma notation



Laws of total probability

$$P(A) = \sum_{i=1}^{L} P(A \mid B_i)P(B_i) = \sum_{i=1}^{L} P(A \cap B_i)$$

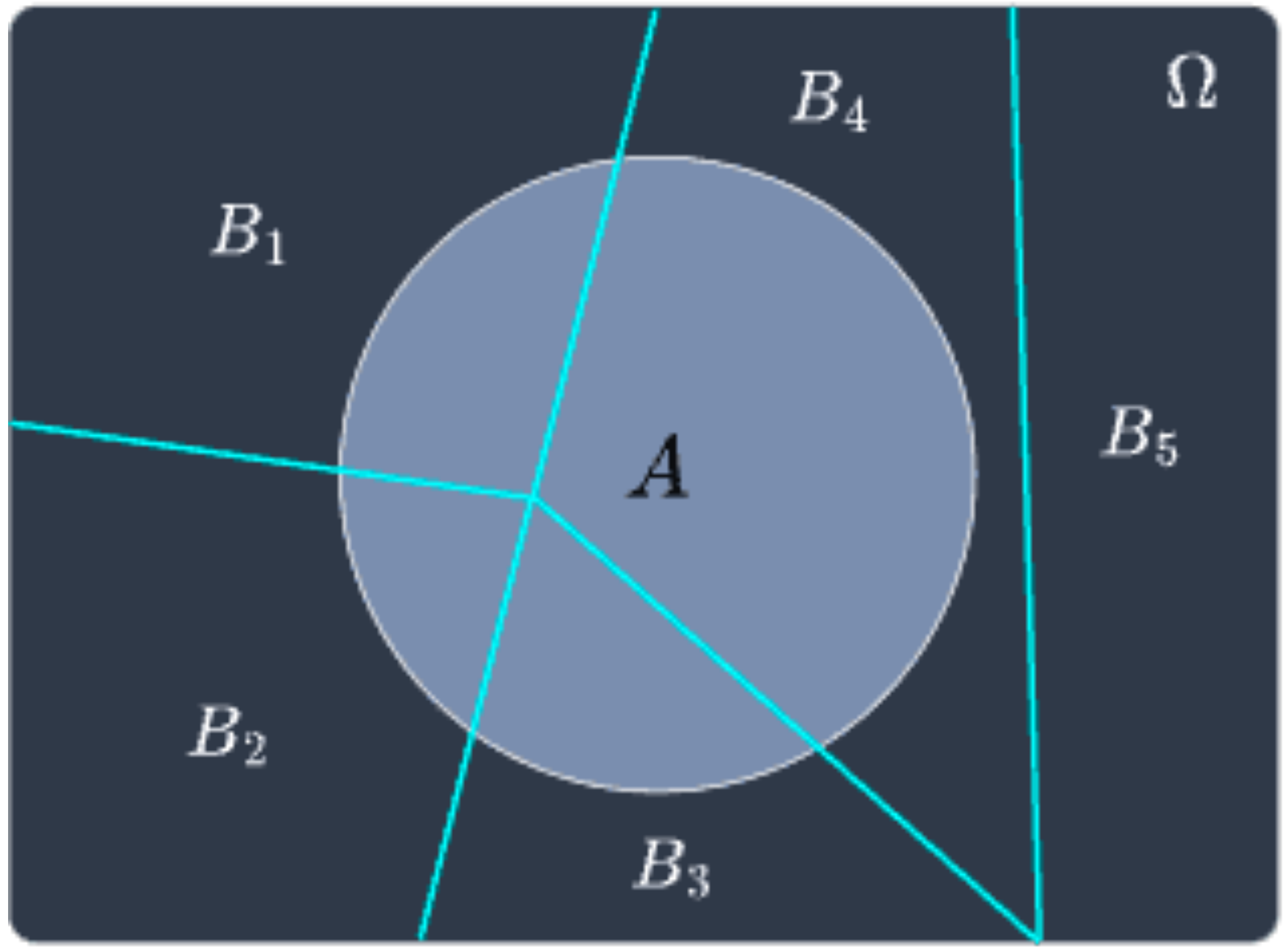
- Exercise:
  - Suppose that two factories supply light bulbs to the market. Factory X's bulbs work for over 5000 hours in 99% of cases, whereas factory Y's bulbs work for over 5000 hours in 95% of cases.
  - It is known that factory X supplies 60% of the total bulbs available and Y supplies 40% of the total bulbs available. What is the chance that a purchased bulb will work for longer than 5000 hours?
  - $P(A) = \frac{99}{100} \cdot \frac{6}{10} + \frac{95}{100} \cdot \frac{4}{10} = \frac{594 + 380}{1000} = \frac{974}{1000}$



Laws of total probability

$$P(A) = \sum_{i=1}^{L} P(A \mid B_i)P(B_i) = \sum_{i=1}^{L} P(A \cap B_i)$$

- What is the sample space?
-



Independence

- Two events are independent of each other if
  - $P(A \cap B) = P(A)P(B)$
  - P of flip a coin twice, get tails both times = 1/2 x 1/2 = 1/4

Expectation

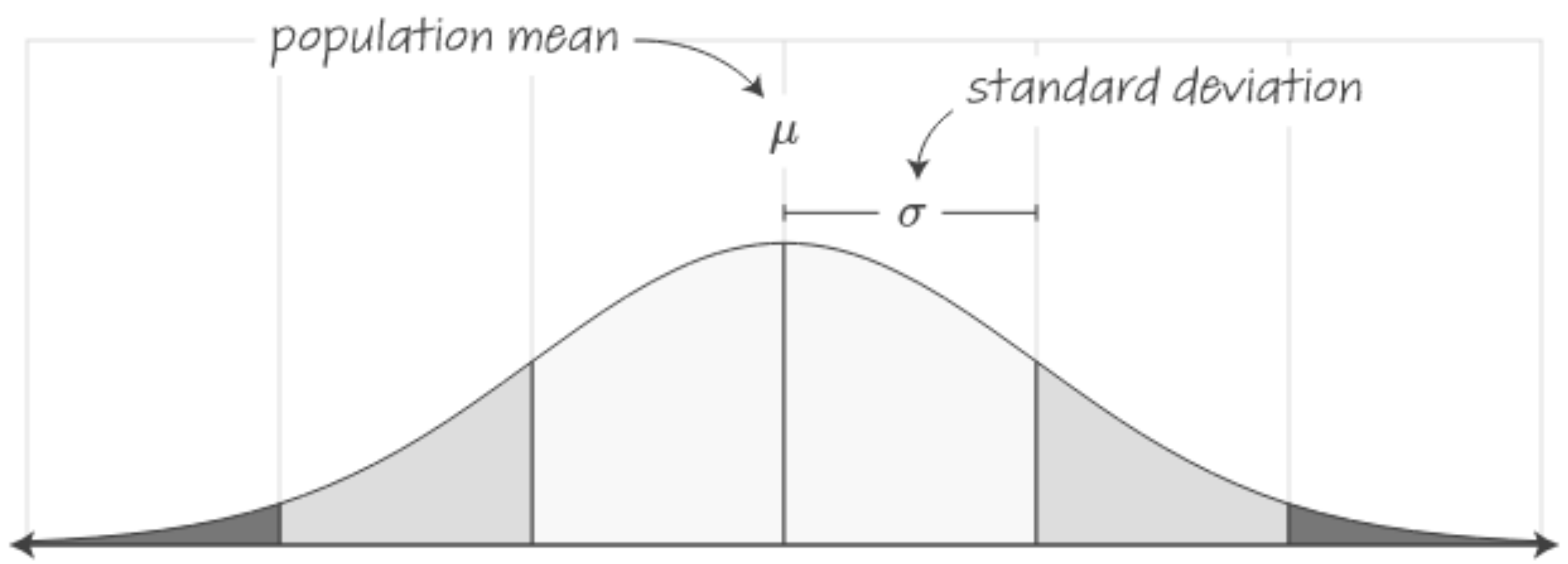
- Expected value is a *weighted average* with weights given by the probability distribution
- Informally, the expected value is the arithmetic mean of a large number of independently selected outcomes of a random variable.
- Can be defined for any function f of the random variable x
- Discrete random variable: $\mathbb{E}[f(x)] = \sum_x f(x)p(x)$
- Continuous random variable: $\mathbb{E}[f(x)] = \int f(x)p(x)dx$

Expectation

- **Moments**: expectations of special functions of the random variable x
  - <u>First moment</u> (mean) μ:
  - $\mathbb{E}[x] = \int xp(x)dx$ or $\mathbb{E}[x] = \sum_x xp(x)$
  - <u>Second</u> *centered* moment (variance) σ^2:
  - $\mathbb{E}[(x-\mu)^2] = \int (x-\mu)^2 p(x)dx$ or $\mathbb{E}[(x-\mu)^2] = \sum_x (x-\mu)^2 p(x)$
- Property:

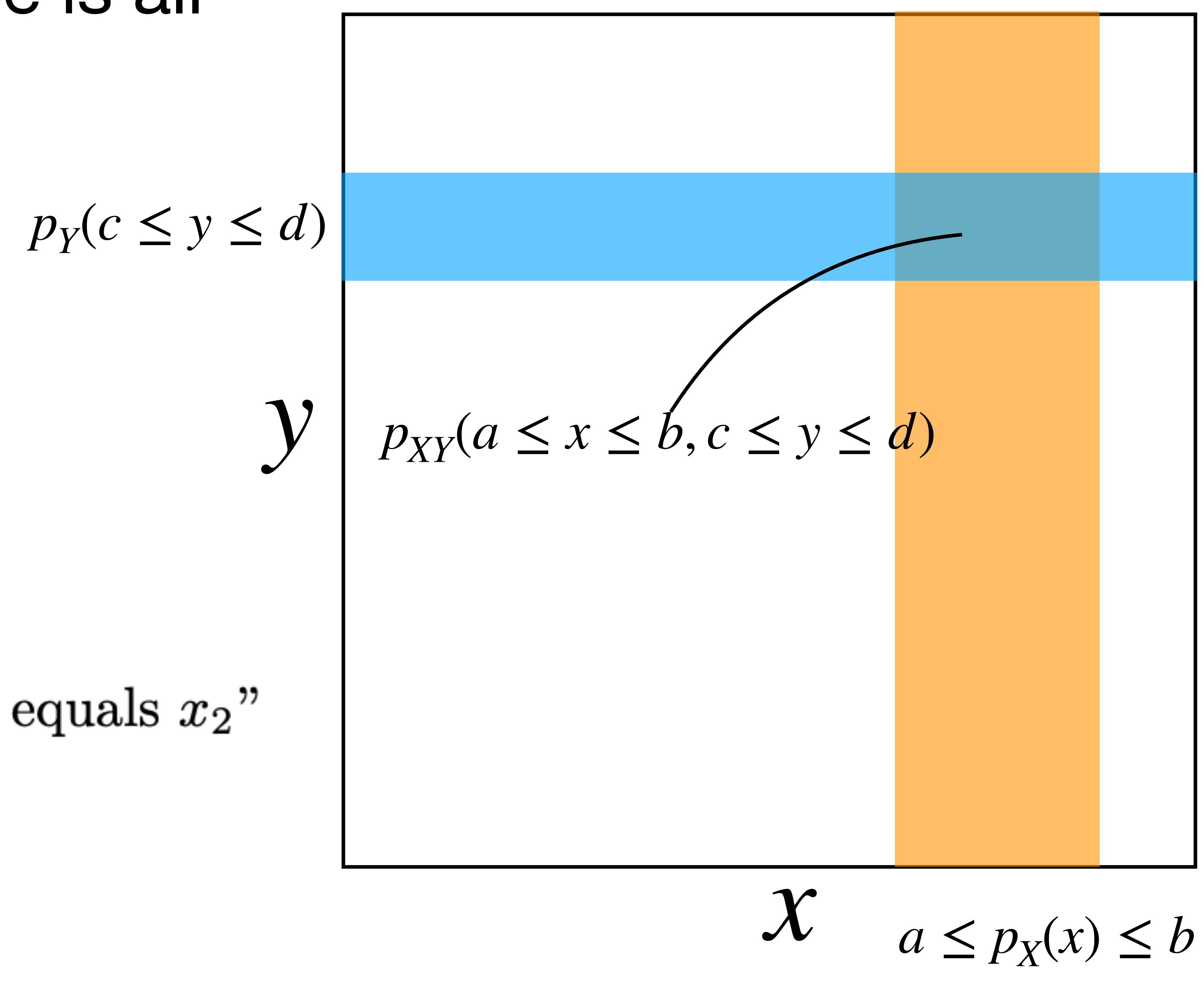
$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

$$\mathbb{E}[(x-\mu)^2] = \mathbb{E}[x^2] + \mathbb{E}[x]^2$$



Multivariate distributions

- Given two random variables *X* and *Y*, the sample space is all ordered pairs *(x, y)*
- Probability distribution
  - Assume independence: $p(x,y) = p(x)p(y)$
  - Non-independence: $p(x,y) = p_{XY}(x,y)$
- Discrete joint distribution:
  - $p_{i,j} = \mathbb{P}[X = x_i, Y = x_j]$ " the probability that X equals x_1 and Y equals x_2"
- Continuous joint distribution:
- $$\mathbb{P}[a \leq X \leq b, c \leq Y \leq d] = \int_a^b \int_c^d p(x,y)\,dx\,dy.$$



Multivariate distributions

- Examples:
  - the outcome of flipping three dice (X_1, X_2, X_3)
  - the body mass of a parent and a child $(M_{\text{parent}}, M_{\text{child}})$
  - If a patient is a smoker, and if they have cancer $(I_{\text{smoker}}, I_{\text{cancer}})$
  - An individual's two homologous alleles at a locus $(A_{\text{mom}}, A_{\text{dad}})$
- Which examples are (likely) independent?

Multivariate distributions

- Expectation
  - Discrete: $\mathbb{E}[f(X, Y)] = \sum_{x,y} f(x, y)p(x, y)$
  - Continuous: $\mathbb{E}[f(X, Y)] = \int\int f(x, y)p(x, y)dxdy$
- Covariance
  - $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \sum_{x,y} (x - \mu_X)(y - \mu_Y)p(x, y)$
  - $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \int\int (x - \mu_X)(y - \mu_Y)p(x, y)dxdy$
- Variance is the covariance of X with itself: $\mathbb{E}[(X - \mu_X)(X - \mu_X)] = \sum_{x,y} (x - \mu_X)^2 p(x)$

Some discrete probability distributions

- Bernoulli
  - Flipping a coin: $P(\text{heads}) = p, P(\text{tails}) = 1 - p$
- Binomial
  - Flipping n coins: $P(k \text{ heads}) = p^k(1-p)^{n-k}\binom{n}{k}$
- Geometric
  - Number of tosses until the coin lands on heads
  - $P(X = k) = (1-p)^{k-1}p$
- Poisson
  - $P(X = k) = \dfrac{\lambda^k \exp^{-\lambda}}{k!}$
  - Example: number of mutations in time interval $[0,T]$
  - *Limit of binomial as* $n \to \infty$, $p \to 0$

Poisson distribution as a limit of binomial

- https://en.wikipedia.org/wiki/Poisson_limit_theorem

Continuous probability distributions

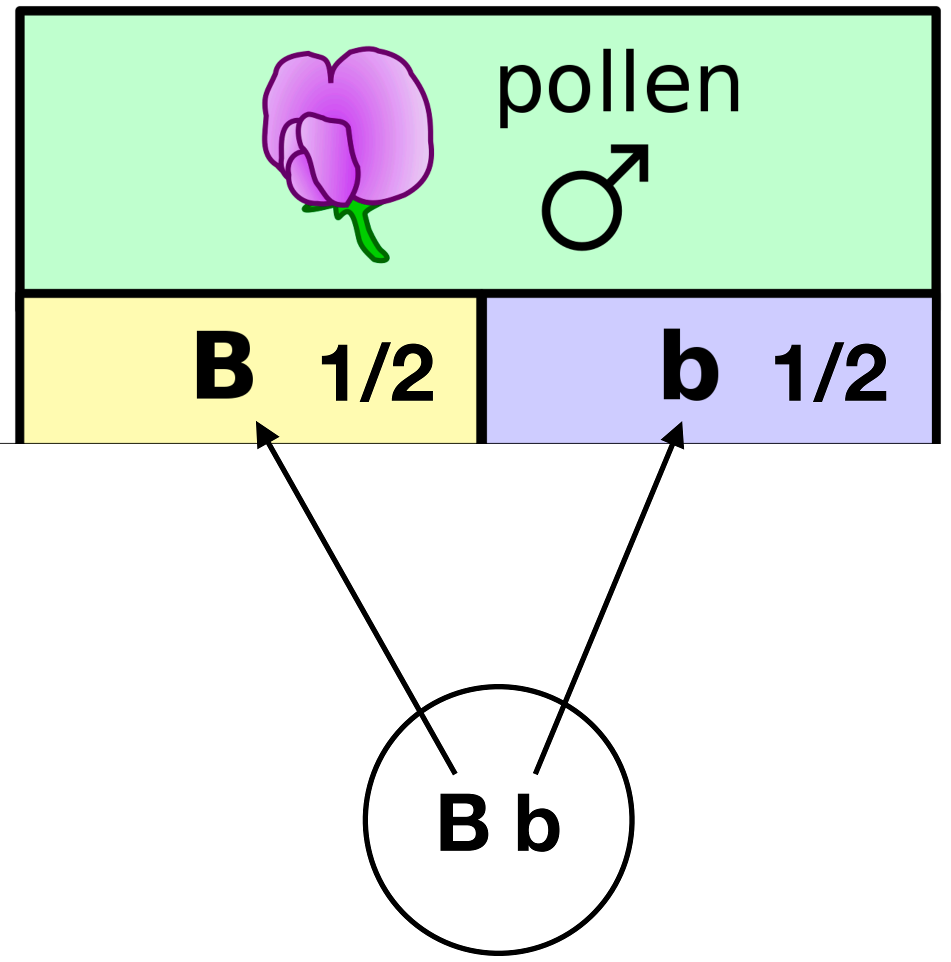
- Uniform:
  - $f(x) = \dfrac{1}{b-a}$
- Normal
- $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
- Exponential
  - $f(x) = \lambda \exp^{-\lambda x}$

Exponential distribution

- Dividing the interval $[0,t]$ in to n small intervals, each with width $\delta = 1/n$
- Assume an outcome (e.g. a light bulb burns out) happens in a small interval at rate $= \lambda \times 1/n$
- Probability that the outcome doesn't happen between 0 and t $= (1 - \frac{\lambda}{n})^{n \times t}$
- Use the fact that $(1 + \frac{x}{n})^n$ approximates e^x as x goes to zero
- $P(\text{outcome does not happen}) \rightarrow e^{-\lambda t}$
- $P(T < t) = 1 - e^{-\lambda t}$, Cumulative distribution function (CDF)
- $f(t) = \lambda e^{-\lambda t}$, Density function (PDF)

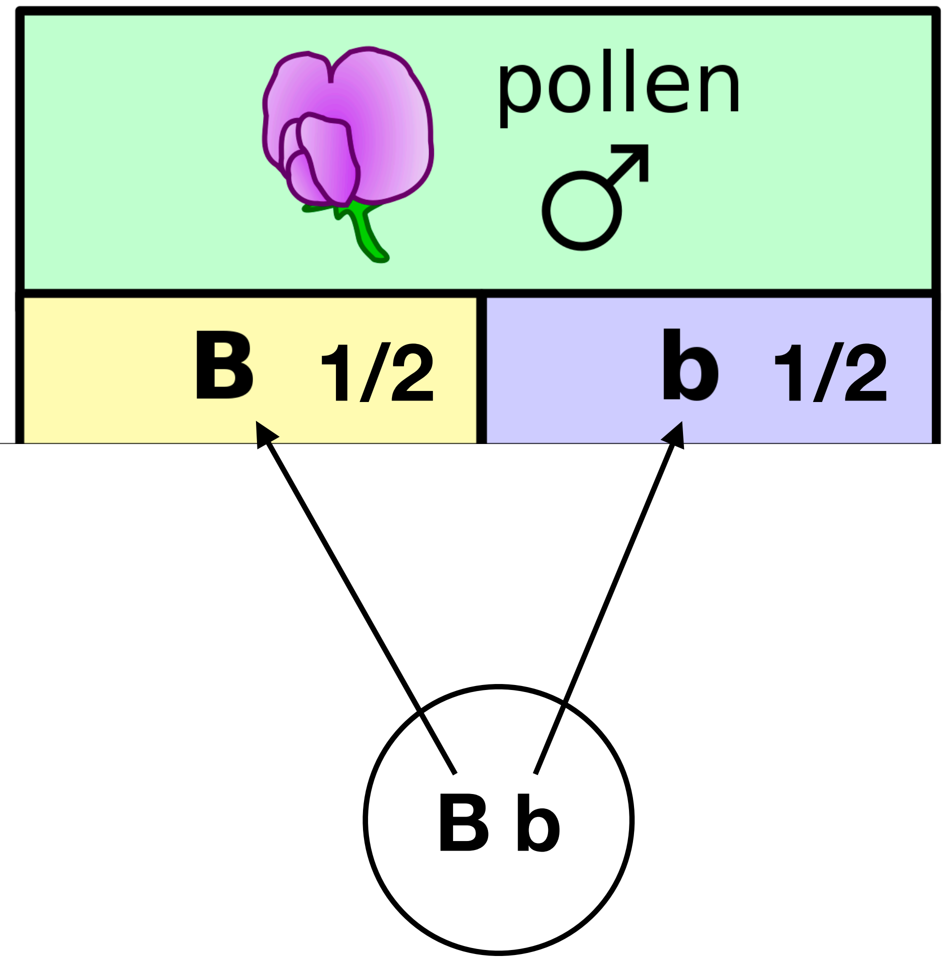
Mendelian inheritance and probability

- Law of segregation
  - During gamete formation, the alleles for each gene segregate from each other so that each gamete carries only one allele for each gene.



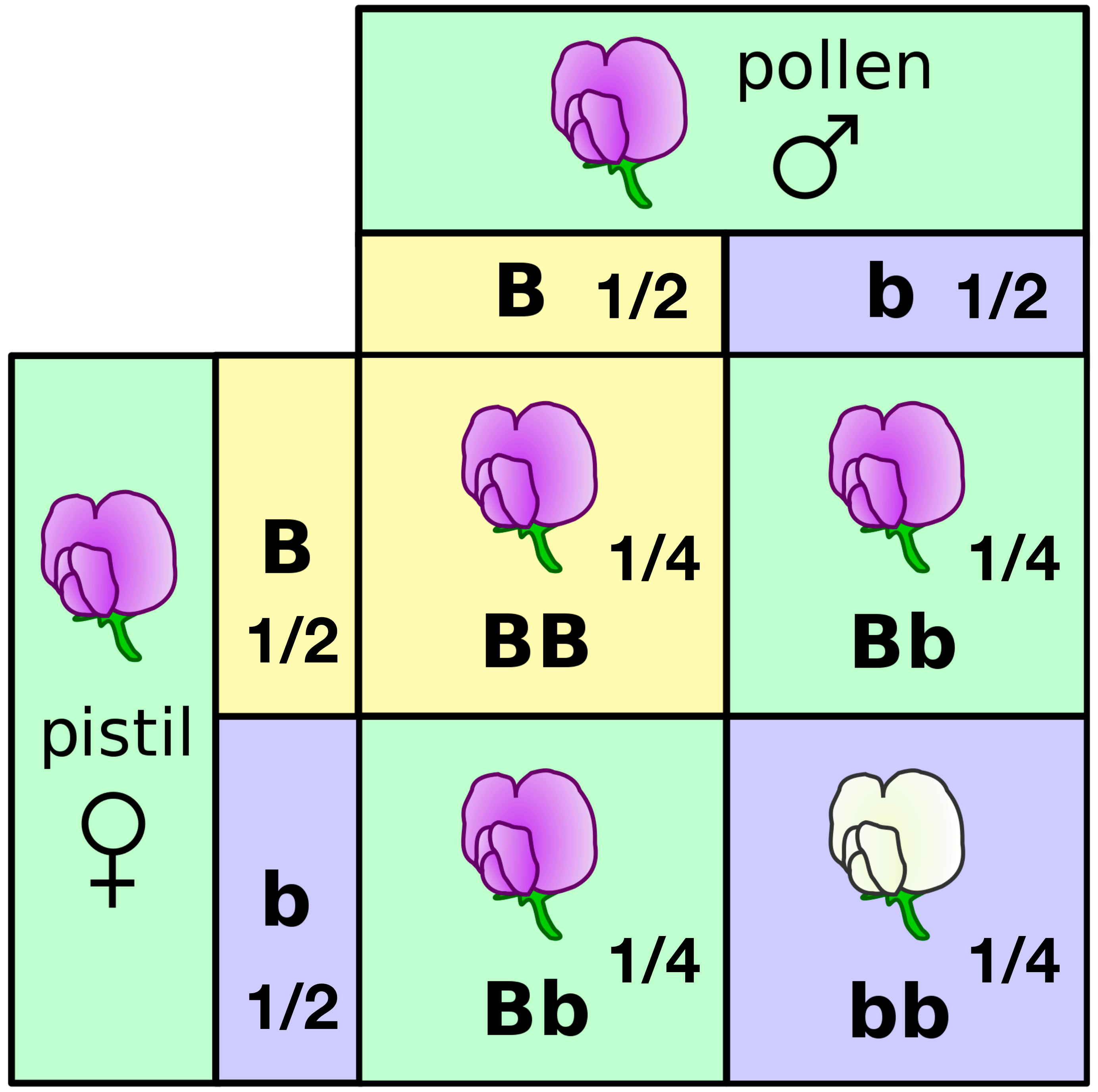
Mendelian inheritance and probability

- Law of segregation
  - During gamete formation, the alleles for each gene segregate from each other so that each gamete carries only one allele for each gene.
  - Sample space: {B, b}
  - P(B) = P(b) = 1/2



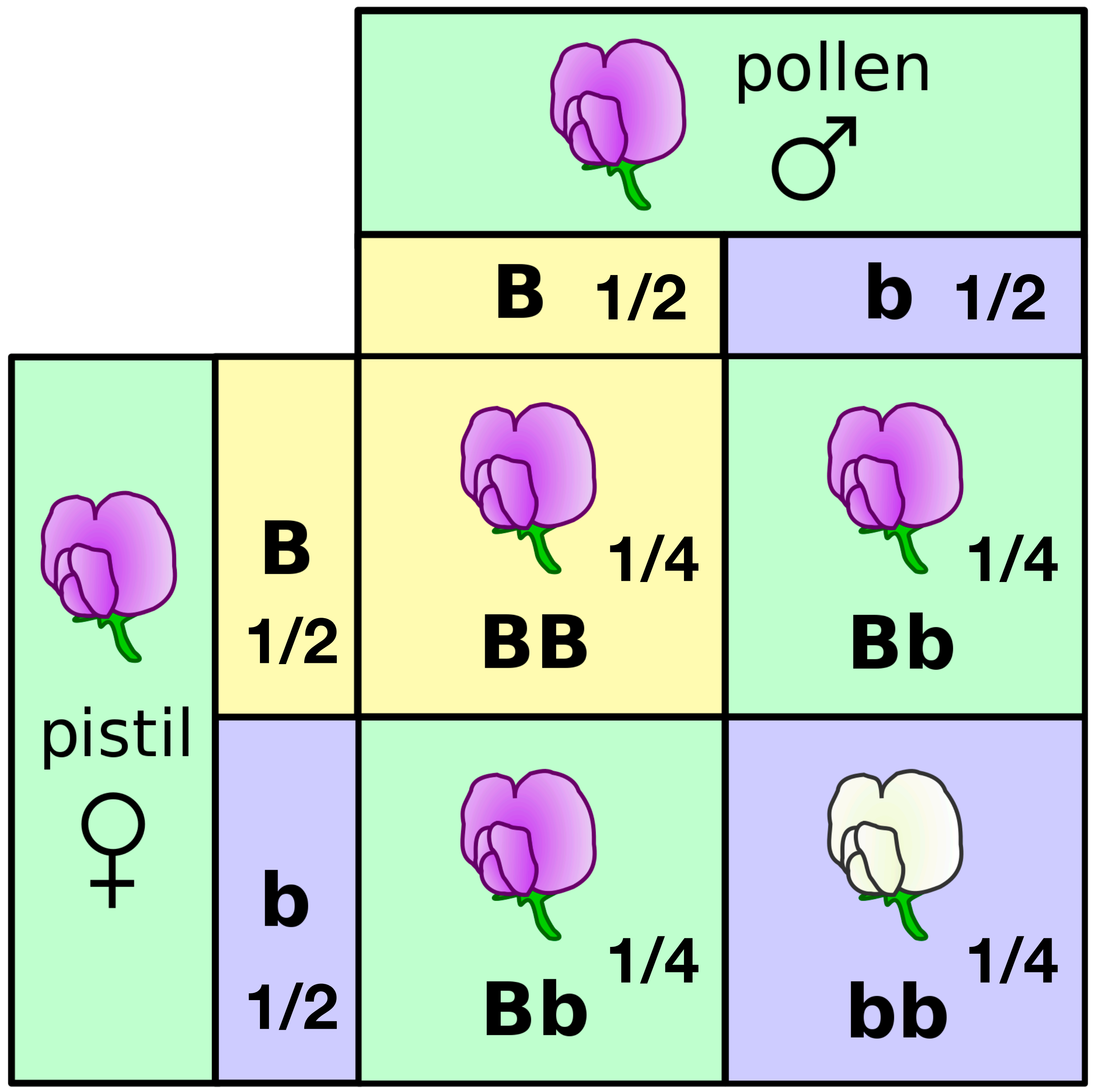
Mendelian inheritance and probability

- Law of segregation and mating
  - Sample space: {(B, B), (B, b), (b, B), (b, b)}
  - P(gamete♀, gamete♂) = P(gamete♀) P(gamete♂) = 1/4

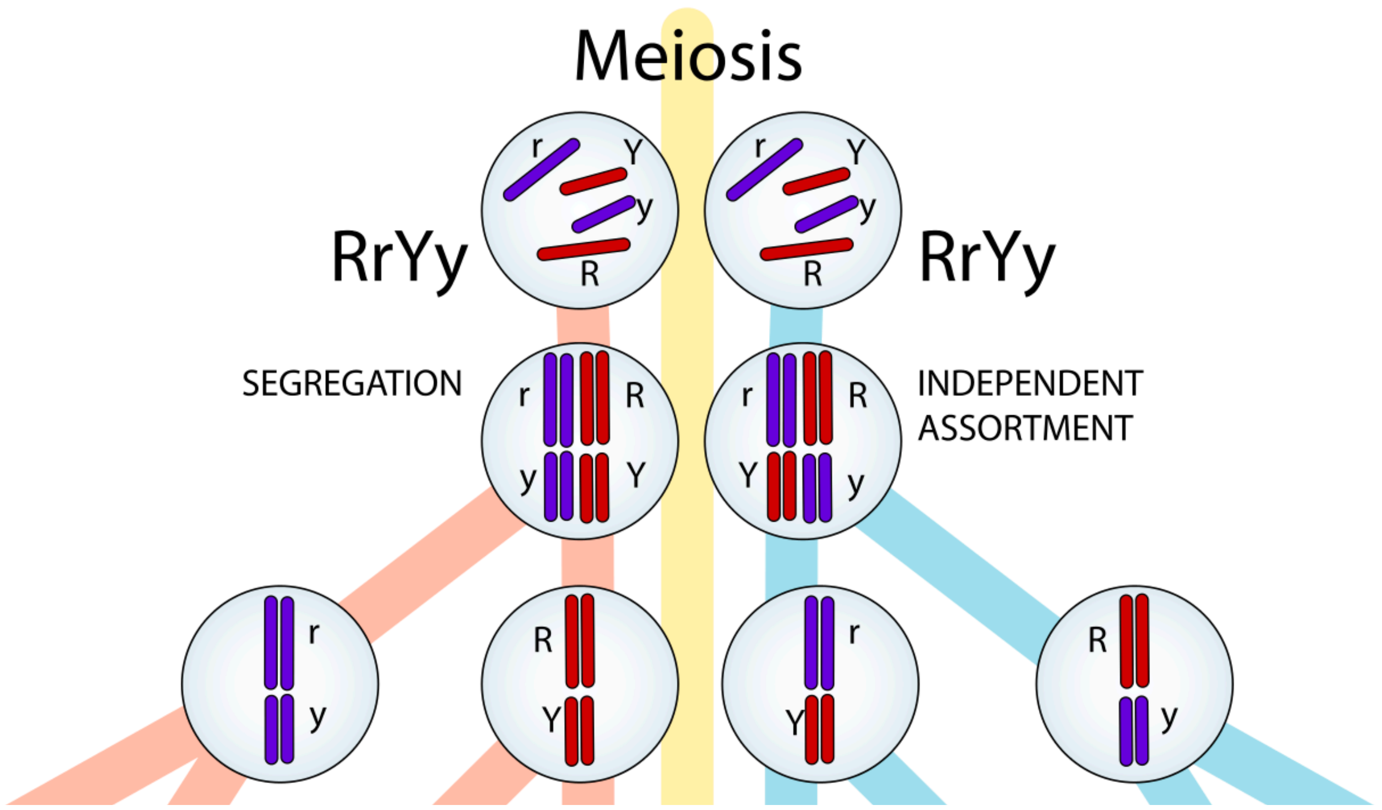


Mendelian inheritance and probability

- Law of dominance and uniformity
  - Some alleles are dominant while others are recessive; an organism with at least one dominant allele will display the effect of the dominant allele.
  - Event: “a seedling has purple flower” = {BB, Bb, bB}
  - P “a seedling has purple flower” = 3/4
  - Conditional probability
    - P(homozygous for B | purple flower)
      = P(BB)/P({BB, Bb, bB} = 1/3



Mendelian inheritance and probability

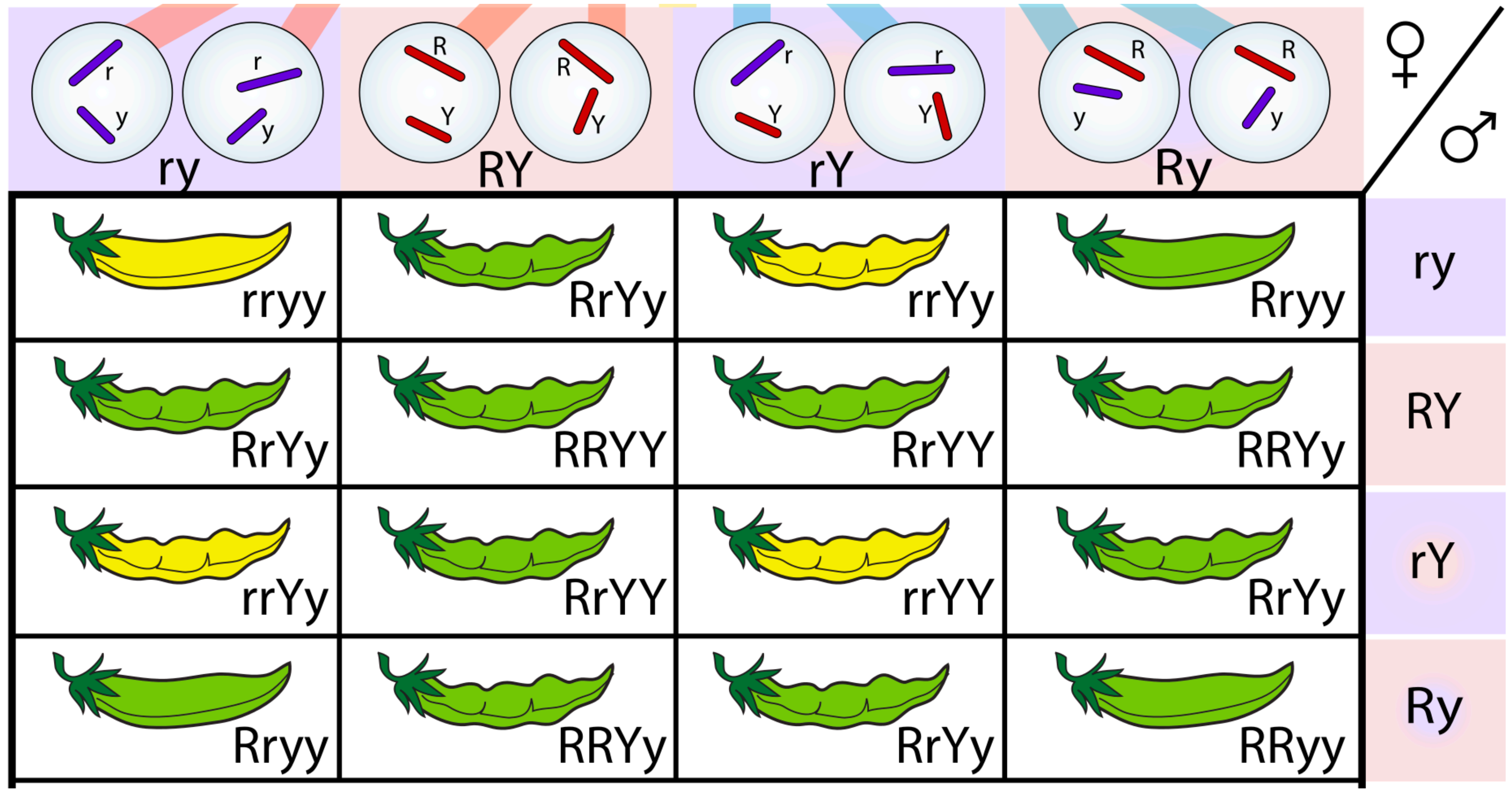


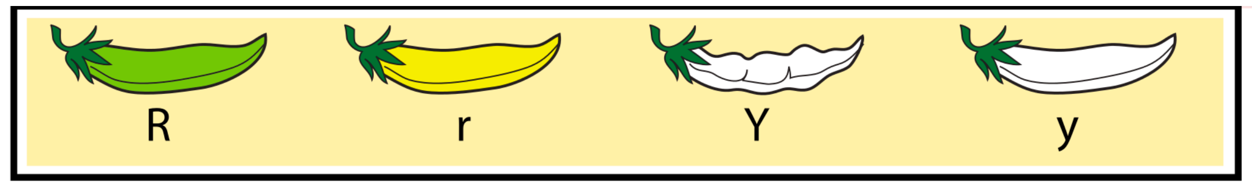
- Law of independent assortment
  - Genes of different traits can segregate independently during the formation of gametes.
  - Locus 1 {r, R}; Locus 2{y, Y}
  - Sample space: {ry,Ry, rY, RY}
  - P(allele at locus 1, allele at locus 2)
    =P(allele at locus1)*P(allele at locus2)
    =1/4

  - *Rule is violated if two alleles are linked*

Mendelian inheritance and probability

- Two traits and mating
  - Sample space
    = {ry,Ry, rY, RY} x {ry,Ry, rY, RY}
  - R is dominant over r
  - Y is dominant over y
  - Event A = “wrinkled pods”, P(A) = 3/4
  - Event B = “green pods”, P(B) = 3/4
  - Event C = “wrinkled, green pods” = P(A&B) = P(A) x P(B) = 9/16
  - Overall ratio is 9:3:3:1

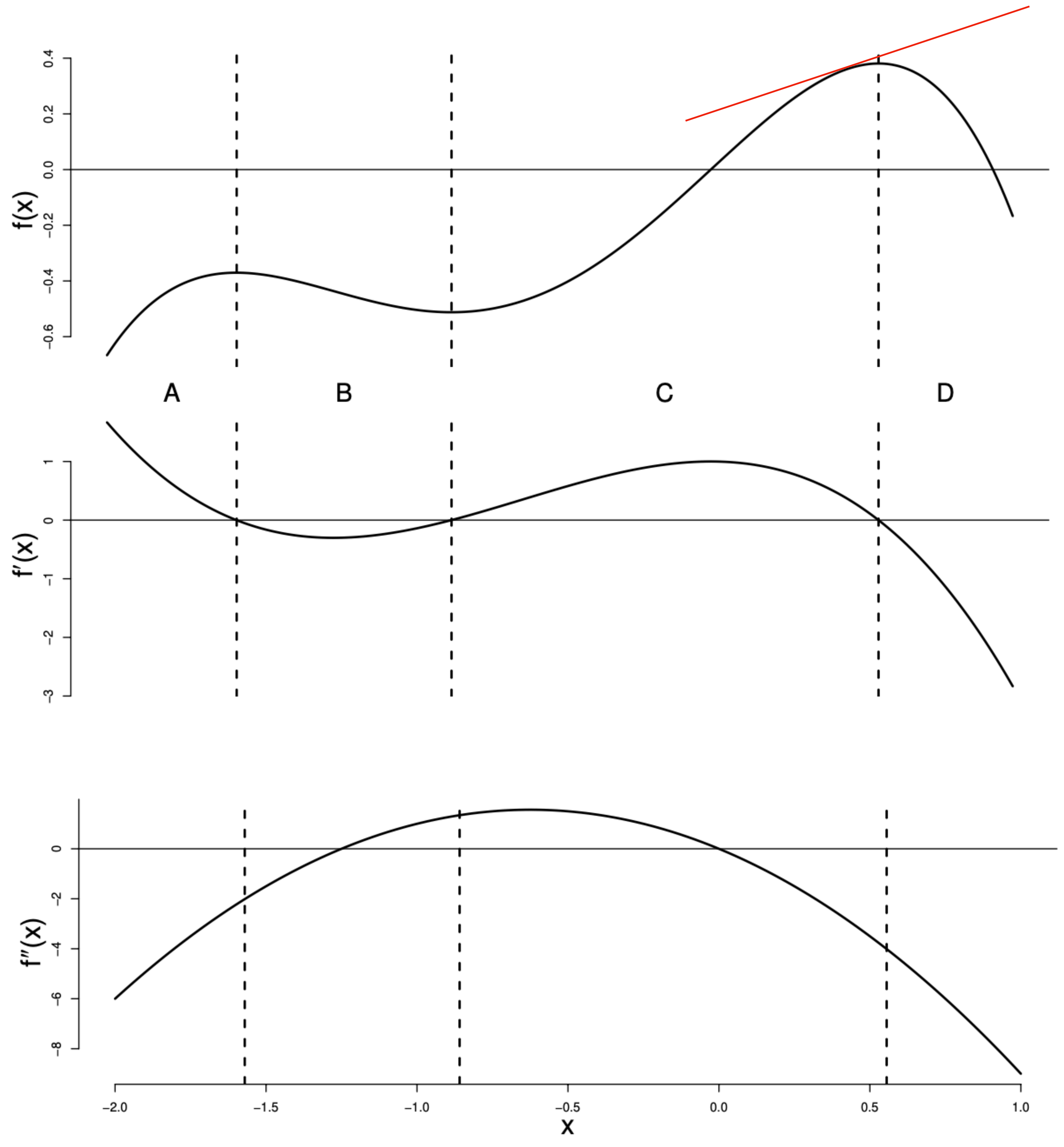




- Derivative
- Integral
- Taylor expansion
- Useful taylor expansion
- Common functions
  - Exponential
  - Polynomial
  - Logrithm

Part 2: Calculus

- <u>Derivative</u>
- The derivative of a function f at point x, is the best linear approximation of the function f centered at x
- $f'(x)$ or $\dfrac{df(x)}{dx}$
- $f(x) \approx f(a) + f'(a)(x - a)$

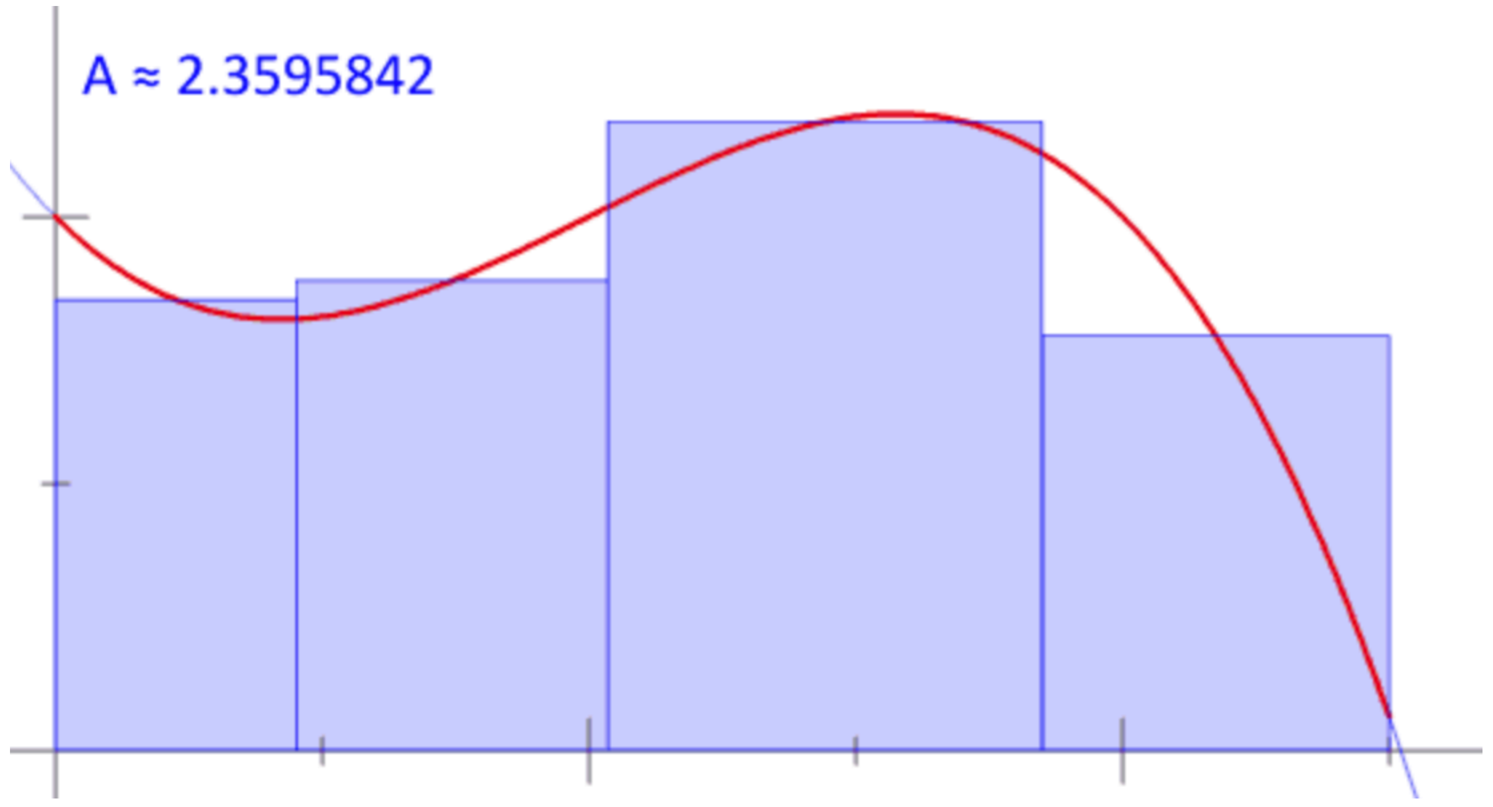


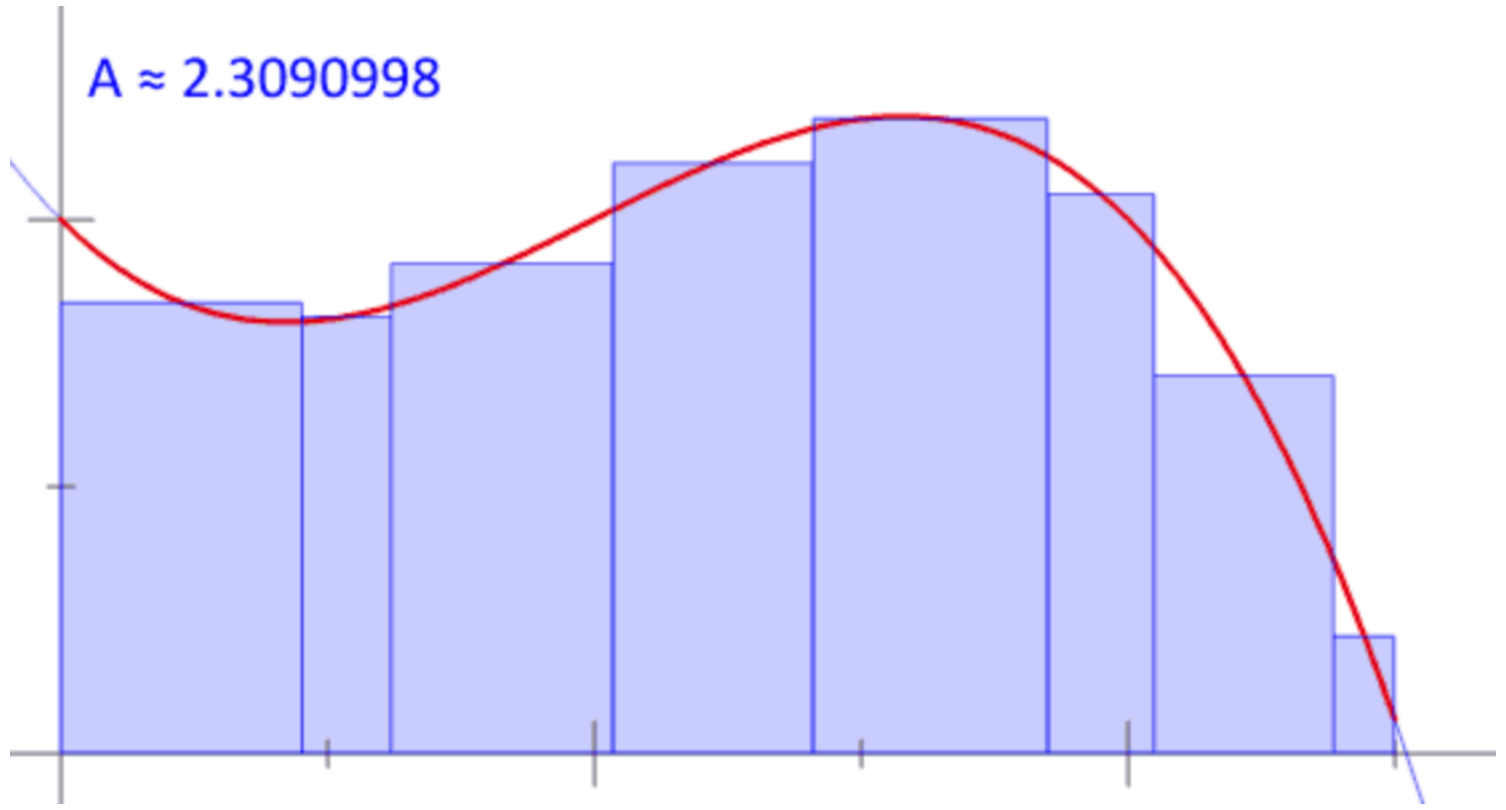
Part 2: Calculus

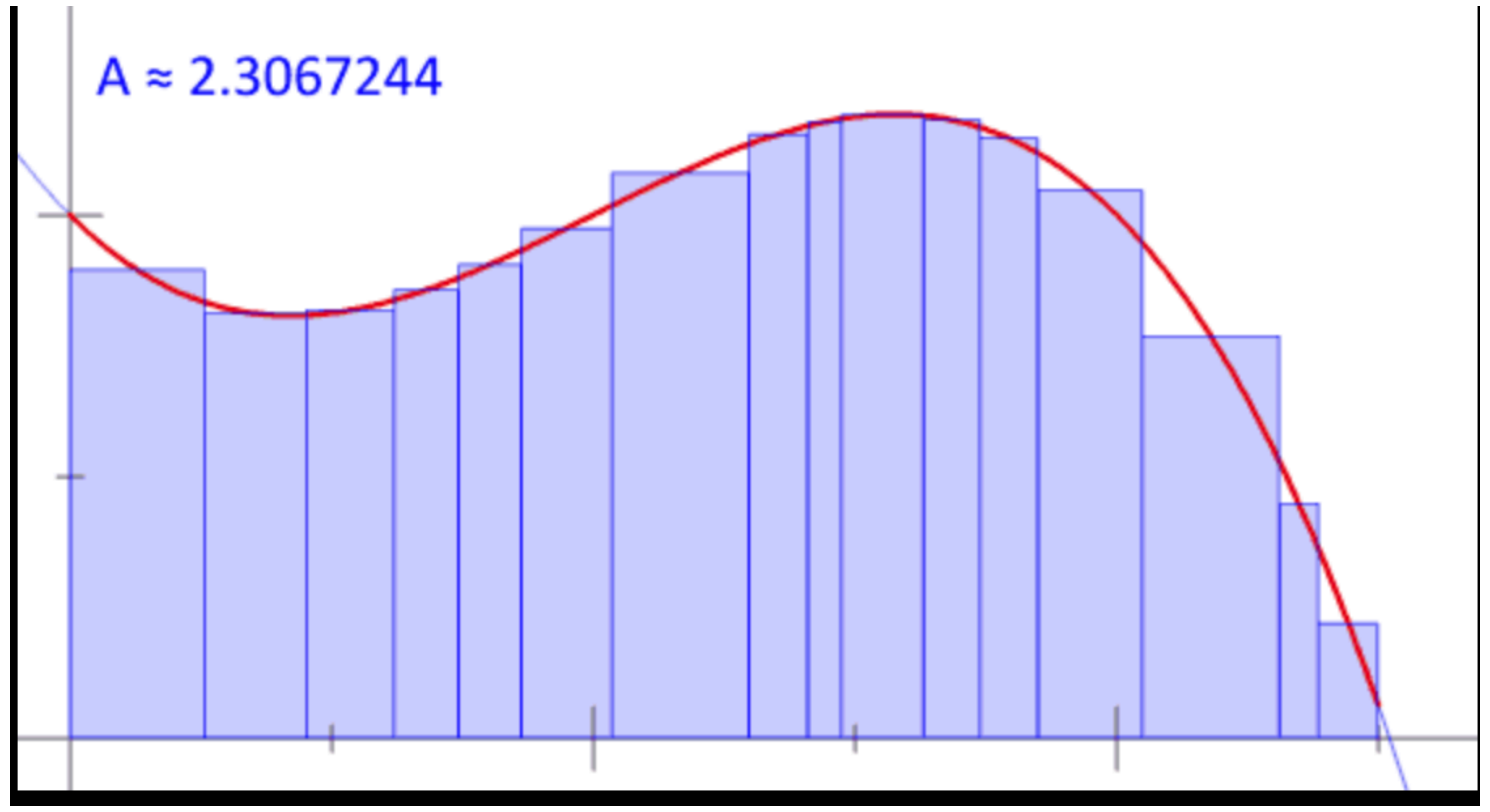
- Commonly used derivatives
  - $\dfrac{d}{dx}x^a = ax^{a-1}$.
  - $\dfrac{d}{dx}e^x = e^x$.
  - $\dfrac{d}{dx}\ln(x) = \dfrac{1}{x}, \qquad x > 0.$

Integral

- The integral of a function on the interval $[a, b]$ is the limit of the sum
- $$\sum_{i=0}^{n-1} f(t_i)\left(x_{i+1} - x_i\right).$$
- where $a = x_0 < x_1 < x_2 < \ldots < x_i < \ldots < x_n = b$
- as $n \to \infty$, that is, if we cut the interval $[a, b]$ into increasing smaller pieces
- The integral is written as $\int_a^b f(x)dx$

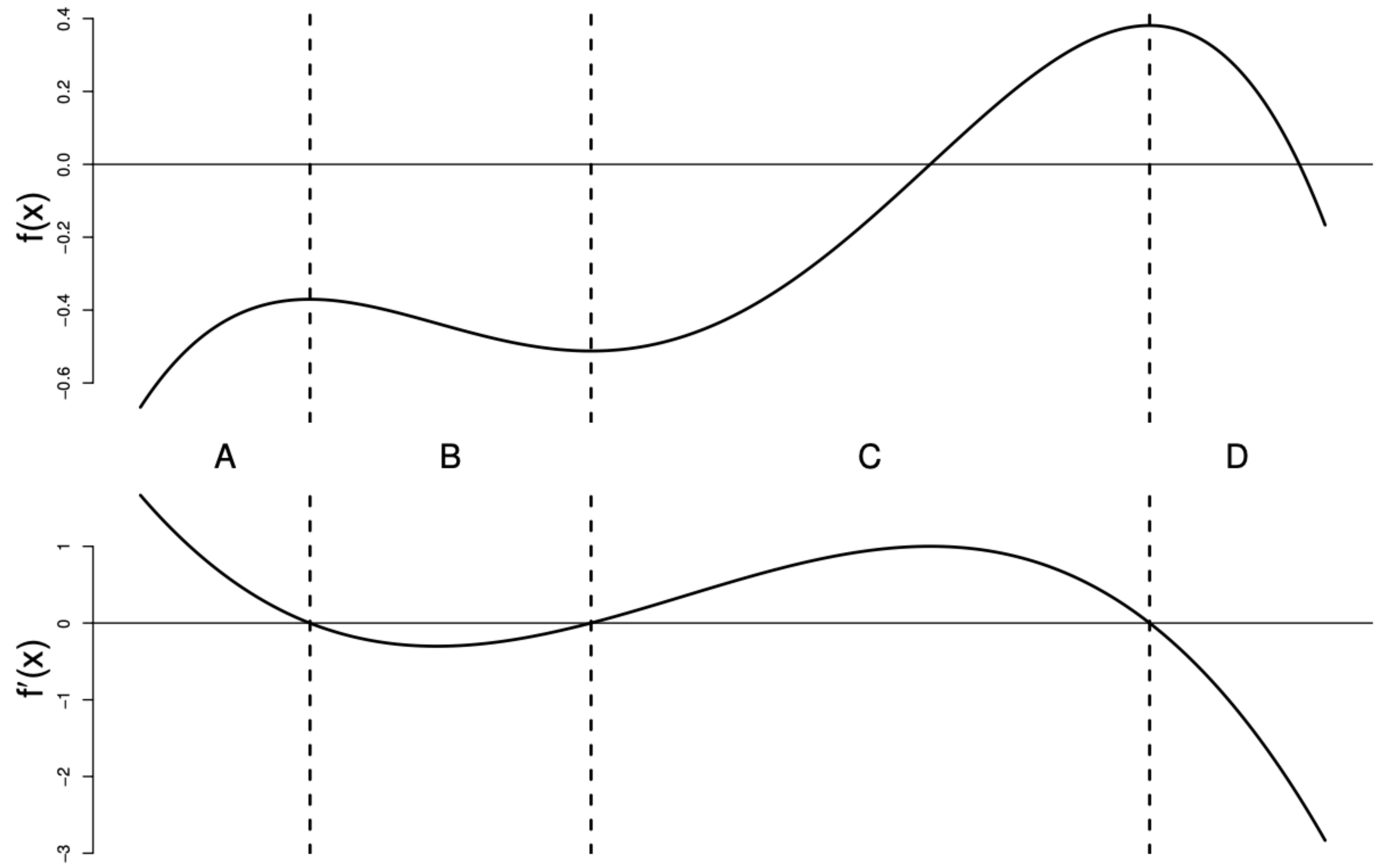






Fundamental theorem of calculus

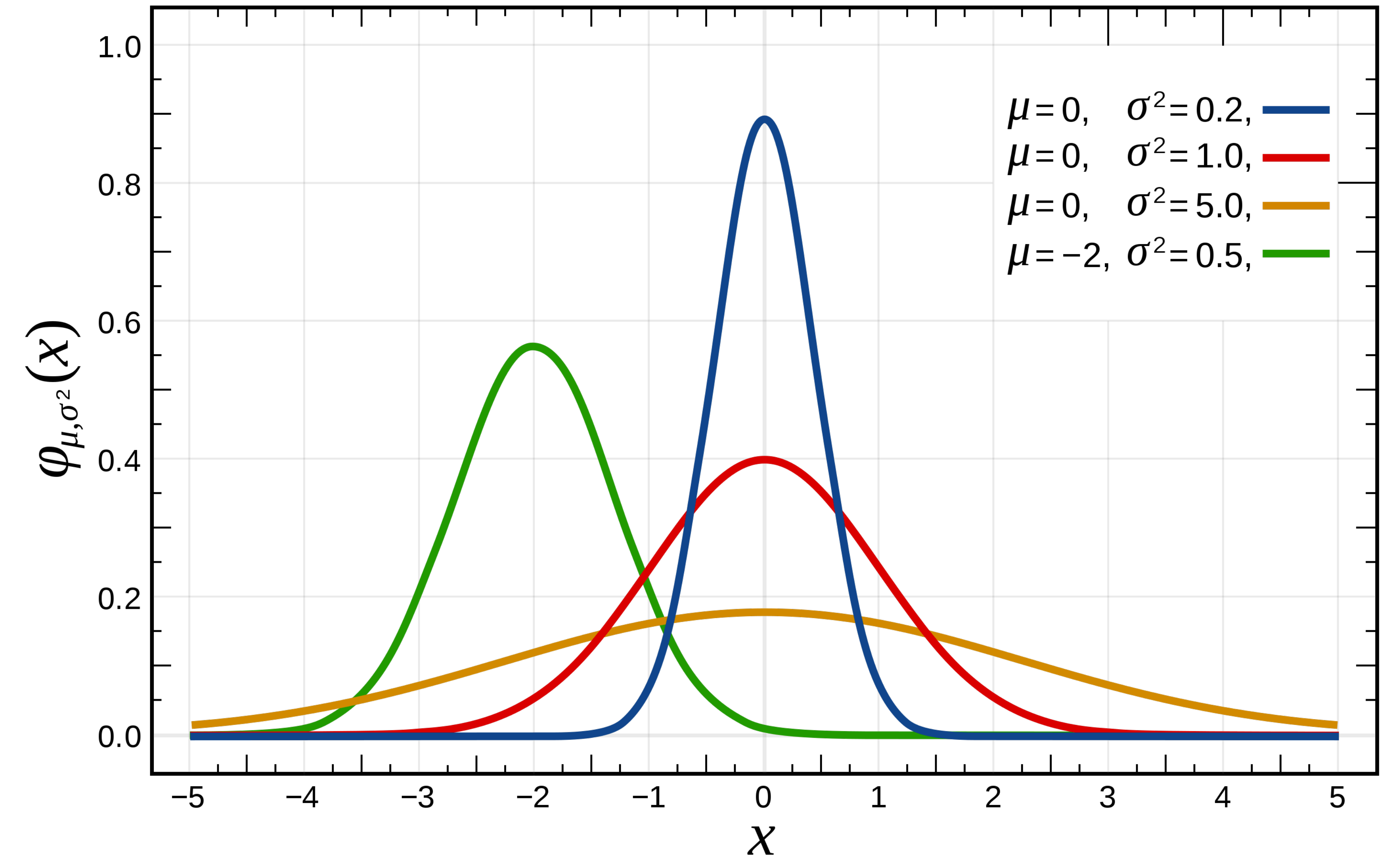
- Let F be a function $F(x) = \int_a^x f(t)\, dt$.
- Then $F'(x) = f(x)$
- $F(x)$ is the *antiderivative* of $f(x)$
- $\int_a^b f(x)\, dx = F(b) - F(a)$.

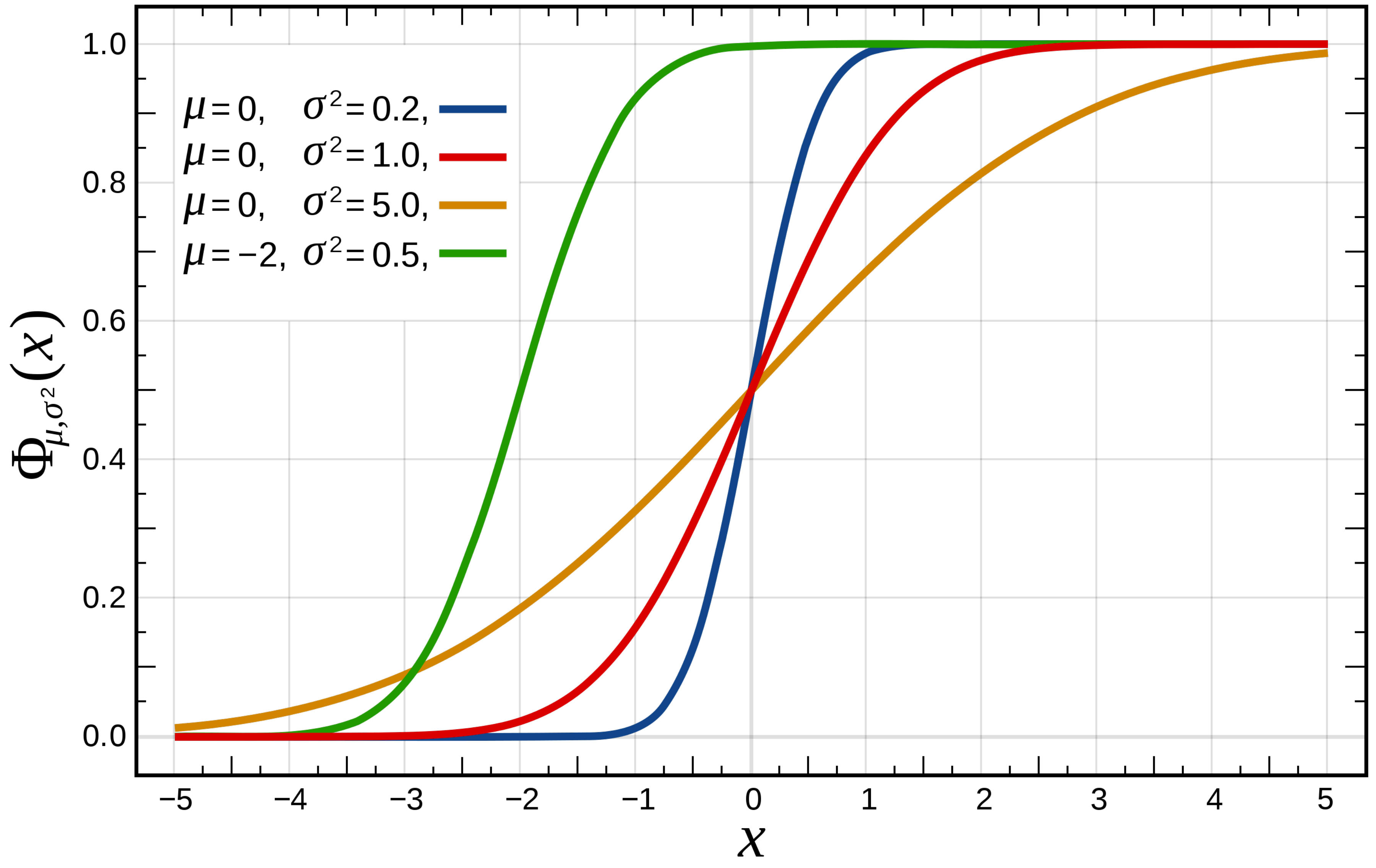


Cumulative and probability density function

- The cumulative distribution function
  - $F_X(x) = \mathrm{P}(X \leq x)$
- The probability density function
  - $f_x(t)$
- We have
  - $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt\,.$
  - $f(x) = \dfrac{dF(x)}{dx}$

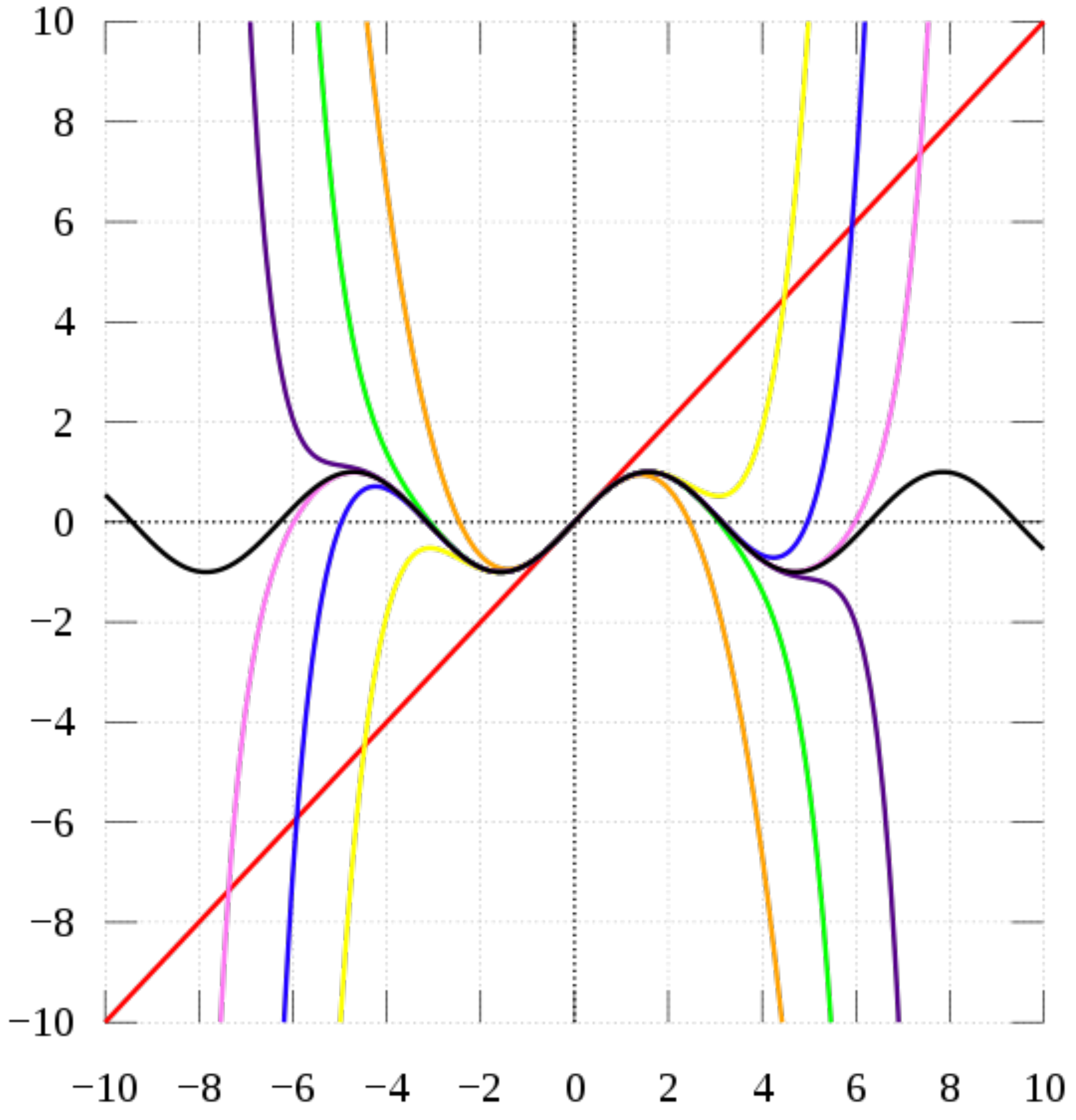
- $\mathrm{P}(a < X \leq b) = \int_a^b f_X(x)\, dx = F_X(b) - F_X(a)$





Approximating functions by Taylor Series

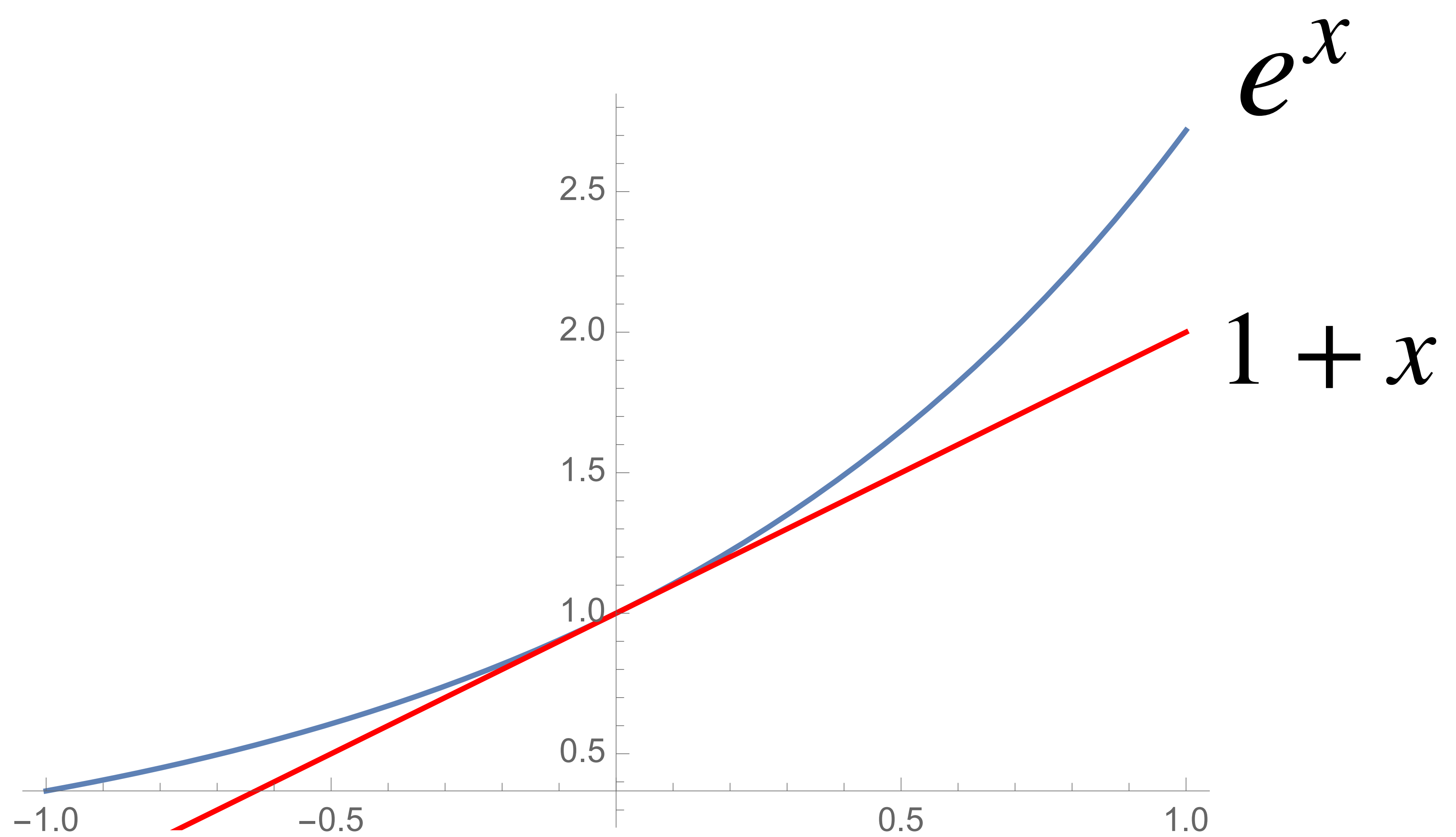
- Recall the derivative of a function provides the best linear approximation of a function at a point
- $f(x) \approx f(a) + f'(a)(x-a)$
- We can generalize this approximation using the Taylor expansion:
- $f(x) = f(a) + \dfrac{f'(a)}{1!}(x-a) + \dfrac{f''(a)}{2!}(x-a)^2 + \dfrac{f'''(a)}{3!}(x-a)^3 + \cdots$
- Taylor polynomials are approximations of a function, which become generally better as n increases.

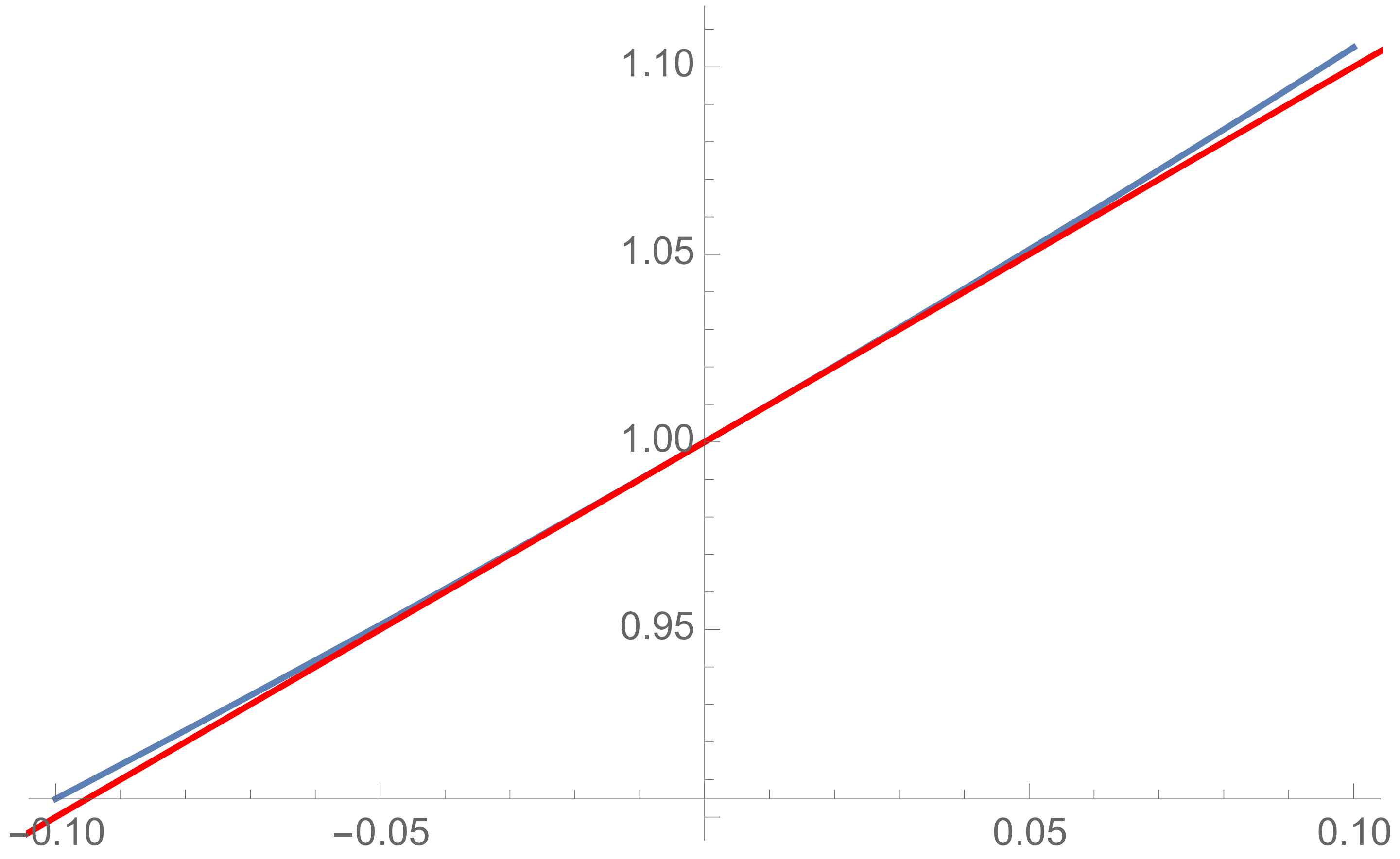


Common approximations

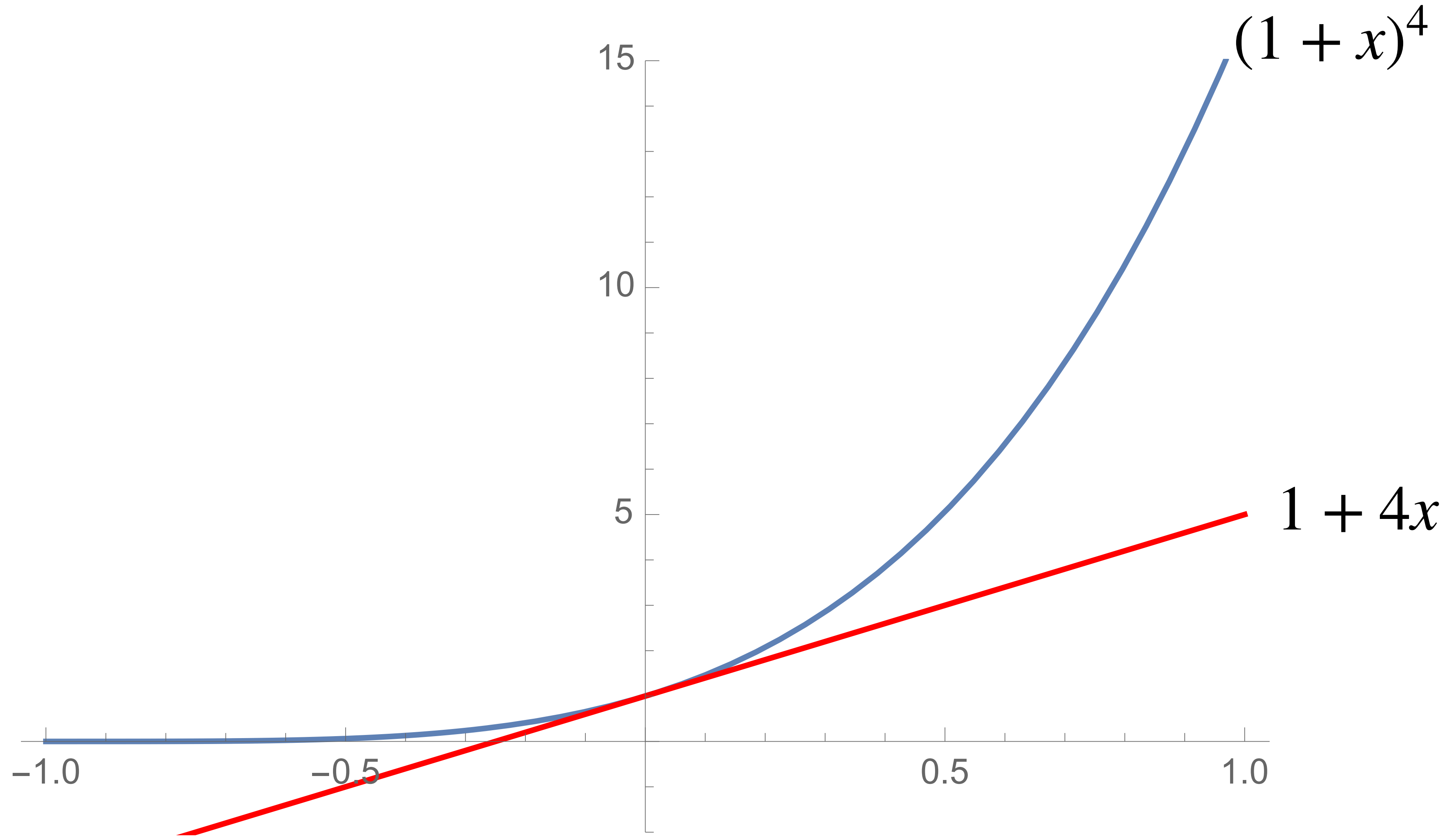
- $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$
- $e^x \approx 1 + x$ for x near 0
- $\dfrac{1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots}{1 + x} = 1 + \dfrac{\frac{x}{2!} + \frac{x^2}{3!} + \cdots}{1 + \frac{1}{x}} \to 1$

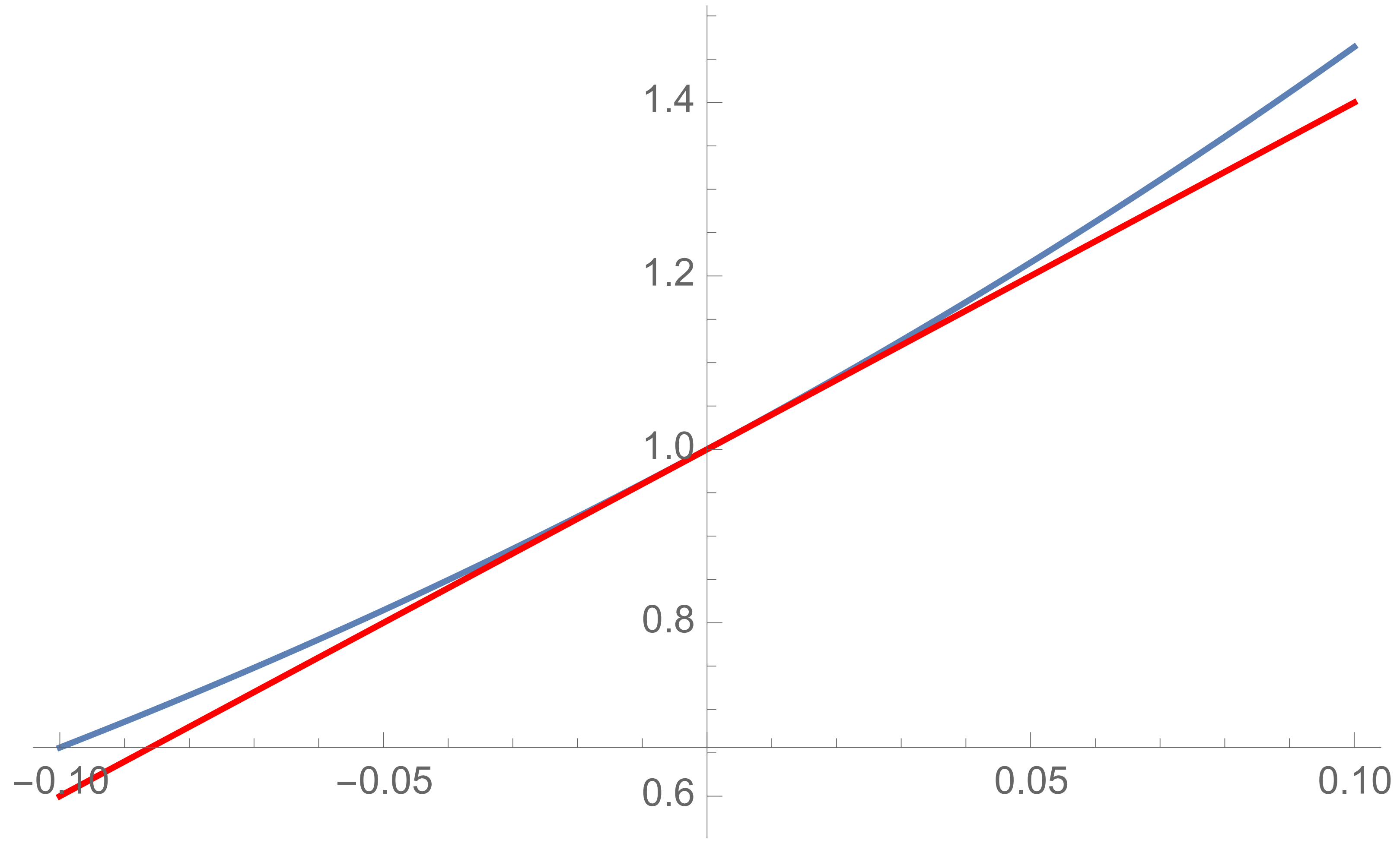
  as $x \to 0$





Common approximations





- $(1+x)^k$
  - $\dfrac{d(1+x)^k}{dx} = k(1+x)^{k-1}$
  - $\dfrac{d(1+x)^k}{dx} = k$, at $x = 0$
  - $(1+x)^k \approx 1 + kx$ near 0

Part3: Basic combinatorics

- Factorial $n! = n \times (n-1) \times (n-2) \times \cdots \times 1$
  - Number of ways to sort n objects
- Binomial coefficient $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$
  - Number of ways to pick k out of n objects
- $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$

Part 4: the exponential function

- $\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$
- $\frac{de^x}{dx} = e^x$
- Connection to e^x
  - $\exp(x)\exp(y) = \exp(xy)$
  - $\exp(n) = \left(\exp(1)\right)^n$
  - $\exp(1) = \sum_{k=0}^{\infty} \frac{1}{k!} = 1 + 1 + \frac{1}{2} + \frac{1}{3!} + \frac{1}{4!} + \cdots = e$
  - $\exp(n) = e^n$
  - $\exp(x) = e^x$

Part 4: the exponential function

- $a^x = (e^{\ln a})^x = e^{x \ln a}$
- Set $y = a^x$, then
- If $\log_a y = x$, then $\ln y = x \ln a$
- Change of base
  - $\log_a y = \dfrac{\ln y}{\ln a}$

Ways the exponential function arises

- Exponential growth
  - $x(t)$ is a function that describes the size of a population at time t
  - Growth rate is equal to population size (everybody in the population makes one offspring)
  - $x'(t) = x(t)$
  - $x(t) = ae^t$
  - $x(0) = a$

Ways the exponential function arises

- Compound interest
  - You have y amount of money in the bank
  - The annual interest rate is r
  - After a year you have $y(1 + r)$
  - Let's say that we divide the whole year into n small intervals of equal sizes, e.g. months, days, minutes, etc. And the interest per unit time is $\frac{r}{n}$
  - So the money you get at the end of year is
    - $y\left(1 + \frac{r}{n}\right)^n$
  - **What happens if $n \to \infty$?**

Ways the exponential function arise

- So the money you get at the end of year is
  - $y\left(1+\frac{r}{n}\right)^n$
- **What happens if $n \to \infty$?**
- $\lim_{n\to\infty}\left(1+\frac{r}{n}\right)^n$
- $\left(1+\frac{r}{n}\right)^n = \sum_{k=0}^{n}\binom{n}{k}\frac{r^k}{n^k} = \sum_{k=0}^{n}\frac{n!}{(n-k)!k!}\frac{r^k}{n^k}$
- **This tends to** $\sum_{k=0}^{\infty}\frac{r^k}{k!} = e^r$ **as** $n \to \infty$