Mathematics for Machine Learning

Lecture Outline

- Linear algebra
  - Vectors
  - Matrices
  - Eigen decomposition
- Probability
  - Random variables
  - Probability distributions

Notation

a, b, c	Scalar (integer or real)
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	Vector (bold-font, lower case)
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	Matrix (bold-font, upper-case)
$\boldsymbol{\mathsf{A}}, \boldsymbol{\mathsf{B}}, \boldsymbol{\mathsf{C}}$	Tensor ((bold-font, upper-case)
X, Y, Z	Random variable (normal font, upper-case)
$a \in \mathcal{A}$	Set membership: a is member of set $\mathcal{A}$
$\|\mathcal{A}\|$	Cardinality: number of items in set $\mathcal{A}$
$\\|\mathbf{v}\\|$	Norm of vector $\mathbf{v}$
$\mathbf{u} \cdot \mathbf{v}$ or $\langle \mathbf{u}, \mathbf{v} \rangle$	Dot product of vectors $\mathbf{u}$ and $\mathbf{v}$
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^n$	Real numbers space of dimension n
$y = f(x)$ or $x \mapsto f(x)$	Function (map): assign a unique value $f(x)$ to each input value x
$f: \mathbb{R}^n \to \mathbb{R}$	Function (map): map an n-dimensional vector into a scalar

Notation

$\mathbf{A} \odot \mathbf{B}$	Element-wise product of matrices **A** and **B**
$X \sim P$	Random variable X has distribution P
$P(X \mid Y)$	Probability of X given Y
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2
$\mathbb{E}_{X \sim P}[f(X)]$	Expectation of $f(X)$ with respect to $P(X)$
$\mathrm{Var}\big(f(X)\big)$	Variance of $f(X)$
$\mathrm{Cov}\big(f(X), g(Y)\big)$	Covariance of $f(X)$ and $g(Y)$
$\mathrm{corr}(X, Y)$	Correlation coefficient for X and Y
$D_{KL}(P \| Q)$	Kullback-Leibler divergence for distributions P and Q
$CE(P, Q)$	Cross-entropy for distributions P and Q

Vectors

- ***Vector:*** vector is a quantity possessing both magnitude and direction; it is represented by a one-dimensional array of ordered real-valued scalars. Vectors are written in column form or in row form
  - Denoted by bold-font lower-case letters

$$\mathbf{x} = \begin{bmatrix} 1 \\ 4 \\ 0 \\ 1 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 1 & 4 & 0 & 1 \end{bmatrix}^T$$

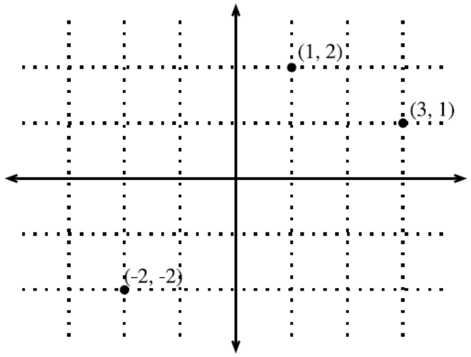
- For a general form vector with n elements, the vector lies in the n-dimensional space $\mathbf{x} \in \mathbb{R}^n$

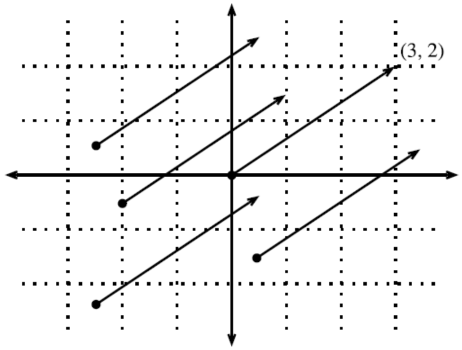
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Geometry of Vectors

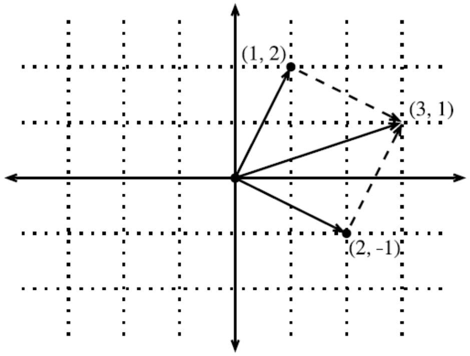
Vectors

- Interpretation of a vector: can be represented as point in space or direction in space
  - For point in space, we can visualize the data points with respect to a coordinate origin
  - For direction in space, the vector $\vec{\mathbf{v}} = [3, 2]^T$ has a direction of 3 steps to the right and 2 steps up
- The geometric interpretation of vectors as points in space allow us to consider a training set of input examples in ML as a collection of points in space





- Vector addition
  - We add the coordinates, and follow the directions given by the two vectors that are added



Picture from: http://d2l.ai/chapter_appendix-mathematics-for-deep-learning/geometry-linear-algebraic-ops.html#geometry-of-vectors

Norm of a Vector

- A vector ***norm*** is a function that maps a vector to a scalar value
  - The norm is a measure of the size of the vector
- The norm f should satisfy the following properties:
  - Scaling: $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$
  - Triangle inequality: $f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$
  - Must be non-negative: $f(\mathbf{x}) \geq 0$
- The general ℓ_p norm of a vector $\mathbf{x}$ is obtained as: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$
  - On next page we will review the most common norms, obtained for $p = 1, 2,$ and ∞

Norm of a Vector

- For $p = 2$, we have ℓ_2 norm
  - Also called **Euclidean norm**
  - It is the most often used norm
  - ℓ_2 norm is often denoted just as $\|\mathbf{x}\|$ with the subscript 2 omitted

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$$

- For $p = 1$, we have ℓ_1 norm
  - Uses the absolute values of the elements
  - Also known as Manhattan Distance or Taxicab norm

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

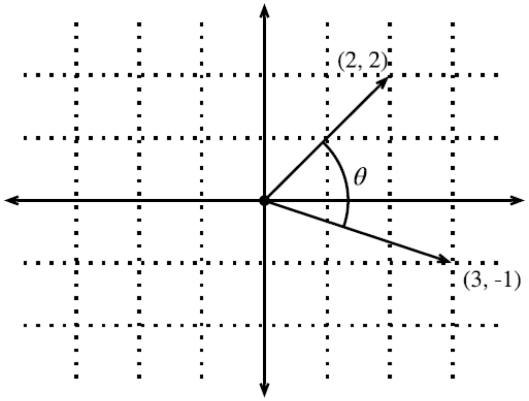
- For $p = \infty$, we have ℓ_∞ norm
  - Known as **infinity norm**, or **max norm**
  - Outputs the absolute value of the largest element

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

- ℓ_0 norm outputs the number of non-zero elements
  - It is not an ℓ_p norm, and it is not really a norm function either (it is incorrectly called a norm)

Dot Product and Angles

- *Dot product* of vectors, $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v} = \sum_i u_i \cdot v_i$
  - It is also referred to as inner product, or scalar product of vectors
  - The dot product $\mathbf{u} \cdot \mathbf{v}$ is also often denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$
- The dot product is a symmetric operation, $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v} = \mathbf{v}^T\mathbf{u} = \mathbf{v} \cdot \mathbf{u}$
- Geometric interpretation of a dot product: angle between two vectors
  - I.e., dot product $\mathbf{v} \cdot \mathbf{w}$ over the norms of the vectors is $\cos(\theta)$



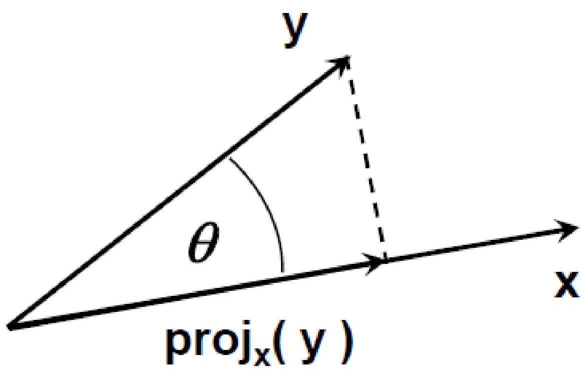
$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| cos(\theta) \qquad \cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

- If two vectors are orthogonal: $\theta = 90°$, i.e., $\cos(\theta) = 0$, then $\mathbf{u} \cdot \mathbf{v} = 0$
- Also, in ML the term $\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$ is sometimes employed as a measure of closeness of two vectors/data instances, and it is referred to as cosine similarity

Picture from: http://d2l.ai/chapter_appendix-mathematics-for-deep-learning/geometry-linear-algebraic-ops.html#geometry-of-vectors

Vector Projection

- *Orthogonal projection* of a vector $\mathbf{y}$ onto vector $\mathbf{x}$
  - The projection can take place in any space of dimensionality ≥ 2
  - The unit vector in the direction of $\mathbf{x}$ is $\frac{\mathbf{x}}{\|\mathbf{x}\|}$
    - A unit vector has norm equal to 1
  - The length of the projection of $\mathbf{y}$ onto $\mathbf{x}$ is $\|\mathbf{y}\| \cdot cos(\theta)$
  - The orthogonal project is the vector $\mathbf{proj_x}(\mathbf{y})$



$$\mathbf{proj_x}(\mathbf{y}) = \frac{\mathbf{x} \cdot \|\mathbf{y}\| \cdot cos(\theta)}{\|\mathbf{x}\|}$$

Slide credit: Jeff Howbert — Machine Learning Math Essentials

Matrices

- ***Matrix*** is a rectangular array of real-valued scalars arranged in m horizontal rows and n vertical columns
  - Each element a_{ij} belongs to the i^{th} row and j^{th} column
  - The elements are denoted a_{ij} or $\mathbf{A}_{ij}$ or $[\mathbf{A}]_{ij}$ or $\mathbf{A}(\boldsymbol{i},\boldsymbol{j})$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

- For the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the size (dimension) is $m \times n$ or (m,n)
  - Matrices are denoted by bold-font upper-case letters

Matrices

- Addition or subtraction $\left(\mathbf{A}\pm\mathbf{B}\right)_{i,j}=\mathbf{A}_{i,j}\pm\mathbf{B}_{i,j}$

$$\begin{bmatrix}1 & 3 & 1\\ 1 & 0 & 0\end{bmatrix}+\begin{bmatrix}0 & 0 & 5\\ 7 & 5 & 0\end{bmatrix}=\begin{bmatrix}1+0 & 3+0 & 1+5\\ 1+7 & 0+5 & 0+0\end{bmatrix}=\begin{bmatrix}1 & 3 & 6\\ 8 & 5 & 0\end{bmatrix}$$

- Scalar multiplication $\left(c\mathbf{A}\right)_{i,j}=c\cdot\mathbf{A}_{i,j}$

$$2\cdot\begin{bmatrix}1 & 8 & -3\\ 4 & -2 & 5\end{bmatrix}=\begin{bmatrix}2\cdot 1 & 2\cdot 8 & 2\cdot -3\\ 2\cdot 4 & 2\cdot -2 & 2\cdot 5\end{bmatrix}=\begin{bmatrix}2 & 16 & -6\\ 8 & -4 & 10\end{bmatrix}$$

- Matrix multiplication $\left(\mathbf{AB}\right)_{i,j}=\mathbf{A}_{i,1}\mathbf{B}_{1,j}+\mathbf{A}_{i,2}\mathbf{B}_{2,j}+\cdots+\mathbf{A}_{i,n}\mathbf{B}_{n,j}$
  - Defined only if the number of columns of the left matrix is the same as the number of rows of the right matrix
  - Note that **AB** ≠ **BA**

$$\begin{bmatrix}\underline{2} & \underline{3} & \underline{4}\\ 1 & 0 & 0\end{bmatrix}\begin{bmatrix}0 & \underline{1000}\\ 1 & \underline{100}\\ 0 & \underline{10}\end{bmatrix}=\begin{bmatrix}3 & \underline{2340}\\ 0 & 1000\end{bmatrix}$$

Matrices

- *Transpose* of the matrix: $\mathbf{A}^T$ has the rows and columns exchanged

$$\left(\mathbf{A}^T\right)_{i,j} = \mathbf{A}_{j,i} \qquad \begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & 7 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 1 & 0 \\ 2 & -6 \\ 3 & 7 \end{bmatrix}$$

  - Some properties

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$

$$(\mathbf{A} + \mathbf{B}\,)^T = \mathbf{A}^T + \boldsymbol{B}^T \qquad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$$

$$(\mathbf{A}^T)^T = \mathbf{A} \qquad (\mathbf{AB}\,)^T = \boldsymbol{B}^T\mathbf{A}^T$$

- *Square matrix*: has the same number of rows and columns
- *Identity matrix* ($\mathbf{I}_n$): has ones on the main diagonal, and zeros elsewhere
  - E.g.: identity matrix of size 3×3 : $\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Matrices

- ***Determinant*** of a matrix, denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$, is a real-valued scalar encoding certain properties of the matrix
  - E.g., for a matrix of size 2×2: $$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc$$
  - For larger-size matrices the determinant of a matrix id calculated as $$\det(\mathbf{A}) = \sum_j a_{ij}(-1)^{i+j} det\left(\mathbf{A}_{(i,j)}\right)$$
  - In the above, $\mathbf{A}_{(i,j)}$ is a minor of the matrix obtained by removing the row and column associated with the indices i and j
- ***Trace*** of a matrix is the sum of all diagonal elements $$\mathrm{Tr}(\mathbf{A}) = \sum_i a_{ii}$$
- A matrix for which $\mathbf{A} = \mathbf{A}^T$ is called a ***symmetric matrix***

Matrices

- Elementwise multiplication of two matrices **A** and **B** is called the ***Hadamard product*** or ***elementwise product***
  - The math notation is $\odot$

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2n}b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \dots & a_{mn}b_{mn} \end{bmatrix}$$

Matrix-Vector Products

- Consider a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^{n}$
- The matrix can be written in terms of its row vectors (e.g., $\mathbf{a}_1^T$ is the first row)

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix}$$

- The matrix-vector product is a column vector of length m, whose i^{th} element is the dot product $\mathbf{a}_i^T\mathbf{x}$

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix} \quad \mathbf{Ax}= \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_m^\top \mathbf{x} \end{bmatrix}$$

- Note the size: $\mathbf{A}(m \times n) \cdot \mathbf{x}(n \times 1) = \mathbf{Ax}(m \times 1)$

Matrix-Matrix Products

- To multiply two matrices $\mathbf{A} \in \mathbb{R}^{n \times k}$and $\mathbf{B} \in \mathbb{R}^{k \times m}$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{km} \end{bmatrix}$$

- We can consider the matrix-matrix product as dot-products of rows in **A** and columns in **B**

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_n^\top \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_m \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{b}_1 & \mathbf{a}_1^\top \mathbf{b}_2 & \cdots & \mathbf{a}_1^\top \mathbf{b}_m \\ \mathbf{a}_2^\top \mathbf{b}_1 & \mathbf{a}_2^\top \mathbf{b}_2 & \cdots & \mathbf{a}_2^\top \mathbf{b}_m \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_n^\top \mathbf{b}_1 & \mathbf{a}_n^\top \mathbf{b}_2 & \cdots & \mathbf{a}_n^\top \mathbf{b}_m \end{bmatrix}$$

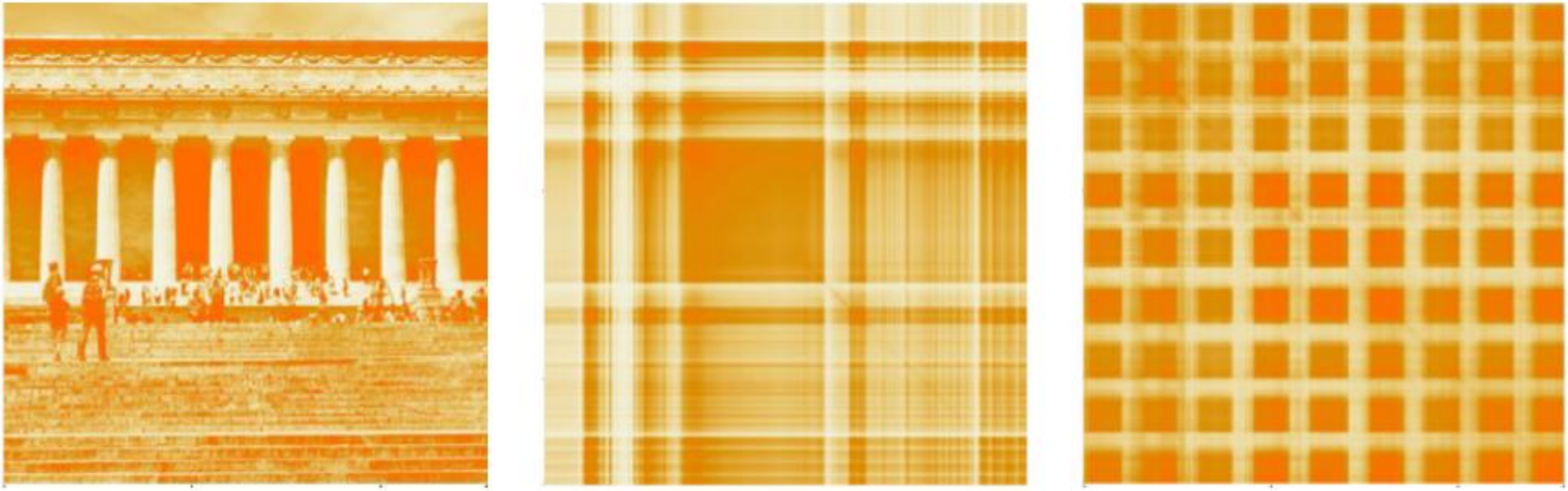
- Size: $\mathbf{A}(n \times k) \cdot \mathbf{B}(k \times m) = \mathbf{C}(n \times m)$

Covariance matrix

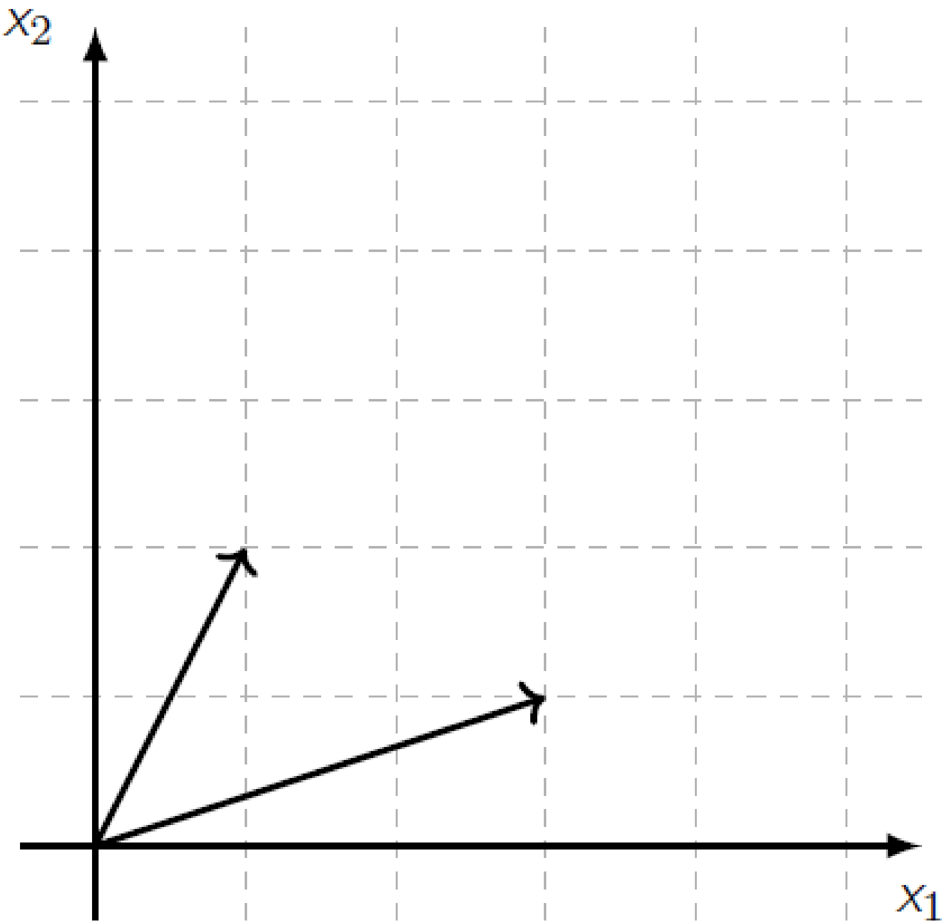
- Multiplication by transpose matrix is common ie. $A \cdot A^T$
- Both $A \cdot A^T$ and $A^T \cdot A$ are compatible for multiplications
- Let $A_{n \times d}$ be a feature matrix, each row represents an item and each column denotes a feature
- $C = AA^T$ is a $n \times n$ matrix dot product
    - C_{ij} is a measure how similar item i is to item j (in syncness)
- $D = A^T A$ is a $d \times d$ dot products in syncness among the features
    - D_{ij} represents the similarity between feature i and feature j

Covariance matrix

- $A, \; A \cdot A^T, \; A^T \cdot A$

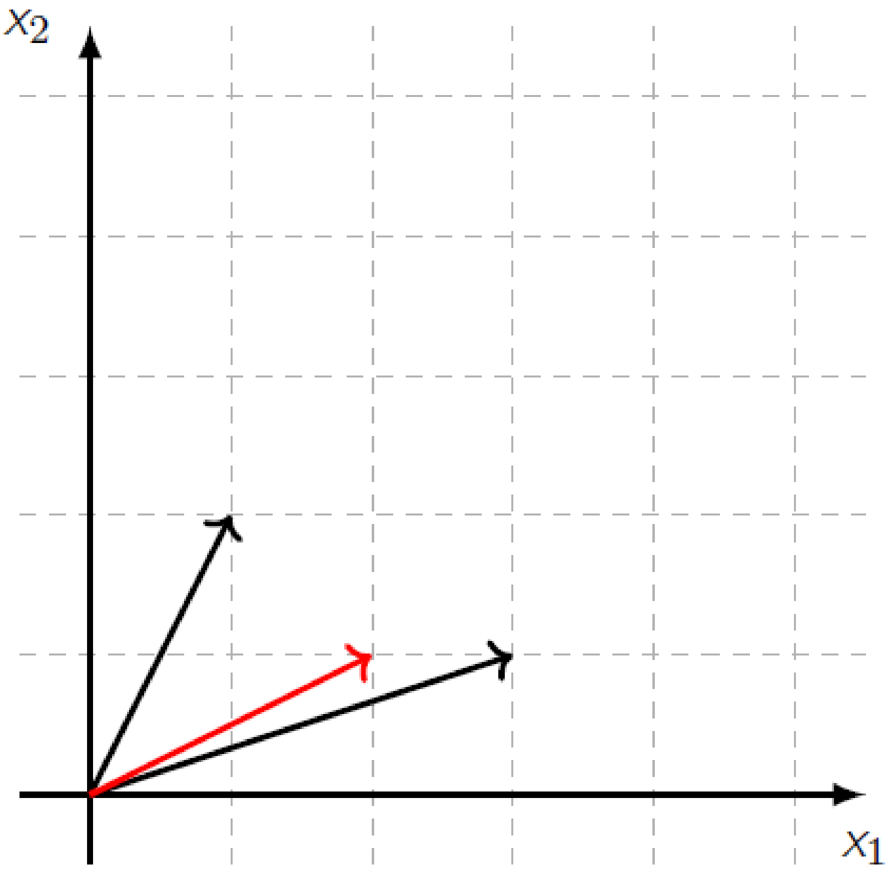


Linear transformation



$$A = \begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix} \quad x = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

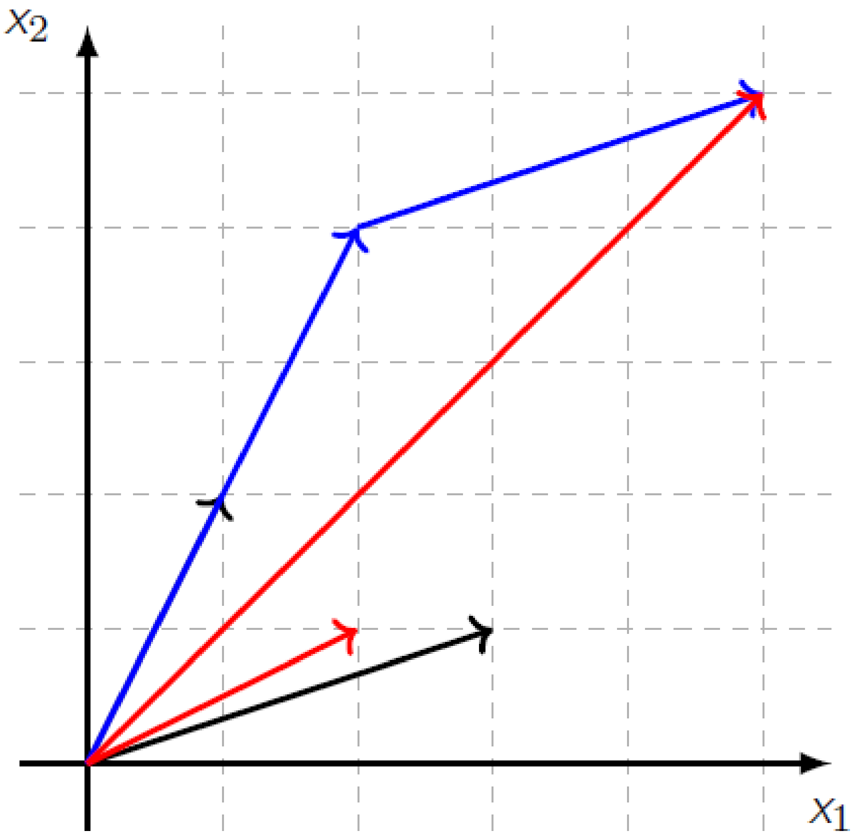
Linear transformation



$$A = \begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix} \quad x = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$Ax = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \times 2 + \begin{bmatrix} 3 \\ 1 \end{bmatrix} \times 1 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$

Linear transformation



$$A = \begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix} \quad x = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$Ax = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \times 2 + \begin{bmatrix} 3 \\ 1 \end{bmatrix} \times 1 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$

Linear Dependence

- For the following matrix $\mathbf{B} = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix}$
- Notice that for the two columns $\mathbf{b}_1 = [2, 4]^T$ and $\mathbf{b}_2 = [-1, -2]^T$, we can write $\mathbf{b}_1 = -2 \cdot \mathbf{b}_2$
  - This means that the two columns are linearly dependent
- The weighted sum $a_1\mathbf{b}_1 + a_2\mathbf{b}_2$ is referred to as a linear combination of the vectors $\mathbf{b}_1$ and $\mathbf{b}_2$
  - In this case, a linear combination of the two vectors exist for which $\mathbf{b}_1 + 2 \cdot \mathbf{b}_2 = \mathbf{0}$
- A collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ are ***linearly dependent*** if there exist coefficients $a_1, a_2, \dots, a_k$ not all equal to zero, so that

$$\sum_{i=1}^{k} a_i \mathbf{v_i} = 0$$

- If there is no linear dependence, the vectors are ***linearly independent***

Matrix Rank

- For an $n \times m$ matrix, the ***rank*** of the matrix is the largest number of linearly independent columns
- The matrix **B** from the previous example has $rank(\mathbf{B}) = 1$, since the two columns are linearly dependent

$$\mathbf{B} = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix}$$

- The matrix **C** below has $rank(\mathbf{C}) = 2$, since it has two linearly independent columns
  - I.e., $\mathbf{c}_4 = -1 \cdot \mathbf{c}_1$, $\mathbf{c}_5 = -1 \cdot \mathbf{c}_3$, $\mathbf{c}_2 = 3 \cdot \mathbf{c}_1 + 3 \cdot \mathbf{c}_3$

$$\mathbf{C} = \begin{bmatrix} 1 & 3 & 0 & -1 & 0 \\ -1 & 0 & 1 & 1 & -1 \\ 0 & -3 & 1 & 0 & -1 \\ 2 & 3 & -1 & -2 & 1 \end{bmatrix}$$

Inverse of a Matrix

- For a square $n \times n$ matrix $\mathbf{A}$ with rank n, $\mathbf{A}^{-1}$ is its ***inverse matrix*** if their product is an identity matrix $\mathbf{I}$

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

- Properties of inverse matrices

$$\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$$

$$\left(\mathbf{AB}\right)^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

- If $\det(A) = 0$ (i.e., $\text{rank}(A) < n$), then the inverse does not exist
  - A matrix that is not invertible is called a singular matrix
- Note that finding an inverse of a large matrix is computationally expensive
- If the inverse of a matrix is equal to its transpose, the matrix is said to be orthogonal matrix

$$\mathbf{A}^{-1} = \mathbf{A}^{T}$$

Special Matrices

- Diagonal matrices – Non-zero diagonal elements and rests are zero. Formally $D_{i,j} = 0, i \neq j$
  - Identity matrix
  - diag(v) – vectors using diagonal elements
  - diag(v)x – x_i is scaled by v_i
  - Inversion is easy diag(v)$^{-1}$=diag($[1/v_1, 1/v_2, \ldots, 1/v_n]^T$)
- In R^n, at most n vectors may be mutually orthogonal with non-zero norm
- Vectors orthogonal and have unit norm is known as orthonormal
- Orthogonal matrix – Square matrix, rows are mutually orthonormal, columns are mutually orthonormal
  - $A^T A = A A^T = I$

Tensors

- ***Tensors*** are n-dimensional arrays of scalars
  - Vectors are first-order tensors, $\mathbf{v} \in \mathbb{R}^n$
  - Matrices are second-order tensors, $\mathbf{A} \in \mathbb{R}^{m \times n}$
  - E.g., a fourth-order tensor is $\mathsf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$
- Tensors are denoted with upper-case letters of a special font face (e.g., X, Y, Z)
- RGB images are third-order tensors, i.e., as they are 3-dimensional arrays
  - The 3 axes correspond to width, height, and channel
  - E.g., 224 × 224 × 3
  - The channel axis corresponds to the color channels (red, green, and blue)

Eigen Decomposition

- *Eigen decomposition* is decomposing a matrix into a set of eigenvalues and eigenvectors
- *Eigenvalues* of a square matrix $\mathbf{A}$ are scalars λ and *eigenvectors* are non-zero vectors $\mathbf{v}$ that satisfy

$$\mathbf{Av} = \lambda\mathbf{v}$$

- Eigenvalues are found by solving the following equation

$$\det(\mathbf{A} - \lambda\mathbf{I}\ ) = 0$$

- If a matrix $\mathbf{A}$ has n linearly independent eigenvectors $\{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ with corresponding eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, the eigen decomposition of $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{V\Lambda V}^{-1}$$

  - Columns of the matrix $\mathbf{V}$ are the eigenvectors, i.e., $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^n]$
  - $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues, i.e., $\mathbf{\Lambda} = [\lambda_1, \ldots, \lambda_n]$
- To find the inverse of the matrix A, we can use $\mathbf{A}^{-1} = \mathbf{V\Lambda}^{-1}\mathbf{V}^{-1}$
  - This involves simply finding the inverse $\mathbf{\Lambda}^{-1}$ of a diagonal matrix

Eigen Decomposition

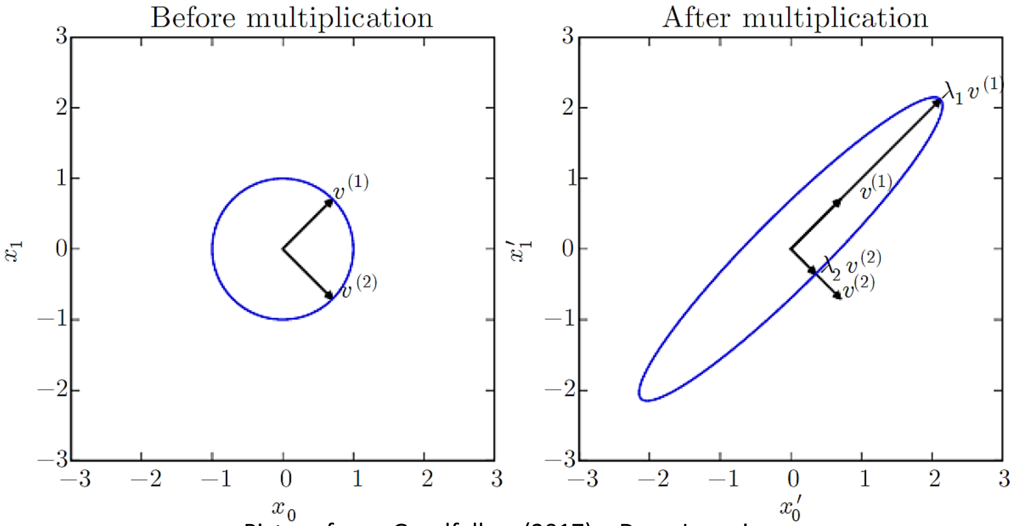
- Decomposing a matrix into eigenvalues and eigenvectors allows to analyze certain properties of the matrix
  - If all eigenvalues are positive, the matrix is positive definite
  - If all eigenvalues are positive or zero-valued, the matrix is positive semidefinite
  - If all eigenvalues are negative or zero-values, the matrix is negative semidefinite
- Eigen decomposition can also simplify many linear-algebraic computations
  - The determinant of A can be calculated as

$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$$

  - If any of the eigenvalues are zero, the matrix is singular (it does not have an inverse)
- However, not every matrix can be decomposed into eigenvalues and eigenvectors
  - Also, in some cases the decomposition may involve complex numbers
  - Still, every real symmetric matrix is guaranteed to have an eigen decomposition according to $\mathbf{A} = \mathbf{V\Lambda V}^{-1}$, where $\mathbf{V}$ is an orthogonal matrix

Eigen Decomposition

- Geometric interpretation of the eigenvalues and eigenvectors is that they allow to stretch the space in specific directions
  - Left figure: the two eigenvectors $\mathbf{v}^1$ and $\mathbf{v}^2$ are shown for a matrix, where the two vectors are unit vectors (i.e., they have a length of 1)
  - Right figure: the vectors $\mathbf{v}^1$ and $\mathbf{v}^2$ are multiplied with the eigenvalues λ_1 and λ_2
    - We can see how the space is scaled in the direction of the larger eigenvalue λ_1
- E.g., this is used for dimensionality reduction with PCA (principal component analysis) where the eigenvectors corresponding to the largest eigenvalues are used for extracting the most important data dimensions



Picture from: Goodfellow (2017) – Deep Learning

Singular Value Decomposition

- ***Singular value decomposition*** (SVD) provides another way to factorize a matrix, into singular vectors and singular values
  - SVD is more generally applicable than eigen decomposition
  - Every real matrix has an SVD, but the same is not true of the eigen decomposition
    - E.g., if a matrix is not square, the eigen decomposition is not defined, and we must use SVD
- SVD of an $m \times n$ matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{T}$$

  - $\mathbf{U}$ is an $m \times m$ matrix, $\mathbf{D}$ is an $m \times n$ matrix, and $\mathbf{V}$ is an $n \times n$ matrix
  - The elements along the diagonal of $\mathbf{D}$ are known as the singular values of A
  - The columns of $\mathbf{U}$ are known as the left-singular vectors
  - The columns of $\mathbf{V}$ are known as the right-singular vectors
- For a non-square matrix $\mathbf{A}$, the squares of the singular values σ_i are the eigenvalues λ_i of $\mathbf{A}^{T}\mathbf{A}$, i.e., $\sigma_i^2 = \lambda_i$ for $i = 1,2, \dots, n$
- Applications of SVD include computing the pseudo-inverse of non-square matrices, matrix approximation, determining the matrix rank

Probability and Random Variable

Sample Space

- Statisticians use the word **experiment** to describe any process that generates a set of data.

- For example: a statistical experiment is the tossing of a coin. In this experiment, there are only two possible outcomes, heads or tails.

- The set of all possible outcomes of a statistical experiment is called the **sample space** and is represented by the symbol *S*.
- Each outcome in a sample space is called an **element** or a **member** of the sample space, or simply a **sample point**.

- If the sample space has a finite number of elements, we may *list* the members separated by commas and enclosed in braces. Thus, the sample space *S*, of possible outcomes when a coin is flipped, may be written

- where *H* and *T* correspond to heads and tails, respectively.

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

- Consider the experiment of tossing a die. If we are interested in the number that shows on the top face, the sample space is

$$S_1 = \{1, 2, 3, 4, 5, 6\}.$$

- Suppose that three items are selected at random from a manufacturing process. Each item is inspected and classified defective, *D*, or nondefective, *N*.
- What will be the sample space?

S = {DDD, DDN, DND, DNN, NDD, NDN, NND, NNN}.

- Sample spaces with a large or infinite number of sample points are best described by a **statement** or **rule method**.
- For example, if the possible outcomes of an experiment are the set of cities in the world with a population over 1 million, our sample space is written
- $S = \{x \mid x$ is a city with a population over 1 million$\}$,

- Similarly, if S is the set of all points (x, y) on the boundary or the interior of a circle of radius 2 with center at the origin, we write the **rule**
- $S = \{(x, y) \mid x^2 + y^2 \leq 4\}$.

Events

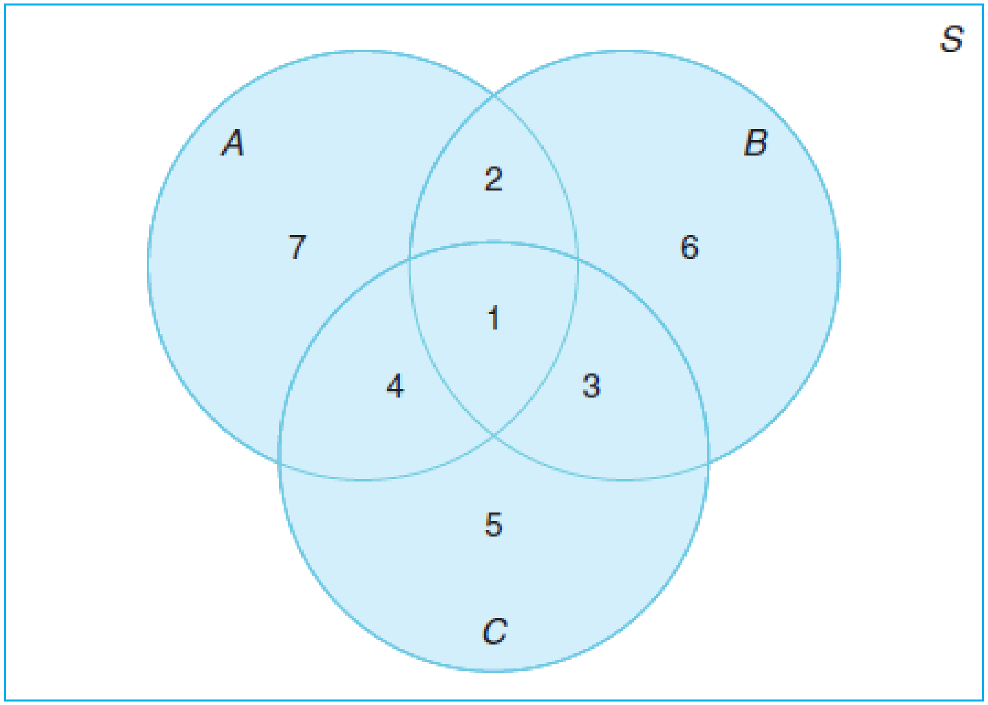
- For any given experiment, we may be interested in the occurrence of certain **events** rather than in the occurrence of a specific element in the sample space. For instance, we may be interested in the event *A* that the outcome when a die is tossed is divisible by 3.
- This will occur if the outcome is an element of the subset *A* = *{*3, 6*}* of the sample space *S*1={1,2,3,4,5,6}.

- Given the sample space $S = \{t \mid t \geq 0\}$, where t is the life in years of a certain electronic component, then the event A that the component fails before the end of the fifth year is the subset $A = \{t \mid 0 \leq t < 5\}$.

- An event is a collection of sample points, which constitute a subset of the sample space where some desired condition(s) is satisfied.

- The **complement** of an event A with respect to S is the subset of all elements of S that are not in A. We denote the complement of A by the symbol A'.
- The **intersection** of two events A and B, denoted by the symbol $A \cap B$, is the event containing all elements that are common to A and B.

- Two events A and B are **mutually exclusive**, or **disjoint**, if $A \cap B = \varphi$, that is, if A and B have no elements in common.
- The **union** of the two events A and B, denoted by the symbol $A \cup B$, is the event containing all the elements that belong to A or B or both.



- $A \cup C$ = regions 1, 2, 3, 4, 5, and 7,
- $B' \cap A$ = regions 4 and 7,
- $A \cap B \cap C$ = region 1,
- $(A \cup B) \cap C'$ = regions 2, 6, and 7,

Counting Sample Points

- One of the problems that the statistician must consider and attempt to evaluate is the element of chance associated with the occurrence of certain events when an experiment is performed.

- Multiplication rule: If an operation can be performed in $n1$ ways, and if for each of these ways a second operation can be performed in $n2$ ways, then the two operations can be performed together in $n1n2$ ways.

- How many sample points are there in the sample space when a pair of dice is thrown once?
- If a 22-member club needs to elect a chair and a treasurer, how many different ways can these two to be elected?

- If an operation can be performed in $n1$ ways, and if for each of these a second operation can be performed in $n2$ ways, and for each of the first two a third operation can be performed in $n3$ ways, and so forth, then the sequence of k operations can be performed in $n1n2 \cdot \cdot \cdot nk$ ways.

- Sam is going to assemble a computer by himself. He has the choice of chips from two brands, a hard drive from four, memory from three, and an accessory bundle from five local stores. How many different ways can Sam order the parts?

- A **permutation** is an arrangement of all or part of a set of objects.
- Consider the three letters *a*, *b*, and *c*. The possible permutations are *abc*, *acb*, *bac*, *bca*, *cab*, and *cba*.
- We can use multiplication rule for counting possible permutation.

- There are $n1 = 3$ choices for the first position. No matter which letter is chosen, there are always $n2 = 2$ choices for the second position. No matter which two letters are chosen for the first two positions, there is only $n3 = 1$ choice for the last position, giving a total of
- $n1n2n3 = (3)(2)(1) = 6$ permutations. For any non-negative integer n, $n!$, called "n factorial," is defined as $n! = n(n - 1) \cdot \cdot \cdot (2)(1)$,

- The number of permutations of the four letters *a*, *b*, *c*, and *d* will be 4! = 24.
- Now consider the number of permutations that are possible by taking two letters at a time from four.
- n1=4, n2=3
- $n1n2 = (4)(3) = 12$

- permutations. In general, *n* distinct objects taken *r* at a time can be arranged in
- $n(n-1)(n-2)\cdots(n-r+1)$ ways. We represent this product by the symbol

$$ {}_nP_r = \frac{n!}{(n-r)!}. $$

Probability of an Event

- The **probability** of an event A is the sum of the weights of all sample points in A. Therefore,

$0 \leq P(A) \leq 1$, $P(\varphi) = 0$, and $P(S) = 1$.

- Furthermore, if $A1$, $A2$, $A3$, . . . is a sequence of mutually exclusive events, then

$P(A1 \cup A2 \cup A3 \cup \cdot\cdot\cdot) = P(A1) + P(A2) + P(A3) + \cdot \cdot \cdot$.

- A coin is tossed twice. What is the probability that at least 1 head occurs?
- The sample space for this experiment is

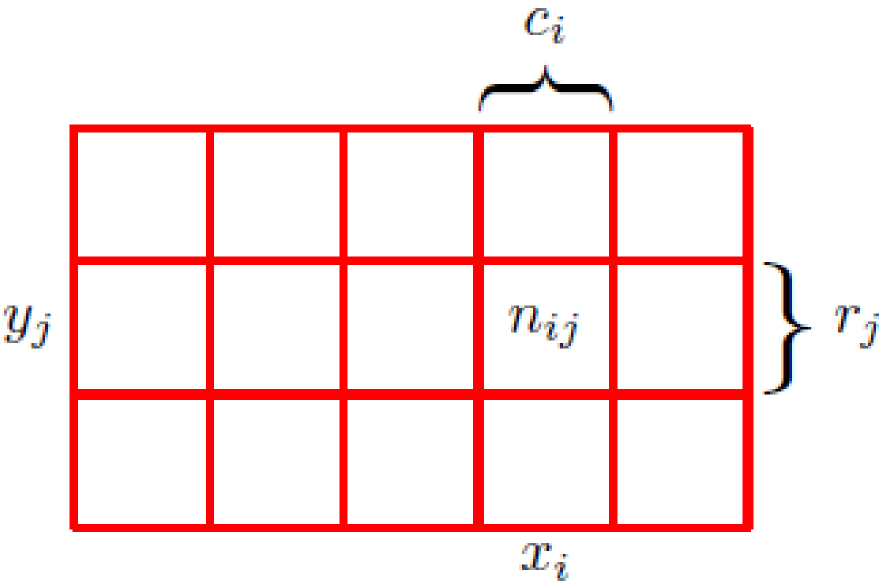
  $S = \{HH, HT, TH, TT\}$.
- If the coin is balanced, each of these outcomes is equally likely to occur. Therefore,
- we assign a probability of ω to each sample point. Then $4\omega = 1$, or $\omega = 1/4$. If A represents the event of at least 1 head occurring, then
- $A = \{HH, HT, TH\}$ and $P(A) = 1/4+1/4+1/4=3/4$.

- A die is loaded in such a way that an even number is twice as likely to occur as an odd number. If E is the event that a number less than 4 occurs on a single toss of the die, find $P(E)$.
- The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. We assign a probability of w to each odd number and a probability of $2w$ to each even number. Since the sum of the probabilities must be 1, we have $9w = 1$ or $w = 1/9$. Hence, probabilities of 1/9 and 2/9 are assigned to each odd and even number, respectively. Therefore,

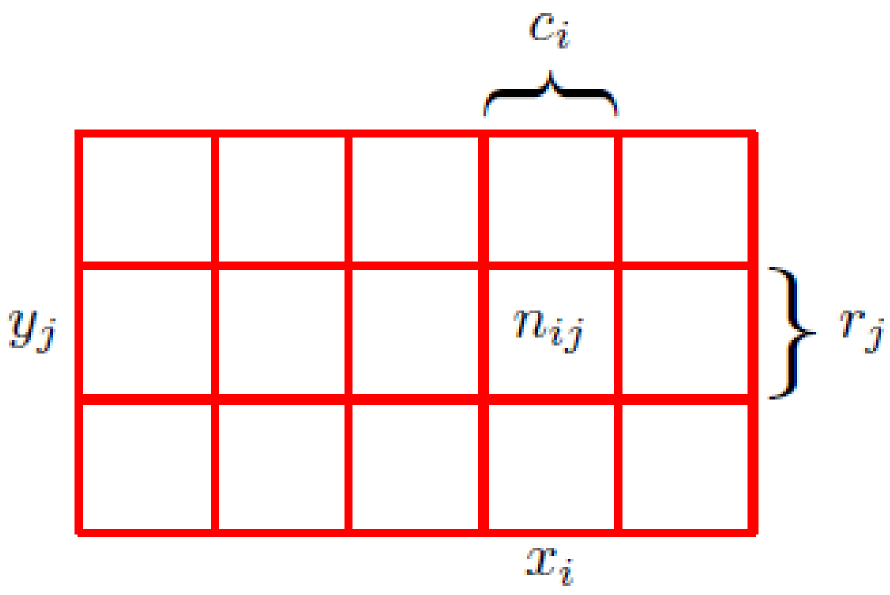
$E = \{1, 2, 3\}$ and $P(E) =1/9+2/9+1/9=4/9$

- let A be the event that an even number turns up and let B be the event that a number divisible by 3 occurs. Find $P(A \cup B)$ and $P(A \cap B)$. Use sample space and probability of last example.
- For the events $A = \{2, 4, 6\}$ and $B = \{3, 6\}$, we have $A \cup B = \{2, 3, 4, 6\}$ and $A \cap B = \{6\}$.
- $P(A \cup B) = 7/9$ $P(A \cap B) =2/9$

- We can derive the sum and product rules of probability by considering two random variables, X, which takes the values $\{x_i\}$ where $i = 1, . . . , M$, and Y , which takes the values $\{y_j\}$ where $j = 1, . . . , L$. In this illustration we have $M = 5$ and $L = 3$.



- If we consider a total number N of instances of these variables, then we denote the number of instances where $X = x_i$ and $Y = y_j$ by n_{ij} , which is the number of points in the corresponding cell of the array. The number of points in column i, corresponding to $X = x_i$, is denoted by c_i, and the number of points in row j, corresponding to $Y = y_j$, is denoted by r_j .



- The probability that X will take the value x_i and Y will take the value y_j is written $p(X = x_i, Y = y_j)$ and is called the *joint* probability of $X = x_i$ and $Y = y_j$. It is given by the number of points falling in the cell i,j as a fraction of the total number of points, and hence

  $$p(X = x_i, Y = y_j) = n_{ij}/N$$

- the probability that X takes the value x_i irrespective of the value of Y is written as $p(X = x_i)$ and is given by the fraction of the total number of points that fall in column i, so that

  $p(X = x_i) = c_i/N$
- Because the number of instances in column i in the grid Figure is just the sum of the number of instances in each cell of that column, we have

  $c_i = \sum_j n_{ij}$

and therefore $p(X = xi) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$ which is the *sum rule* of probability. Note that $p(X = xi)$ is sometimes called the *marginal* probability, because it is obtained by marginalizing, or summing out, the other variables (in this case Y).

Conditional Probability

- If we consider only those instances for which $X = x_i$, then the fraction of such instances for which $Y = y_j$ is written $p(Y = y_j \mid X = x_i)$ and is called the conditional probability of $Y = y_j$ given $X = x_i$. It is obtained by finding the fraction of those points in column i that fall in cell i,j and hence is given by

$$p(Y = y_j \mid X = x_i) = n_{ij}/c_i$$

$$p(X = x_i, Y = y_j) = n_{ij}/N= n_{ij}/c_i \cdot c_i/N$$

$$= p(Y = y_j \mid X = x_i)p(X = x_i)$$

which is the *product rule* of probability

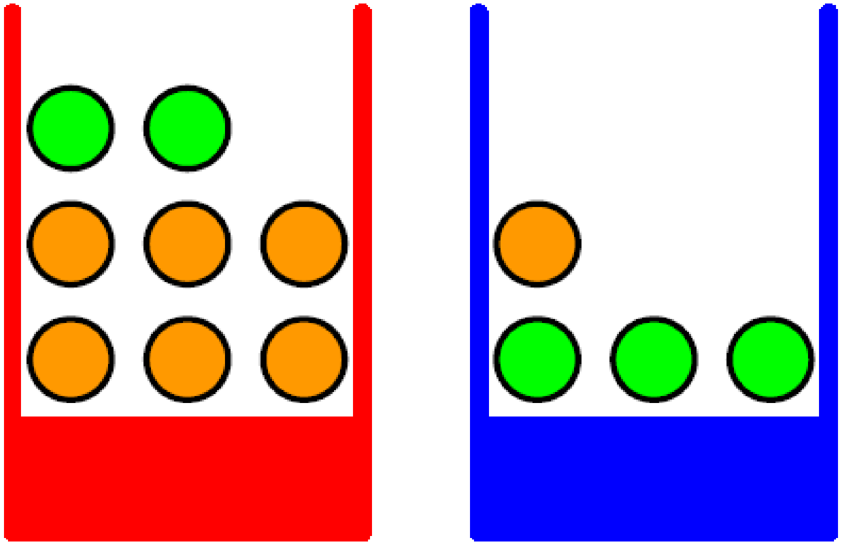
The Rules of Probability

- **sum rule** p(X) =$\sum_{Y}$ p(X, Y)
- **product rule** p(X, Y) = p(Y |X)p(X).
- From the product rule, together with the symmetry property p(X, Y) = p(Y,X), we immediately obtain the following relationship between conditional probabilities p(Y |X).p(X) = p(X|Y)p(Y)
- p(Y |X) =$\frac{\text{p(X|Y)p(Y)}}{\text{p(X)}}$ $(Bay's\ theorem)$

Total Probability from sum rule

- $p(X) = \sum_Y p(X, Y)$
- $= \sum_Y p(X|Y)\, p(Y)$

- Imagine we have two boxes, one red and one blue, and in the red box we have 2 apples and 6 oranges, and in the blue box we have 3 apples and 1 orange. Let us also assume probability that we pick the red box is 0.4 and the blue box is 0.6
- $p(B = r) = 4/10$
- $p(B = b) = 6/10$



- Now suppose that we pick a box at random, and it turns out to be the blue box. Then the probability of selecting an apple is just the fraction of apples in the blue box which is 3/4, and so $p(F = a | B = b) = 3/4$. In fact, we can write out all four conditional probabilities for the type of fruit, given the selected box
- $p(F = a | B = r) = 1/4$
- $p(F = o | B = r) = 3/4$
- $p(F = a | B = b) = 3/4$
- $p(F = o | B = b) = 1/4$.

- We can now use the sum and product rules of probability to evaluate the overall probability of choosing an apple
- $p(F = a) = p(F = a | B = r)p(B = r) + p(F = a | B = b)p(B = b)$ $[p(X) = \sum_{Y} p(X|Y)\, p(Y)]$
- $=1/4\times 4/10+3/4\times 6/10=11/20$
- So $p(F=o)=1-11/20=9/20$

- Suppose instead we are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from.
- $p(B = r|F = o) = p(F = o|B = r)p(B = r)/p(F = o)$ $[p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}]$
- $= \frac{3/4 \times 4/10}{9/20} = 2/3$
- $p(B = b|F = o) = 1 - 2/3 = 1/3$

- We can provide an important interpretation of Bayes’ theorem as follows. If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $p(B)$. We call this the ***prior probability*** because it is the probability available *before* we observe the identity of the fruit.
- Once we are told that the fruit is an orange, we can then use Bayes’ theorem to compute the probability p(B|F), which we shall call the ***posterior probability*** because it is the probability obtained *after* we have observed *F*.

- Note that in this example, the prior probability of selecting the red box was 4/10, so that we were more likely to select the blue box than the red one. However, once we have observed that the piece of selected fruit is an orange, we find that the posterior probability of the red box is now 2/3, so that it is now more likely that the box we selected was in fact the red one.

- if the joint distribution of two variables factorizes into the product of the marginals, so that $p(X, Y) = p(X)p(Y)$, then X and Y are said to be *independent*. From the product rule, we see that $p(Y \mid X) = p(Y)$, and so the conditional distribution of Y given X is indeed independent of the value of X. For instance, in our boxes of fruit example, if each box contained the same fraction of apples and oranges, then $p(F \mid B) = P(F)$, so that the probability of selecting, say, an apple is independent of which box is chosen.

Random Variable and Probability Distribution

- In a statistical experiment 3 electronic items are tested, D signifies the item is defective while N signifies the item is non-defective.
- *S* = *{NNN,NND,NDN,DNN,NDD,DND,DDN,DDD}*
- Anyone might be interested to know the number of defective items.
- With each sample point there is an associated number of defective items and it can be 0, 1, 2, 3.
- These values are, of course, random quantities *determined by the outcome of the experiment*.

- They may be viewed as values assumed by the *random variable X*, the number of defective items when three electronic components are tested.
- A **random variable** is a function that associates a real number with each element in the sample space.
- We shall use a capital letter, say X, to denote a random variable and its corresponding small letter, x in this case, for one of its values.

- In the electronic component testing illustration, we notice that the random variable *X* assumes the value 2 for all elements in the subset *E* = *{DDN,DND,NDD}* of the sample space *S*. That is, each possible value of *X* represents an event that is a subset of the sample space for the given experiment.

- Two balls are drawn in succession without replacement from an urn containing 4 red balls and 3 black balls. The possible outcomes and the values y of the random variable Y , where Y is the number of red balls, are

Sample Point	y
RR	2
RB	1
BR	1
BB	0

- Statisticians use **sampling plans** to either accept or reject batches or lots of material. Suppose one of these sampling plans involves sampling independently 10 items from a lot of 100 items in which 12 are defective.
- Let *X* be the random variable defined as the number of items found defective in the sample of 10. In this case, the random variable takes on the values 0, 1, 2, . . . , 9, 10.

Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, outcomes when a die is rolled, rain/not-rain
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

- **Discrete Probability Distributions:** A discrete random variable assumes each of its values with a certain probability.
- In the case of tossing a coin three times, the variable *X*, representing the number of heads,

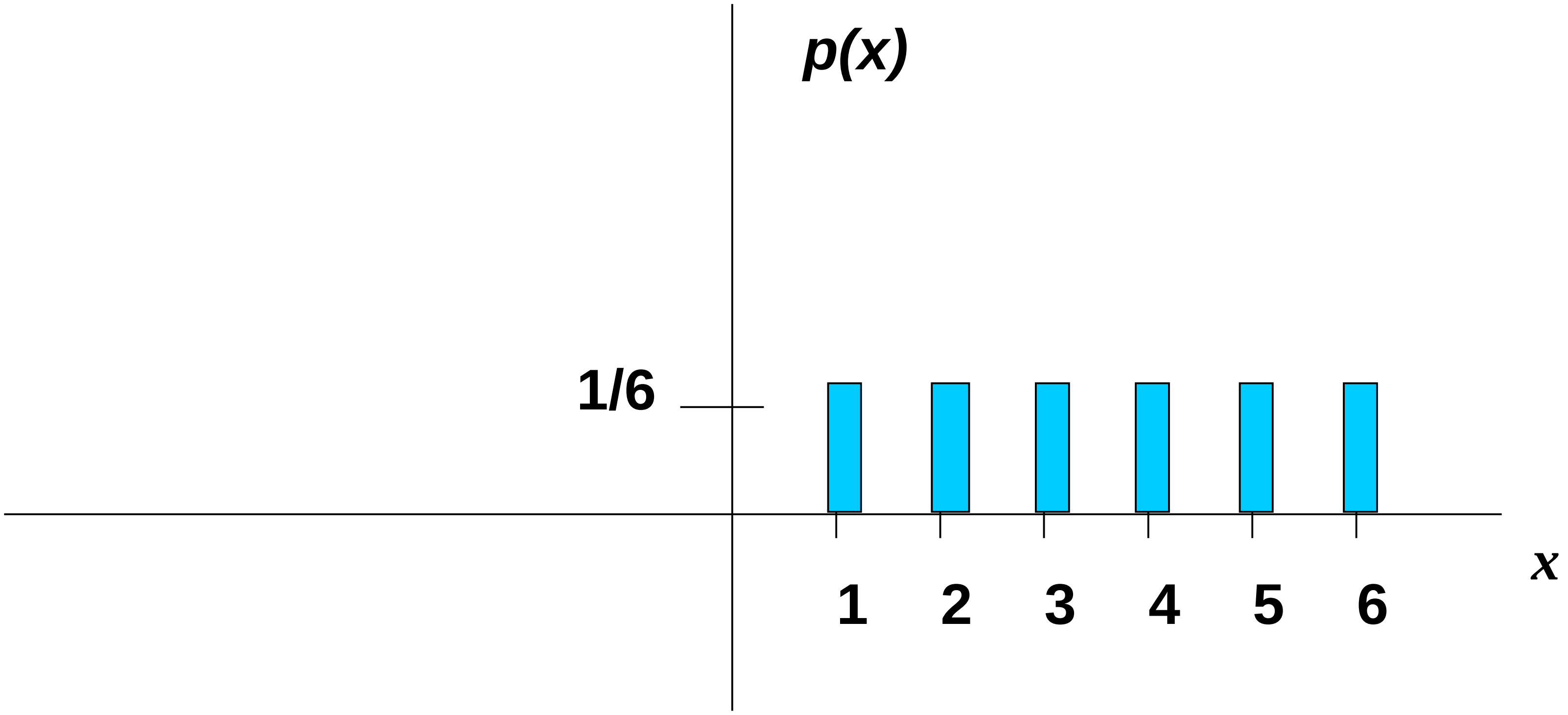
Sample Points	x
HHH	3
HHT	2
HTH	2
HTT	1
THH	2
THT	1
TTH	1
TTT	0

x	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

- Frequently, it is convenient to represent all the probabilities of a random variable X by a formula. Therefore, we write $f(x) = P(X = x)$.
- The set of ordered pairs $(x, f(x))$ is called the **probability function**, **probability mass function**, or **probability distribution** of the discrete random variable X.

- The set of ordered pairs (x, $f(x)$) is a p**robability function**, **probability mass function**, or **probability distribution** of the discrete random variable X if, for each possible outcome x,
- *1.* $f(x) \geq 0$,
- 2.$\sum f(x) = 1$
- 3. $P(X = x) = f(x)$.

Discrete example: roll of a die



$$\sum_{\text{all } x} P(x) = 1$$

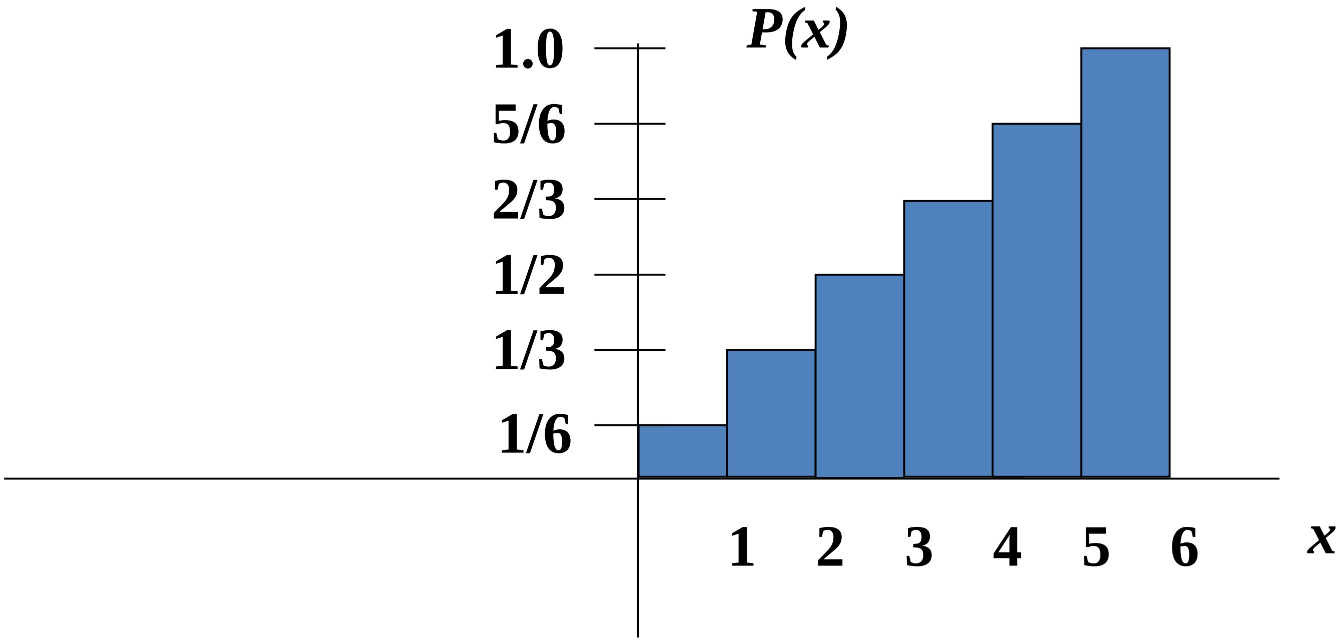
Probability mass function (pmf)

x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$\underline{p(x=6)=1/6}$

- A shipment of 20 similar laptop computers to a retail outlet contains 3 that are defective. If a school makes a random purchase of 2 of these computers, find the probability distribution for the number of defectives.
- Let X be a random variable whose values x are the possible numbers of defective computers purchased by the school. Then x can only take the numbers 0, 1, and 2. Now
- $f(0) = P(X = 0) = {}^{3}C_{0} * {}^{17}C_{2} / {}^{20}C_{2} = 68/95$
- $f(1) = P(X = 1) = {}^{3}C_{1} * {}^{17}C_{1} / {}^{20}C_{2} = 51/190$
- $f(2) = P(X = 2) = {}^{3}C_{2} * {}^{17}C_{0} / {}^{20}C_{2} = 3/190$
- Thus, the probability distribution of X is

x	0	1	2
f(x)	68/95	51/190	3/190

Cumulative distribution function (CDF)



Cumulative distribution function

x	$P(x\leq A)$
1	$P(x\leq 1)=1/6$
2	$P(x\leq 2)=2/6$
3	$P(x\leq 3)=3/6$
4	$P(x\leq 4)=4/6$
5	$P(x\leq 5)=5/6$
6	$P(x\leq 6)=6/6$

- **Continuous Probability Distributions:** A continuous random variable has a probability of 0 of assuming *exactly* any of its values. Consequently, its probability distribution cannot be given in tabular form.

- Consider a random variable whose values are the heights of all people over 21 years of age. Between any two values, say 163.5 and 164.5 centimeters, there are an infinite number of heights, one of which is 164 centimeters. The probability of selecting a person at random who is exactly 164 centimeters tall is close to 0 and we assign it 0.

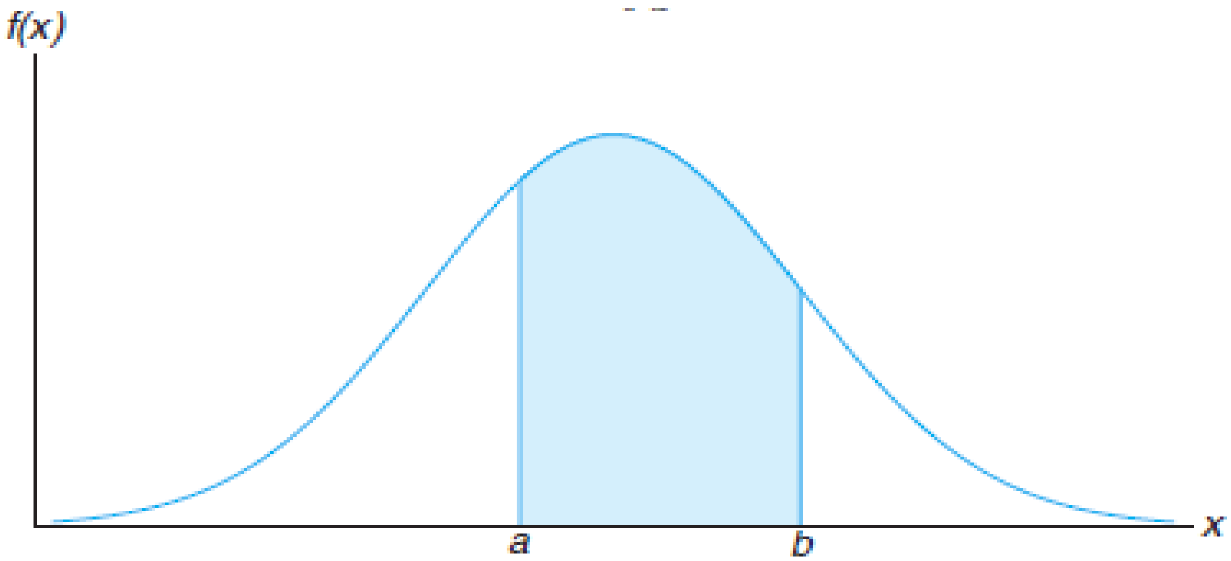
- However, if we talk about the probability of selecting a person who is at least 163 centimeters but not more than 165 centimeters tall. Now we are dealing with an interval rather than a point value of our random variable.

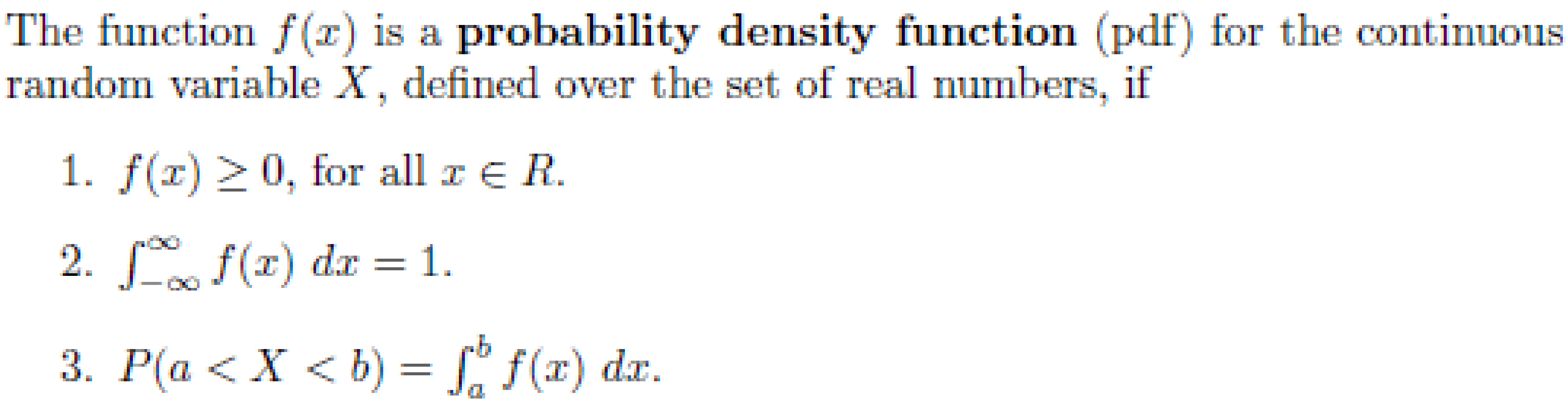
- We shall concern ourselves with computing probabilities for various intervals of continuous random variables such as $P(a < X < b)$,

- Although the probability distribution of a continuous random variable cannot be presented in tabular form, it can be stated as a formula.
- Such a formula would necessarily be a function of the numerical values of the continuous random variable X and as such will be represented by the functional notation $f(x)$.

  $f(x)$ is usually called the **probability density function.**

- In Figure , the probability that X assumes a value between a and b is equal to the shaded area under the density function between the ordinates at $x = a$ and $x = b$, and from integral calculus is given by
- $P(a < X < b) = \int_a^b f(x)dx$

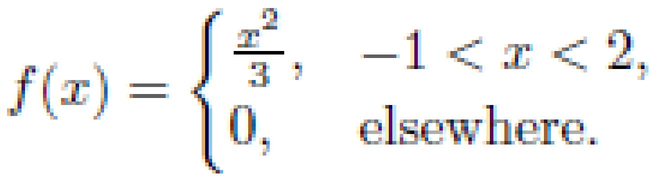




The function $f(x)$ is a **probability density function** (pdf) for the continuous random variable X, defined over the set of real numbers, if

1. $f(x) \geq 0$, for all $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\, dx$.

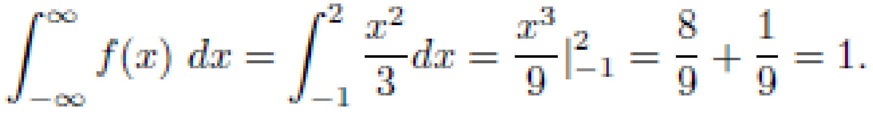
- Suppose that the error in the reaction temperature, in ◦C, for a controlled laboratory experiment is a continuous random variable *X* hav ity function



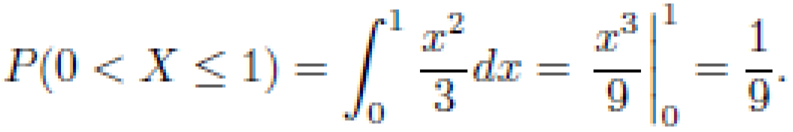
$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

- (a) Verify that *f*(*x*) is a density function.
- (b) Find *P*(0 < *X* ≤ 1).

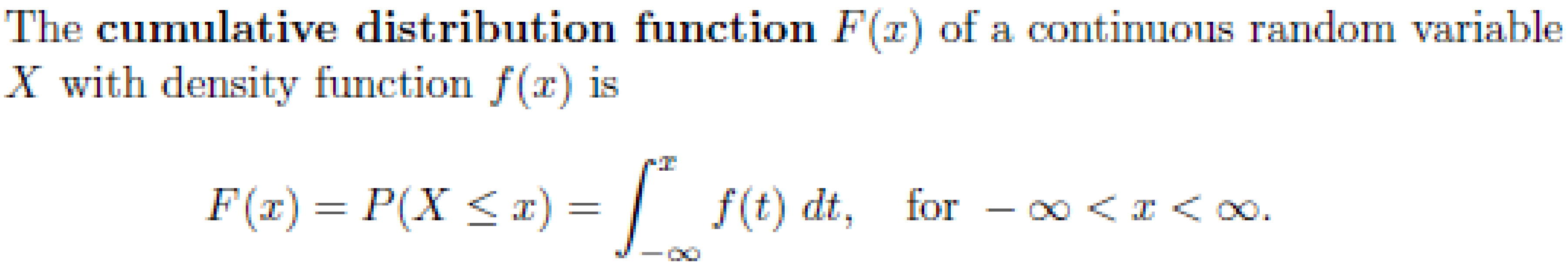
- a)
  - as per the pdf f(x)>=0
- b)



$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-1}^{2} \frac{x^2}{3} dx = \frac{x^3}{9}|_{-1}^{2} = \frac{8}{9} + \frac{1}{9} = 1.$$



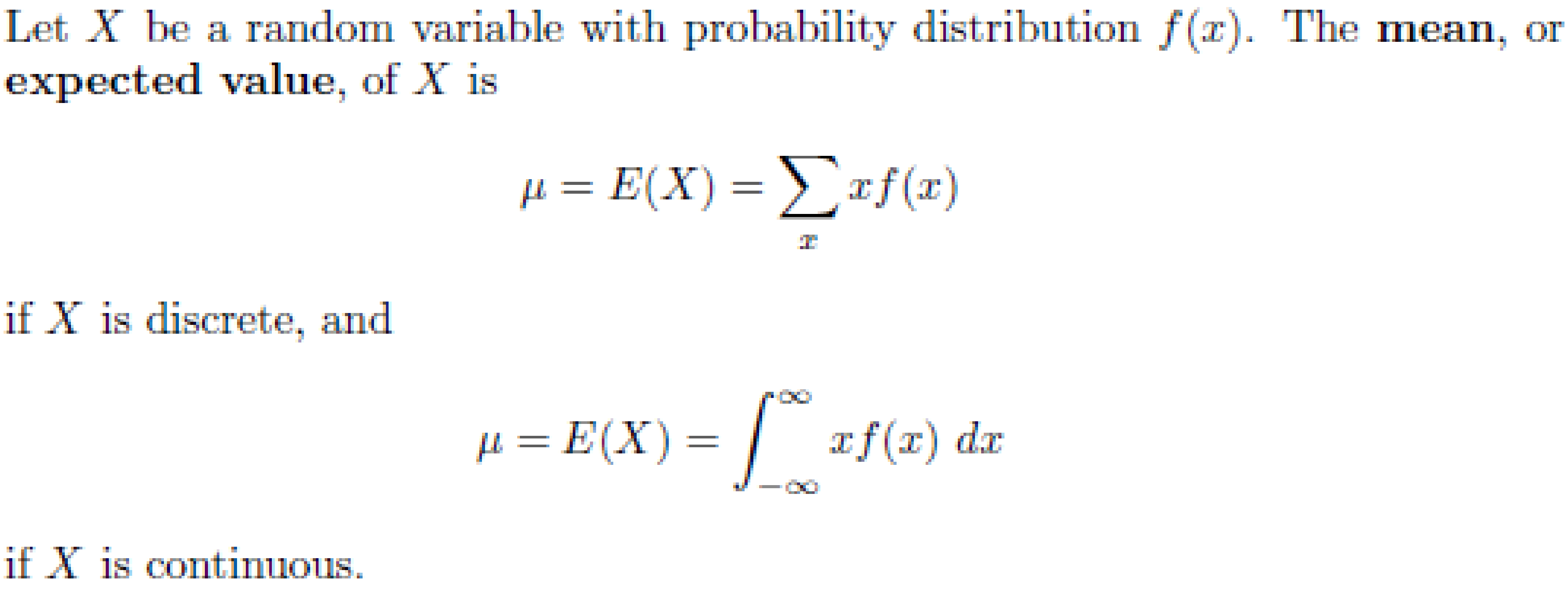
$$P(0 < X \leq 1) = \int_0^1 \frac{x^2}{3} dx = \left.\frac{x^3}{9}\right|_0^1 = \frac{1}{9}.$$



The **cumulative distribution function** $F(x)$ of a continuous random variable X with density function $f(x)$ is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt, \quad \text{for } -\infty < x < \infty.$$

- **Mean of a Random Variable:** If a pair of coins are tossed 16 times and *X* is the number of heads that occur per toss, then the values of *X* are 0, 1, and 2. Suppose that the experiment yields no heads, one head, and two heads a total of 4, 7, and 5 times, respectively. The average number of heads per toss of the two coins is then
- (0*4 + 1*7 + 2*5)/16 = 1.06
- Let us now restructure our computation for the average number of heads so as to have the following equivalent form:
- 0*(4/16) + 1*(7/16) + 2*(5/16) = 1.06
- The numbers 4/16, 7/16, and 5/16 are the fractions of the total tosses resulting in 0, 1, and 2 heads, respectively. These fractions are also the relative frequencies for the different values of *X* in our experiment.



Let X be a random variable with probability distribution $f(x)$. The **mean**, or **expected value**, of X is

$$\mu = E(X) = \sum_{x} x f(x)$$

if X is discrete, and

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\ dx$$

if X is continuous.

- A lot containing 7 components is sampled by a quality inspector; the lot contains 4 good components and 3 defective components. A sample of 3 is taken by the inspector. Find the expected value of the number of good components in this sample.
- Let *X* represent omponents in the sample. The pro of *X* is

$$f(x) = \frac{\binom{4}{x}\binom{3}{3-x}}{\binom{7}{3}}, \qquad x = 0, 1, 2, 3.$$

- Simple calculations yield *f*(0) = 1/35, *f*(1) = 12/35, *f*(2) = 18/35, and *f*(3) =4/35. Therefore,
- $\mu = E(X) =0*(1/35)+ 1*(12/35) + 2*(18/35) + 3(4/35) = 12/7=1.7$

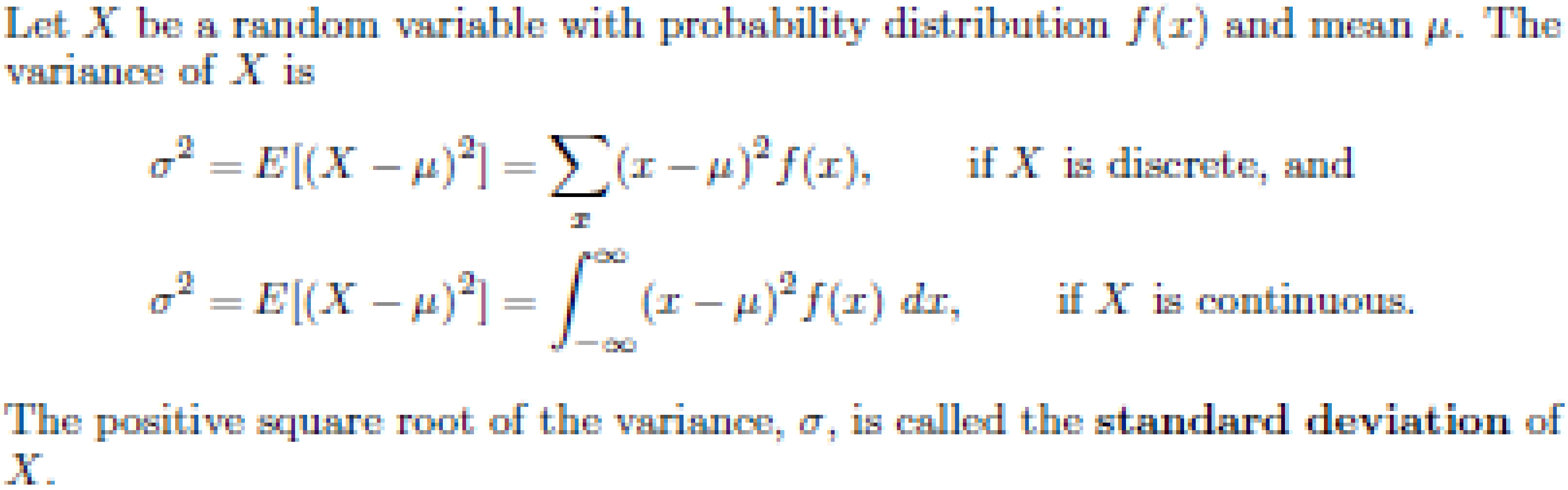
- Let *X* be the random variable that denotes the life in hours of a certain electronic device. The probability density function is

$$f(x) = \begin{cases} \frac{20{,}000}{x^3}, & x > 100, \\ 0, & \text{elsewhere.} \end{cases}$$

$$\mu = E(X) = \int_{100}^{\infty} x \frac{20,000}{x^3}\, dx = \int_{100}^{\infty} \frac{20,000}{x^2}\, dx = 200.$$

- Therefore, we can expect this type of device to last, *on an average*, 200 hours.

Variance of Random Variables



Let X be a random variable with probability distribution $f(x)$ and mean μ. The variance of X is

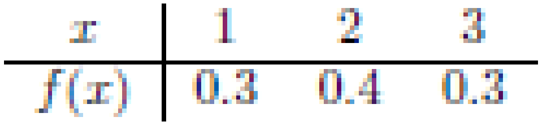
$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x), \qquad \text{if } X \text{ is discrete, and}$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx, \qquad \text{if } X \text{ is continuous.}$$

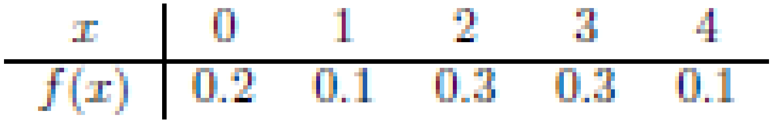
The positive square root of the variance, σ, is called the **standard deviation** of X.

- Let the random variable *X* represent the number of automobiles that are used for official business purposes on any given workday. The probability distribution for company *A*

- and that for comp



x	1	2	3
$f(x)$	0.3	0.4	0.3



x	0	1	2	3	4
$f(x)$	0.2	0.1	0.3	0.3	0.1

- Show that the vari ion for company *B* is greater than that for company *A*.

- $\mu_A = E(X) = (1)(0.3) + (2)(0.4) + (3)(0.3) = 2.0,$

- $\sigma_A^2 = \sum_{x=1}^{3} (x-2)^2$ f(x) $= (1-2)^2(0.3) + (2-2)^2(0.4) + (3-2)^2(0.3) = 0.6.$

- $\mu_B = E(X) = (0)(0.2) + (1)(0.1) + (2)(0.3) + (3)(0.3) + (4)(0.1) = 2.0$

$$\begin{aligned}\sigma_B^2 &= \sum_{x=0}^{4}(x-2)^2 f(x)\\ &= (0-2)^2(0.2) + (1-2)^2(0.1) + (2-2)^2(0.3)\\ &\quad + (3-2)^2(0.3) + (4-2)^2(0.1) = 1.6.\end{aligned}$$

The variance of a random variable X is

$$\sigma^2 = E(X^2) - \mu^2.$$

$$\begin{aligned}\sigma^2 &= \sum_x (x-\mu)^2 f(x) = \sum_x (x^2 - 2\mu x + \mu^2) f(x) \\ &= \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x).\end{aligned}$$

Since $\mu = \sum_x x f(x)$ by definition, and $\sum_x f(x) = 1$ for any discrete probability distribution, it follows that

$$\sigma^2 = \sum x^2 f(x) - \mu^2 = E(X^2) - \mu^2.$$

Discrete Probability Distribution

Discrete Uniform Probability Distribution

We may have a situation where the probabilities of each event are the same. For example, if we roll a fair die, we assume that the probability of obtaining each number is $\frac{1}{6}$

If X is the random variable (r.v.) " the number showing" , the probability distribution table is

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The probability distribution function (p.d.f.) is

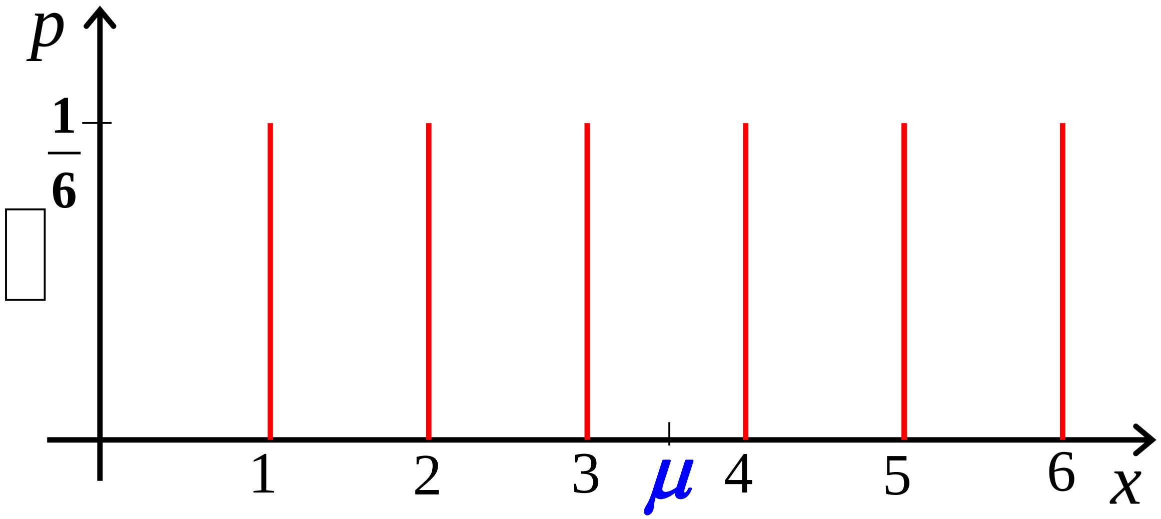
$$P(X = x) = \frac{1}{6}, \qquad x = 1,\ 2,\ 3,\ 4,\ 5,\ 6$$

The distribution with equal probabilities is called "uniform"

A diagram for the distribution

$$P(X=x)=\frac{1}{6}, \qquad x=1,\ 2,\ 3,\ 4,\ 5,\ 6$$

looks like this:



The mean value of X is given by the average of the 1^{st} and last values of x, so,

$$\mu=\frac{1+6}{2}=3\cdot 5$$

However, we could also use the formula for the mean of any discrete distribution of a random variable:

$$\mu = E(x) = \sum xf(x)$$

For $f(x) = P(X = x) = \frac{1}{6}, \qquad x = 1, 2, 3, 4, 5, 6$

we would get

$$\begin{aligned}\boldsymbol{\mu} &\mathbf{= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + . . . + 6 \times \frac{1}{6}} \\ &\mathbf{= 21 \times \frac{1}{6}} \\ &\mathbf{= 3 \cdot 5}\end{aligned}$$

$$
\begin{aligned}
\mathrm{var}(X) &= E[(X-\mu)^2] \\
&= \sum (X-\mu)^2 f(x) \\
&= \sum X^2 f(x) - 2\mu \sum X f(x) + \mu^2 \sum f(x) \\
&= \sum X^2 f(x) - 2\mu^2 + \mu^2 \\
&= \sum X^2 f(x) - \mu^2
\end{aligned}
$$

The Variance of the Uniform Distribution

We can find the variance for any discrete random variable X using

$$\mathrm{Var}(X) = \sigma^2 = \sum x^2 f(x) - \mu^2$$

e.g. The random variable X has p.d.f. given by

$$f(x) = P(X = x) = \frac{1}{6}, \qquad x = 1, 2, 3, 4, 5, 6$$

So, $\mathbf{Var}(X) = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \ldots + 6^2 \times \frac{1}{6} - \mu^2$

We found earlier that $\mu = 3{\cdot}5$, so

$$\mathrm{Var}(X) = \frac{91}{6} - 3{\cdot}5^2 = 2{\cdot}92$$

The Bernoulli Process

Bernoulli process must possess the following properties:

1. The experiment consists of repeated trials.

2. Each trial results in an outcome that may be classified as a success or a failure.

3. The probability of success, denoted by *p*, remains constant from trial to trial.

4. The repeated trials are independent.

Binomial Distribution

The number *X* of successes in *n* Bernoulli trials is called a **binomial random variable**.

The probability distribution of this discrete random variable is called the binomial distribution and its values will be denoted by b(x; n, p) where n is number of trials and p is probability of success at each trial

A Bernoulli trial can result in a success with probability p and a failure with probability q = 1−p. Then the probability distribution of the binomial random variable X, the number of successes in n independent trials, is

$$f(x) = \binom{n}{x} p^{x} (1 - p)^{n-x}$$

Examples

- A coin is flipped 10 times. What is the probability that exactly we will get 4 head?
- Flip a fair coin 10 times. X=4 heads

$$f(x)=\binom{10}{4}\left(\frac{1}{2}\right)^{4}\left(\frac{1}{2}\right)^{6}=\binom{10}{4}\left(\frac{1}{2}\right)^{10}=210*.00097=0.205$$

- Die rolled for 3 times. What is the probability that at least once 6 will result?

$$f(x)=1-\binom{3}{0}\left(\frac{1}{6}\right)^{0}\left(\frac{5}{6}\right)^{3}=1-0.5787=0.4213$$

Examples

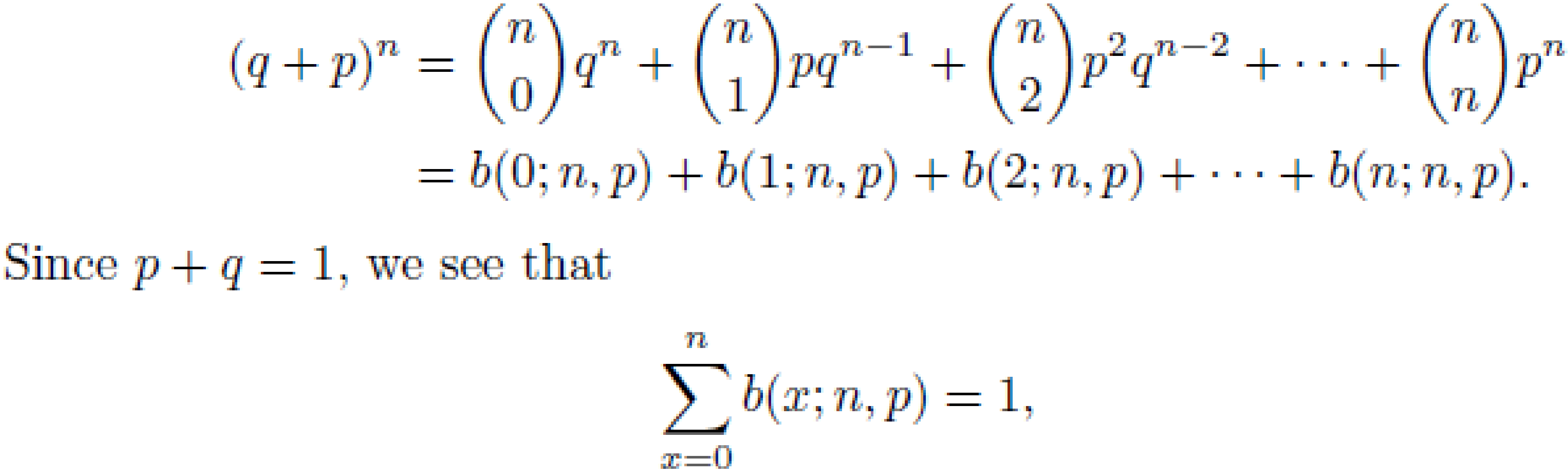
- Twelve pregnant women selected at random, take a home pregnancy test. This test give correct result with 0.8 probability. What is the probability that 10 women find a correct result?

$$f(x)=\binom{12}{10}(0.8)^{10}(0.2)^{2}=66*0.107*0.04=0.283$$

- Random guessing on a multiple choice exam. 25 questions. 4 answers per question. A person get pass marks if he correctly guesses at least 15. What is the probability that a person who does not know correct answer of any question will get a pass marks?

$$f(x)=\binom{25}{15}(0.25)^{15}(0.75)^{10}+\ldots+\binom{25}{25}(0.25)^{25}(0.75)^{0}$$

- The binomial distribution derives its name from the fact that the $n + 1$ terms in the binomial expansion of $(q+p)^n$ correspond to the various values of $b(x; n, p)$ for $x = 0, 1, 2, \ldots, n$. That is



$$(q+p)^n = \binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \binom{n}{2}p^2q^{n-2} + \cdots + \binom{n}{n}p^n$$
$$= b(0; n, p) + b(1; n, p) + b(2; n, p) + \cdots + b(n; n, p).$$

Since $p + q = 1$, we see that

$$\sum_{x=0}^{n} b(x; n, p) = 1,$$

Finding Mean and Variance

- Let the outcome of j^{th} trial be represented by indicator variable I_j which assumes the value 0 and 1 with probabilities q and p
- In binomial experiment number of success can be written as the sum of the n independent indicator variable
- $X = I_1 + I_2 + \ldots + I_n$
- Mean of $I_j = \Sigma\, [x \bullet P(x)] = 0.q + 1.p = p$

- μ = $E(X) = E(I_1)+E(I_2) +\ldots +E(I_n)=p+p+\ldots+p=np$
- $\sigma^2\ I_j = E[(I_j - p)^2]=E(I_j^2) - p^2 = (0^2)q+(1^2)p- p^2 =p(1-p)=pq$
- For n independent variable variance would be $pq+pq+\ldots+pq=npq$

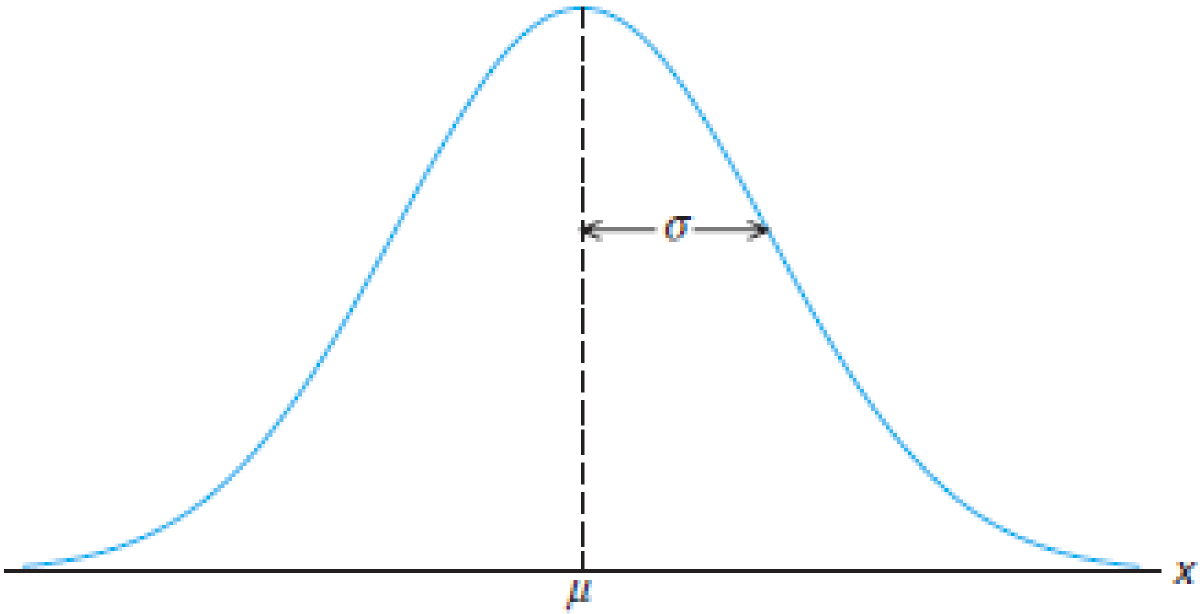
Continuous Probability Distributions

Normal Distribution

- The most important continuous probability distribution in the entire field of statistics

- Approximately describes many phenomena that occur in nature, industry, and research. For example, physical measurements in areas such as meteorological experiments, rainfall studies, and measurements of manufactured parts

- The normal distribution is often referred to as the **Gaussian distribution.**



Normal Curve

- A continuous random variable X having the bell-shaped distribution of like above Figure is called a **normal random variable**.

- The mathematical equation for the probability distribution of the normal variable depends on the two parameters μ and σ, its mean and standard deviation, respectively. Hence, we denote the values of the density of X by $n(x; \mu, \sigma)$.

The Normal Distribution: as mathematical function (pdf)

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

This is a bell shaped curve with different centers and spreads depending on μ and σ

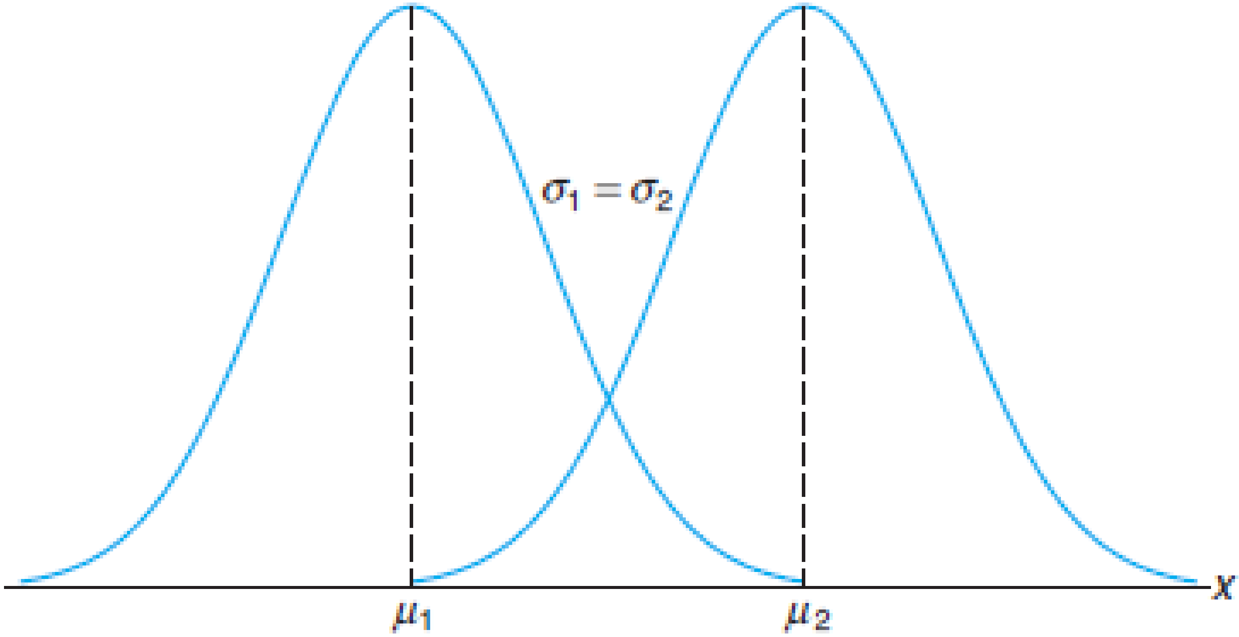
<u>Note constants:</u>

π=3.14159

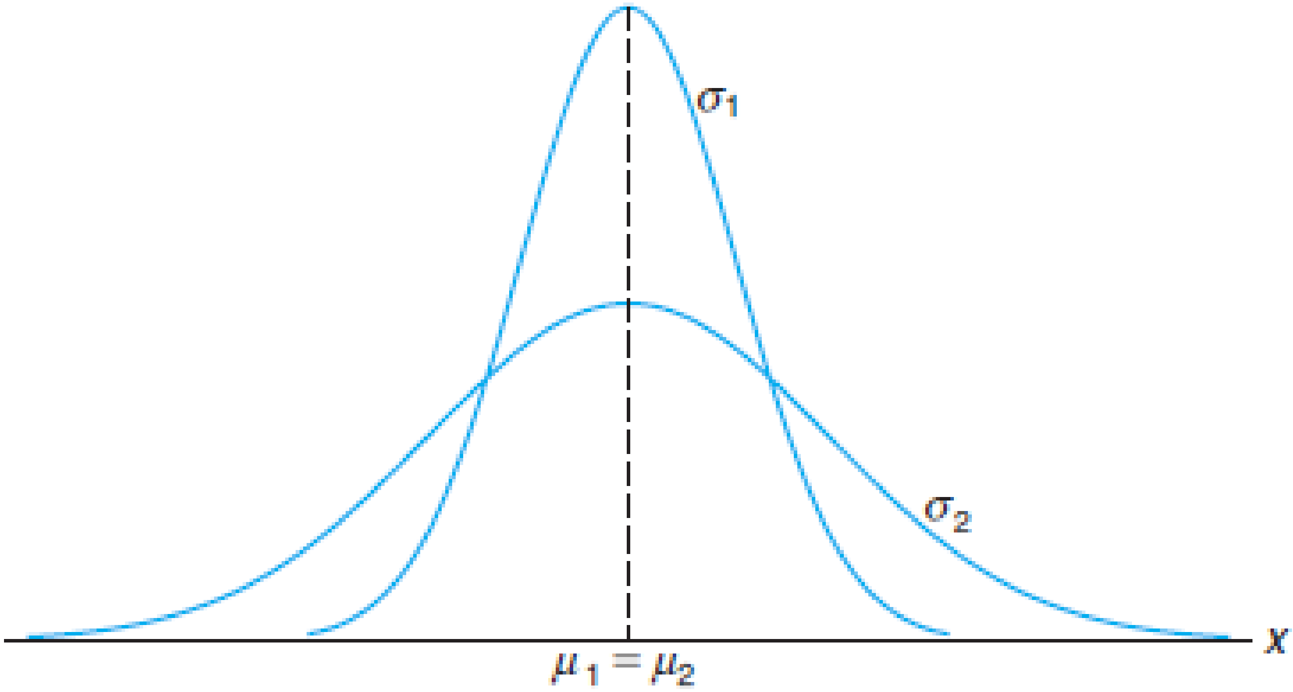
e=2.71828

Impact of μ and σ *on shape and location*

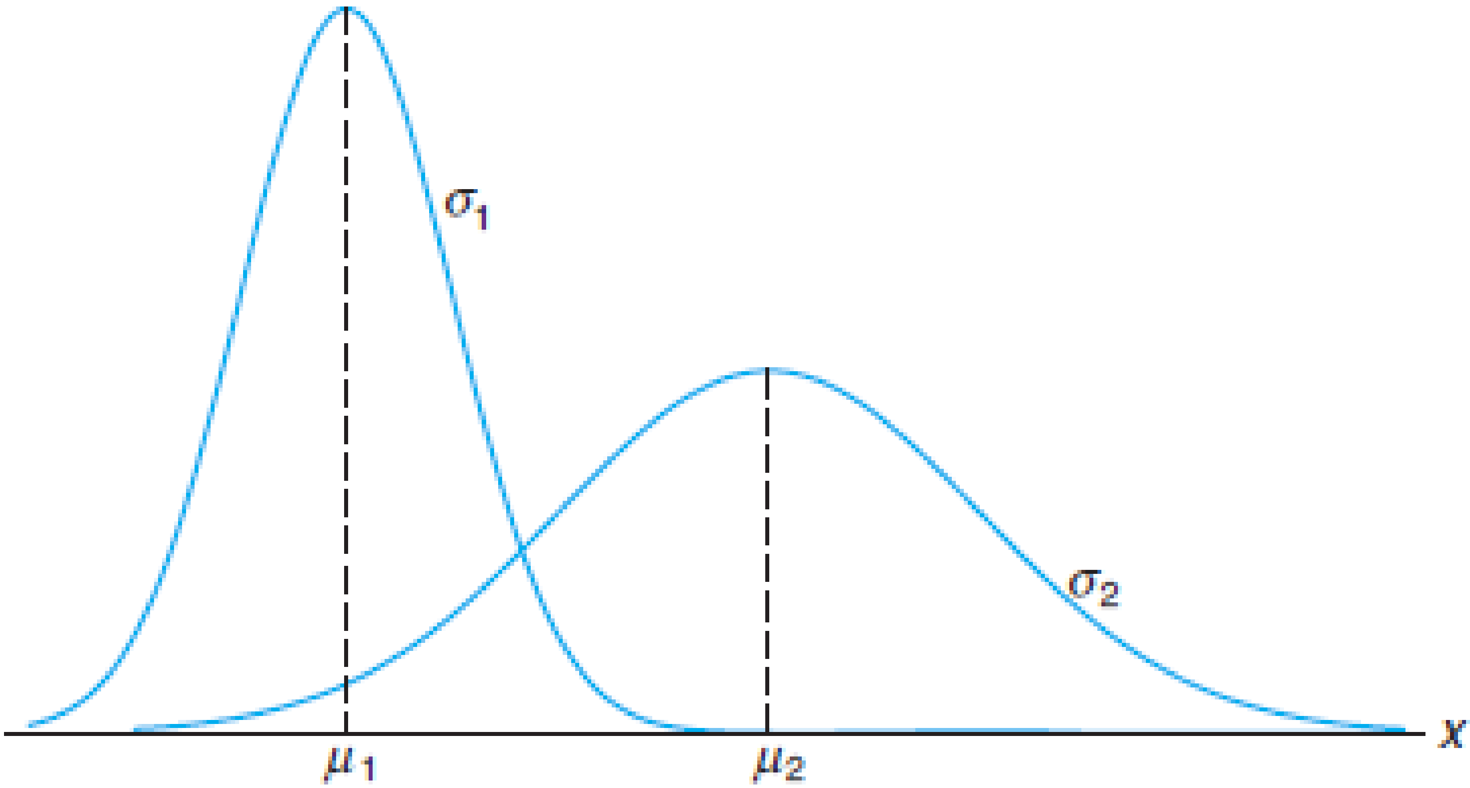
- Normal curves with $\mu1 < \mu2$ and $\sigma1 = \sigma2$.



- Normal curves with $\mu 1 = \mu 2$ and $\sigma 1 < \sigma 2$.



- Normal curves with $\mu 1 < \mu 2$ and $\sigma 1 < \sigma 2$



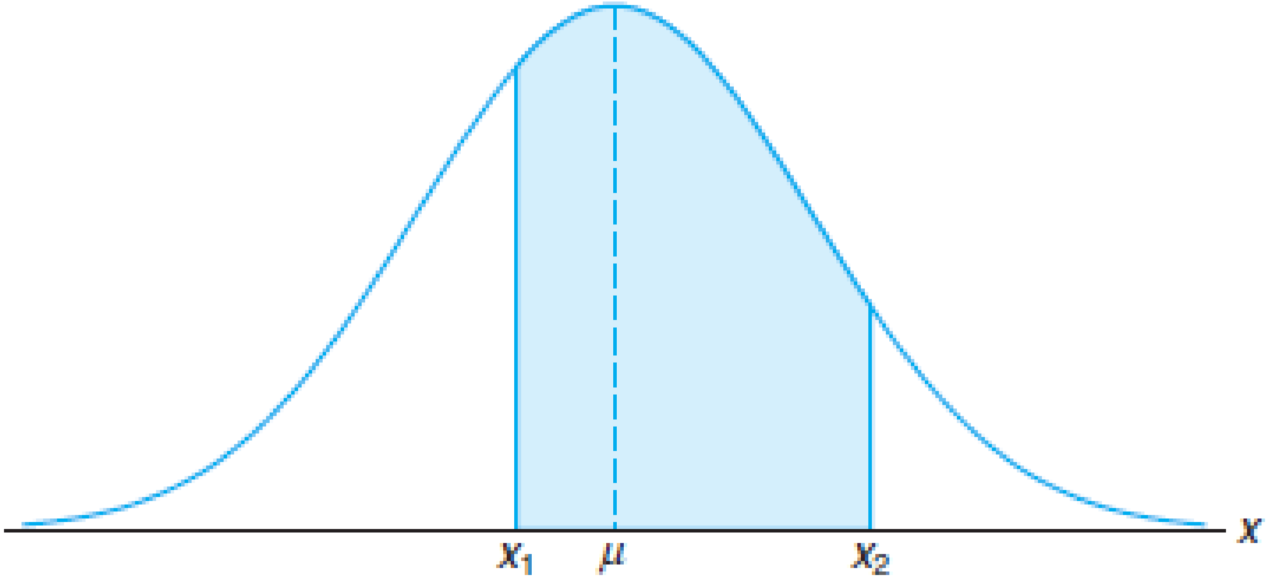
Properties of normal curve

1. The mode, which is the point on the horizontal axis where the curve is a maximum, occurs at $x = \mu$.
2. The curve is symmetric about a vertical axis through the mean μ.
3. The curve has its points of inflection at $x = \mu \pm \sigma$; it is concave downward if $\mu - \sigma < X < \mu + \sigma$ and is concave upward otherwise.
4. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.
5. The total area under the curve and above the horizontal axis is equal to 1.

Areas under the Normal Curve

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x;\mu,\sigma)\,dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx$$

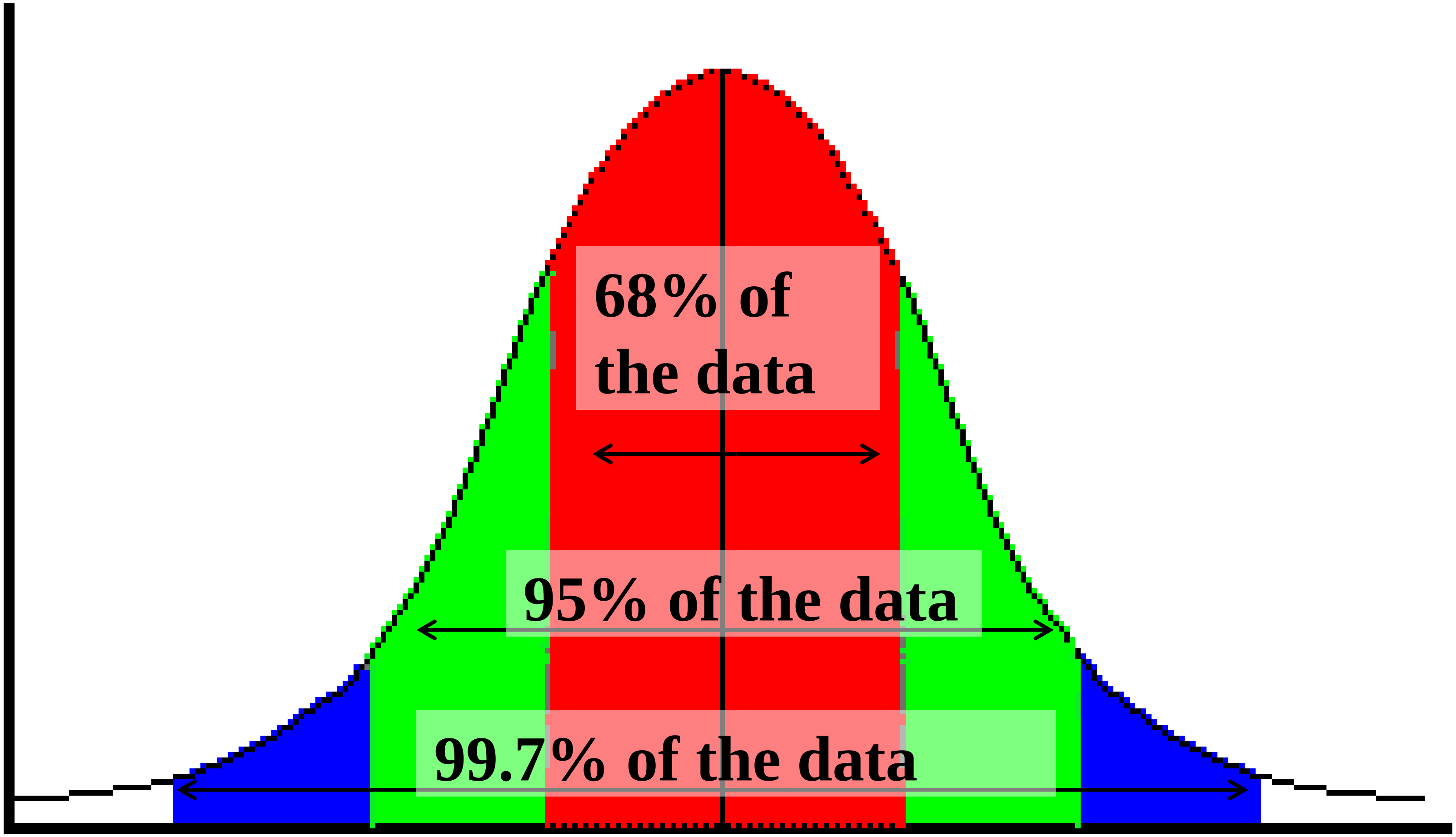
is represented by the area of the shaded region.



The beauty of the normal curve:

No matter what μ and σ are, the area between μ-σ and μ+σ is about 68%; the area between μ-2σ and μ+2σ is about 95%; and the area between μ-3σ and μ+3σ is about 99.7%. Almost all values fall within 3 standard deviations.

68-95-99.7 Rule



68-95-99.7 Rule
in Math terms…

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = 0.68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = 0.95$$

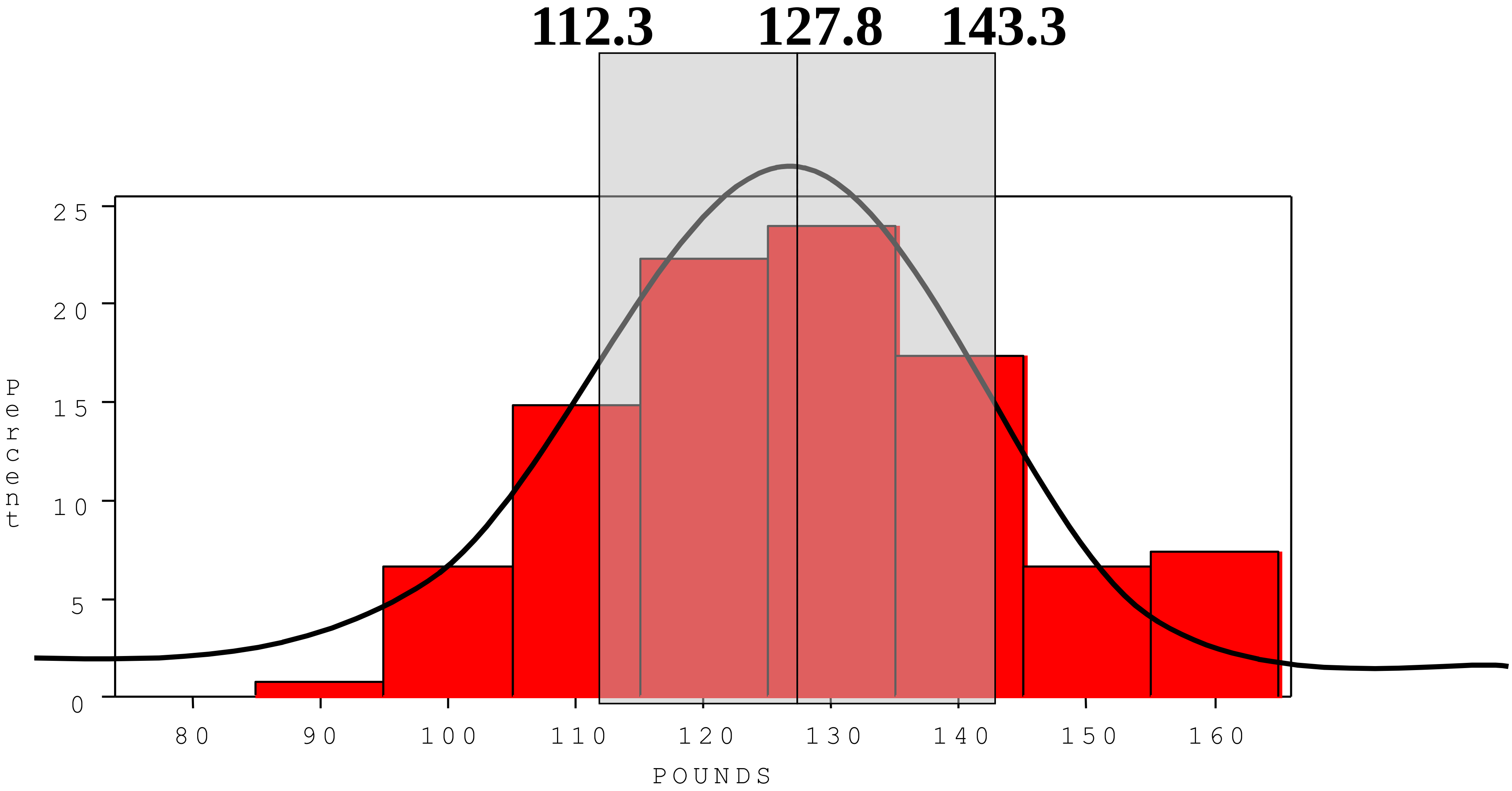
$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = 0.997$$

How good is rule for real data?

Check some example data:
The mean of the weight of the women = 127.8
The standard deviation (SD) = 15.5
Total participant = 120

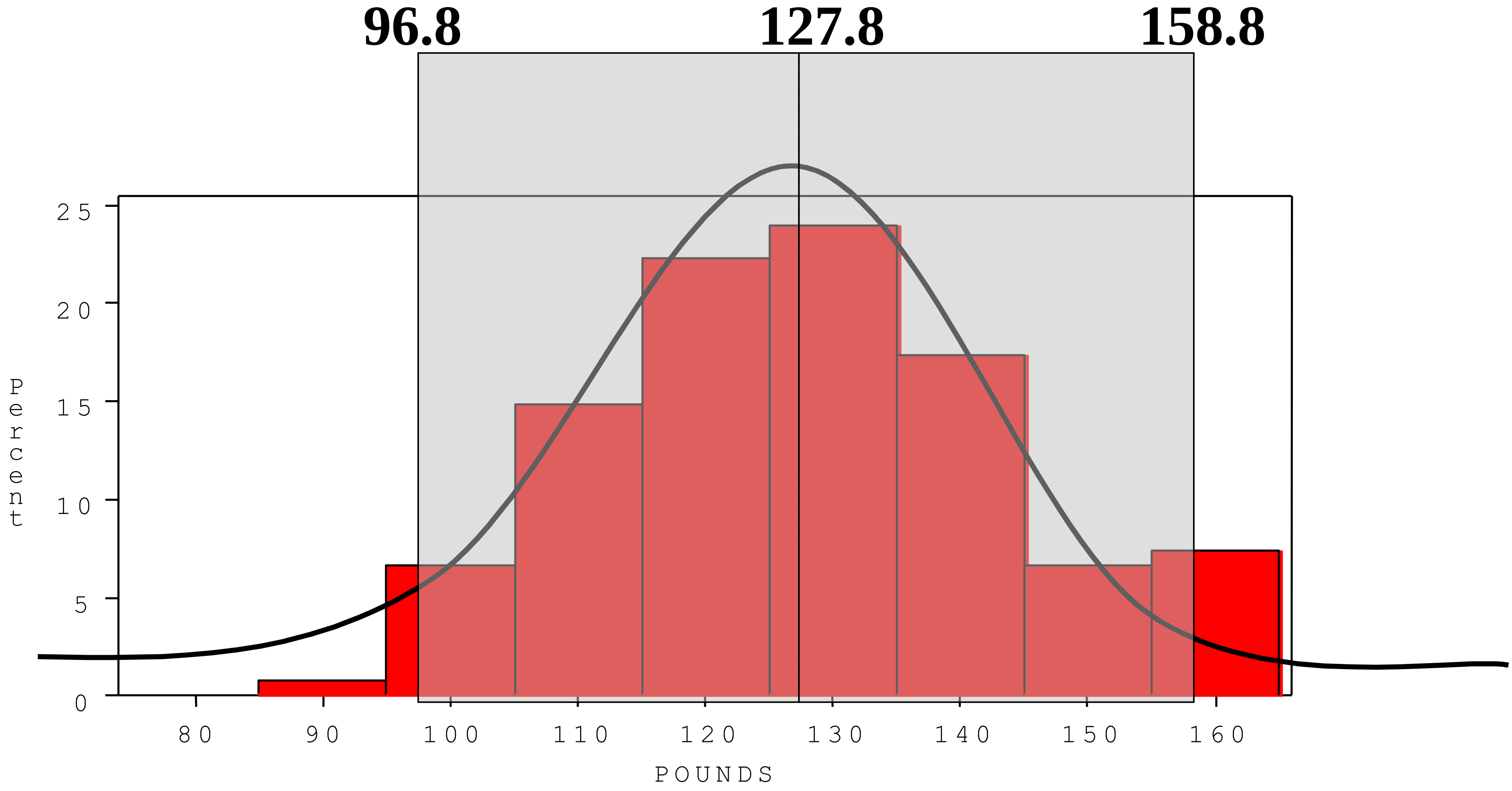
68% of 120 = 0.68x120 = ~ 82 runners

In fact, 79 runners fall within 1-SD of the mean.



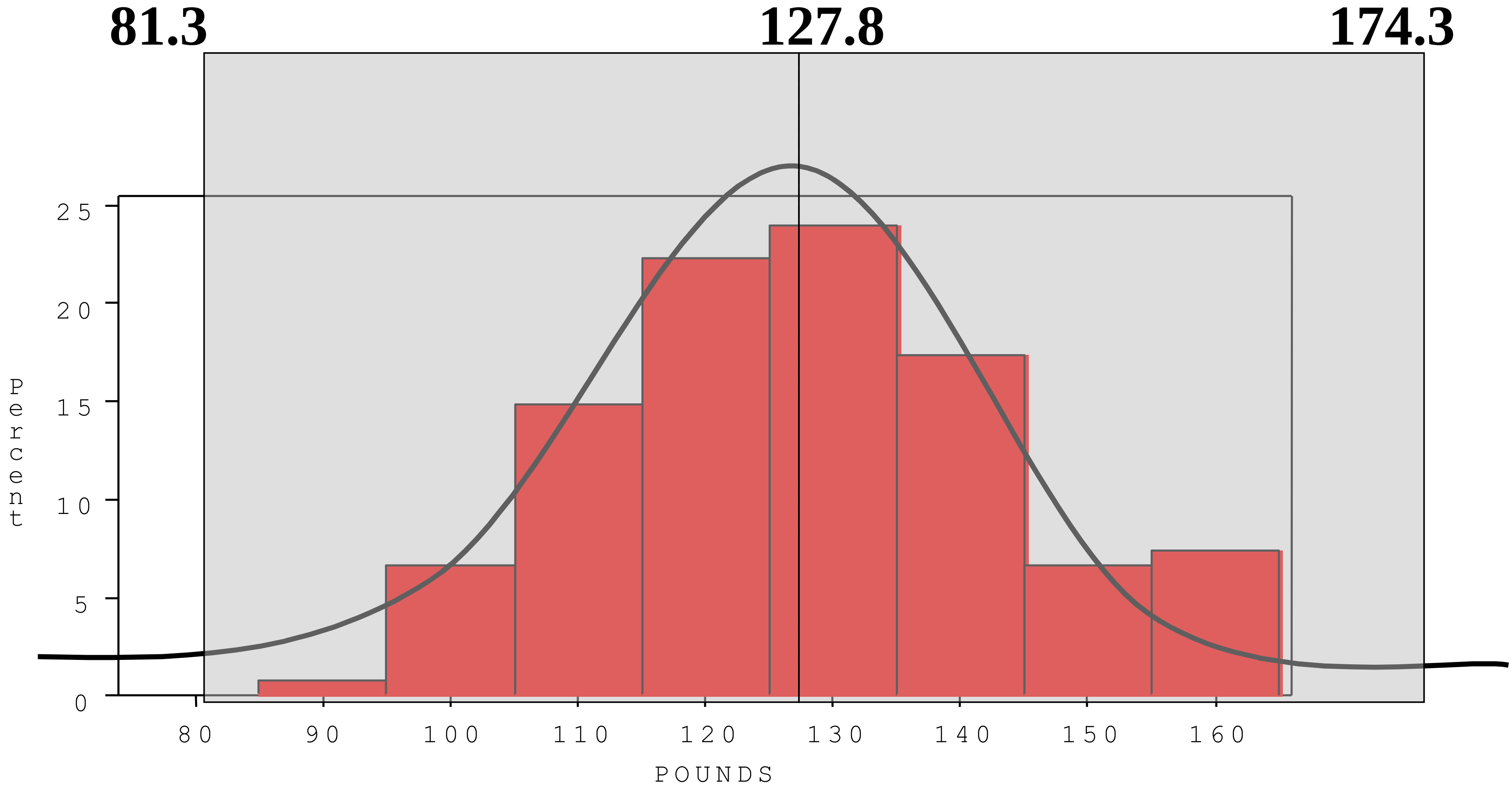
95% of 120 = 0.95 x 120 = ~ 114 runners

In fact, 115 runners fall within 2-SD's of the mean.



99.7% of 120 = 0.997 x 120 = 119.6 runners

In fact, all 120 runners fall within 3-SD's of the mean.



Example

- Suppose SAT scores roughly follows a normal distribution in the U.S. population of college-bound students (with range restricted to 200-800), and the average math SAT is 500 with a standard deviation of 50, then:
  - 68% of students will have scores between 450 and 550
  - 95% will be between 400 and 600
  - 99.7% will be between 350 and 650

Example

- BUT…
- What if you wanted to know the number of students who scores less than equal to 575

P(X≤575)

$$\int_{200}^{575} \frac{1}{(50)\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-500}{50})^2} dx$$

The Standard Normal (Z):

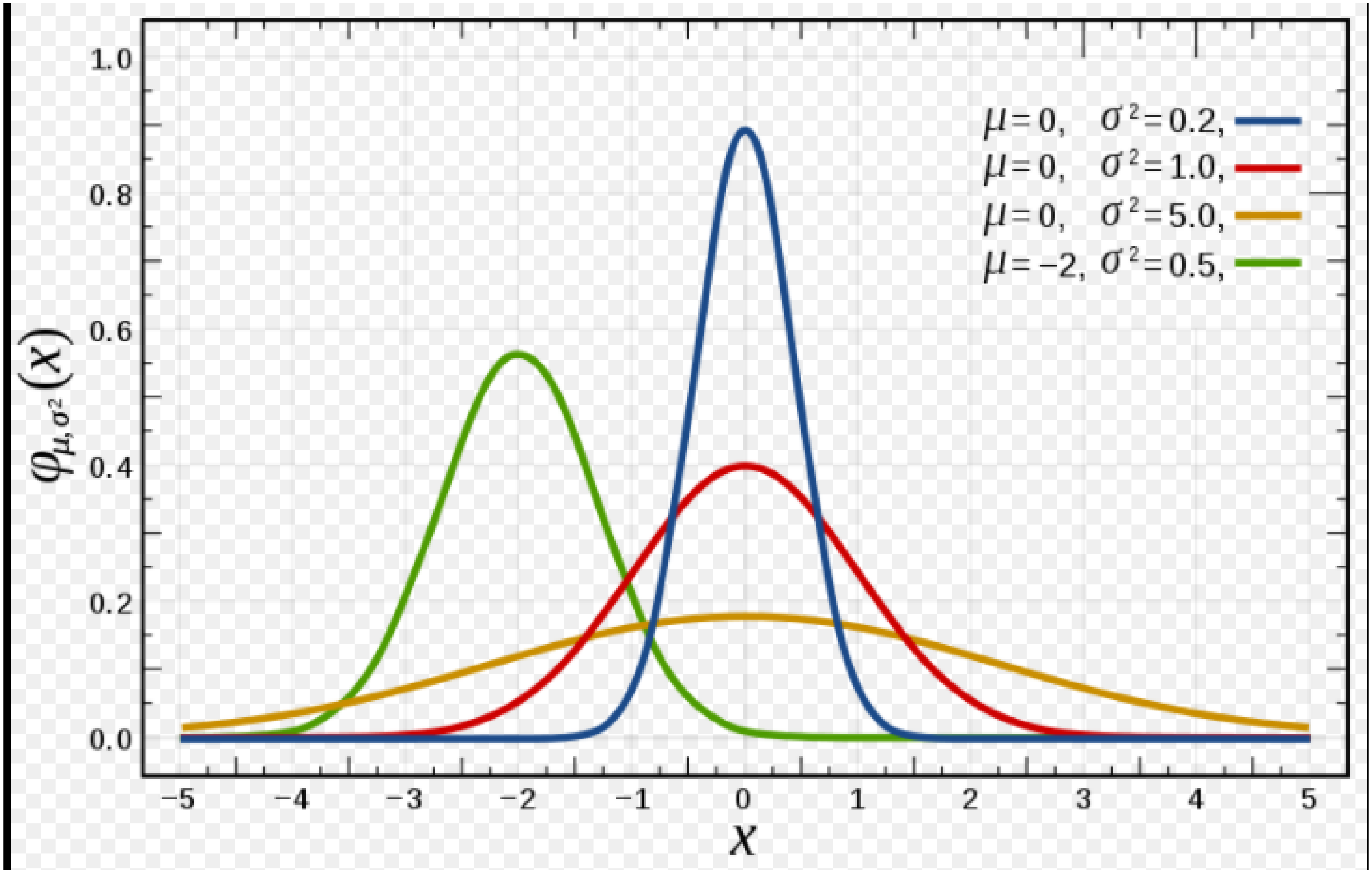
The formula for the standardized normal probability density function is

$$p(Z) = \frac{1}{(1)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{Z-0}{1})^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(Z)^2}$$

The Standard Normal Distribution (Z)

All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation:

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

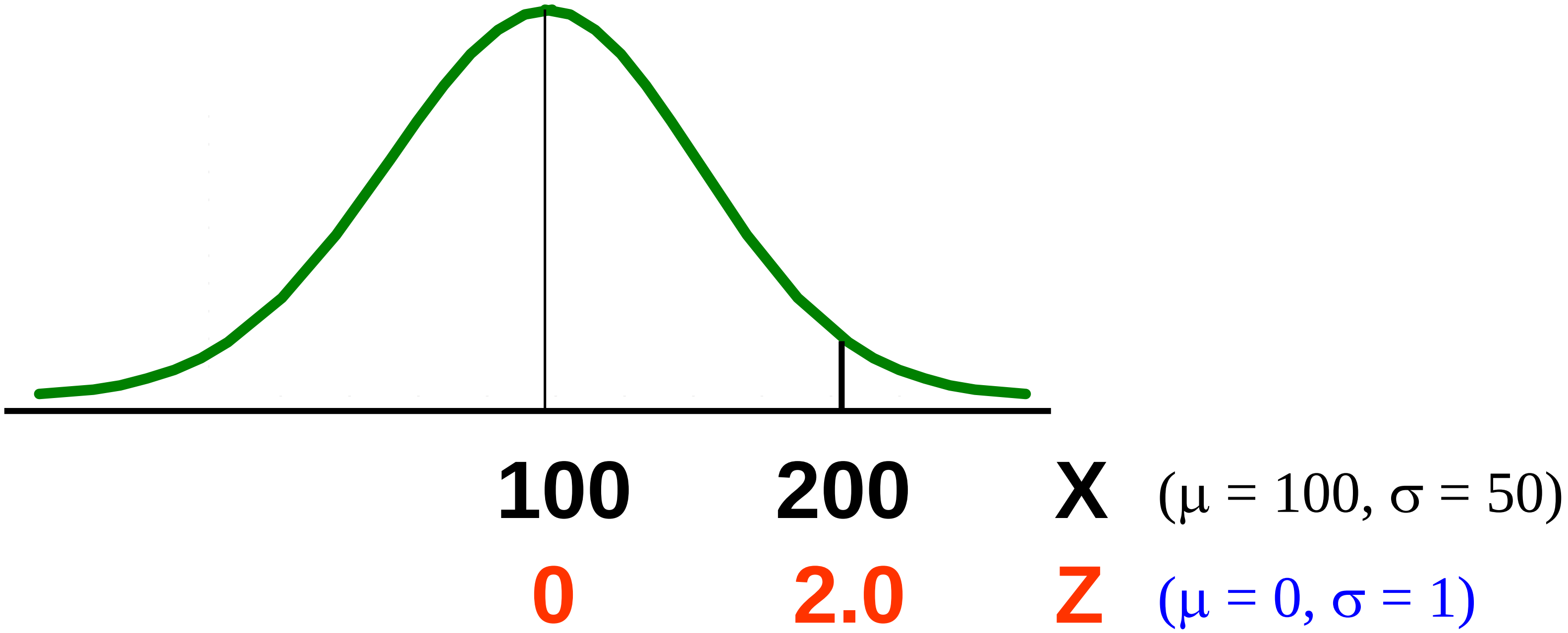

$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \sigma^2=0.5,$
$\varphi_{\mu,\sigma^2}(x)$
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5
x

- Whenever X assumes a value x, the corresponding value of Z is given by $z = (x - \mu)/\sigma$. Therefore, if X falls between the values $x = x1$ and $x = x2$, the random variable Z will fall between the corresponding values $z1 = (x1 - \mu)/\sigma$ and $z2 = (x2 - \mu)/\sigma$. Consequently, we may write

$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi}\sigma}\int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{z_1}^{z_2} e^{-\frac{1}{2}z^2}\,dz$$

$$= \int_{z_1}^{z_2} n(z;0,1)\,dz = P(z_1 < Z < z_2),$$

Comparing X and Z units



Example

- For example: What's the probability of getting a math SAT score of 575 or less, μ=500 and σ=50?

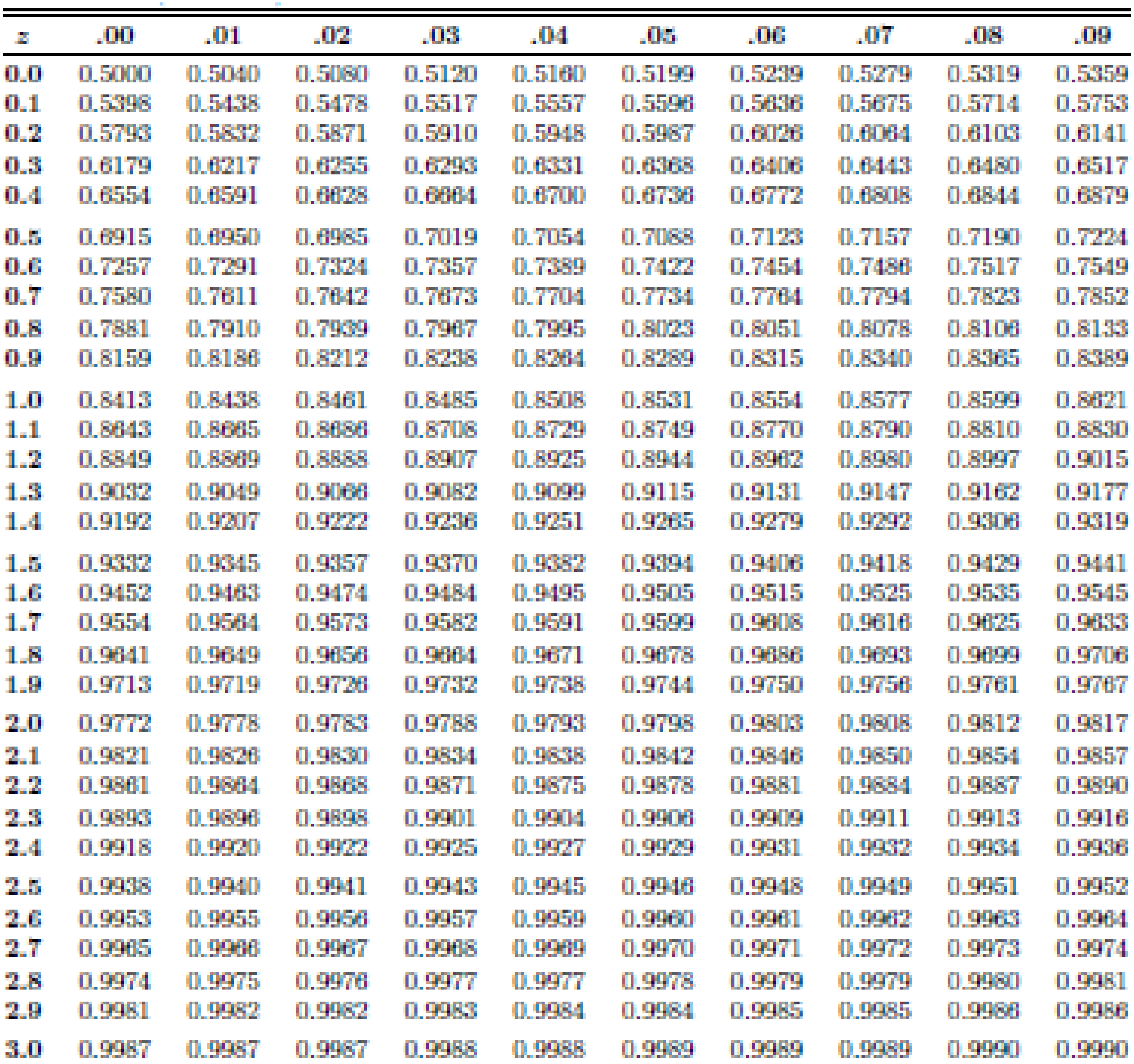
$$\boxed{Z = \frac{575-500}{50} = 1.5}$$

- i.e., A score of 575 is 1.5 standard deviations above the mean

$$\boxed{\therefore P(X \le 575) = \int_{200}^{575} \frac{1}{(50)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-500}{50})^2} dx \longrightarrow \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}Z^2} dz}$$

Yikes!

But to look up Z= 1.5 in standard normal chart = 0.9332



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

References

1. A. Zhang, Z. C. Lipton, M. Li, A. J. Smola, *Dive into Deep Learning*, https://d2l.ai, 2020.
2. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2017.
3. M. P. Deisenroth, A. A. Faisal, C. S. Ong, *Mathematics for Machine Learning*, Cambridge University Press, 2020.
4. Jeff Howbert — Machine Learning Math Essentials presentation
5. Brian Keng – Manifolds: A Gentle Introduction blog
6. Martin J. Osborne – Mathematical Methods for Economic Theory (link)