FIT2086 Lecture 2
Expectations and Probability Distributions

Daniel F. Schmidt

Faculty of Information Technology, Monash University

August 4, 2022

Outline

Today's Relevant Figure (N/A)

Johann Carl Friedrich Gauss (1777 - 1855). Born in the Duchy of Brauchschweig-Wolfenbüttel (now Germany). Highly influential mathematician who made profound contributions to numerous fields of mathematics and science, including the fledging field of statistics. Sometimes called "The foremost (or Prince) of mathematicians".

Outline

Expected Values (1)

- Given a distribution, we can define the expected value of the RV:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, p(x)$$

  recalling that $p(x) \equiv \mathbb{P}(X = x)$.
- The expected value is the average value over $\mathcal{X}$, weighted by the probability of each particular $x \in \mathcal{X}$ appearing.
- For continuous RVs, replace the sum with an integral:

$$\mathbb{E}[X] = \int x\, p(x) dx$$

- Example: $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = 2) = 0.4$, $\mathbb{P}(X = 3) = 0.1$:

$$\mathbb{E}[X] = 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 = 1.6$$

Expected Values (2)

- More generally:

$$\mathbb{E}\left[f(X)\right] = \sum_{x \in \mathcal{X}} f(x)p(x)$$

  where $f(x)$ is any function of x.

- Example: $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = 2) = 0.4$, $\mathbb{P}(X = 3) = 0.1$:

$$\mathbb{E}\left[\log X\right] = \log 1 \cdot 0.5 + \log 2 \cdot 0.4 + \log 3 \cdot 0.1 = 0.3871$$

  where $\log x$ is the natural logarithm (sometimes called $\ln$).

Variance (1)

- This lets us define important properties such as the variance:

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 p(x)\end{aligned}$$

  $\Rightarrow$ The expected squared deviation around the mean
- The larger $\mathbb{V}[X]$ the more variation around the mean
- The standard deviation is equal to $\sqrt{\mathbb{V}[X]}$.

- Example: $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(X = 2) = 0.4$, $\mathbb{P}(X = 3) = 0.1$; recall that in this case, $\mathbb{E}[X] = 1.6$, so:

$$\mathbb{V}[X] = (1 - 1.6)^2 \cdot 0.5 + (2 - 1.6)^2 \cdot 0.4 + (3 - 1.6)^2 \cdot 0.1 = 0.44$$

Variance (2)

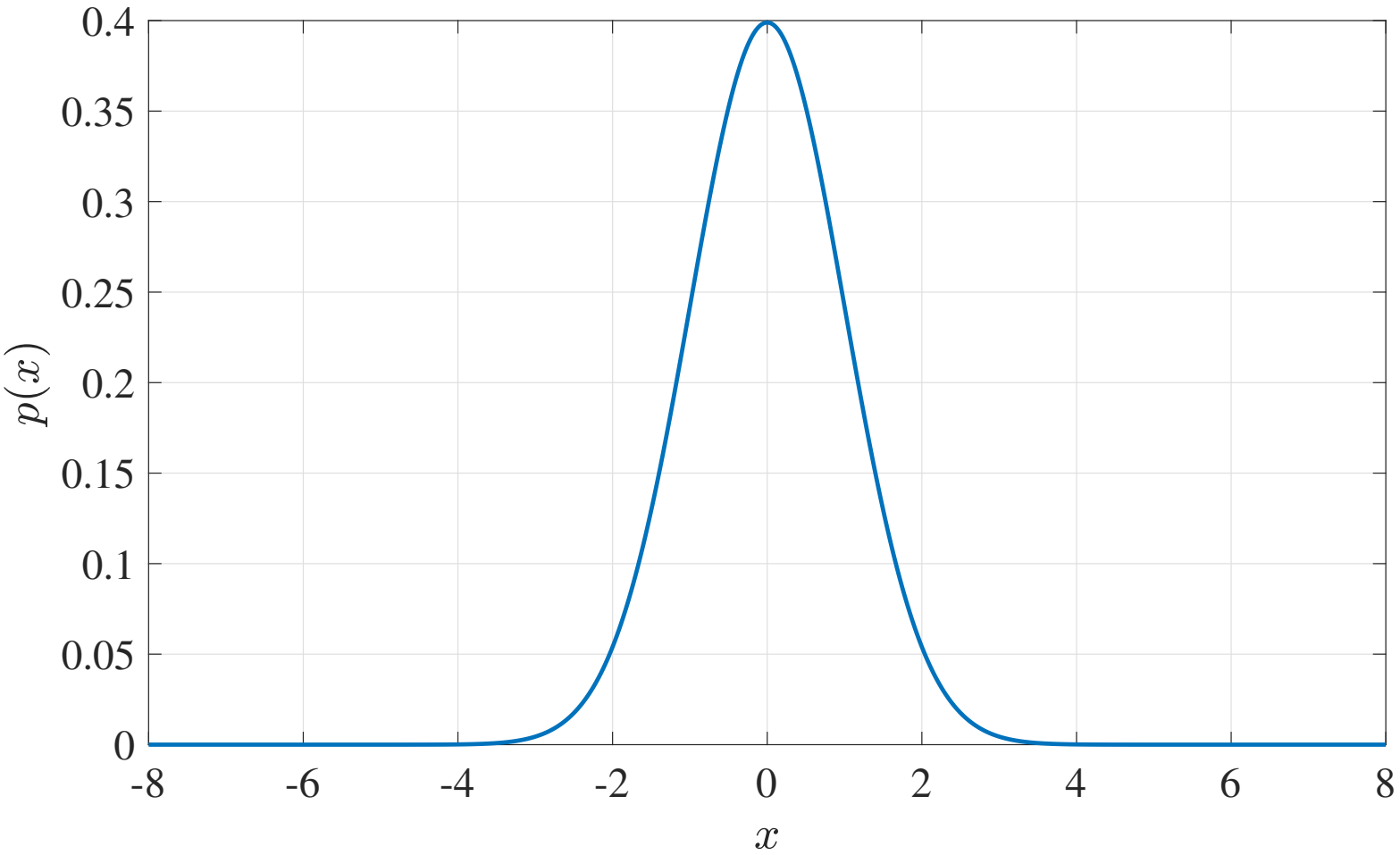
- A useful alternative expression for variance is:

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}$$

  where the third step follows from properties of sums/integrals

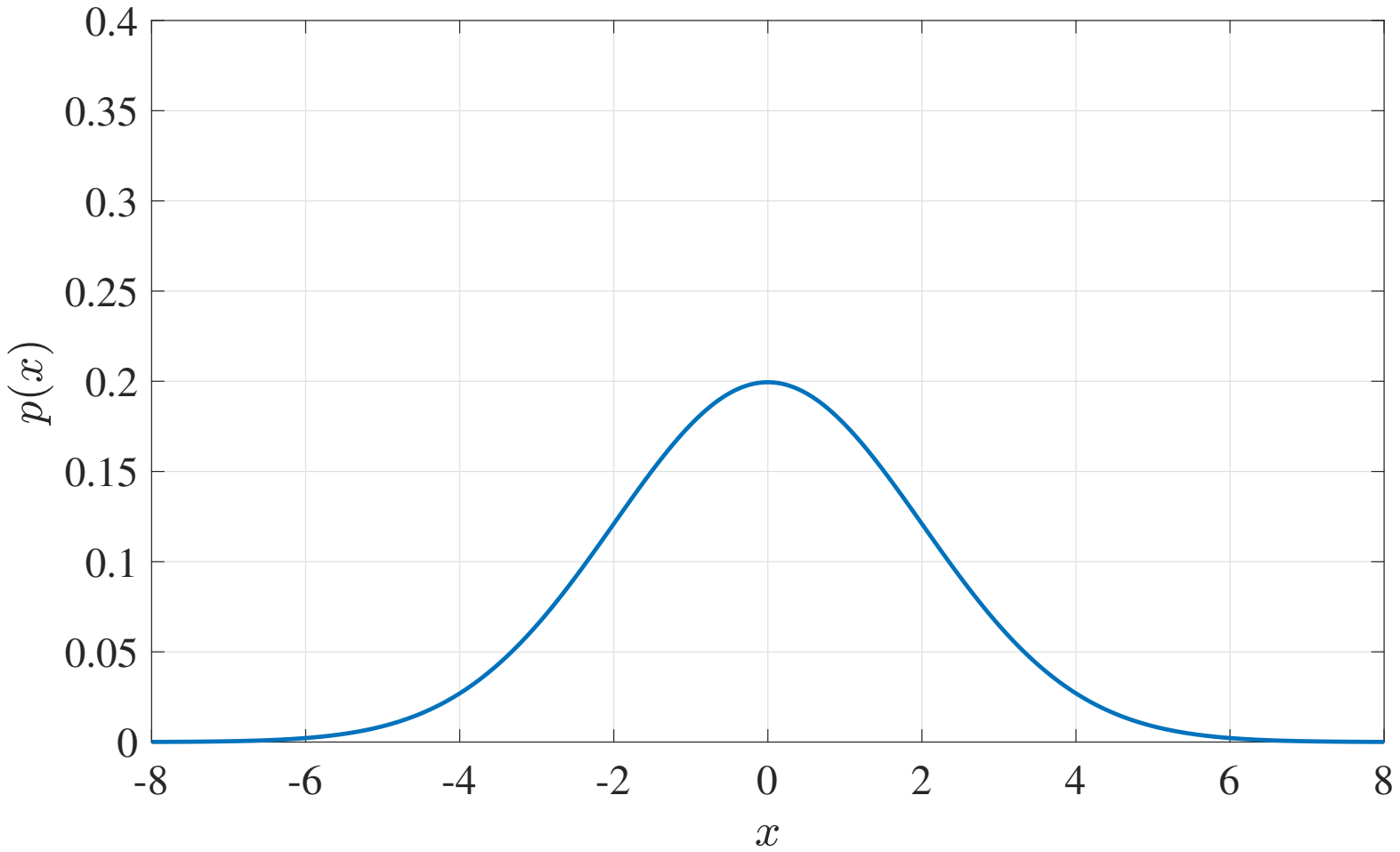
- Variance is sum of expected squared value of X, minus square of expected value of X
  $\Rightarrow$ Use this to find variance for our example on previous slide

Variance Example (1)



$$\mathbb{E}[X] = 0, \mathbb{V}[X] = 1$$

Variance Example (2)



$\mathbb{E}[X] = 0$, $\mathbb{V}[X] = 4$. Notice how the probability distribution is spread more thinly across the x-axis

Covariance/Correlation (1)

- For two variables X and Y we can define the covariance:

$$\begin{aligned} \operatorname{cov}(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \end{aligned}$$

and from this, we can define the correlation:

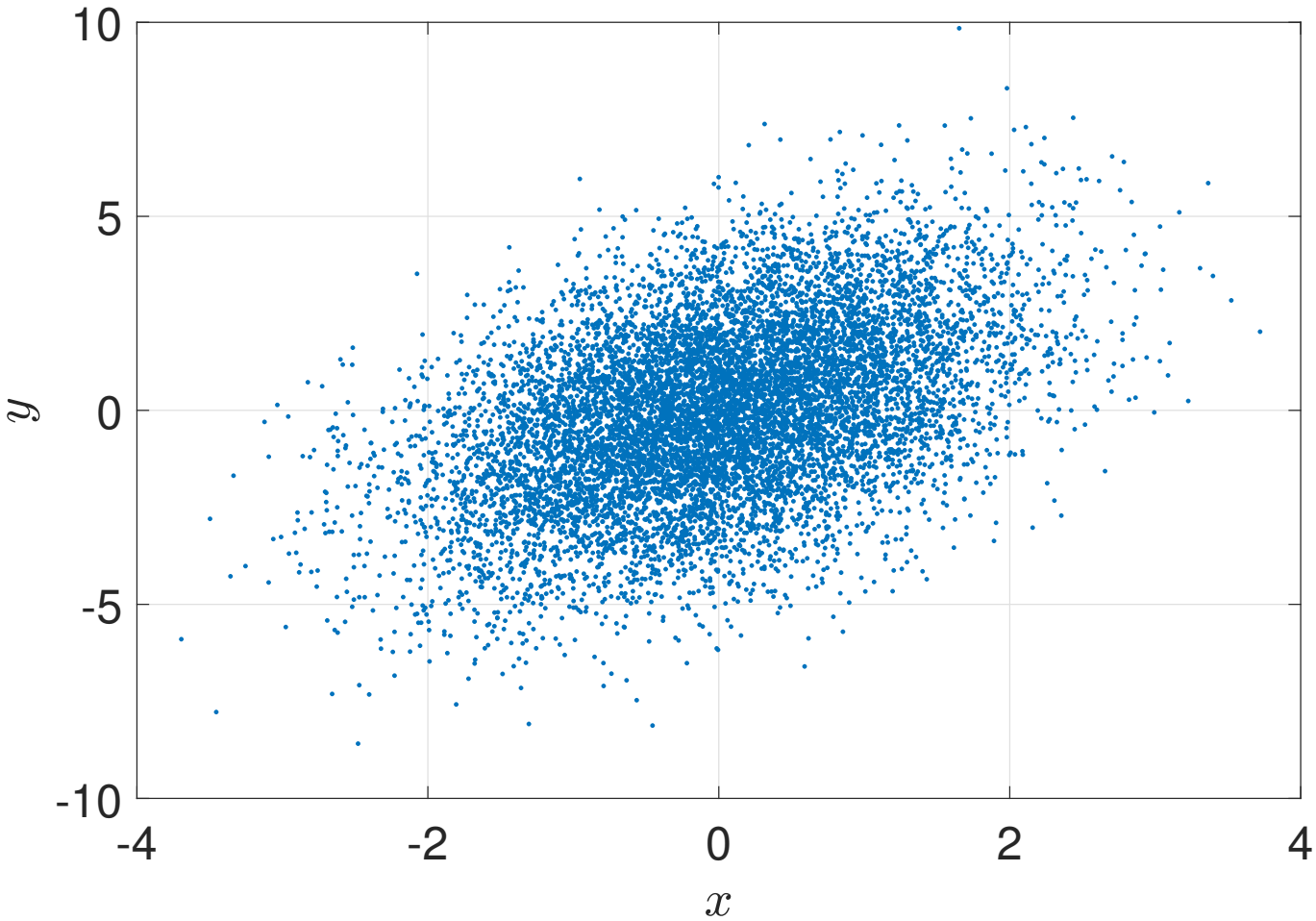
$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}}$$

$\Rightarrow$ compare to the sample correlation formula in Lecture 0

Covariance/Correlation (2)

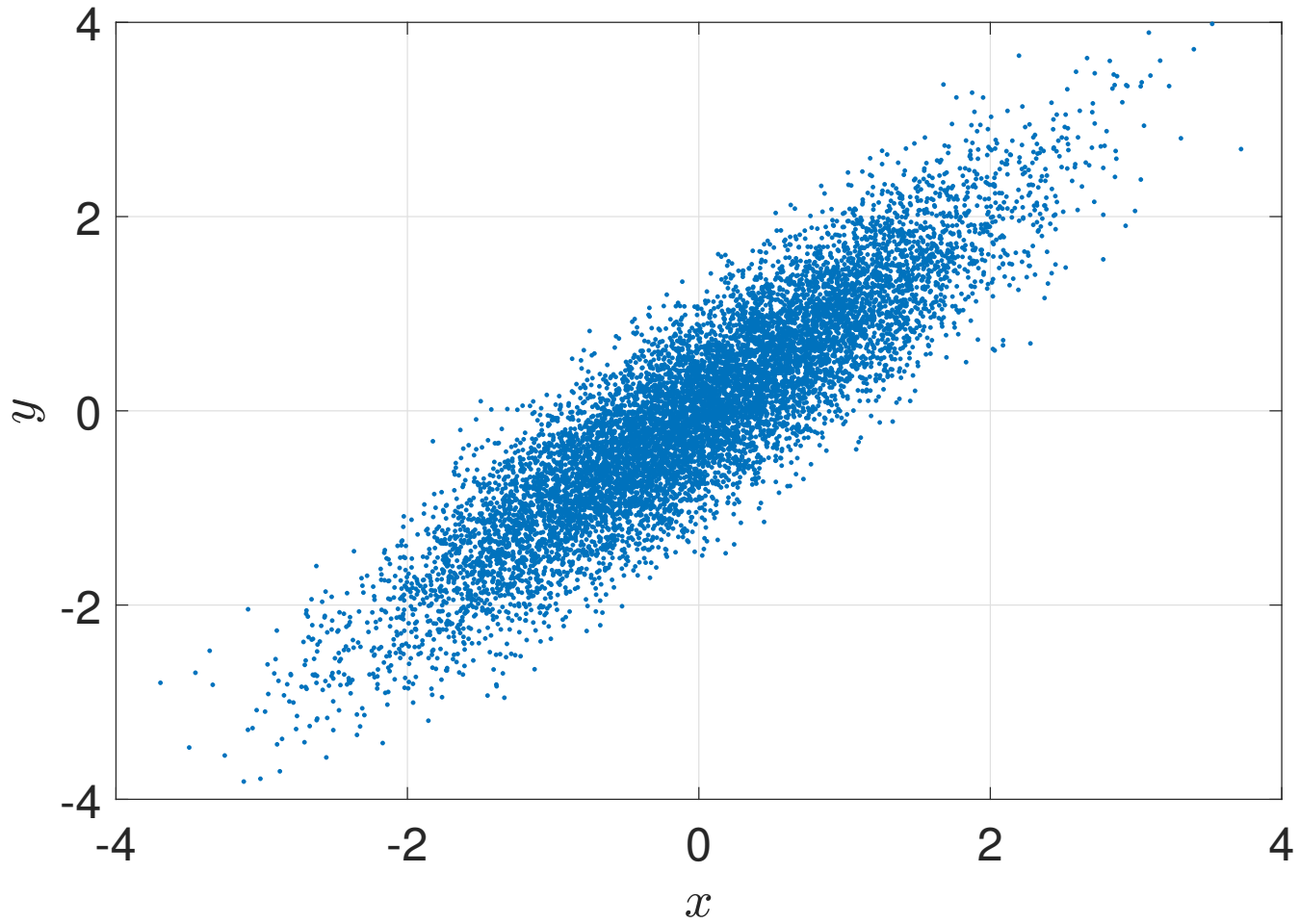
- Let x and y be realisations of the RVs X and Y
  - Positive covariance/correlation:
    $\Rightarrow$ if x greater than $\mathbb{E}[X]$ then likely y is *greater* than $\mathbb{E}[Y]$
  - Negative covariance/correlation:
    $\Rightarrow$ if x greater than $\mathbb{E}[X]$ then likely y is *less* than $\mathbb{E}[Y]$
- Covariance between $(-\infty, \infty)$,
  - Depends on scale (unit of measurement) of variables X and Y
- Correlation between $[-1, 1]$,
  - Independent of scale of variables
- If X, Y independent, $\text{cov}(X, Y) = \text{corr}(X, Y) = 0$
  $\Rightarrow$ Converse is **not** true!

Correlation/Scatter Plot Example (1)



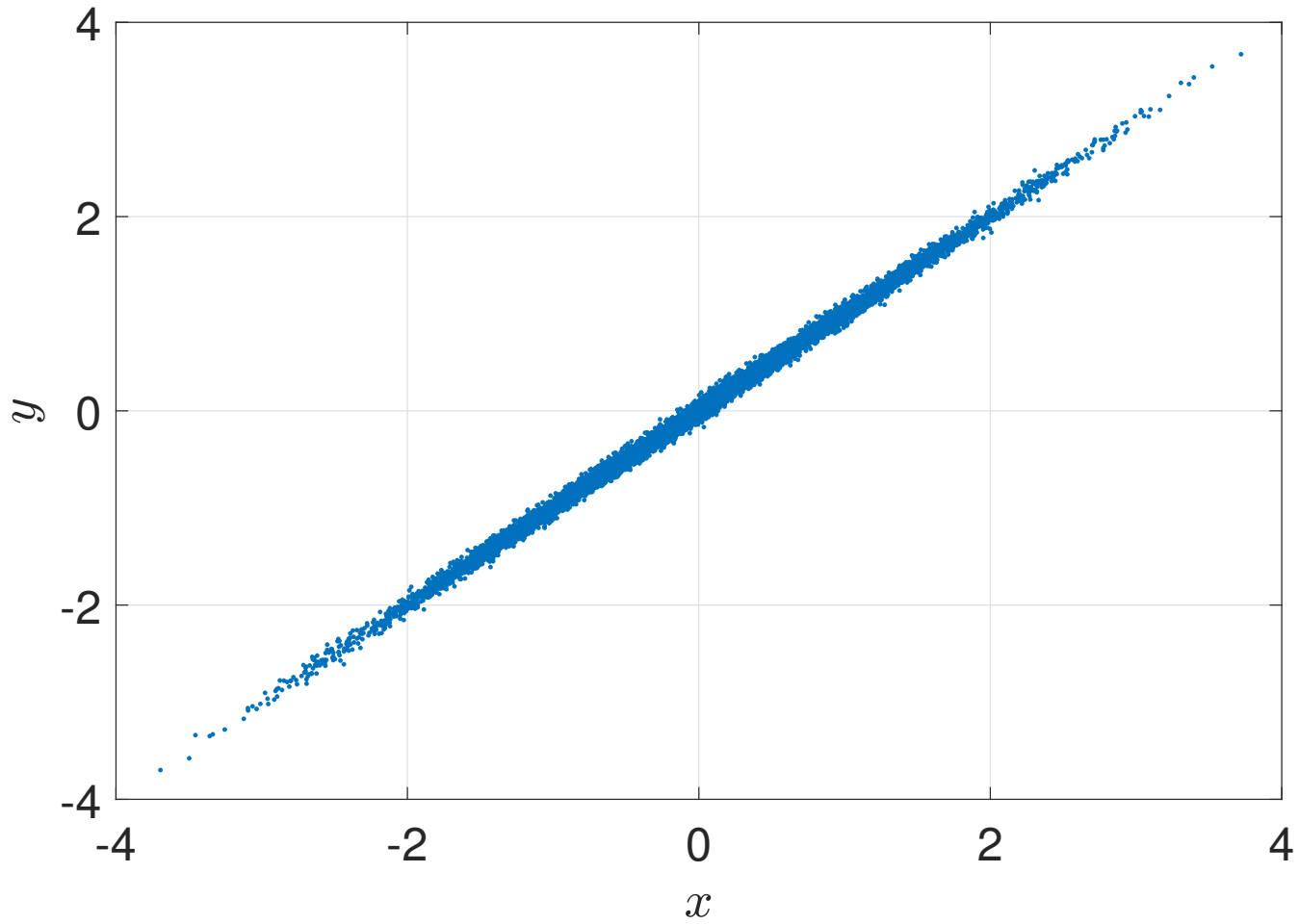
$$\text{corr}(X, Y) = 0.45$$

Correlation/Scatter Plot Example (2)



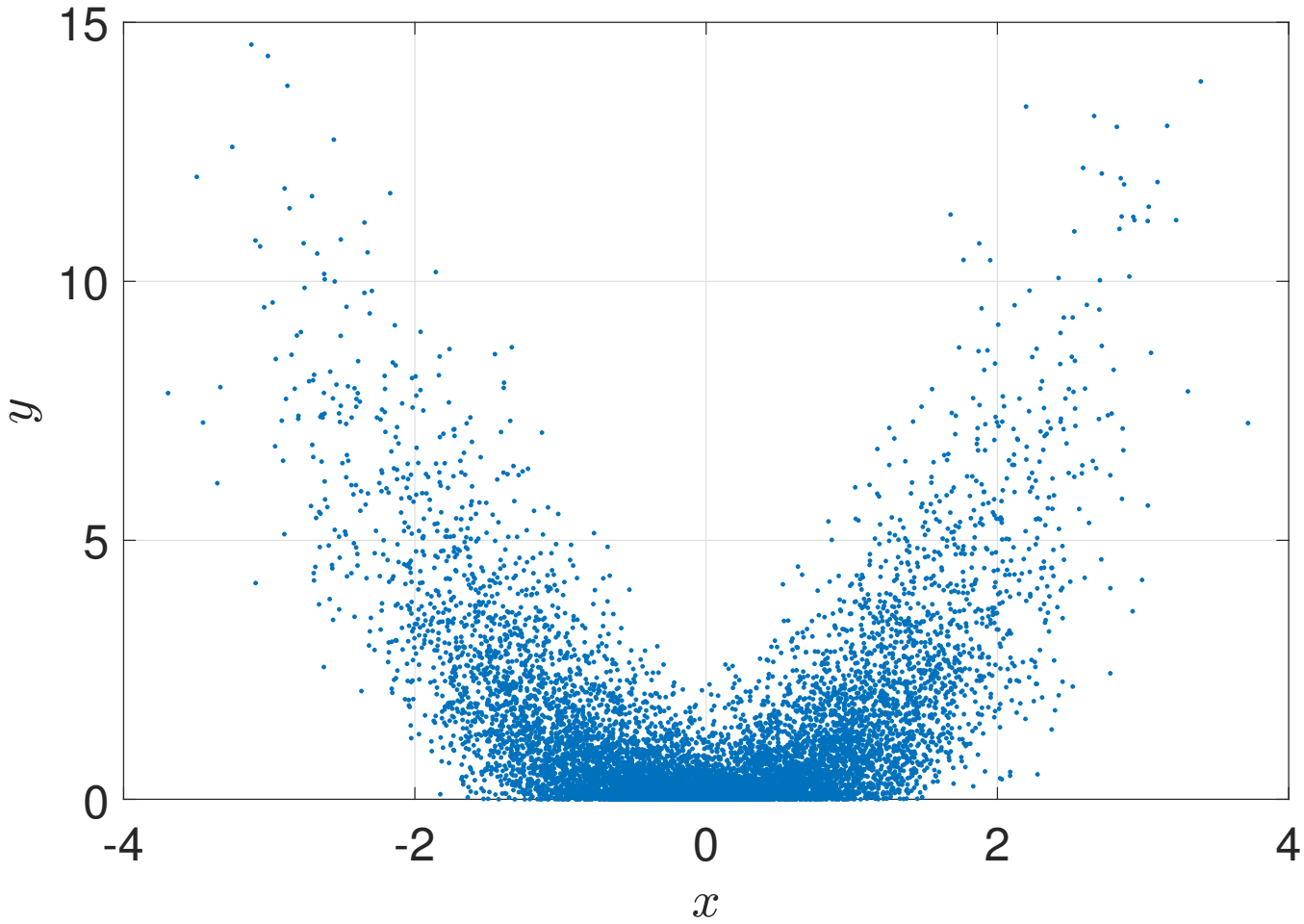
$$\text{corr}(X, Y) = 0.9$$

Correlation/Scatter Plot Example (3)



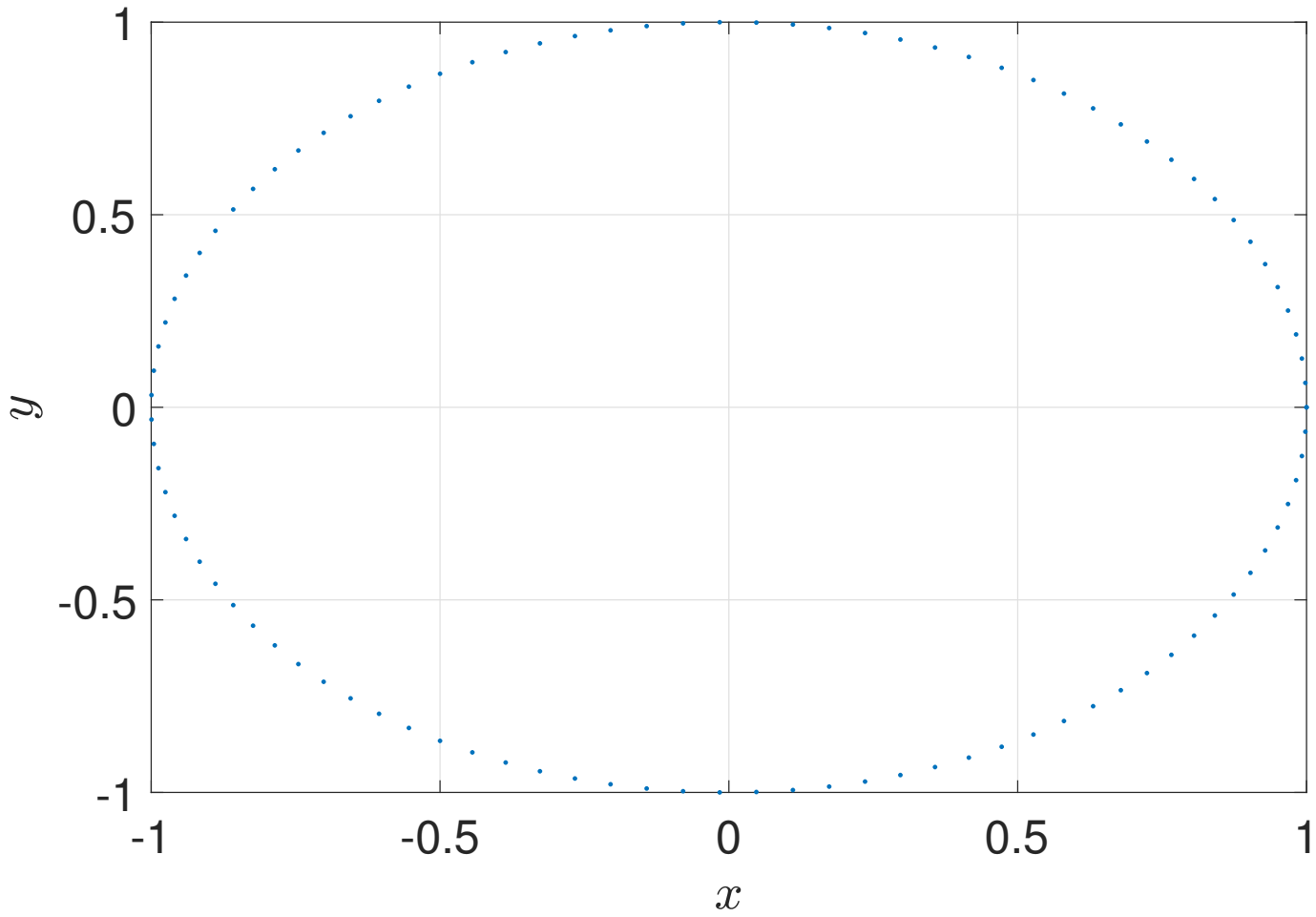
$$\mathrm{corr}(X, Y) = 0.999$$

Correlation/Scatter Plot Example (4)



$\mathrm{corr}(X, Y) = 0$ – though clearly associated, as $y = x^2 + \text{noise}$

Correlation/Scatter Plot Example (5)



$\mathrm{corr}(X, Y) = 0$, though there is a deterministic association between x and y

Expectations and Independent RVs

- In general, expectation of a function of two RVs is

$$\mathbb{E}\left[f(X,Y)\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f(x,y)p(x,y)$$

- Fact 1: Due to linearity of expectation, we have

$$\mathbb{E}\left[f(X)+g(Y)\right] = \mathbb{E}\left[f(X)\right] + \mathbb{E}\left[g(Y)\right]$$

for all RVs X and Y, and

- Fact 2: For independent RVs, we have

$$\mathbb{E}\left[f(X)g(Y)\right] = \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right]$$

implying that

$$\mathbb{V}\left[X+Y\right] = \mathbb{V}\left[X\right] + \mathbb{V}\left[Y\right]$$

for X and Y independent.

Weak Law of Large Numbers

- Consider a sample of n realisations of a RV X, say $x_1, \ldots, x_n$
- Then, we know the sample mean is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- How is this related to the theoretical mean (expectation) of X?
  - The weak law of large numbers connects these quantities

- Let $X_1, \ldots, X_n$ be RVs with $\mathbb{E}[X_i] = \mu$; then for any $\varepsilon > 0$

$$\mathbb{P}\left\{ \left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| > \varepsilon \right\} \to 0 \text{ as } n \to \infty.$$

- Informally, you can think of this result as saying that the (sample) mean of a realisation of random variables converges to the expected value as the number of realisations grows larger and larger.

Existence of Expected Values

- Expected values do not always exist
- If $\mathcal{X}$ is *finite*, then $\mathbb{E}[X]$ always exists
- However, in general, $\mathcal{X}$ will not be finite
- $\mathcal{X}$ is usually the set of integers $\mathbb{Z}$ or real numbers $\mathbb{R}$
  $\Rightarrow$ In this case, expectations are not guaranteed to exist
- In contrast, the quantiles (such as median) *always* exist
- A simple example of a density for which $\mathbb{E}[X]$ does not exist is:

$$p(x) = \frac{2}{\pi(1+x^2)}$$

Approximate Expectations of Functions of RVs (1)

- Consider a function $f(x)$ and random variable X
- In general it is the case that

$$\mathbb{E}\left[f(X)\right] \neq f(\mathbb{E}\left[X\right])$$

  unless $f(\cdot)$ is a linear function of X.
- However, frequently we need to find expectations of this kind
- Can we say anything about $\mathbb{E}\left[f(X)\right]$ and $\mathbb{V}\left[f(X)\right]$?
  - The answer is yes, if we make certain assumptions

Approximate Expectations of Functions of RVs (2)

- Let us assume that
  1. The quantities $\mu_X = \mathbb{E}[X]$ and $\sigma_X^2 = \mathbb{V}[X]$ are finite
  2. The function $f(x)$ is twice differentiable in x
- Then, we have the following results

$$\begin{aligned} \mathbb{E}[f(X)] &\approx f(\mu_X) + \frac{\sigma_X^2}{2} f''(\mu_X) \\ \mathbb{V}[f(X)] &\approx \sigma_X^2 (f'(\mu_X))^2 \end{aligned}$$

- These formulas are the result of a Taylor series expansion

Approximate Expectations of Functions of RVs (3)

- Example: Consider the transformation $f(X) = aX^2 + c$; we have

$$\frac{df(x)}{dx} = 2ax, \quad \frac{d^2 f(x)}{dx^2} = 2a$$

- Then, applying the previous formulae yields

$$\begin{aligned} \mathbb{E}\left[f(X)\right] &\approx a\mu_X^2 + c + a\sigma_X^2 \\ \mathbb{V}\left[f(X)\right] &\approx 4(a\mu_X)^2\sigma_X^2 \end{aligned}$$

$\Longrightarrow$ so the variance of the square of a RV grows with the square of the mean of the random variable

Outline

Probability Distributions as Models

- We can use probability distributions as models of reality
- For example, imagine we knew the distribution of heights of people in Australia, say $p(h)$
- We could then make predictions/statements about heights
  - For example, the proportion of people taller than $1.7m$

$$\mathbb{P}(H > 1.7) = \int_{1.7}^{\infty} p(h)dh$$

- Then a clothing company could use this information to determine how much product of different sizes to stock

Parametric Probability Distributions (1)

- So far we have built probability distributions by directly specifying the probabilities for each element $x \in \mathcal{X}$
- This is fine if $\mathcal{X}$ is a small finite set
- But if $\mathcal{X}$ is large, or infinite (for example, all the integers), this approach no longer works
- Instead it is usual to use parametric probability distributions
- We will look at several important distributions:
  - The Gaussian distribution;
  - The Bernoulli distribution;
  - The binomial distribution;
  - The uniform distribution;
  - The Poisson distribution.

Parametric Probability Distributions (2)

- We specify the probability density function by

$$p(x \,|\, \boldsymbol{\theta}), \quad x \in \mathcal{X}, \; \boldsymbol{\theta} \in \Theta$$

or, for discrete RVs we use the shorthand notation:

$$\mathbb{P}(X = x \,|\, \boldsymbol{\theta}) \equiv p(x \,|\, \boldsymbol{\theta}), \quad x \in \mathcal{X}, \; \boldsymbol{\theta} \in \Theta$$

where

  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ are the parameters that control distribution of the probabilities;
  - Θ is the set of valid parameters for the model.

$\Rightarrow$ by changing $\boldsymbol{\theta}$ we can change the distribution.

- Usually, the number of parameters $k \ll |\mathcal{X}|$

Parametric Probability Distributions (3)

- The properties of the RV are determined by $p(x \mid \boldsymbol{\theta})$
- For example, the mean is

$$\mathbb{E}[X] = f(\boldsymbol{\theta}),$$

  where $f(\cdot)$ is a function that depends on $\boldsymbol{\theta}$ and $p(x \mid \boldsymbol{\theta})$.

- The same applies to the variance, cdf, quantiles, etc.

- The parameterisation will not be unique
  $\Rightarrow$ there are often several common parameterisations for the same distribution

Gaussian Distribution (1)

- Let's begin with the case that $\mathcal{X} = \mathbb{R}$
  $\Rightarrow$ that is, we want a distribution over all the real numbers
- Probably the most important distribution for real numbers is the Gaussian (normal) distribution
  $\Rightarrow$ named after Carl Friedrich Gauss (1777-1855)
- The pdf for a Gaussian distribution is given by

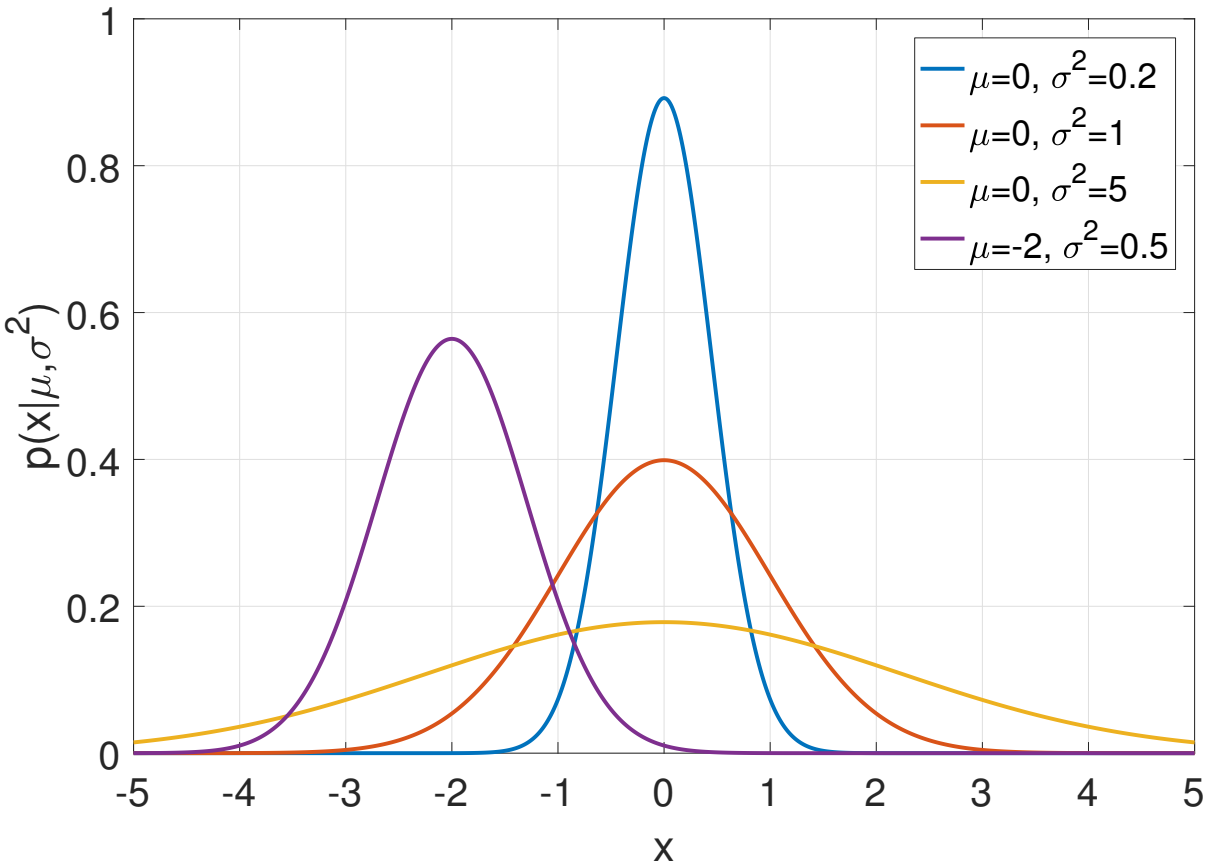
$$p(x \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where

- μ is the mean of the distribution;
- σ^2 is the variance of the distribution;

so that $\boldsymbol{\theta} = (\mu, \sigma^2)$ for the Gaussian distribution.

Gaussian Distribution (2)



Probability density functions for several normal (Gaussian) distributions. The orange curve is the *standard normal distribution*. Note that the normal distribution is symmetric and tails off to zero as $|x| \to \infty$.

Gaussian Distribution (3)

- If X follows a Gaussian distribution, we write that

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

  where "$\sim$" is read as "is distributed per a"
- An important property of Gaussian RVs is self-similarity
- Every Gaussian distribution is a translated and scaled version of the standard normal distribution $N(0, 1)$
- If $Z \sim N(0, 1)$, then

$$X = \sigma Z + \mu$$

  is distributed as per $N(\mu, \sigma^2)$

Gaussian Distribution (4)

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

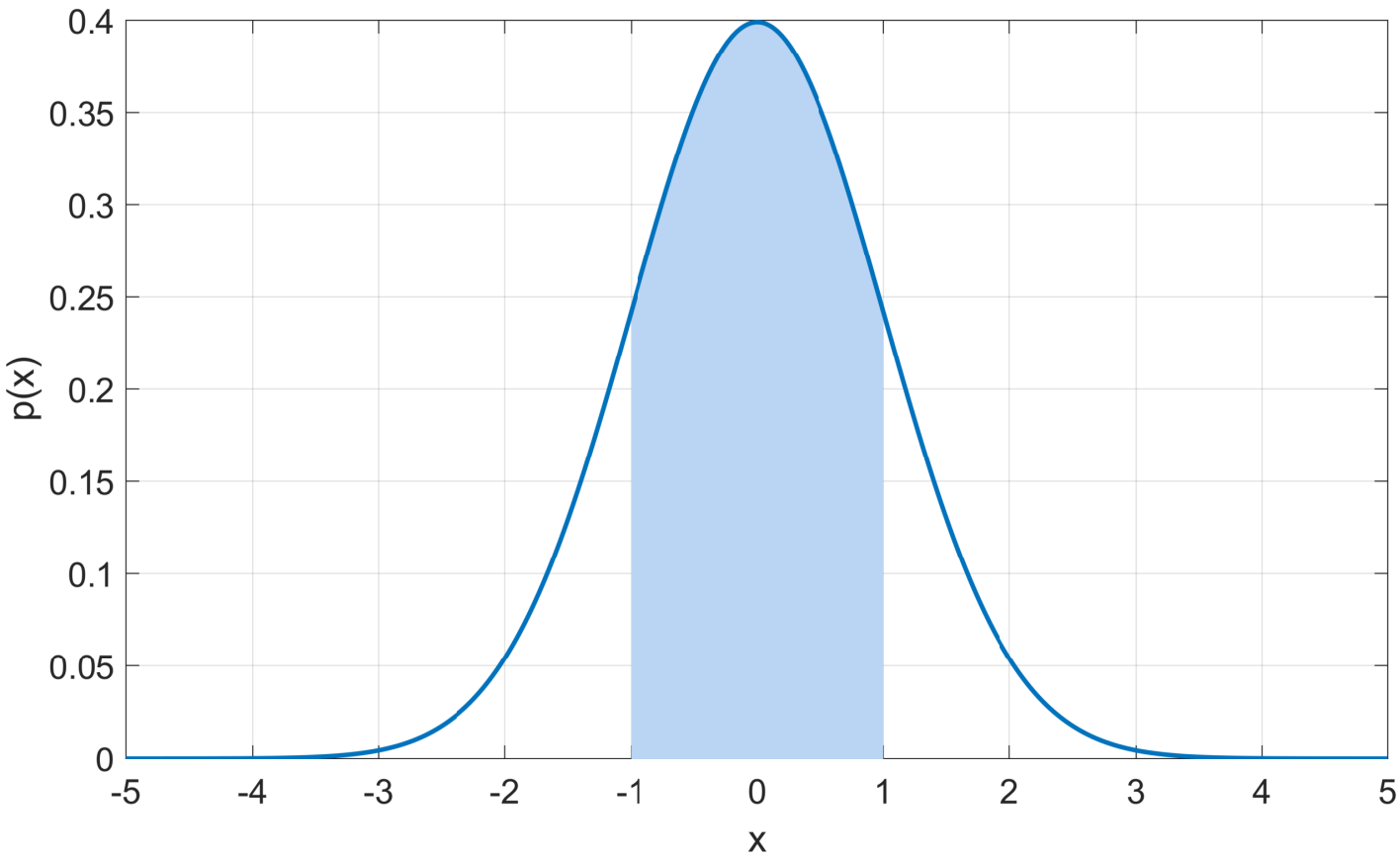
$$\begin{aligned} \mathbb{E}[X] &= \mu, \\ \mathbb{V}[X] &= \sigma^2. \end{aligned}$$

- The Gaussian distribution is symmetric around μ, so that:
  - its mode is μ;
  - its median is μ.

- The cdf for the Gaussian has no closed form
  - Most packages have algorithms to evaluate it numerically
  - There are some well known rules regarding the cdf
  - They are useful as they apply for all μ and σ

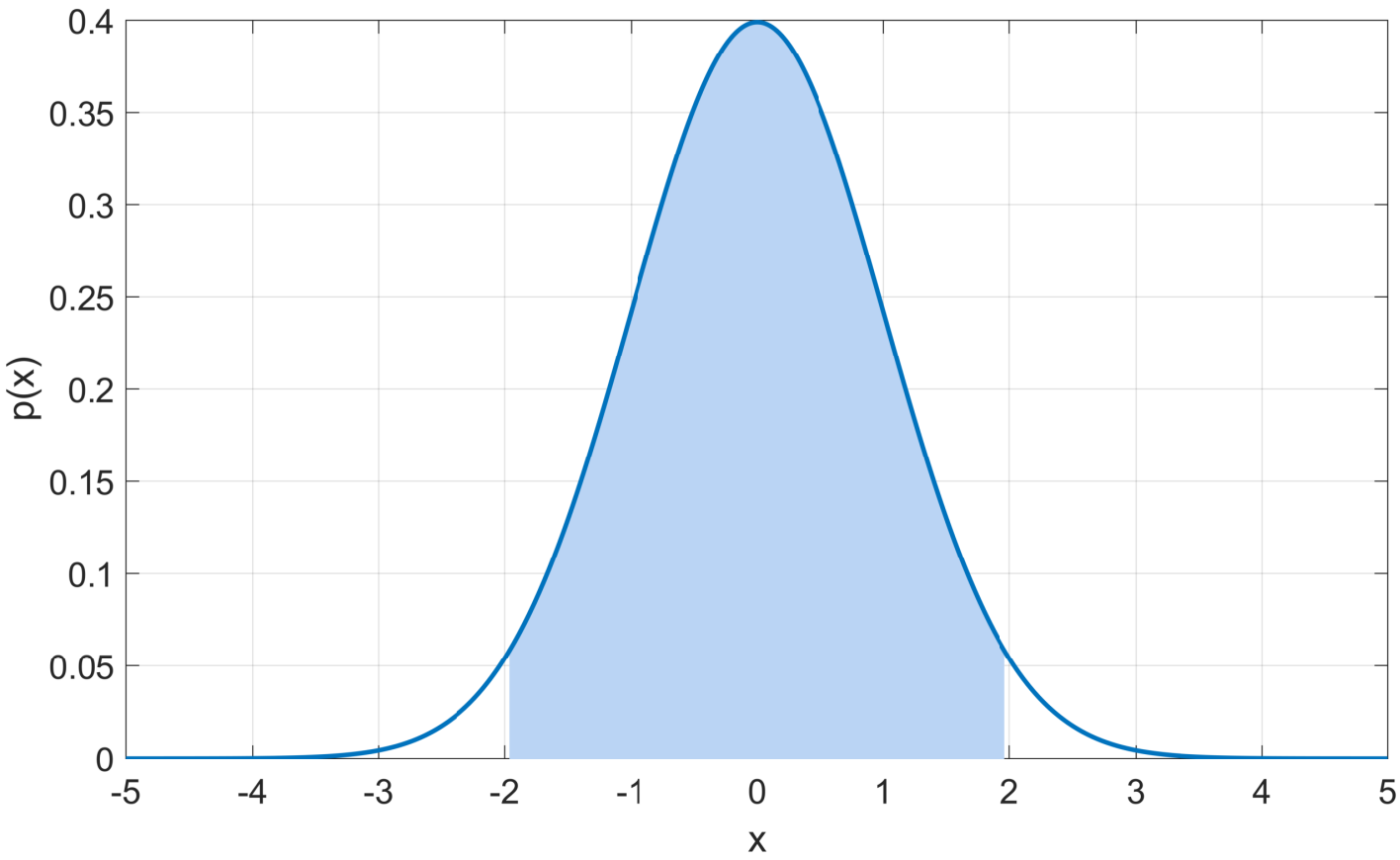
Gaussian Distribution (5)

- For any $N(\mu, \sigma^2)$:
  - 68.27% of probability falls within $(\mu - \sigma, \mu + \sigma)$



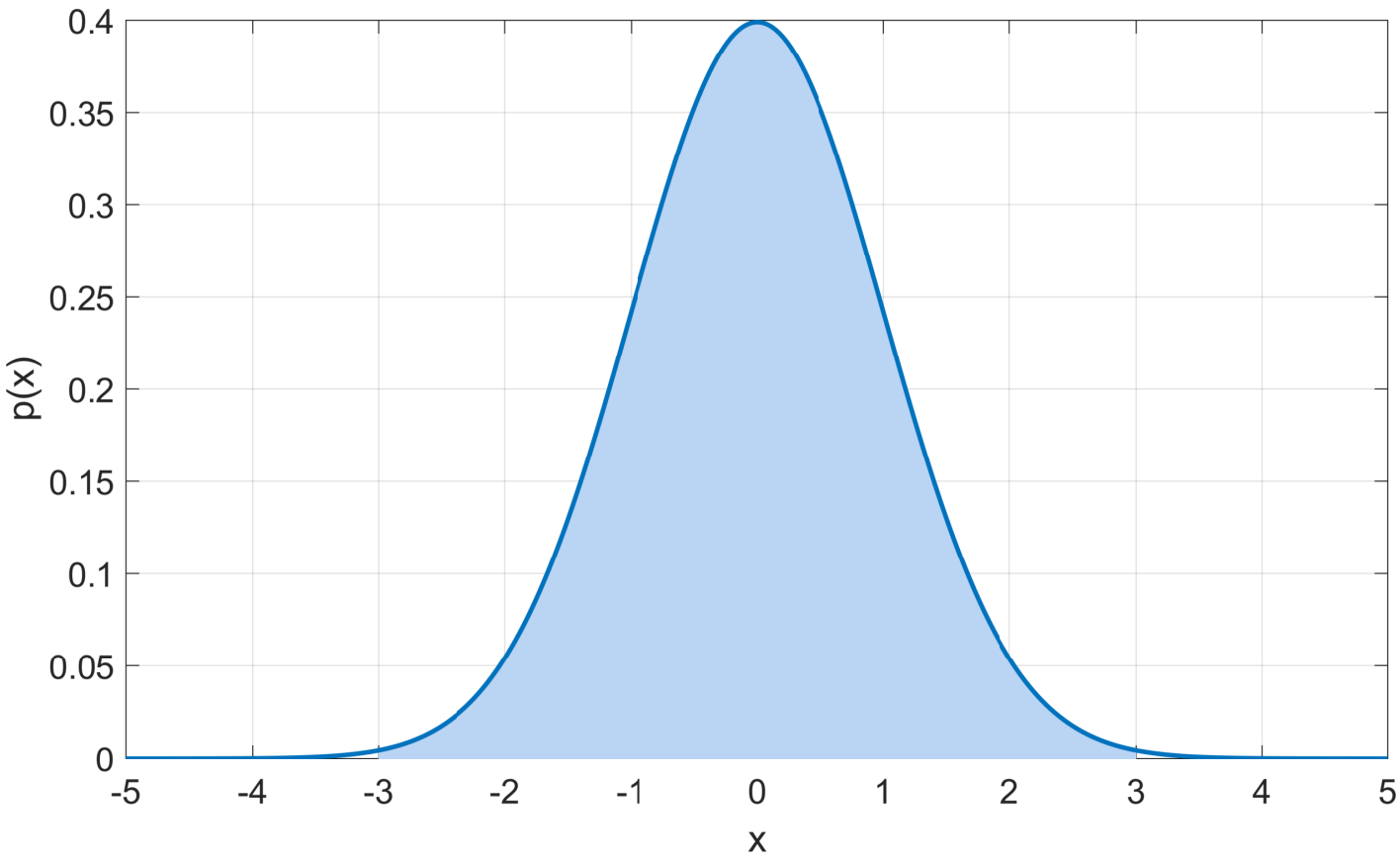
Gaussian Distribution (6)

- For any $N(\mu, \sigma^2)$:
  - 95.45% of probability falls within $(\mu - 2\sigma, \mu + 2\sigma)$



Gaussian Distribution (7)

- For any $N(\mu, \sigma^2)$:
  - 99.73% of probability falls within $(\mu - 3\sigma, \mu + 3\sigma)$



Gaussian Distribution (8)

- The final property we will examine is additivity
- If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ then

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

- More generally, if $X \sim N(\mu, \sigma^2)$, for all $n \geq 1$, then

$$X = \sum_{i=1}^{n} X_i$$

where $X_i \sim N(\mu_i, \sigma_i^2)$ and

$$\sum_{i=1}^{n} \mu_i = \mu, \quad \sum_{i=1}^{n} \sigma_i^2 = \sigma^2.$$

$\Longrightarrow$ we can decompose a normal RV into a sum of an arbitrary number of appropriate normally distributed RVs

Bernoulli Distribution (1)

- Let's consider the case of discrete, binary RVs, i.e., $\mathcal{X} = \{0, 1\}$
- The Bernoulli distribution models these variables

$$\mathbb{P}(X = 1 \,|\, \theta) = \theta, \; \theta \in [0, 1]$$

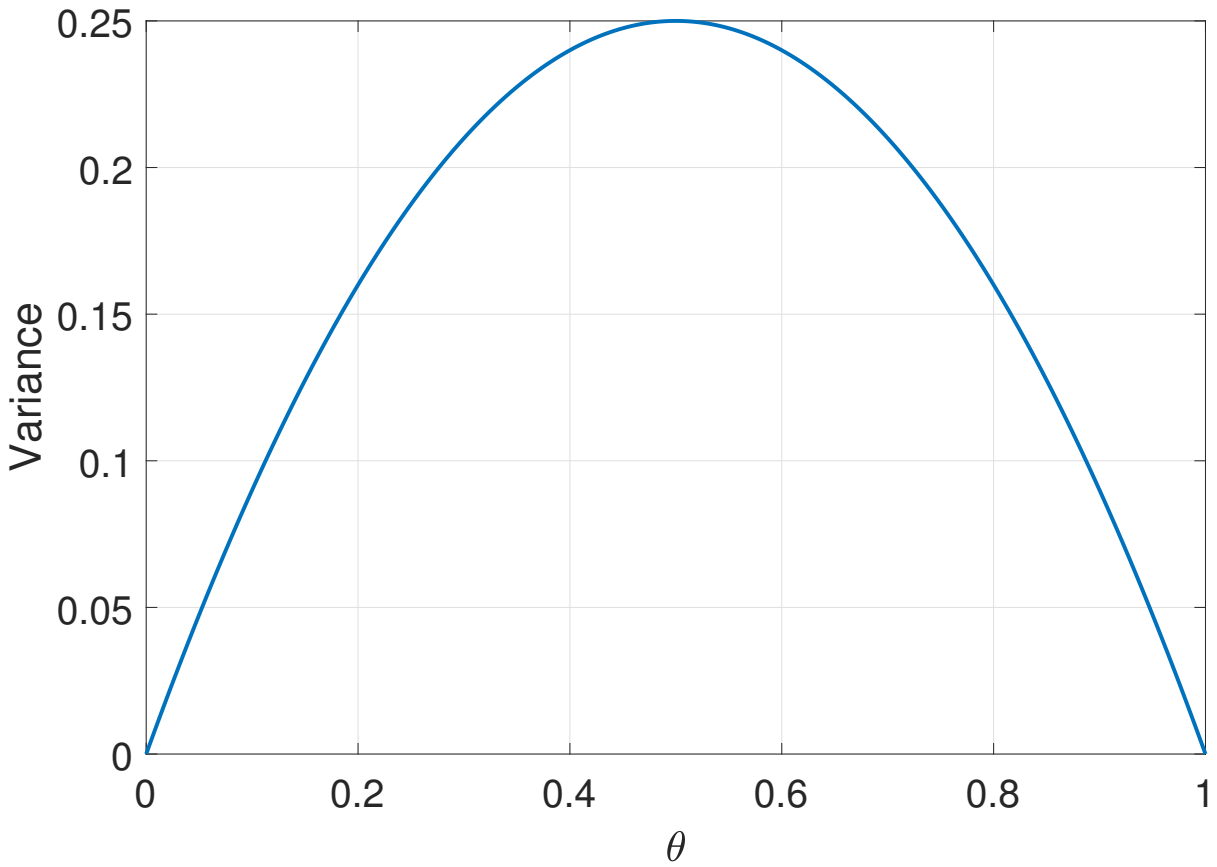
so that the parametric probability distribution follows:

$$p(x \,|\, \theta) = \theta^x (1 - \theta)^{(1-x)}$$

- The parameter θ is the probability of observing a "success"
- If X follows a Bernoulli distribution, we write $X \sim \mathrm{Be}(\theta)$
- It is easy to see that

$$\begin{aligned} \mathbb{E}[X] &= \theta \\ \mathbb{V}[X] &= \theta(1 - \theta) \end{aligned}$$

Bernoulli Distribution (2)



Variance of a Bernoulli random variable as a function of θ. The variance is maximum when $\theta = 1/2$ and smallest for $\theta = 0$ and $\theta = 1$.

Binomial Distribution (1)

- Now consider n binary RVs $\mathbf{X} = (X_1, \ldots, X_n)$.
  - Example realisation: $\mathbf{x} = (0, 1, 1, 1, 0, 1, 0, 0, 1, 1)$
- The sum

$$m(\mathbf{x}) \equiv m = \sum_{j=1}^{n} x_j$$

  counts the number of "successes"
  $\Rightarrow$ in our example, $m = 6$
- Given n, the count is a RV, say M, over the sample space $\{0, 1, 2, \ldots, n\}$

Binomial Distribution (2)

- The binomial distribution describes the probability that M takes a particular value m

$$p(m \,|\, \theta) = \binom{n}{m} \prod_{i=1}^{n} p(x_i \,|\, \theta) = \binom{n}{m} \theta^m (1-\theta)^{(n-m)}$$

where

$$\binom{n}{m} = \frac{n!}{(n-m)!m!}$$

is the number of ways of choosing m objects out of n identical objects (the binomial coefficient)
$\Rightarrow m! = 1 \times 2 \times 3 \times \ldots \times m$ is the factorial function

- This captures the fact that, for $1 \leq m \leq (n-1)$ there are multiple sequences with m successes out of n trials

Binomial Distribution (3)

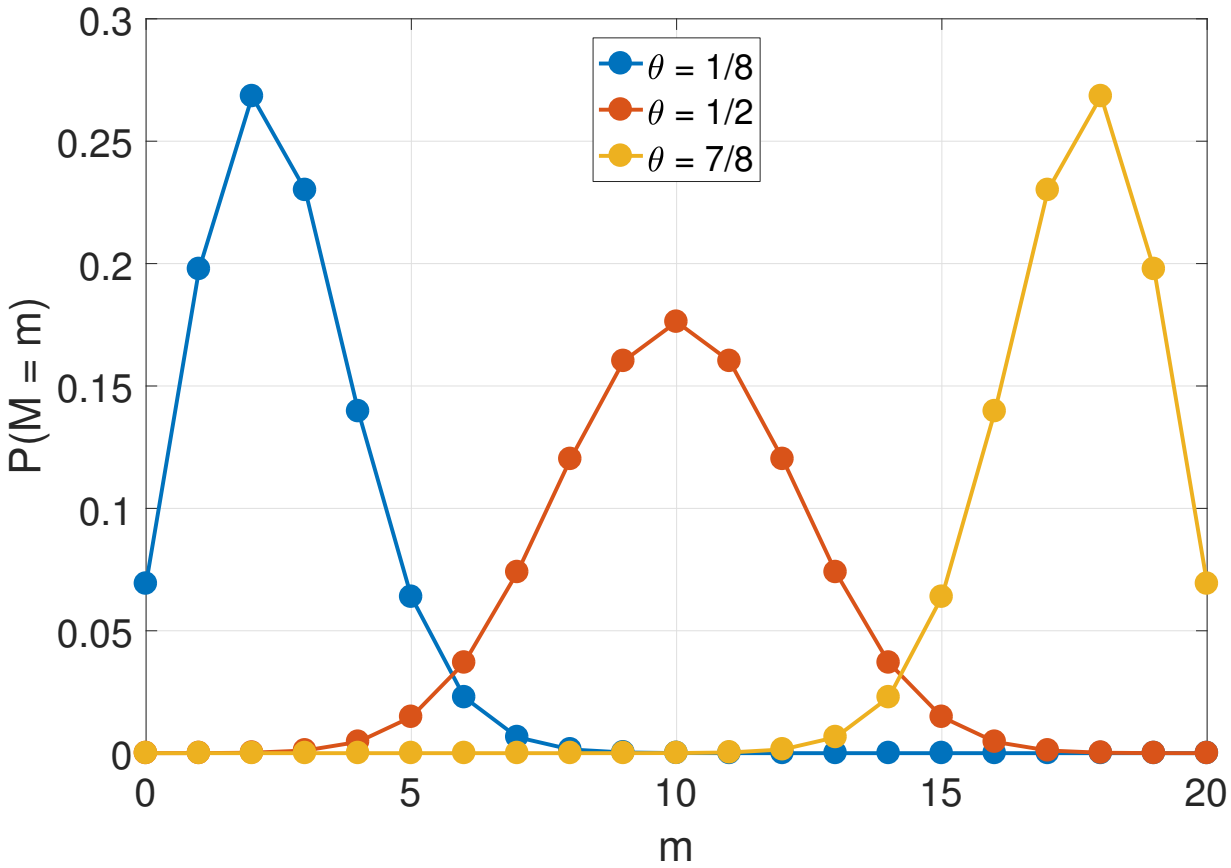
- Example: The following six sequences have $m = 2$ successes out of $n = 4$ trials:

$$1, 1, 0, 0$$
$$1, 0, 1, 0$$
$$1, 0, 0, 1$$
$$0, 1, 1, 0$$
$$0, 1, 0, 1$$
$$0, 0, 1, 1$$

so that

$$p(m = 2 \,|\, \theta) = \binom{n = 4}{m = 2} \theta^2 (1 - \theta)^{(4-2)}$$

Binomial Distribution (4)



Binomial distribution for $n = 20$ and $\theta = 1/8$, $\theta = 1/12$, $\theta = 7/8$. The distribution is defined only on the integers – the connecting lines are only guides for the eye. Note that $\theta = 7/8$ is a mirror of $\theta = 1/8$.

Binomial Distribution (5)

- If M follows a binomial distribution, we write

$$M \sim \text{Bin}(\theta, n)$$

- As m is a sum of independent Bernoulli RVs we have

$$\begin{aligned} \mathbb{E}[M] &= n\theta \\ \mathbb{V}[M] &= n\theta(1-\theta) \end{aligned}$$

- If $M_1 \sim \text{Bin}(\theta, n_1)$ and $M_2 \sim \text{Bin}(\theta, n_2)$ then

$$M_1 + M_2 \sim \text{Bin}(\theta, n_1 + n_2)$$

This follows from the definition of the binomial distribution as the sum of n Bernoulli variates with success probability θ

Discrete Uniform Distribution (1)

- The binomial distribution gives a distribution over a bounded set of integers $\{0, 1, \ldots, n\}$
  - Probability is skewed towards $n\theta$
- What if we believe all outcomes equally likely?
  - Classic example would be roll of a fair six-sided dice
- Then we can use the uniform distribution

$$\mathbb{P}(X = k \,|\, a, b) = \frac{1}{b - a + 1}$$

where $X \in \{a, \ldots, b\}$ with $b \geq a$

Discrete Uniform Distribution (2)

- If X follows a uniform distribution we write

$$X \sim U(a, b)$$

- The mean and variance of such a RV is

$$\begin{aligned} \mathbb{E}[X] &= \frac{a+b}{2} \\ \mathbb{V}[X] &= \frac{(b-a+1)^2 - 1}{12} \end{aligned}$$

  - Note that $\mathbb{E}[X]$ may not be an integer

- Discrete uniform can be generalized to any set of integers $\mathcal{X} \subseteq \mathbb{Z}$

$$\mathbb{P}(X = k \,|\, \mathcal{X}) = \frac{1}{|\mathcal{X}|}$$

where $|\mathcal{X}|$ is the number of elements in $\mathcal{X}$

Continuous Uniform Distribution (1)

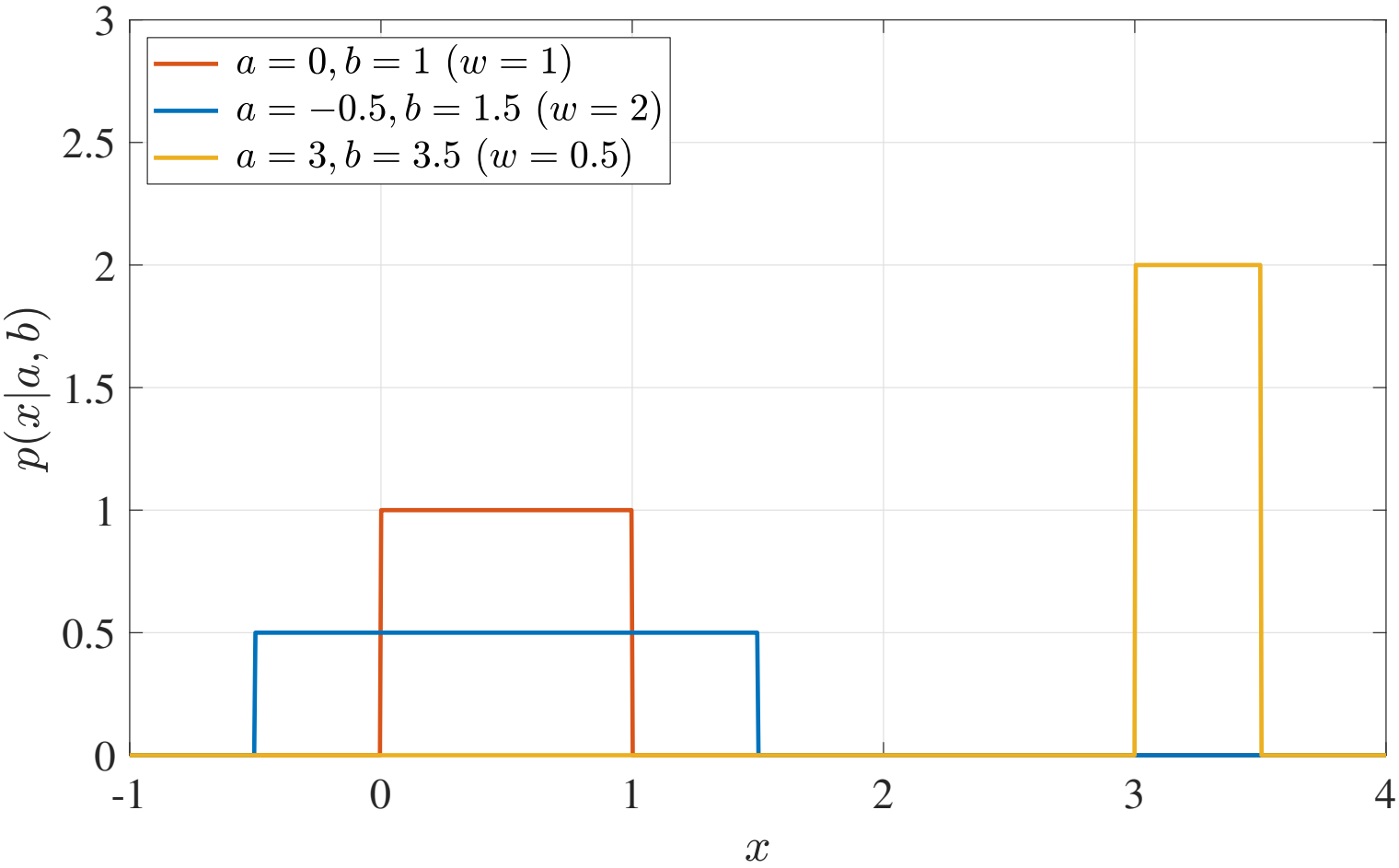
- What if X is a continuous random variable?
- We can define a continuous version of the uniform with pdf

$$p(x \mid a, b) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x > b \end{cases} .$$

where $a > b$

  - The quantity a determines the start of the distribution
  - The quantity $w = b - a$ is the width of the distribution

Continuous Uniform Distribution (2)



Uniform distributions for $a = 0$, $b = 1$ $(w = 1)$; $a = -0.5$, $b = 1.5$ $(w = 2)$, and $a = 3$, $b = 3.5$ $(w = 0.5)$. The connecting vertical lines are guides for the eye only.

Continuous Uniform Distribution (3)

- If X follows a uniform distribution between a and b we say

$$X \sim U(a, b)$$

- The mean of the distribution is given by

$$\mathbb{E}[X] = \frac{a+b}{2} = a + \frac{w}{2}$$

- The variance of the distribution is given by

$$\mathbb{V}[X] = \frac{(b-a)^2}{12} = \frac{w^2}{12}$$

Poisson Distribution (1)

- What if our data is non-negative integers; for example:
  - number of telephone calls made in an hour
  - number of people kicked to death by horses in a year
  - Sample space is then $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$
- One suitable distribution is the Poisson distribution
  - Named after Simeon Poisson (1781–1840)
- Has the form

$$p(k \mid \lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

where λ is often called the *rate*.

Poisson Distribution (2)

- If X is distributed per a Poisson distribution we write

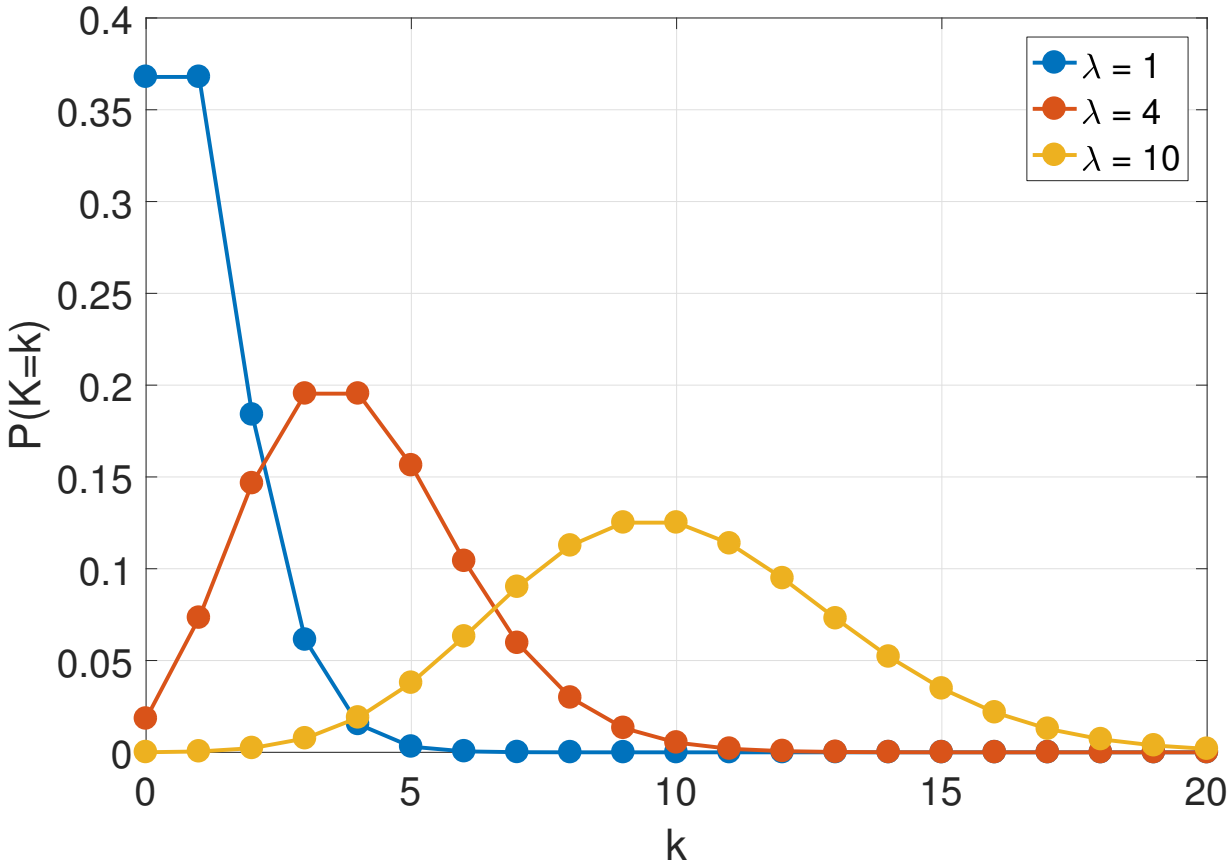
$$X \sim \text{Pois}(\lambda)$$

- The Poisson distribution has

$$\begin{aligned} \mathbb{E}[X] &= \lambda \\ \mathbb{V}[X] &= \lambda \end{aligned}$$

- The Poisson distribution is an example of a distribution in which the variance grows with the mean

Poisson Distribution (3)



Poisson distribution for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$. The distribution is defined only on the integers – the connecting lines are only guides for the eye.

Poisson Distribution (4)

- The Poisson distribution models the number of events in an interval of time
- When is the Poisson appropriate?
  - The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
  - The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals.
  - Two events cannot occur at exactly the same instant.
  - The probability of an event in a small interval is proportional to the length of the interval.

(taken from Wikipedia)

Poisson Distribution (5)

- If $X_1 \sim \text{Poi}(\lambda_1)$ and $X_2 \sim \text{Poi}(\lambda_2)$ then

$$X_1 + X_2 \sim \text{Poi}(\lambda_1 + \lambda_2)$$

and if $X \sim \text{Poi}(\lambda)$, then $X = \sum_{i=1}^{n} X_i$ where

$$X_i \sim \text{Poi}(\lambda_i), \quad \sum_{i=1}^{n} \lambda_i = \lambda$$

- Let $X_T \sim \text{Poi}(\lambda)$ be the distribution of events occuring in a time period T
  - What is the distribution of events $X_{T/k}$ occuring in the period T/k?
- From the result above we have

$$X_{T/k} \sim \text{Poi}(\lambda/k)$$

Reading/Terms to Revise

- Reading for this week: Chapters 4 and 5 of Ross.
- Terms you should know:
  - Random variable;
  - Conditional Probability;
  - Probability density function;
  - Expectations;
  - Variance and co-variance;
  - Normal, Bernoulli, binomial and Poisson distributions
  - Law of large numbers