Chapter 5: Distributions

DSCC 462
Computational Introduction to Statistics

Anson Kahng
Fall 2022

Random Variables

- **Random variable**: A variable that can take a number of different values and whose outcome is determined by chance
- **Discrete random variable**: A random variable whose possible outcomes are a list of discrete values (finite or countably infinite sample space)
  - Example: Coin flip (heads/tails)
- **Continuous random variable**: A random variable whose possible outcomes are any value in an interval (uncountable sample space)
  - Examples: Time required to run a mile

Notation

- Random variable: Uppercase letters (e.g., X, Y)
- Outcome of a random variable: Lowercase letters (e.g., x, y)
- Example: Let $X =$ the number of surgeries a person has had
  - $\Pr(X = 1)$: Probability of having 1 surgery
  - $\Pr(X = x)$: Probability of having x surgeries

Probability Distribution

- **Probability distribution**: List of all possible values that a random variable can take, along with their corresponding probabilities
    - Discrete: Probability mass function (PMF)
    - Continuous: Probability density function (PDF)
- Let X be a random variable defined over sample space S_X
- For any $E \subseteq S_X$, we can define $p_X(E) = \Pr(X \in E)$

Discrete Probability Distribution

- For a discrete random variable X with sample space S_X, a probability mass function (PMF) p_X maps $x \in S_X$ to a number in $[0,1]$ such that:

$$0 \leq p_X(s) = \Pr(X = x) \leq 1$$

$$\sum_{x \in S_X} p_X(x) = \sum_{x \in S_X} \Pr(X = x) = 1$$

- The support S_X consists of all x for which $p_X(x) > 0$ (all achievable outcomes)

Discrete Probability Distribution: Example

- Setup: A fair coin is flipped 3 times. Let X be a random variable that counts the number of heads observed
- Fill in the following table:
- Probability distribution tables resemble relative frequency distribution tables: probability of each outcome is the relative frequency distribution of each outcome in a large number of trials

x	Pr(X = x)
0	
1	
2	
3	

Continuous Probability Distribution

- Specify continuous probability distributions through a *density function*, $f(x)$
- Properties:

$$f(x) \geq 0, \text{ for all } x \in S_X \text{ (nonnegative density)}$$

$$\int f(x)\,dx = 1 \text{ (total probability is 1)}$$

- X is continuous iff there is a density function f_X such that the following holds:

$$\begin{aligned} \Pr(a \leq X \leq b) &= \int_a^b f_X(x)\,dx \\ &= \text{Area under } f \text{ between } a \text{ and } b \end{aligned}$$

- The support S_X consists of all x for which $f_X(x) > 0$

Normalization

- We must ensure that probability distributions sum / integrate to 1 (i.e., total probability must equal 1)

- **Normalization**: Scalar adjustment in order to ensure that $\Pr(S_X) = 1$

- If $g(x) > 0$ for all $x \in S_X$, then

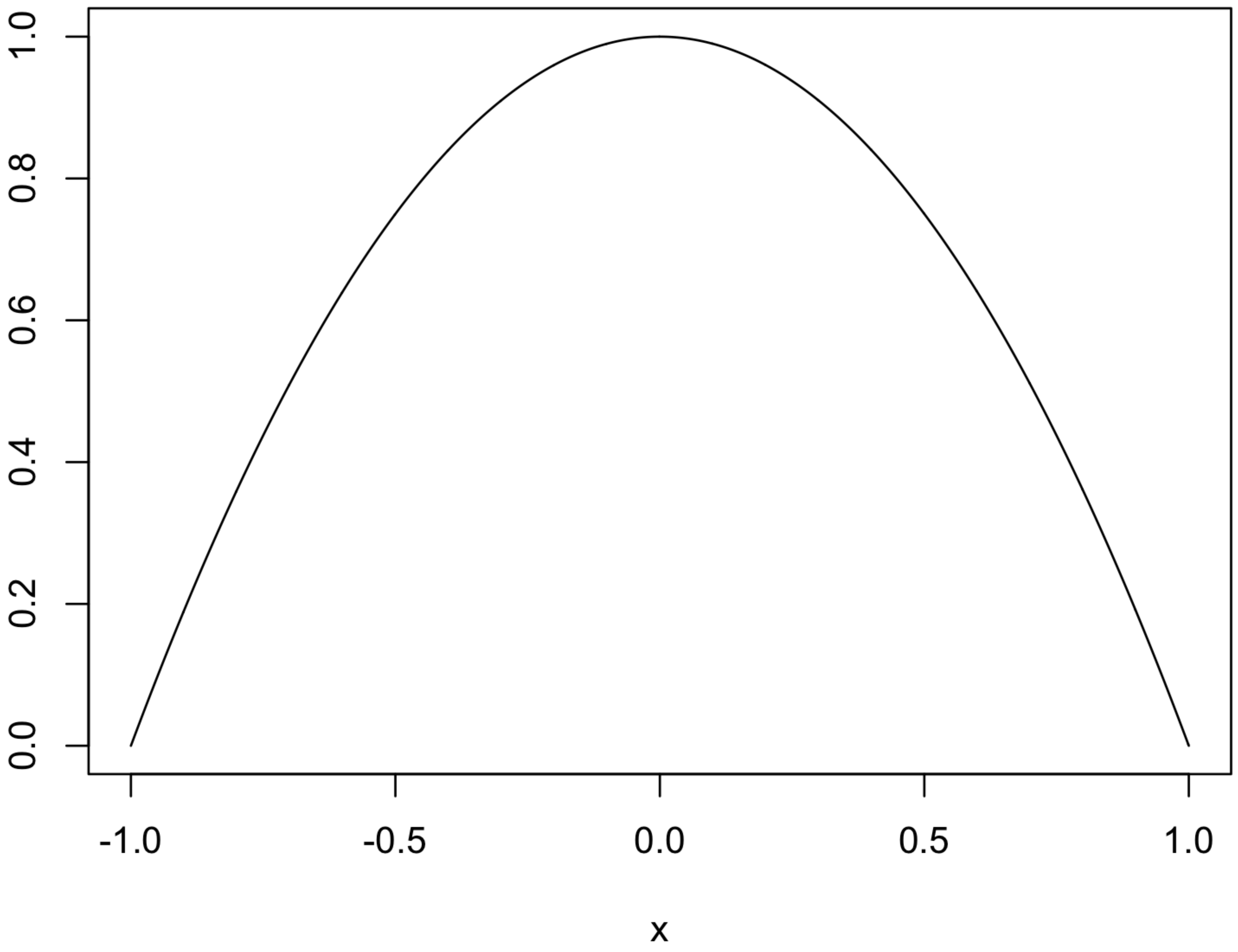
$$\text{Discrete: } p(x) = \frac{g(x)}{\sum_{x^* \in S_X} g(x^*)}$$

$$\text{Continuous: } f(x) = \frac{g(x)}{\int_{x^* \in S_X} g(x^*)\, dx^*}$$

- *Normalization constant*: 1/denominator

Normalization: Example

- Suppose that we generally know that probability is distributed according to the following curve:



Normalization: Example

- We can generally define the shape of this curve as

$$g(x) = 1 - x^2, \quad -1 \leq x \leq 1$$

- Is this a proper density?

integrate g(x) = [x - x^3/3] = 4/3 (1 <= x <= 1)
it is not a proper density

- What's the normalization constant?

normalization constant (nc) = 3/4

- What is $f(x)$?

f(x) = 3/4 x g(x)

Cumulative Distribution Functions (CDFs)

- The **cumulative distribution function (CDF)** of random variable X is

$$F_X(x) = \Pr(X \leq x) \text{ for all } x \in (-\infty, \infty)$$

- If X is discrete with support S_X, then the CDF is defined as

$$F_X(x) = \Pr(X \leq x) = \sum_{u \in S_X : u \leq x} \Pr(X = u)$$

- If X is continuous, then the CDF is defined as

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_X(u)\, du$$

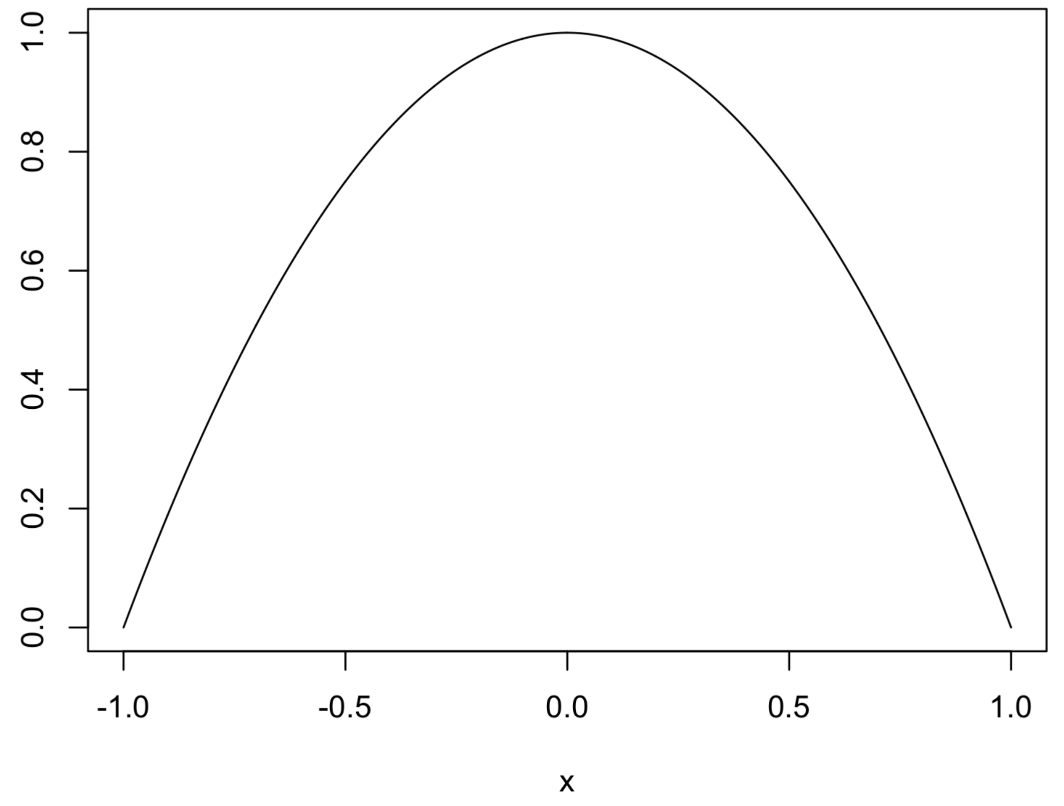
PDF is the differential function of CDF

Cumulative Distribution Functions (CDFs)

- Consider the parabolic density

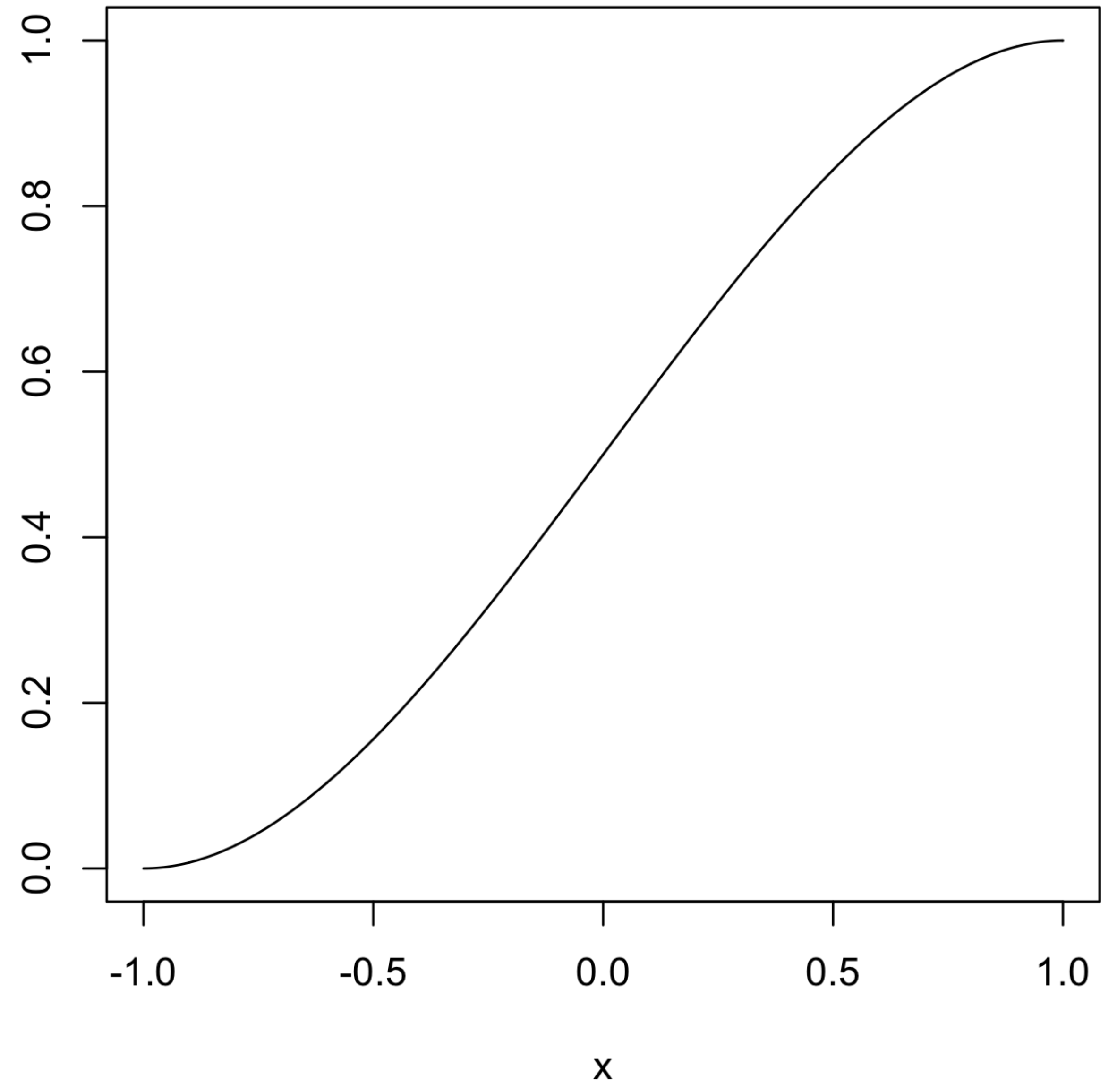
$$f(x) = \begin{cases} 0 & x \in (-\infty, -1) \\ \frac{3}{4}(1 - x^2) & x \in [-1,1] \\ 0 & x \in (1,\infty) \end{cases}$$

PDF curve



- Over the range $x \in (-\infty, -1)$, we have $F(x) = 0$

CDF curve



- Over the range $x \in [-1,1]$, we have

need to review calculus….

$$F(x) = \int_{-1}^{x} \frac{3}{4}(1 - u^2)\, du = -x^3/4 + 3x/4 + 1/2$$

- Over the range $x \in (1,\infty)$, we have $F(x) = 1$

if x = a, then F(a) = 1 = integral of f(x) when x <= a

Quantiles and Percentiles

- Suppose that a student with an 85 on an exam scored higher than 72% of their classmates
- Then, $\Pr(X \leq 85) = 0.72$
- We say that $q = 85$ is the $p = 0.72$ quantile of this distribution (also called the 72nd percentile)

Quantiles and Percentiles

- More generally: For a random variable X, q is the p-quantile of X if

$$\Pr(X < q) \leq p \text{ and } \Pr(X > q) \leq 1 - p$$

- The quantile function of X is then defined as

$$Q(p) = \min\{x \in S_X : \Pr(X \leq x) \geq p\}$$

- If the CDF $F_X(x)$ is continuous and strictly increasing on S_X, then

$$Q(p) = F_x^{-1}(p)$$

- Although $Q(p)$ is uniquely defined, the p-quantile may not be unique

qnorm() in R

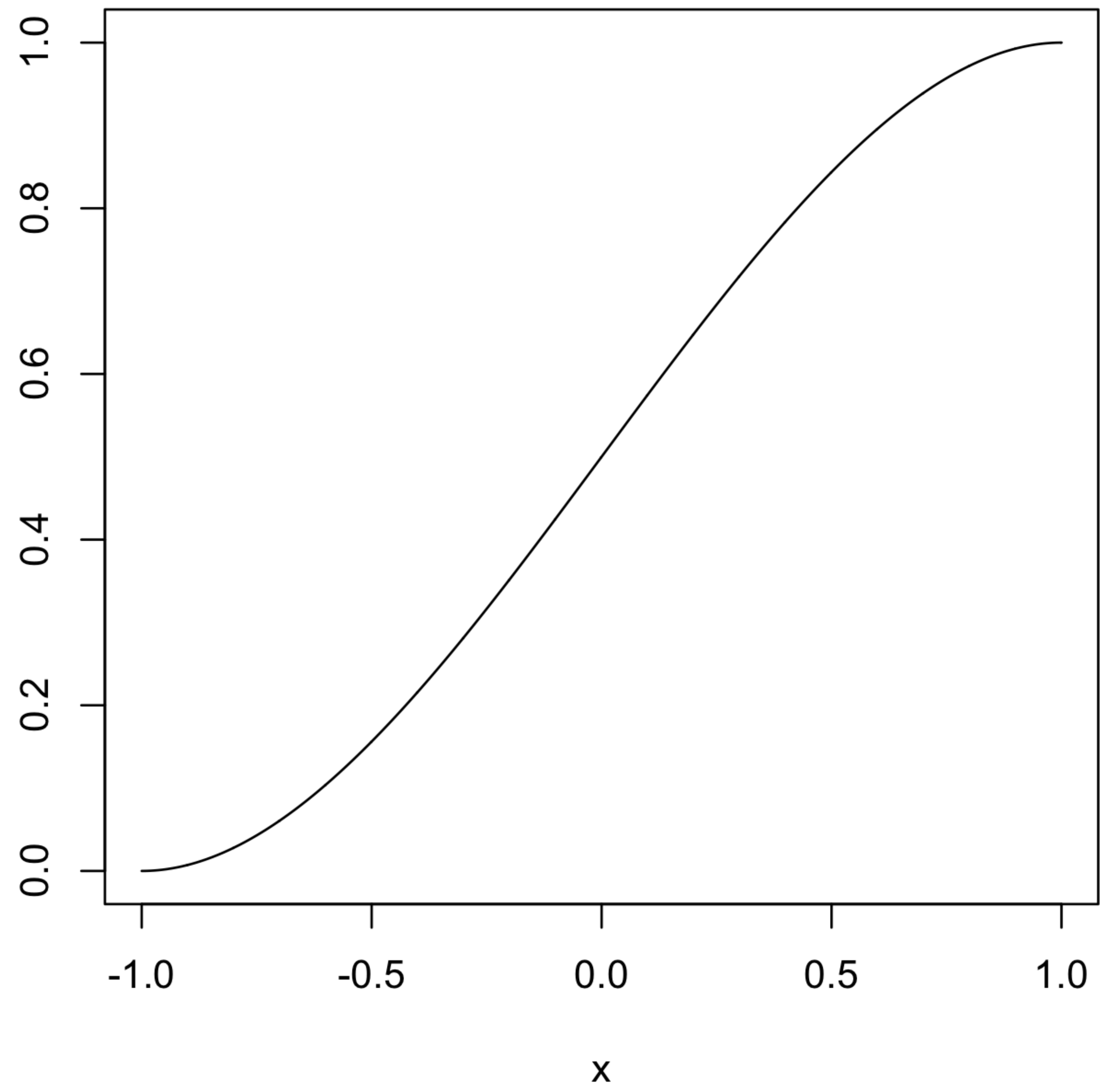
Quantiles and Percentiles: Example

- Consider the parabolic density, $f(x) = \frac{3}{4}(1 - x^2)$
- What is the 0.25-quantile?

integral of f(x) = 3/4 (x - x^3/3) + 1/2 = 0.25
x = 0.25-quantile
x = -1.53 or -0.35, since -1.53 is out of range (f(x)
must larger than 0), x = -0.35

how to estimate by eye?
what is the x value when y = 0.25

CDF



Summarizing Probability Distributions

- Many random variables can take a large number of values, so an explicit probability distribution may be quite complicated
- We can describe a probability distribution with measures of central tendency and dispersion
- *Population mean*: Average value that a random variable takes
- *Population variance*: Dispersion of the values relative to the population mean
- *Population standard deviation*: The square root of the population variance

Expected Value

- **Expected value** of X, denoted $E(X)$, represents a theoretical average of an infinitely large sample
  - $E(X)$ is what we "expect" X to equal; the population mean of X
- We use the notation $\mu = \mu_X = E(X)$

Expected Value

- If X is a discrete random variable:

$$\mu_X = E(X) = \sum_{x \in S_X} x \cdot \Pr(X = x)$$

- If X is a continuous random variable:

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

the range of x

- If c is a constant, then

$$E(c) = c$$

Linearity of Expectation

- For any random variables X and Y:

$$E(X + Y) = E(X) + E(Y)$$

- This holds even if X and Y are *not* independent

whether X and Y are correlated or not does not matter

- In general, for random variables $X_1, \ldots, X_n$:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

good way to swap the order of calculation

Mean of a Random Variable: Example

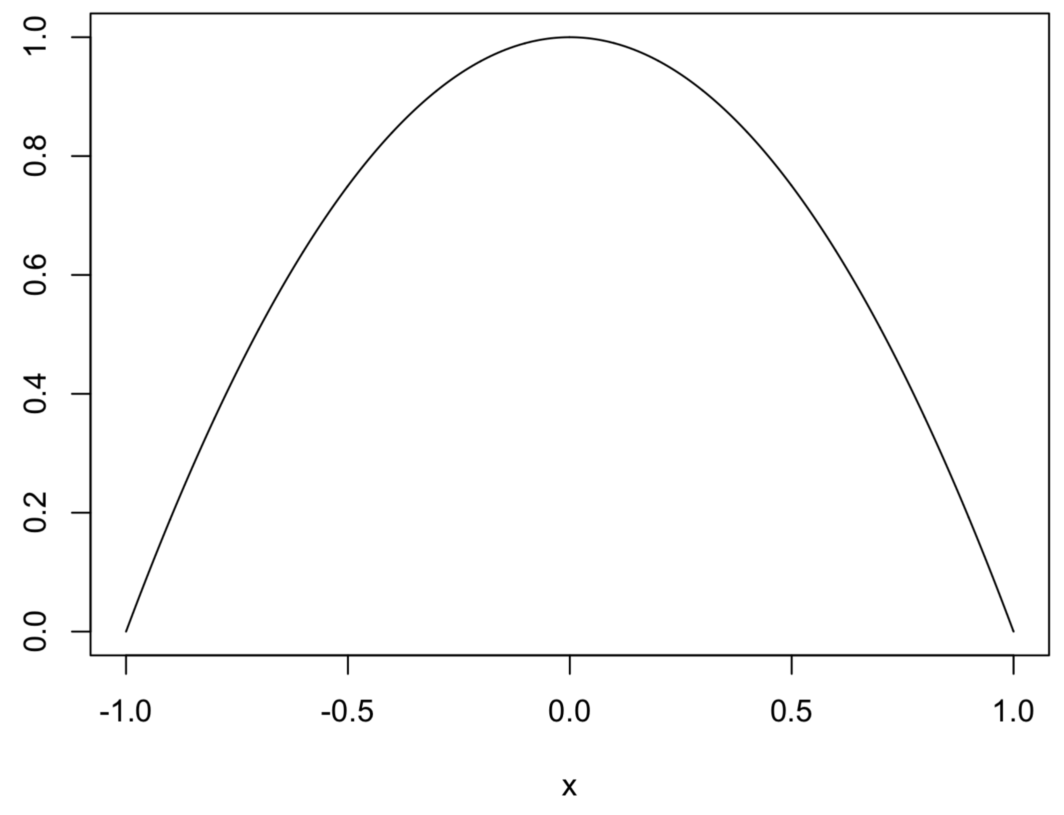
- Q1: What is the expected number of heads when flipping a fair coin (1/2 H, 1/2 T)?

- Q2: What is the expected number of heads when flipping an unfair coin (2/3 H, 1/3 T)?

- Q3: What is the expected number of heads when flipping three fair coins (1/2 H, 1/2 T) and two unfair coins (2/3 H, 1/3 T)?

$$E\left(\sum^n X_i\right) = \sum^n E(X_i)$$

Mean of a Random Variable: Example



- Consider the parabolic density, $f(x) = \frac{3}{4}(1 - x^2)$
- What is $E(X)$?
- Intuition: By symmetry, $E(X) = 0$

Variance

- The variance of X, denoted $var(X)$, measures the tendency of X to deviate from $E(X)$ and is defined as follows

$$var(X) = E\left(\left(X - E(X)\right)^2\right)$$
$$= E(X^2) - E(X)^2$$

consider E(x) as constant

= E(x^2) - E(2xE(x)) + E((E(x))^2)

= E(x^2) - 2E(x)E(x) + (E(x))^2

- We use the notation $\sigma^2 = \sigma_X^2 = var(X)$
- The standard deviation is the square root of the variance: $\sigma = \sigma_X = \sqrt{var(X)}$

Variance

- Recall: $var(X) = E\left(\left(X - E(X)\right)^2\right)$

- Let X be a discrete random variable with mean μ_X:

$$\sigma_X^2 = \sum_{S_X} (x - \mu_X)^2 \Pr(X = x)$$

- Let X be a continuous random variable with mean μ_X

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx$$

Variance: Example

- Setup: Flip two fair coins; let X be the number of heads
- Q: What is $var(X)$?
- Q: What is the standard deviation of X?

Functions of Random Variables

- Take random variable X and function $g(\,.\,)$
- We can obtain a new random variable: $Y = g(X)$
- This is what is called a *transformation of variables*
- In general, to get the distribution of Y, we have that for any event $E \subseteq S_Y$, we have $p_Y(E) = p_X(g^{-1}(E))$

Linear Transformations: Mean and Variance

- Let g be a linear function of the form $g(X) = aX + b$
- Let X be a random variable with mean μ_X and variance σ_X^2
- Define a new random variable $Y = g(X) = aX + b$
- Finding the mean of Y:

$$\mu_Y = E(Y) = E(aX + b) = aE(X) + b = a\mu_X + b$$

- Finding the variance of Y:

$$\begin{aligned}
\sigma_Y^2 &= var(Y) = E((aX + b - E(aX + b))^2) \\
&= E((aX + b - aE(X) - b)^2) = E((aX - aE(X))^2) \\
&= E(a^2(X - E(X))^2) = a^2E((X - E(X))^2) \\
&= a^2 \cdot var(X) = a^2 \cdot \sigma_X^2
\end{aligned}$$

General Transformations: Mean

- If we have $Y = g(X)$ for general $g(X)$, then we have:

$$\mu_Y = E(Y) = E(g(X))$$

- We do **not** necessarily have that:

$$E(g(X)) = g(E(X))$$

- Example: Consider X = the outcome of rolling a fair six-sided die, and let $g(X) = X^2$

$$E(g(X)) = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} = \frac{91}{6} \approx 15.17$$

$$g(E(X)) = \left(\frac{1 + 2 + 3 + 4 + 5 + 6}{6}\right)^2 = (3.5)^2 = 12.25$$

Independence

- Two random variables X_1 and X_2 are independent if the following holds, for any two events E_1 and E_2:

$$P(X_1 \in E_1 \cap X_2 \in E_2) = P(X_1 \in E_1) \cdot P(X_2 \in E_2)$$

- Notation:

$$X_1 \perp X_2 \text{ means } X_1 \text{ and } X_2 \text{ are independent}$$

- If a collection of random variables $X_1, X_2, \ldots, X_n$ are all independent and have the same distribution, we say that they are i.i.d. (independent and identically distributed)
  - Example: Roll two dice, or flip three fair coins

Covariance

- If two variables are not independent, we measure their dependency through their **covariance**
- Let X and Y be two random variables with means μ_X and μ_Y, respectively
- The covariance of X and Y is defined as follows:

$$cov(X, Y) = \sigma_{XY} = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - E(X)E(Y)$$

- Correlation (essentially normalized covariance):

$$corr(X, Y) = \rho = \rho_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Properties of Covariance

- Given random variables X and Y, the following hold:
  - If either X or Y is a constant, then $cov(X, Y) = 0$ and $corr(X, Y)$ is undefined
  - If $X \perp Y$, then $cov(X, Y) = corr(X, Y) = 0$
  - $cov(X, X) = var(X)$
  - $cov(X, Y) = cov(Y, X)$

Linear Combinations

- Suppose you have random variables X and Y with means μ_X, μ_Y and variances σ_X^2, σ_Y^2
- Let $Z = aX + bY$
- The mean of Z is

$$\mu_Z = E(Z) = E(aX + bY) = E(aX) + E(bY) = a\mu_X + b\mu_Y$$

- The variance of Z is

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$$

- The standard deviation of Z is

$$\sigma_Z = \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}}$$

Theoretical Distributions

- Theoretical probability distributions describe what we expect to happen based on populations on a theoretical level
- We will consider the following theoretical distributions (D = discrete, C = continuous):
  - Bernoulli distribution (D)
  - Binomial distribution (D)
  - Poisson distribution (D)
  - Geometric distribution (D)
  - Uniform distribution (C)
  - Exponential distribution (C)
  - Normal distribution (C)

Bernoulli Distribution

- Let Y be a dichotomous random variable (takes one of two mutually exclusive values)
    - Classic example: Coin flip
- Successes (= 1) occur with probability p and failures (= 0) occur with probability $1 - p$, for constant $p \in [0,1]$
- Notation: $\underset{\text{follows}}{Y \sim} Bern(p)$

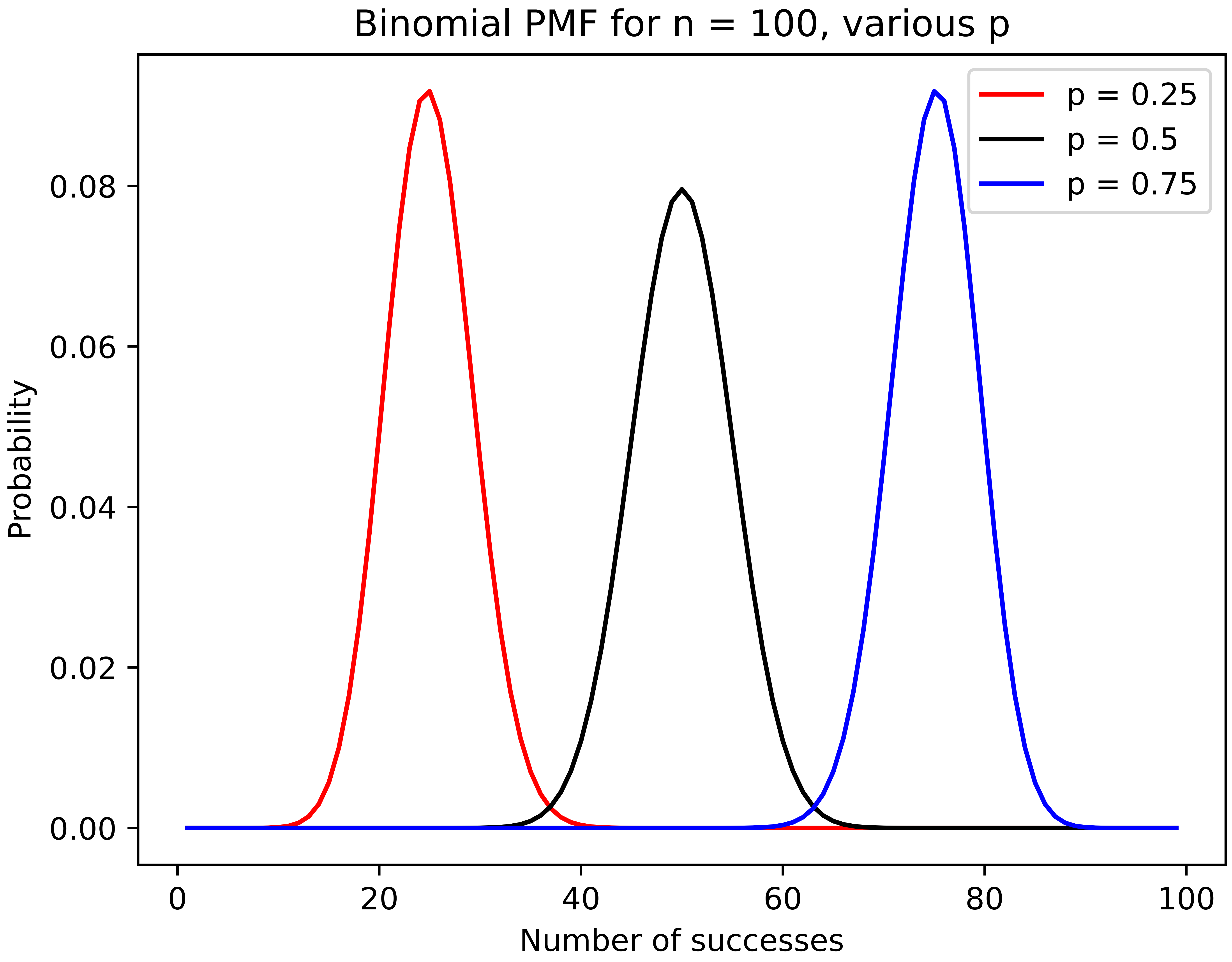
Bernoulli Distribution: Example

- Let Y be a dichotomous random variable representing a coin flip
  - $Y = 1$: heads
  - $Y = 0$: tails
- If the coin is fair, then $p =$ and $1 - p =$
- If the coin has a 60% chance of landing heads, then $p =$ and $1 - p =$
- What is $E(Y)$ in terms of p?
- What is $var(Y)$ in terms of p?

Binomial Distribution

- Suppose we flip n i.i.d. coins instead of just one coin
- Let $X = \sum_{i=1}^{n} X_i$ be the number of heads we observe
- X is binomially distributed
- **Binomial distribution**: If we have a sequence of n Bernoulli random variables, each with a probability of success p, then the total number of successes is a binomial random variable
  - Assumptions: fixed number of trials, independent, constant p
- Notation: $X \sim Bin(n, p)$

Binomial Distribution



Binomial Coefficients

- Let $X = \sum_{i=1}^{n} X_i$ be the number of heads we observe when we flip n i.i.d. coins
- Each coin has probability of heads p, and flips are independent
- Q: What is the probability of getting exactly x out of n successes?
  - Choose which x trials succeed: choose(n, x)
  - Probability that these x trials succeed: p^x
  - Probability that the other $n - x$ trials fail: (1-p)^(n - x)
- In general, choose(n, x) * p^x * (1-p)^(n - x)

Binomial Probabilities in R

- Calculate probabilities in R:
  - Calculate $\Pr(X = x)$ using `dbinom(x,n,p)`
  - Calculate $\Pr(X \leq x)$ using `pbinom(x,n,p)`
  - Calculate $\Pr(X \geq x)$ using `1-pbinom(x-1,n,p)`

Binomial Distribution: Summary Measures

- Note that a binomial distribution with parameters n and p is the sum of n independent Bernoulli distributions with parameter p

$$E(X) = \mu_X = np$$

$$var(X) = \sigma_X^2 = np(1-p)$$

$$stdev(X) = \sigma_X = \sqrt{np(1-p)}$$

- Q: How does $var(X)$ change with $p \in [0,1]$?

Binomial Distribution: Summary

- Main take-away points from the binomial distribution:
  - Fixed number of independent Bernoulli trials, n
  - Constant probability of success, p (Bernoulli parameter)
  - Interested in the total number of successes in n trials (not order)
  - Mean: $\mu_X = np$
  - Variance: $\sigma_X^2 = np(1-p)$
- Examples:
  - Number of heads in 15 flips of a fair coin

Poisson Distribution

- **Poisson distribution**: Probability of observing a certain number of discrete events within a known interval
  - Models discrete events that occur infrequently in time or space
- Example:
  - The number of babies born at Strong Memorial Hospital between 10 am and 2 pm today
  - The number of students who enter River Campus today

Poisson Distribution

- Let $X \in [0,\infty)$ be the number of occurrences of some event over a given interval

- Let $\lambda > 0$ be the average number of occurrences of the event over the specified interval

- In this case, we say that $X \sim Pois(\lambda)$

- The probability function is given by $P(X = x) = \dfrac{e^{-\lambda}\lambda^{x}}{x!}$

- If $X \sim Pois(\lambda)$, then $\mu_X = \sigma_X^2 = \lambda$
  - For a Poisson distribution, both the mean and the variance are equal to λ

Poisson Distribution

- Poisson distribution assumptions:
  - The probability of an event occurring is proportional to the length of the interval
  - Within an interval, an infinite number of events is theoretically possible
  - Events occur independently
  - The number of events that occur must be non-negative

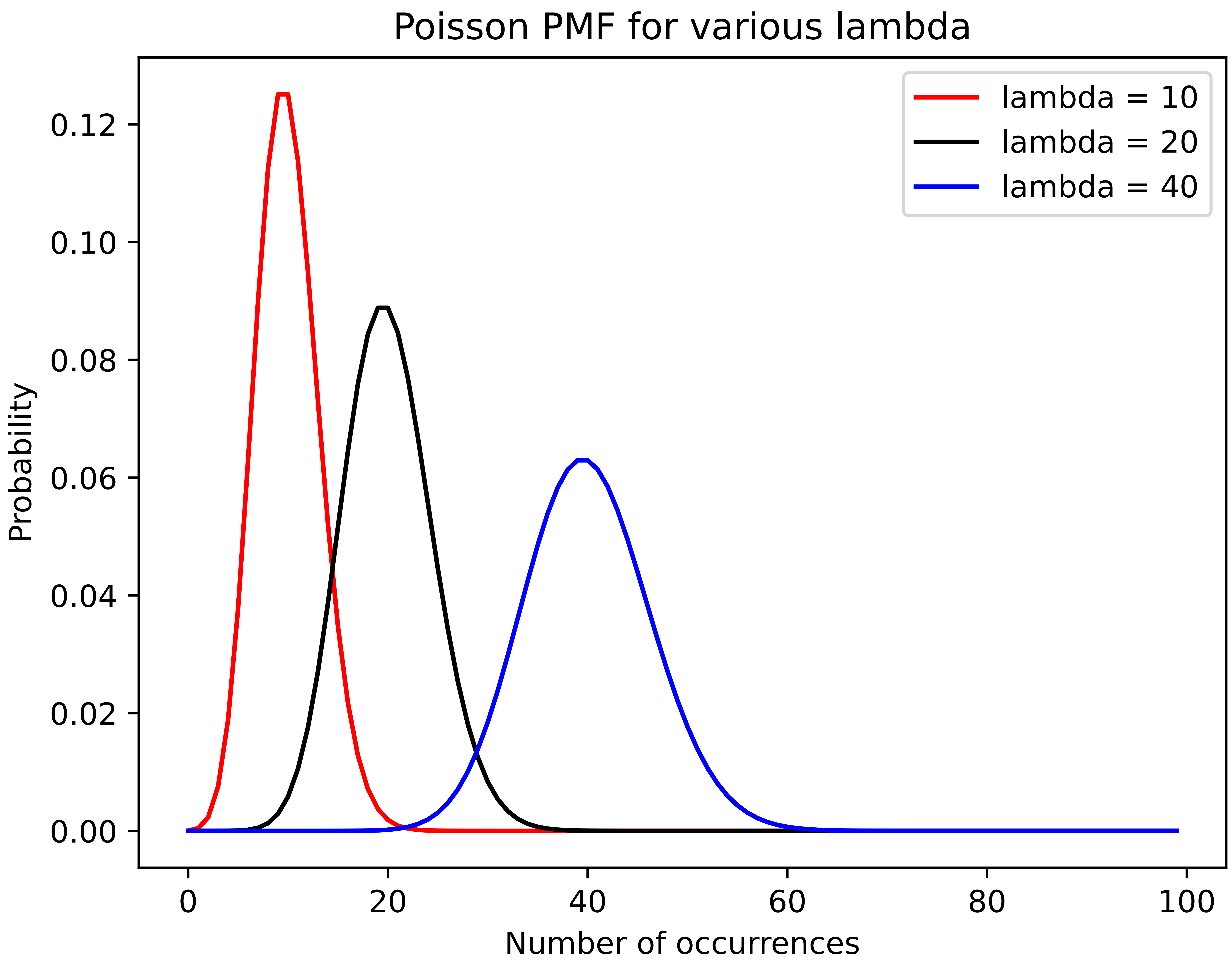
Poisson Distribution: Example

- Setup: We want to examine the probability of certain numbers of people developing a rare disease in the next year. On average, 1.95 people develop the disease per year

- Q: What is the probability of no one developing the disease in the next year?

- Q: What is the probability of one person developing the disease in the next year?

Poisson Distribution: Visualized



Poisson Probabilities in R

- Calculate probabilities in R:
  - Calculate $\Pr(X = x)$ using `dpois(x,lambda)`
  - Calculate $\Pr(X \leq x)$ using `ppois(x,lambda)`
  - Calculate $\Pr(X \geq x)$ using `1-ppois(x-1,lamda)`

Poisson Distribution: Summary

- Main take-away points from the Poisson distribution:
  - Fixed interval, independent events, interested in number of events in interval
  - Unlimited number of events is theoretically possible
  - Mean: $\mu_X = \lambda$
  - Variance: $\sigma_X^2 = \lambda$
- Examples:
  - Number of calculators the book store sells this year
  - Number of babies born today

Geometric Distribution

- Suppose $Y_1, Y_2, \ldots$ is an *infinite* sequence of independent Bernoulli random variables with parameter p
- Let X be the first index i for which $Y_i = 1$ (location of first success)
- The probability mass function (PMF) is given by

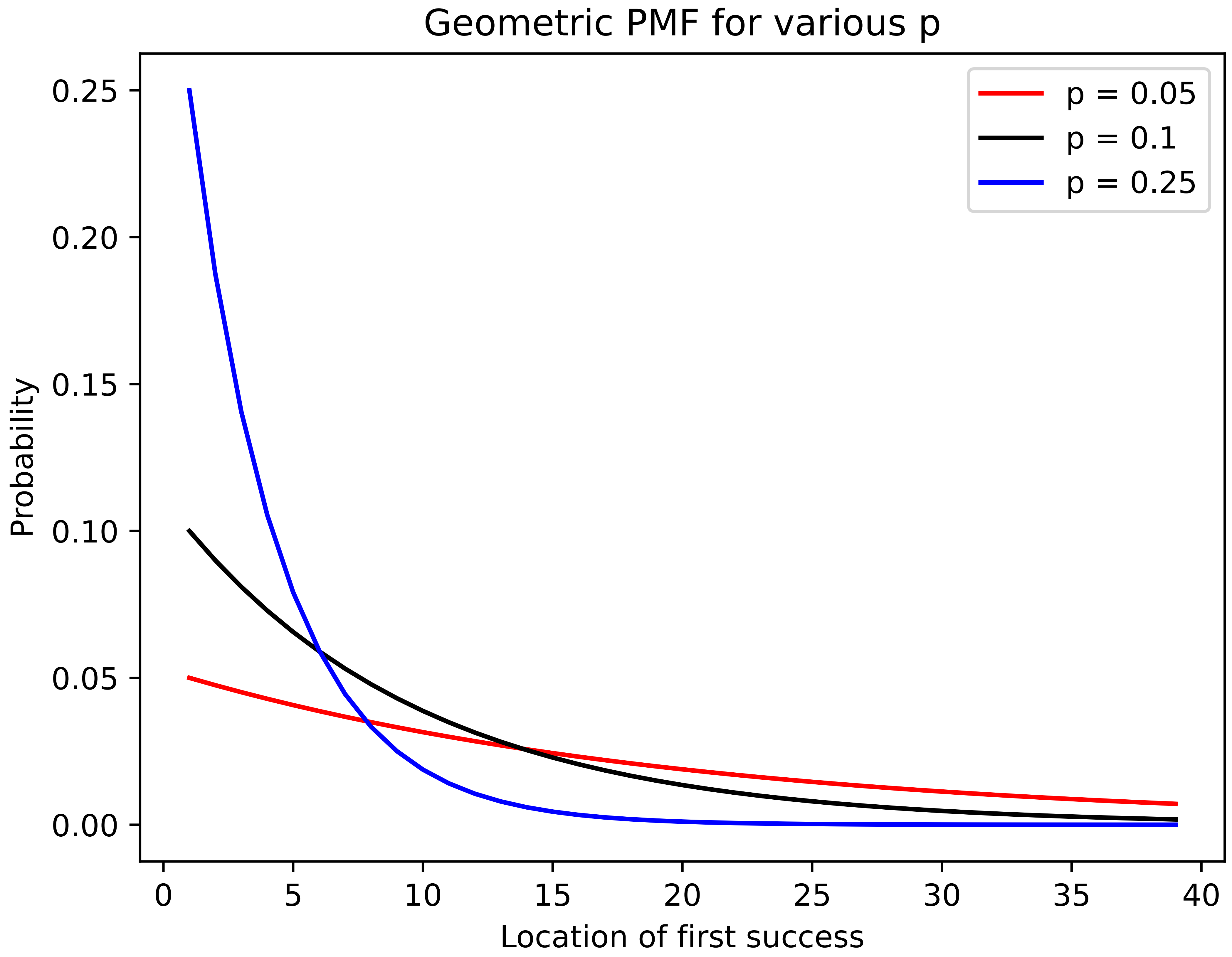
$$P(X = x) = p(1-p)^{x-1}$$

- Notation: $X \sim Geom(p)$
- Mean and variance:

$$E(X) = \mu_X = \frac{1}{p}$$

$$var(X) = \sigma_X^2 = \frac{1-p}{p^2}$$

- CDF: $P(X \le x) = 1 - (1-p)^x$



Continuous Distributions

- Continuous random variables: Intuitively, discrete random variables that are infinitesimally close together
- Instead of having discrete bars for the density at each discrete value, we now have a continuous density curve
- The area under the curve equals 1 (law of total probability)
- Since the random variable X can take on an infinite number of values, the probability associated with any single outcome equals 0
- The probability that $X \in (x_1, x_2)$ is equal to the area under the curve that lies between these two values

Uniform Distribution

- Let X be a continuous random variable which can take on any value between a and b with equal probability

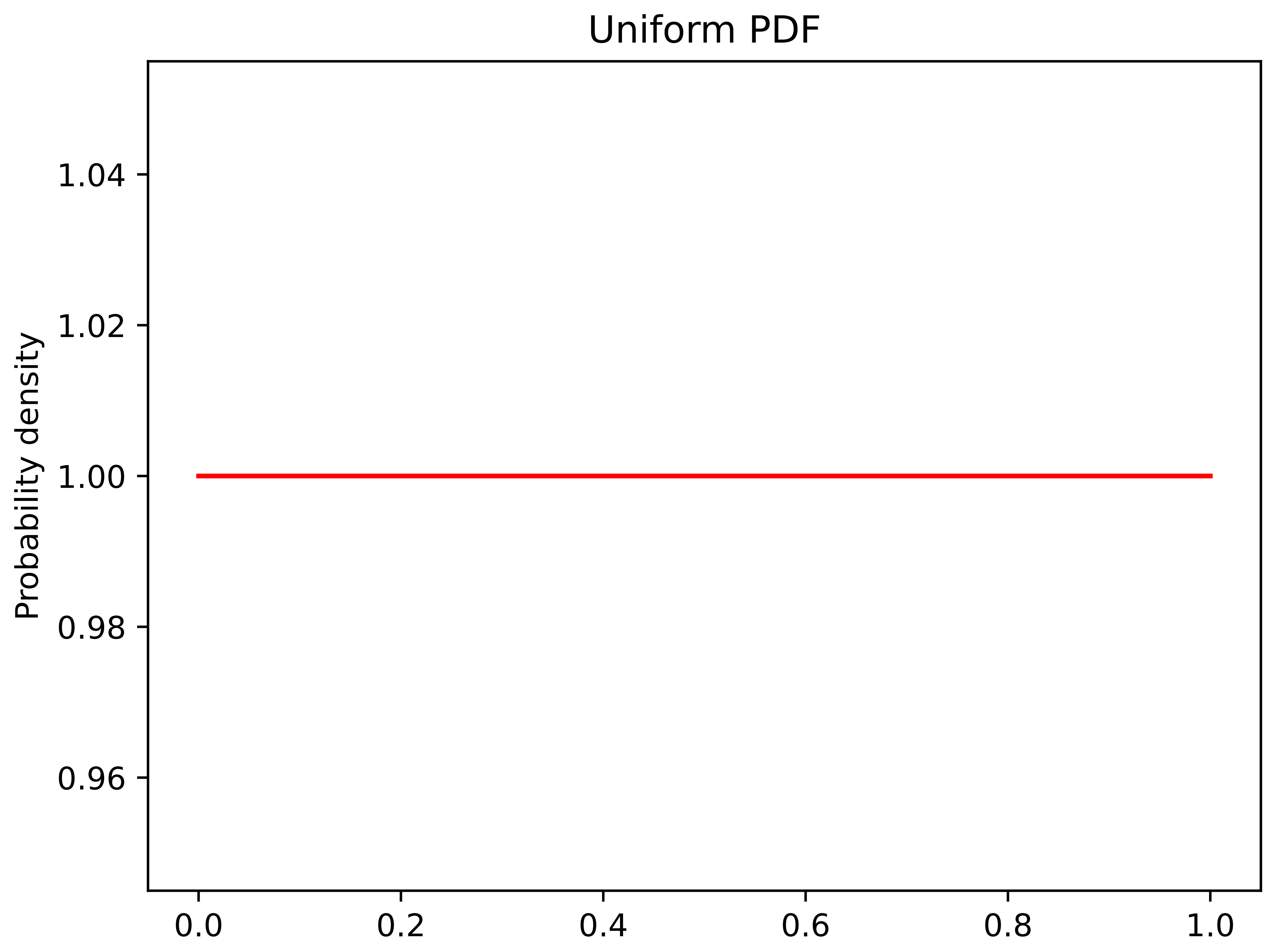
- Any value outside the range $[a, b]$ cannot occur

- The uniform probability density is

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

- Notation: $X \sim Unif(a, b)$

- $\mu_X = \dfrac{a+b}{2}$ and $\sigma_X^2 = \dfrac{(b-a)^2}{12}$



Exponential Distribution memoryless

- A continuous random variable X is exponentially distributed if it follows the following density:

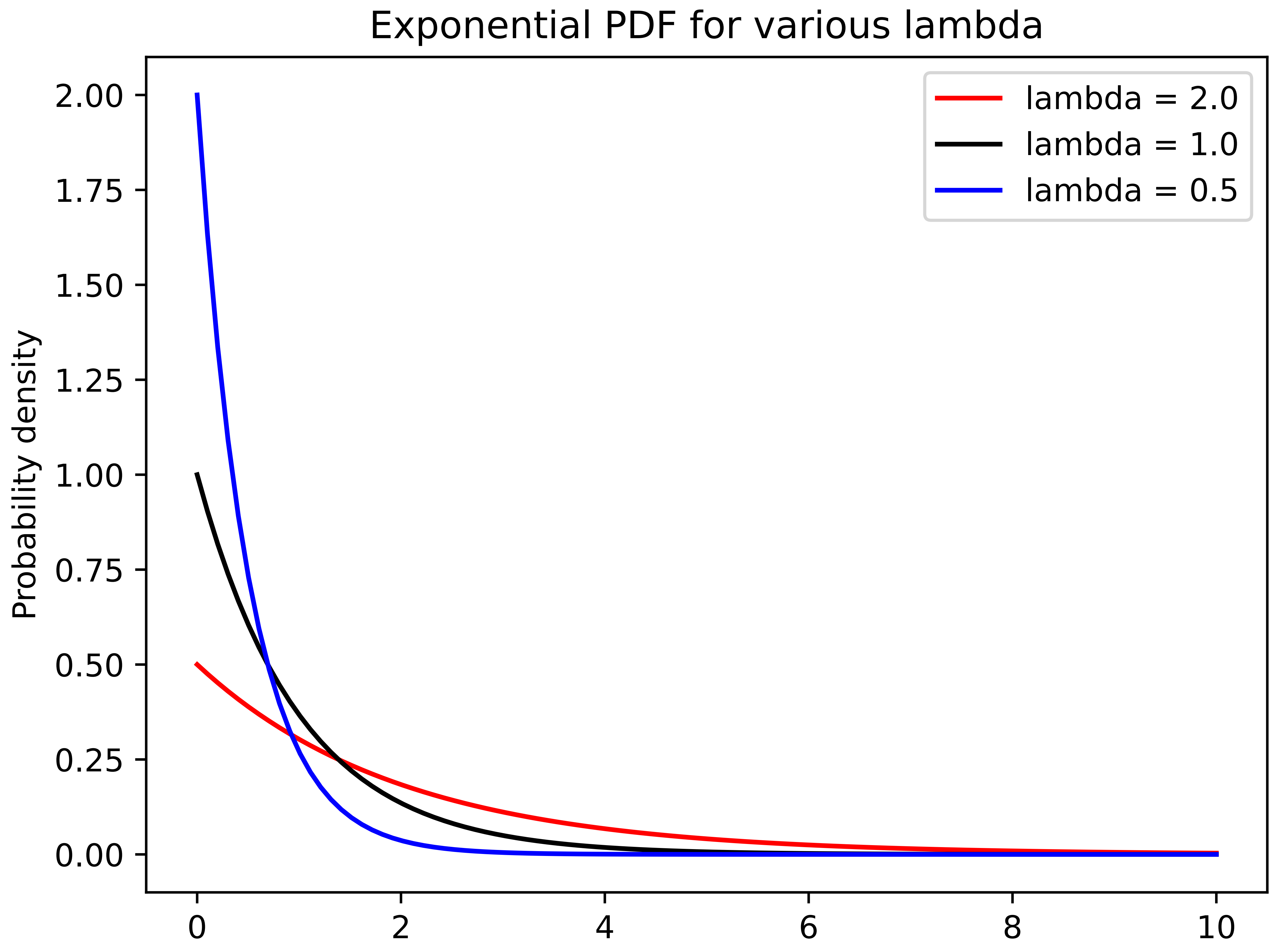
$$f_X(x) = \lambda e^{-\lambda x}, \lambda > 0$$

- Notation: $X \sim Exp(\lambda)$

- Generalizes the geometric distribution

- $\mu_X = \frac{1}{\lambda}$ and $\sigma_X^2 = \frac{1}{\lambda^2}$

- CDF: $F_X(x) = 1 - e^{-\lambda x}$



Normal Distribution

- The most common continuous distribution is the normal distribution (also called a Gaussian distribution or bell-shaped curve)
  - Shape of the binomial distribution when p is constant but $n \to \infty$
  - Shape of the Poisson distribution when $\lambda \to \infty$
- Given μ, σ, the density function is

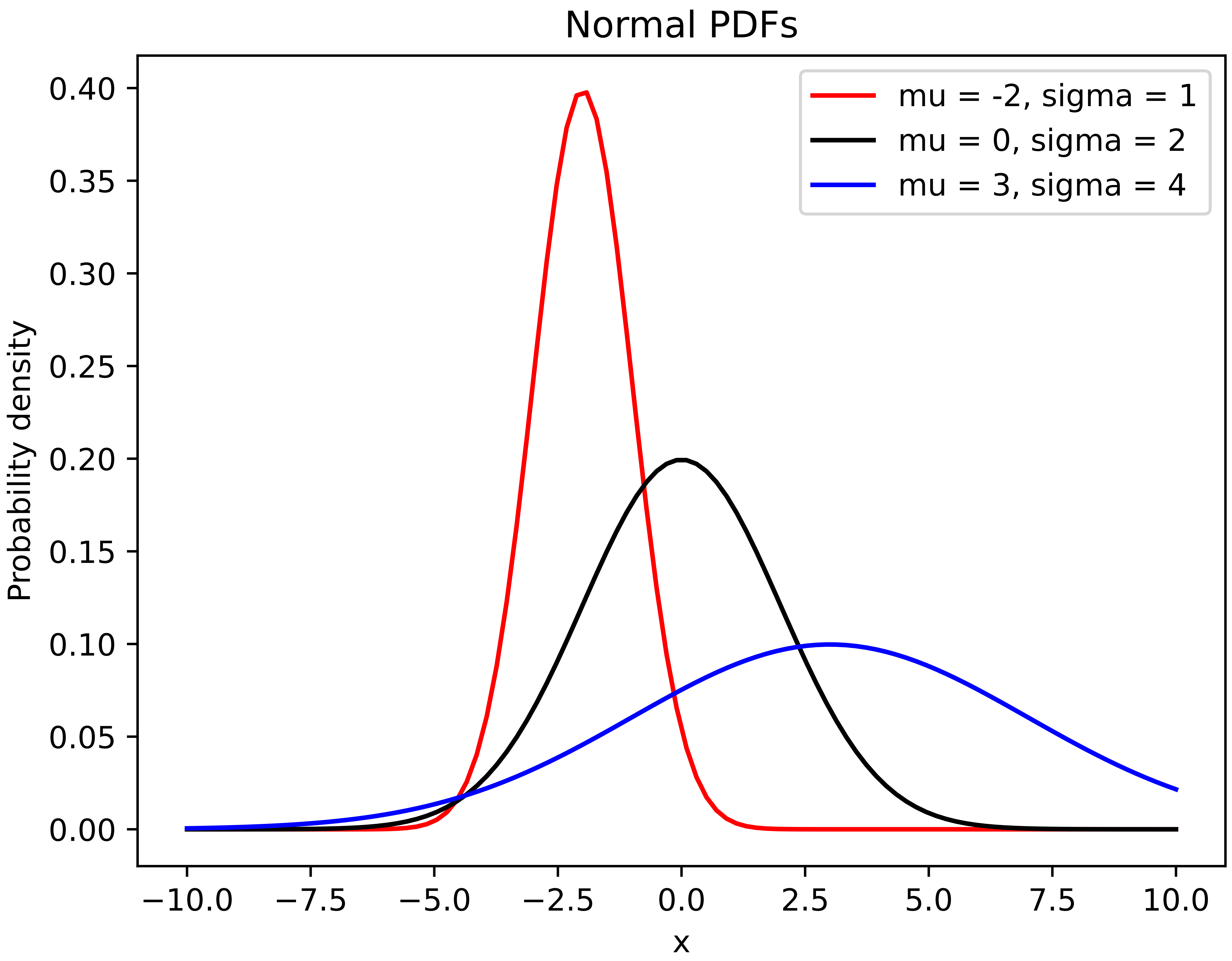
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

- Notation: $X \sim N(\mu, \sigma^2)$ – but in R, use standard deviation instead of variance
- Mean = median = mode = μ, variance = σ^2, standard deviation = σ

Normal Distribution: Visualization

- μ (center) and σ^2 (spread) fully define the normal distribution
- Always symmetric
- When $\mu = 0$ and $\sigma^2 = 1$, we have the *standard normal distribution*

sigma larger - distribution wider



Normal Distribution: z-scores

- Recall from Chapter 2 that a z-score tells us how many standard deviations an observation is from its mean:

$$z = \frac{x - \mu}{\sigma}$$

- Z has the nice property that it will always be $N(0,1)$
- Given $X \sim N(\mu, \sigma)$, we can calculate a z-score, which will be $Z \sim N(0,1)$
- Standardizes the procedures for all normal distribution problems

Normal Distribution

- Recall that for continuous distributions, we are interested in determining the probability of being in an interval:

$$\Pr(X \leq a), \Pr(X \geq b), \text{ or } \Pr(a \leq X \leq b)$$

- We can look at the plot of the normal distribution and determine the probability (= area under the curve between endpoints)
- In general, we will use R to calculate area under the curve (i.e., probabilities)
- By default, R works in terms of z-scores:

$\Pr(Z \leq z)$: `pnorm(z)`

$\Pr(Z \geq z)$: `1-pnorm(z)`

$\Pr(z_1 \leq Z \leq z_2)$: `pnorm(`z_2`)-pnorm(`z_1`)`

Normal Distribution

- The general process for calculating probabilities based on a normal distribution is as follows:
  - Calculate appropriate z-scores: $z = \dfrac{X - \mu}{\sigma}$
  - Use R to calculate the probability based on this z-score (`pnorm(z)`)

Normal Distribution: Example

- Suppose that test scores are normally distributed with mean 78 and standard deviation 9
- Q: What is the probability that a person scored below 60?
- Q: What is the probability that a person scored between 80 and 90?

pnorm(z)

Normal Probabilities in R (Shortcut)

- We can let R do the entire process of calculating a z-score and probability for us
- Let $X \sim N(\mu, \sigma)$

another way to use pnorm()

$\Pr(X \leq x)$: `pnorm(x,mean,sd)`

$\Pr(X \geq x)$: `1-pnorm(x,mean,sd)`

$\Pr(x_1 \leq X \leq x_2)$: `pnorm(`x_2`,mean,sd)-pnorm(`x_1`,mean,sd)`

Normal Distribution

- Suppose that BMI is normally distributed with mean 26.6 and standard deviation 3.2
- What is the probability that a person has a BMI in the range (18.5, 24.9)?
- Find $\Pr(18.5 \leq X \leq 24.9)$:

$$z_1 = \frac{18.5 - 26.6}{3.2} = -2.53$$

$$z_2 = \frac{24.9 - 26.6}{3.2} = -0.53$$

$$\Pr(-2.53 \leq Z \leq -0.53) = 0.2924$$

- `pnorm(-0.53)-pnorm(-2.53)=0.2924`
- `pnorm(24.9,26.6,3.2)-pnorm(18.5,26.6,3.2)=0.2919`

Revisiting the Empirical Rule

- How well does the empirical rule approximate the normal distribution?
  - $\Pr(-1 \leq Z \leq 1) = 0.683$
  - $\Pr(-2 \leq Z \leq 2) = 0.954$
  - $\Pr(-3 \leq Z \leq 3) = 0.997$
- Empirical rule (68%, 95%, 99.7%) appears to be quite good

Normal Distribution: Percentiles

- Given data $x_1, \ldots, x_n$, what value of x corresponds to a probability of the p^{th} percentile?
- Strategy:
  - Find z value such that $\Pr(Z \leq z) = p$ (lower tail probability is p)
  - Solve for x by inverting z-score: $x = z \cdot \sigma + \mu$
- Directly in R: `qnorm(p, mean, sd)`; `qnorm(p)` for z value

Normal Distribution: Example

- Setup: Let X be a random variable that represents weights of patients in American hospital EDs; X is normally distributed with $\mu = 160$ and $\sigma = 15$
- Q1: Find the probability that a randomly selected patient in the ED weighs between 140 pounds and 210 pounds

z1 = (140 - 160)/15

z2 = (210 - 160)/15

- Q2: Find the value that cuts off the upper 10% of the curve in American ED patient weights

qnorm(0.9, 160, 15)

Sampling Distributions

- Suppose we want to estimate the mean value of some continuous random variable of interest

- We can take a sample from the population and use the sample mean as an estimate of the population mean: $\bar{x}$ is an estimate for μ

- For a normally distributed population, $\bar{x}$ is the *maximum likelihood estimator* for μ

  - Value of the parameter that is most likely to have produced the observed sample data

- Different samples will have different means

Sampling Distributions

- What if you continued sampling m times?
  - You take one random sample and get mean $\bar{x}_1$, take another random sample and get mean $\bar{x}_2$, and repeat until you have $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$
  - Take m random samples for a total of m sample means
- These m means form a distribution with mean μ and variance $\frac{\sigma^2}{n}$ where n is the sample size
- Key idea: $\bar{X}$ has its own distribution
- Standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$; this is known as the **standard error**

Sampling Normal Distributions

- If the population we are sampling from is normal, then the distribution of $\overline{X}$ will also be normal
- If $X \sim N(\mu, \sigma)$, then $\overline{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$

Central Limit Theorem (CLT)

- If the population we are sampling from is not normal, then we can use the *Central Limit Theorem (CLT)* to get the distribution of $\overline{X}$
- **Central Limit Theorem**: If the population we are sampling from is not normal, then the shape of the distribution of $\overline{X}$ will be normal as long as n is sufficiently large (typically $n \geq 30$ suffices)

Central Limit Theorem (CLT)

- In particular, given that the distribution of an underlying population has mean μ and standard deviation σ, the distribution of the sample means computed for samples of size n has three important properties:
  - The mean of the sampling distribution equals the population mean μ
  - The standard deviation of the distribution of sample means is equal to $\frac{\sigma}{\sqrt{n}}$, which is the standard error of the mean
  - Given that n is sufficiently large, the shape of the sampling distribution is approximabely normal
- In notation: $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Central Limit Theorem (CLT)

- The further the underlying population is from normal, the larger the sample size you need to ensure normality of the sampling distribution

- However, if the underlying population is normal, you do not need the central limit theorem to ensure normality of the sampling distribution – normality will hold regardless of the sample size if the underlying population is normal

- Since $\bar{X} \sim N\left(\mu, \sigma/\sqrt{n}\right)$, we can standardize $\bar{X}$ to a standard normal distribution as follows:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

small sample size and not normal distribution - nonparametric
large sample size but not normal distribution - CLT

Sampling Distributions Example

- Setup: Suppose house prices have a distribution with a mean of $\mu = \$450{,}000$ and standard deviation $\sigma = \$100{,}000$. You draw a random sample of $n = 64$ houses and determine their prices

- Q1: What is the mean of the distribution of sample means?

  450k

- Q2: What is the standard error of the sample mean?

  sigma' = sigma/sqrt(n) = 100000/sqrt(64) = 12.5k

- Q3: What distribution does the sample mean follow?

  x bar ~ N(450k, 12.5k)

- Q4: What is the probability that the sample mean of $n = 64$ house prices is greater than \$500,000?

  1 - pnorm(500000, 450000, 12500) = 3.167124e-05

Sampling Distribution of a Proportion

- Suppose we are interested in the proportion of the time that an event occurs
- If we take a sample of size n and observe x successes, then we could estimate the population proportion p by $\hat{p} = x/n$
- We can do this sampling process m times for a total of m different values of $\hat{p}_i$ for $i \in \{1,2,\ldots,m\}$
- These m proportions form a distribution with mean p
- Standard deviation of $\hat{p}$, known as standard error, is $\sqrt{\dfrac{p(1-p)}{n}}$

Sampling Distribution of a Proportion

- The shape of the distribution of $\hat{p}$ will be approximately normal as long as two conditions are met:
  - $np \geq 5$
  - $n(1-p) \geq 5$
- If both of these conditions are met, then $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$