DSI Summer Prep · Brown University · Samuel S. Watson

## Sets and Functions: Sets

1 A **set** is an unordered collection of objects. The objects in a set are called *elements*.

2 The **cardinality** of a set is the number of elements it contains. The **empty set** $\varnothing$ is the set with no elements.

3 If every element of $A$ is also an element of $B$, then we say $A$ is a **subset** of $B$ and write $A \subset B$. If $A \subset B$ and $B \subset A$, then we say that $A = B$.

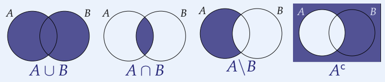
4 Set operations:

(i) An element is in the **union** $A \cup B$ of two sets $A$ and $B$ if it is in $A$ or $B$.

(ii) An element is in the **intersection** $A \cap B$ of two sets $A$ and $B$ if it is in $A$ and $B$.

(iii) An element is in the **set difference** $A \setminus B$ if it is in $A$ but not $B$.

(iv) Given a set $\Omega$ and a set $A \subset \Omega$, the **complement** of $A$ with respect to $\Omega$ is $A^c = \Omega \setminus A$.



5 Two sets $A$ and $B$ are **disjoint** if $A \cap B = \varnothing$ (in other words, if they have no elements in common).

6 A **partition** of a set is a collection of nonempty disjoint subsets whose union is the whole set.

7 The **Cartesian product** of $A$ and $B$ is

$$A \times B = \{(a,b) : a \in A \text{ and } b \in B\}.$$

8 (**De Morgan's laws**) If $A, B \subset \Omega$, then

(i) $(A \cap B)^c = A^c \cup B^c$, and

(ii) $(A \cup B)^c = A^c \cap B^c$.

9 A **list** is an ordered collection of finitely many objects. Lists differ from sets in that (i) order matters, (ii) repetition matters, and (iii) the cardinality is restricted.

## Sets and Functions: Functions

1 If $A$ and $B$ are sets, then a **function** $f : A \to B$ is an assignment of some element of $B$ to each element of $A$.

2 The set $A$ is called the **domain** of $f$ and $B$ is called the **codomain** of $f$.

3 Given a subset $A'$ of $A$, we define the **image** of $f$—denoted $f(A')$—to be the set of elements which are mapped to from some element in $A'$.

4 The **range** of $f$ is the image of the domain of $f$.

5 The **composition** of two functions $f : A \to B$ and $g : B \to C$ is the function $g \circ f$ which maps $a \in A$ to $g(f(a)) \in C$.

6 The **identity function** on a set $A$ is the function $f : A \to A$ which maps each element to itself.

7 A function $f$ is **injective** if no two elements in the domain map to the same element in the codomain.

8 A function $f$ is **surjective** if the range of $f$ is equal to the codomain of $f$.

9 A function $f$ is **bijective** if it is both injective and surjective. If $f$ is bijective, then the function from $B$ to $A$ that maps $b \in B$ to the element $a \in A$ that satisfies $f(a) = b$ is called the **inverse** of $f$.

10 If $f : A \to B$ is bijective, then the function $f^{-1} \circ f$ is equal to the identity function on $A$, and $f \circ f^{-1}$ is the identity function on $B$.

## Programming in Python

1 A **value** is a fundamental entity that may be manipulated by a program. Values have **types**; for example, `5` is an `int` and `"Hello world!"` is a `str`.

2 A **variable** is a name used to refer to a value. We can **assign** a value `5` to a variable `x` using `x = 5`.

3 A **function** performs a particular task. You prompt a function to perform its task by **calling** it. Values supplied to a function are called **arguments**. For example, in the function call `print(1,2)`, `1` and `2` are arguments.

4 An **operator** is a function that can be called in a special way. For example, `*` is an operator since we can call the multiplication function with the syntax `3 * 5`.

5 A **statement** is an instruction to be executed (like `x = -3`).

6 An **expression** is a combination of values, variables, operators, and function calls that a language interprets and **evaluates** to a value.

7 A numerical value can be either an **integer** or a **float**. The basic operations are `+`,`-`,`*`,`/`,`**`, and expressions are evaluated according to the order of operations.

8 Numbers can be compared using `<`,`>`,`==`,`<=` or `>=`.

9 Textual data is represented using **strings**. `len(s)` returns the number of characters in `s`. The `+` operator concatenates strings.

10 A **boolean** is a value which is either `True` or `False`. Booleans can be combined with the operators `and`, `or`, or `not`.

11 Code blocks can be executed conditionally:

```
if x > 0:
    print("x is positive")
elif x == 0:
    print("x is zero")
else:
    print("x is negative")
```

12 Functions may be defined using `def` (`show_temp` is a **keyword argument**):

```
def fahrenheit_to_celsius(F, show_temp = False):
    if show_temp:
        print("Original temp is " + str(F))
    return 5/9 * (F - 32)
```

13 The **scope** of a variable is the region in the program where it is accessible. Variables defined in the body of a function are not accessible outside the body of the function.

14 `list` is a compound data type for storing lists of objects. Entries of a list may be accessed with square bracket syntax using an index (starting from 0) or using a **slice** `a:b`.

```
A = [-5,3,11,1]
A[0] # first element (-5)
A[2:] # sublist from 2 to end ([11,1])
A[:3] # sublist from beginning to 2 ([-5,3,11])
```

15 A **list comprehension** can be used to generate new lists:

```
[k**2 for k in range(10) if k % 2 == 0]
```

16 A **dictionary** encodes a discrete function by storing input-output pairs and looking up input values when indexed.

```
colors = {"blue": [0,0,1.0], "red": [1.0,0,0]}
colors["blue"] # returns [0,0,1.0]
```

17 A `while` loop takes a conditional expression and a body and evaluates them alternatingly until the conditional expression returns false. A `for` loop evaluates its body once for each entry in a given *iterator* (for example, a range, list, or dictionary). Each value in the iterator is assigned to a loop variable which can be referenced in the body of the loop.

```
while x > 0:              for i in range(10):
    x -= 1                    print(i)
```

## Programming Design Principles

1 **Program design.** A well-structured program is easier to read and maintain than a poorly structured one. It will also run more reliably and require less debugging.

(i) **Don't repeat yourself**. Use abstractions (loops, functions, objects, etc.) to avoid repetition. A given piece of information or functionality should live in one place only.

(ii) **Separation of concerns**. Distinct functionality should be supplied by distinct sections of code.

(iii) **Simplify**. Don't introduce unnecessary complexity.

(iv) **Use informative names**. Choose names for variables and functions which elucidate their role in the program.

(v) **Comment**. Document any features of your program which are not immediately apparent from the code.

2 **Test-first design** is an approach to developing code which aims to improve productivity and reliability. When writing a function:

(i) Begin by writing the **signature**, that is, the function name, parameters, and docstring.

(ii) *Before* writing the body of the function, write **tests** for the function. Think carefully about the desired behavior, including degenerate and corner cases.

(iii) Write the body of the function.

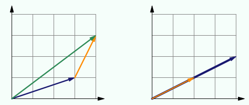
(iv) Run the tests. If some of them fail, address the failures and run all of the tests again.

`pytest` is a package for implementing test-first design in Python.

## Linear Algebra: Vector Spaces

1 A **vector** in $\mathbb{R}^n$ is a column of $n$ real numbers, also written as $[v_1, \ldots, v_n]$. A vector may be depicted as an arrow from the origin in $n$-dimensional space. The **norm** of a vector $\mathbf{v}$ is the length $\sqrt{v_1^2 + \cdots + v_n^2}$ of its arrow.

2 The fundamental vector space operations are **vector addition** and **scalar multiplication**.



3 A **linear combination** of a list of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is an expression of the form

$$c_1\mathbf{v_1} + c_2\mathbf{v_2} + \cdots + c_k\mathbf{v_k},$$

where $c_1, \ldots, c_k$ are real numbers. The $c$'s are called the **weights** of the linear combination.

4 The **span** of a list $L$ of vectors is the set of all vectors which can be written as a linear combination of the vectors in $L$.

5 A list of vectors is **linearly independent** if and only if the only linear combination which yields the zero vector is the one with all weights zero.

6 A **vector space** is a nonempty set of vectors which is closed under the vector space operations.

7 A list of vectors in a vector space is a **spanning list** of that vector space if every vector in the vector space can be written as a linear combination of the vectors in that list.

8 A linearly independent spanning list of a vector space is called a **basis** of that vector space. The number of vectors in a basis of a vector space is called the **dimension** of the space.

9 A **linear transformation** $L$ is a function from a vector space $V$ to a vector space $W$ which satisfies $L(c\mathbf{v} + \beta\mathbf{w}) = cL(\mathbf{v}) + L(\mathbf{w})$ for all $c \in \mathbb{R}$, $\mathbf{u}, \mathbf{v} \in V$. These are "flat maps": equally spaced lines are mapped to equally spaces lines or points. Examples: scaling, rotation, projection, reflection.

10 Given two vector spaces $V$ and $W$, a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $V$, and a list $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ of vectors in $W$, there exists one and only one linear transformation which maps $\mathbf{v}_1$ to $\mathbf{w}_1$, $\mathbf{v}_2$ to $\mathbf{w}_2$, and so on.

11 The **rank** of a linear transformation from one vector space to another is the dimension of its range.

12 The **null space** of a linear transformation is the set of vectors which are mapped to the zero vector by the linear transformation.

13 The rank of a transformation plus the dimension of its null space is equal to the dimension of its domain (the **rank-nullity theorem**).

## Linear Algebra: Matrix Algebra

1 The **matrix-vector product** $A\mathbf{x}$ is the linear combination of the columns of $A$ with weights given by the entries of $\mathbf{x}$.

2 Linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ are in one-to-one correspondence with $m \times n$ matrices.

3 The identity transformation corresponds to the **identity matrix**, which has entries of 1 along the diagonal and zero entries elsewhere.

4 **Matrix multiplication** corresponds to composition of the corresponding linear transformations: $AB$ is the matrix for which $(AB)(\mathbf{x}) = A(B\mathbf{x})$ for all $\mathbf{x}$.

5 A $m \times n$ matrix is **full rank** if its rank is equal to $\min(m, n)$

6 $A\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x}$ if and only if $\mathbf{b}$ is in the span of the columns of $A$. If $A\mathbf{x} = \mathbf{b}$ does have a solution, then the solution is unique if and only if the columns of $A$ are linearly independent. If $A\mathbf{x} = \mathbf{b}$ does not have a solution, then there is a unique vector $\mathbf{x}$ which minimizes $|A\mathbf{x} - \mathbf{b}|^2$.

7 If the columns of a square matrix $A$ are linearly independent, then it has a unique **inverse matrix** $A^{-1}$ with the property that $A\mathbf{x} = \mathbf{b}$ implies $\mathbf{x} = A^{-1}\mathbf{b}$ for all $\mathbf{x}$ and $\mathbf{b}$.

8 Matrix inversion satisfies $(AB)^{-1} = B^{-1}A^{-1}$ if $A$ and $B$ are both invertible.

9 The **transpose** $A'$ of a matrix $A$ is defined so that the rows of $A'$ are the columns of $A$ (and vice versa).

10 The transpose is a linear operator: $(cA + B)' = cA' + B'$ if $c$ is a constant and $A$ and $B$ are matrices.

11 The transpose distributes across matrix multiplication but with an order reversal: $(AB)' = B'A'$ if $A$ and $B$ are matrices for which $AB$ is defined.

12 A matrix $A$ is **symmetric** if $A = A'$.

13 A linear transformation $T$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ scales all $n$-dimensional volumes by the same factor: the (absolute value of the) **determinant** of $T$.

14 The sign of the determinant tells us whether $T$ reverses orientations.

15 $\det AB = \det A \det B$ and $\det A^{-1} = (\det A)^{-1}$.

16 A square matrix is invertible if and only if its determinant is nonzero.

## Linear Algebra: Orthogonality

1 The **dot product** of two vectors in $\mathbb{R}^n$ is defined by

$$\mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

2 $\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\| \cos\theta$, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\theta$ is the angle between the vectors.

DSI Summer Prep · Brown University · Samuel S. Watson

**3** $\mathbf{x} \cdot \mathbf{y} = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal.

**4** The dot product is linear: $\mathbf{x} \cdot (c\mathbf{y} + \mathbf{z}) = c\mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}$.

**5** The **orthogonal complement** of a subspace $V \subset \mathbb{R}^n$ is the set of vectors which are orthogonal to every vector in $V$.

**6** The orthogonal complement of the span of the columns of a matrix $A$ is equal to the null space of $A'$.

**7** $\operatorname{rank} A = \operatorname{rank} A'A$ for any matrix $A$.

**8** A list of vectors satisfying $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ for $i \neq j$ is **orthogonal**. An orthogonal list of unit vectors is **orthonormal**.

**9** Every orthogonal list is linearly independent

**10** A matrix $U$ has orthonormal columns if and only if $U'U = I$. A square matrix with orthonormal columns is called **orthogonal**. An orthogonal matrix and its transpose are inverses.

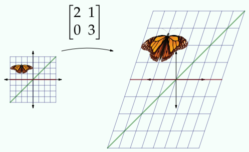
**11** Orthogonal matrices represent **rigid transformations** (ones which preserve lengths and angles).

**12** If $U$ has orthonormal columns, then $UU'$ is the matrix which represents projection onto the span of the columns of $U$.

## Linear Algebra: Spectral Analysis

**1** An **eigenvector** $\mathbf{v}$ of an $n \times n$ matrix $A$ is a nonzero vector with the property that $A\mathbf{v} = \lambda\mathbf{v}$ for some $\lambda \in \mathbb{R}$. We call $\lambda$ an **eigenvalue**.

If $\mathbf{v}$ is an eigenvector of $A$, then $A$ maps the line $\operatorname{span}(\{\mathbf{v}\})$ to itself:



**2** Eigenvectors of $A$ with distinct eigenvalues are linearly independent.

**3** Not every $n \times n$ matrix $A$ has $n$ linearly independent eigenvectors. If $A$ does have $n$ linearly independent eigenvectors, we can make a matrix $V$ with these eigenvectors as columns and get

$$AV = V\Lambda \implies A = V\Lambda V^{-1} \quad (\textbf{diagonalization} \text{ of } A)$$

where $\Lambda$ is a diagonal matrix of eigenvalues.

**4** If $A = V\Lambda V^{-1}$, then $A^n = V\Lambda^n V^{-1}$.

**5** If $A$ is a symmetric matrix, then $A$ is **orthogonally diagonalizable**:

$$A = V\Lambda V',$$

where $V$ is an orthogonal matrix (the **spectral theorem**).

**6** A symmetric matrix is **positive semidefinite** if its eigenvalues are all nonnegative. We define the square root of a positive semidefinite matrix $A = V\Lambda V'$ to be $V\sqrt{\Lambda}V'$, where $\sqrt{\Lambda}$ is obtained by applying the square root function elementwise.

## Multivariable calculus

**1** A **sequence** of real numbers $(x_n)_{n=1}^\infty = x_1, x_2, \ldots$ **converges** to a number $x \in \mathbb{R}$ if the distance from $x_n$ to $x$ on the number line can be made as small as desired by choosing $n$ sufficiently large. We say $\lim_{n\to\infty} x_n = x$ or $x_n \to x$.

**2** (**Squeeze theorem**) If $a_n \leq b_n \leq c_n$ for all $n \geq 1$ and if $\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = b$, then $b_n \to b$ as $n \to \infty$.

**3** (**Comparison test**) If $\sum_{n=1}^\infty b_n$ converges and if $|a_n| \leq b_n$ for all $n$, then $\sum_{n=1}^\infty a_n$ converges.

Conversely, if $\sum_{n=1}^\infty b_n$ does not converge and $0 \leq b_n < a_n$, then $\Sigma_{n=1}^\infty a_n$ also does not converge.

**4** The series $\sum_{n=1}^\infty n^p$ converges if and only if $p < -1$. The series $\sum_{n=1}^\infty a^n$ converges if and only if $-1 < a < 1$.

**5** The **Taylor series**, centered at $c$, of an infinitely differentiable function $f$ is defined to be

$$f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \frac{f'''(c)}{3!}(x-c)^3 + \cdots$$

**6** We can multiply or add Taylor series term-by-term, we can integrate or differentiate a Taylor series term-by-term, we can substitute one Taylor series into another to obtain a Taylor series for the composition.

**7** The **partial derivative** $\frac{\partial f}{\partial x}(x_0, y_0)$ of a function $f(x,y)$ at a point $(x_0, y_0)$ is the slope of the graph of $f$ in the $x$-direction at the point $(x_0, y_0)$.

**8** Given $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, we define $\partial\mathbf{f}/\partial\mathbf{x}$ to be the matrix whose $(i,j)$th entry is $\partial f_i/\partial x_j$. Then

$$\text{(i)}\ \frac{\partial}{\partial \mathbf{x}}(A\mathbf{x}) = A \qquad \text{(ii)}\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'A) = A'$$
$$\text{(iii)}\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{u}'\mathbf{v}) = \mathbf{u}'\frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}'\frac{\partial \mathbf{u}}{\partial \mathbf{x}}.$$

**9** A function of two variables is **differentiable** at a point if its graph looks like a plane when you zoom in sufficiently around the point. More generally, a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{x}$ if it is well-approximated by its derivative near $\mathbf{x}$:

$$\lim_{\Delta\mathbf{x}\to\mathbf{0}} \frac{\mathbf{f}(\mathbf{x}+\Delta\mathbf{x}) - \left(\mathbf{f}(\mathbf{x}) + \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x})\Delta\mathbf{x}\right)}{|\Delta\mathbf{x}|} = 0.$$

**10** The **Hessian** $\mathcal{H}$ of $f : \mathbb{R}^n \to \mathbb{R}$ is the matrix of its second order derivatives: $\mathcal{H}_{i,j}(\mathbf{x}) = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}f(\mathbf{x})$.

The **quadratic approximation** of $f$ at the origin is $f(\mathbf{0}) + \frac{\partial f}{\partial \mathbf{x}}(\mathbf{0})\mathbf{x} + \frac{1}{2}\mathbf{x}'\mathcal{H}(\mathbf{0})\mathbf{x}$.

**11** Suppose that $f$ is a continuous function defined on a closed and bounded subset $D$ of $\mathbb{R}^n$. Then:

(i) $f$ realizes an absolute maximum and absolute minimum on $D$ (the **extreme value theorem**).

(ii) Any point where $f$ realizes an extremum is either a critical point—meaning that $\nabla f = 0$ or $f$ is non-differentiable at that point—or at a point on the boundary.

(iii) (**Lagrange multipliers**) If $f$ realizes an extremum at a point on a portion of the boundary which is the level set of a differentiable function $g$ with non-vanishing gradient $\nabla g$, then either $f$ is non-differentiable at that point or the equation

$$\nabla f = \lambda \nabla g$$

is satisfied at that point, for some $\lambda \in \mathbb{R}$.

**12** If $\mathbf{r} : \mathbb{R}^1 \to \mathbb{R}^2$ and $f : \mathbb{R}^2 \to \mathbb{R}^1$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \mathbf{r}) = \frac{\partial f}{\partial \mathbf{r}}(\mathbf{r}(t))\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}(t). \quad (\textbf{chain rule})$$

**13** Integrating a function is a way of totaling up its values. $\iint_D f(x,y)\,\mathrm{d}x\,\mathrm{d}y$ can be interpreted as the mass of an object occupying the region $D$ and having mass density $f(x,y)$ at each point $(x,y)$.

**14** Double integration over $D$: the bounds for the outer integral are the smallest and largest values of $y$ for any point in $D$, and the bounds for the inner integral are the smallest and largest values of $x$ for any point in a given "$y =$ constant" slice of the region.

**15** Polar integration over $D$: the outer integral bounds are the least and greatest values of $\theta$ for a point in $D$, and the inner integral bounds are the least and greatest values of $r$ for any point in $D$ along each given "$\theta =$ constant" ray. The area element is $\mathrm{d}A = r\,\mathrm{d}r\,\mathrm{d}\theta$.

## Probability: Probability Spaces

**1** Given a random experiment, the set of possible outcomes is called the **sample space** $\Omega$, like $\{(\mathtt{H},\mathtt{H}), (\mathtt{H},\mathtt{T}), (\mathtt{T},\mathtt{H}), (\mathtt{T},\mathtt{T})\}$.

**2** We associate with each outcome $\omega \in \Omega$ a **probability mass**, denoted $m(\omega)$. For example, $m((\mathtt{H},\mathtt{T})) = \frac{1}{4}$.

**3** In a random experiment, an **event** is a predicate that can be determined based on the outcome of the experiment (like "first flip turned up heads"). Mathematically, an event is a subset of $\Omega$ (like $\{(\mathtt{H},\mathtt{H}), (\mathtt{H},\mathtt{T})\}$).

**4** Basic set operations $\cup$, $\cap$, and $^{\mathsf{c}}$ correspond to disjunction, conjunction, and negation of events:

(i) The event that $E$ happens or $F$ happens is $E \cup F$.

(ii) The event that $E$ happens and $F$ happens is $E \cap F$.

(iii) The event that $E$ does not happen is $E^{\mathsf{c}}$.

**5** If $E$ and $F$ cannot both occur (that is, $E \cap F = \emptyset$), we say that $E$ and $F$ are **mutually exclusive** or **disjoint**.

**6** If $E$'s occurrence implies $F$'s occurrence, then $E \subset F$.

**7** The probability $\mathbb{P}(E)$ of an event $E$ is the sum of the probability masses of the outcomes in that event. The domain of $\mathbb{P}$ is $2^\Omega$, the set of all subsets of $\Omega$.

**8** The pair $(\Omega, \mathbb{P})$ is called a probability space. The fundamental probability space properties are

(i) $\mathbb{P}(\Omega) = 1$ — "something has to happen"

(ii) $\mathbb{P}(E) \geq 0$ — "probabilities are non-negative"

(iii) $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ if $E$ and $F$ are mutually exclusive — "probability is additive".

**9** Other properties which follow from the fundamental ones:

(i) $\mathbb{P}(\emptyset) = 0$

(ii) $\mathbb{P}(E^{\mathsf{c}}) = 1 - \mathbb{P}(E)$

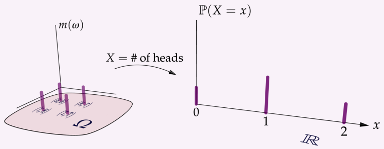
(iii) $E \subset F \implies \mathbb{P}(E) \leq \mathbb{P}(F)$ (monotonicity)

(iv) $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$ (principle of inclusion-exclusion).

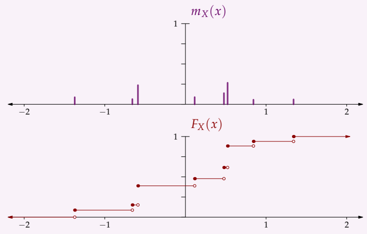
## Probability: Random Variables

**1** A **random variable** is a number which depends on the result of a random experiment (one's lottery winnings, for example). Mathematically, a random variable is a function $X$ from the sample space $\Omega$ to $\mathbb{R}$.

**2** The **distribution** of a random variable $X$ is the probability measure on $\mathbb{R}$ which maps each set $A \subset \mathbb{R}$ to $\mathbb{P}(X \in A)$. The probability mass function of the distribution of $X$ may be obtained by pushing forward the probability mass from each $\omega \in \Omega$:



**3** The **cumulative distribution function** (CDF) of a random variable $X$ is the function $F_X(x) = \mathbb{P}(X \leq x)$.



**4** The **joint distribution** of two random variables $X$ and $Y$ is the probability measure on $\mathbb{R}^2$ which maps $A \subset \mathbb{R}^2$ to $\mathbb{P}((X,Y) \in A)$. The probability mass function of the joint distribution is $m_{(X,Y)}(x,y) = \mathbb{P}(X = x \text{ and } Y = y)$.

## Probability: Conditional Probability

**1** Given a probability space $\Omega$ and an event $E \subset \Omega$, the **conditional probability measure** given $E$ is an updated probability measure on $\Omega$ which accounts for the information that the result $\omega$ of the random experiment falls in $E$:

$$\mathbb{P}(F \mid E) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)}$$

**2** The conditional probability mass function of $Y$ given $\{X = x\}$ is $m_{Y \mid X=x}(y) = m_{X,Y}(x,y)/m_X(x)$.

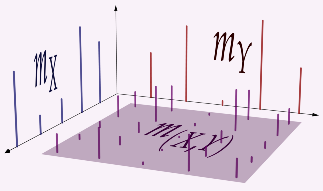
**3** **Bayes' theorem** tells us how to update beliefs in light of new evidence. It relates the conditional probabilities $\mathbb{P}(A \mid E)$ and $\mathbb{P}(E \mid A)$:

$$\mathbb{P}(A \mid E) = \frac{\mathbb{P}(E \mid A)\mathbb{P}(A)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E \mid A)\mathbb{P}(A)}{\mathbb{P}(E \mid A)\mathbb{P}(A) + \mathbb{P}(E \mid A^{\mathsf{c}})\mathbb{P}(A^{\mathsf{c}})}.$$

**4** Two events $E$ and $F$ are **independent** if $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$.

**5** Two random variables $X$ and $Y$ are **independent** if the every pair of events of the form $\{X \in A\}$ and $\{Y \in B\}$ are independent, where $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$.

**6** The PMF of the joint distribution of a pair of independent random variables factors as $m_{X,Y}(x,y) = m_X(x)m_Y(y)$:



## Probability: Expectation and Variance

**1** The **expectation** $\mathbb{E}[X]$ (or **mean** $\mu_X$) of a random variable $X$ is the *probability-weighted average of* $X$:

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)m(\omega)$$

**2** The expectation $\mathbb{E}[X]$ may be thought of as the value of a random game with payout $X$, or as the long-run average of $X$ over many independent runs of the underlying experiment. The **Monte Carlo** approximation of $\mathbb{E}[X]$ is obtained by simulating the experiment many times and averaging the value of $X$.

3 The expectation is the center of mass of the distribution of $X$:

4 The expectation of a function of a discrete random variable (or two random variables) may be expressed in terms of the PMF $m_X$ of the distribution of $X$ (or the PMF $m_{(X,Y)}$ of the joint distribution of $X$ and $Y$):

$$\mathbb{E}[g(X)] = \sum_{x\in\mathbb{R}} g(x)m_X(x)$$
$$\mathbb{E}[g(X,Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x,y)m_{(X,Y)}(x,y).$$

5 Expectation is **linear**: if $c \in \mathbb{R}$ and $X$ and $Y$ are random variables defined on the same probability space, then

$$\mathbb{E}[cX+Y] = c\,\mathbb{E}[X] + \mathbb{E}[Y]$$

6 The **variance** of a random variable is its average squared deviation from its mean. The variance measures how spread out the distribution of $X$ is. The **standard deviation** $\sigma(X)$ is the square root of the variance.

7 Variance satisfies the properties, if $X$ and $Y$ are independent random variables and $a \in \mathbb{R}$:

$$\operatorname{Var}(aX) = a^2 \operatorname{Var} X$$
$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$$

8 The **covariance** of two random variables $X$ and $Y$ is the expected product of their deviations from their respective means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$:

$$\operatorname{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

9 The covariance of two independent random variables is zero, but zero covariance does not imply independence.

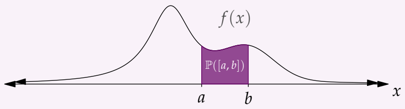
10 The **correlation** of two random variables is their normalized covariance:

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma(X)\sigma(Y)} \in [-1,1]$$

11 The **covariance matrix** of a vector $\mathbf{X} = [X_1,\ldots,X_n]$ of random variables defined on the same probability space is defined to be the matrix $\Sigma$ whose $(i,j)$th entry is equal to $\operatorname{Cov}(X_i,X_j)$. If $\mathbb{E}[\mathbf{X}] = \mathbf{0}$, then $\Sigma = \mathbb{E}[\mathbf{X}\mathbf{X}']$.

## Probability: Continuous Distributions

1 If $\Omega \subset \mathbb{R}^n$ and $\mathbb{P}(A) = \int_A f$, where $f \geq 0$ and $\int_{\mathbb{R}^n} f = 1$, then we call $(\Omega,\mathbb{P})$ a **continuous probability space**.



2 The function $f$ is called a **density**, because it measures the amount of probability mass per unit volume at each point (2D volume = area, 1D volume = length).

3 If $(X,Y)$ is a pair of random variables whose joint distribution has density $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$, then the conditional distribution of $Y$ given the event $\{X=x\}$ has density $f_{Y\,|\,X=x}$ defined by

$$f_{Y\,|\,\{X=x\}}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

where $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$ is the PDF of $X$.

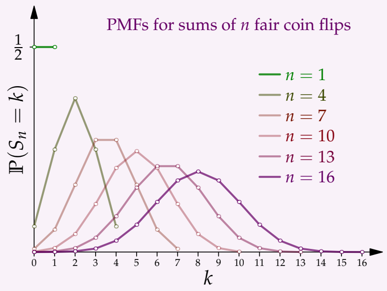
4 If a random variable $X$ has density $f_X$ on $\mathbb{R}$, then

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)f_X(x)\,\mathrm{d}x.$$

5 **CDF sampling**: $F^{-1}(U)$ has CDF $F$ if $f_U = \mathbf{1}_{[0,1]}$.

## Probability: Central Limit Theorem

1 The PDF of a sum of $n$ independent samples from a finite-variance distribution looks increasingly bell-shaped as $n$ increases, *regardless of the distribution being sampled from*.



2 We define the **standardized running sum** of $X_1, X_2, \ldots$ to have zero mean and unit variance for all $n \geq 1$:

$$S_n^* = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

3 **Central limit theorem**: the sequence of standardized sums of an i.i.d. sequence of finite-variance random variables converges in distribution to $\mathcal{N}(0,1)$: for any interval $[a,b]$, we have

$$\mathbb{P}(S_n^* \in [a,b]) \to \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t$$

as $n \to \infty$.

## Statistics: Point estimation

1 The central problem of statistics is to make inferences about a population or data-generating process based on the information in a finite sample drawn from the population.

2 **Parametric estimation** involves an assumption that the distribution of the data-generating process comes from a family of distributions parameterized by finitely many real numbers, while **nonparametric estimation** does not. *Examples: Assuming normality is parametric, while histograms are nonparametric.*

3 **Point estimation** is the inference of a single real-valued feature of the distribution of the data-generating process (such as its mean, variance, or median).

4 A **statistical functional** is any function $T$ from the set of distributions to $[-\infty,\infty]$. An **estimator** $\widehat{\theta}$ is a random variable defined in terms of $n$ i.i.d. random variables, the purpose of which is to approximate some statistical functional of the random variables' common distribution. *Example: Suppose that $T(\nu)$ = the mean of $\nu$, and that $\widehat{\theta} = (X_1 + \cdots + X_n)/n$.*

5 The empirical measure $\widehat{\nu}$ of $X_1,\ldots,X_n$ is the probability measure which assigns mass $\frac{1}{n}$ to each sample's location. The **plug-in estimator** of $\theta = T(\nu)$ is obtained by applying $T$ to the empirical measure: $\widehat{\theta} = T(\widehat{\nu})$.

6 Given a distribution $\nu$ and a statistical functional $T$, let $\theta = T(\nu)$. The **bias** of an estimator of $\theta$ is the difference between the estimator's expected value and $\theta$. *Example: The expectation of the sample mean $\widehat{\theta} = (X_1 + \cdots + X_n)/n$ is $\mathbb{E}(X_1 + \cdots + X_n)/n = \mathbb{E}[\nu]$, so the bias of the sample mean is zero.*

7 The **standard error** $\operatorname{se}(\widehat{\theta})$ of an estimator $\widehat{\theta}$ is its standard deviation.

8 An estimator is **consistent** if $\widehat{\theta} \to \theta$ in probability as $n \to \infty$.

9 The **mean squared error** of an estimator is defined to be

$$\operatorname{MSE}(\theta) = \mathbb{E}[(\widehat{\theta} - \theta)^2].$$

10 MSE is equal to variance plus squared bias. Therefore, MSE converges to zero as the number of samples goes to $\infty$ if and only if variance and bias both converge to zero.

## Statistics: Confidence intervals

1 Consider an unknown probability distribution $\nu$ from which we get $n$ independent samples $X_1,\ldots,X_n$, and suppose that $\theta$ is the value of some statistical functional of $\nu$. A **confidence interval** for $\theta$ is an interval-valued function of the sample data $X_1,\ldots,X_n$. A confidence interval has **confidence level** $1-\alpha$ if it contains $\theta$ with probability at least $1-\alpha$.

2 If $\widehat{\theta}$ is unbiased and approximately normally distributed, then $\left(\widehat{\theta} - 1.96\operatorname{se}(\widehat{\theta}), \widehat{\theta} + 1.96\operatorname{se}(\widehat{\theta})\right)$ is an approximate 95% confidence interval, since 95% of the mass of the standard normal distribution is in the interval $[-1.96, 1.96]$.

## Statistics: Maximum likelihood estimation

1 Maximum likelihood estimation is a general approach for proposing an estimator. Consider a parametric family $\{f_{\theta}(x) : \theta \in \mathbb{R}^d\}$ of PDFs or PMFs. Given $\mathbf{x} \in \mathbb{R}^n$, the **likelihood** $\mathcal{L}_{\mathbf{x}} : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\mathcal{L}_{\mathbf{x}}(\theta) = f_{\theta}(x_1)f_{\theta}(x_2)\cdots f_{\theta}(x_n).$$

If $\mathbf{X}$ is a vector of $n$ independent samples drawn from $f_{\theta}(x)$, then $\mathcal{L}_{\mathbf{X}}(\theta)$ is small or zero when $\theta$ is not in accordance with the observed data.

*Example: Suppose $x \mapsto f(x;\theta)$ is the density of a uniform random variable on $[0,\theta]$. We observe four samples drawn from this distribution: 1.41, 2.45, 6.12, and 4.9. Then the likelihood of $\theta = 5$ is zero, and the likelihood of $\theta = 10^6$ is very small.*

2 The **maximum likelihood estimator** is

$$\widehat{\theta}_{\mathrm{MLE}} = \operatorname{argmax}_{\theta\in\mathbb{R}^d} \mathcal{L}_{\mathbf{X}}(\theta).$$

Equivalently, $\widehat{\theta}_{\mathrm{MLE}} = \operatorname{argmax}_{\theta\in\mathbb{R}^d} \ell_{\mathbf{X}}(\theta)$, where $\ell_{\mathbf{X}}(\theta)$ denotes the logarithm of $\mathcal{L}_{\mathbf{X}}(\theta)$.

*Example: Suppose that $x \mapsto f(x;\mu,\sigma^2)$ is the normal density with mean $\mu$ and variance $\sigma^2$. Then the maximum likelihood estimator is the minimizer of the log-likelihood*

$$-\frac{n}{2}\log 2\pi - n\log\sigma - \frac{(X_1-\mu)^2}{2\sigma^2} - \cdots - \frac{(X_n-\mu)^2}{2\sigma^2}$$

*Setting the derivatives with respect to $\mu$ and $\sigma^2$ equal to zero, we find $\mu = \overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ and $\sigma^2 = \frac{1}{n}((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2)$. So the maximum likelihood estimators agree with the plug-in estimators.*

3 MLE enjoys several nice properties: under certain regularity conditions, we have (stated for $\theta \in \mathbb{R}^1$):

(i) **Consistency**: $\mathbb{E}[(\widehat{\theta} - \theta)^2] \to 0$ as the number of samples goes to $\infty$.

(ii) **Asymptotic normality**: $(\widehat{\theta} - \theta)/\sqrt{\operatorname{Var}\widehat{\theta}}$ converges to $\mathcal{N}(0,1)$ as the number of samples goes to $\infty$.

(iii) **Asymptotic optimality**: the MSE of the MLE converges to 0 approximately as fast as the MSE of any other consistent estimator.

4 Potential difficulties with the MLE:

(i) **Computational challenges**. It might be hard to work out where the maximum of the likelihood occurs, either analytically or numerically.

(ii) **Misspecification**. The MLE may be inaccurate if the distribution of the samples is not in the specified parametric family.

(iii) **Unbounded likelihood**. If the likelihood function is not bounded, then $\widehat{\theta}$ is not well-defined.

## Statistics: Hypothesis testing

1 **Hypothesis testing** is a disciplined framework for adjudicating whether observed data do not support a given hypothesis.

2 Consider an unknown distribution from which we will observe $n$ samples $X_1, \ldots X_n$.

(i) We state a hypothesis $H_0$–called the **null hypothesis**–about the distribution.

(ii) We come up with a **test statistic** $T$, which is a function of the data $X_1, \ldots X_n$, for which we can evaluate the distribution of $T$ assuming the null hypothesis.

(iii) We give an **alternative hypothesis** $H_{\mathrm{a}}$ under which $T$ is expected to be significantly different from its value under $H_0$.

(iv) We give a significance level $\alpha$ (like 5% or 1%), and based on $H_{\mathrm{a}}$ we determine a set of values for $T$—called the *critical region*—which $T$ would be in with probability at most $\alpha$ under the null hypothesis.

(v) **After setting $H_0$, $H_{\mathrm{a}}$, $\alpha$, $T$, and the critical region**, we run the experiment, evaluate $T$ on the samples we get, and record the result as $t_{\mathrm{obs}}$.

(vi) If $t_{\mathrm{obs}}$ falls in the critical region, we reject the null hypothesis. The corresponding ***p*-value** is defined to be the minimum $\alpha$-value which would have resulted in rejecting the null hypothesis, with the critical region chosen in the same way*.

*Example: Muriel Bristol claims that she can tell by taste whether the tea or the milk was poured into the cup first. She is given eight cups of tea, four poured milk-first and four poured tea-first.*

*We posit a null hypothesis that she isn't able to discern the pouring method, under which the number of cups identified correctly is 4 with probability $1/\binom{8}{4} \approx 1.4\%$ and at least 3 with probability $17/70 \approx 24\%$. Therefore, at the 5% significance level, only a correct identification of all the cups would give us grounds to reject the null hypothesis. The p-value in that case would be 1.4%.*

3 Failure to reject the null hypothesis is not necessarily evidence *for* the null hypothesis. The **power** of a hypothesis test is the conditional probability of rejecting the null hypothesis given that the alternative hypothesis is true. A $p$-value may be low either because the null hypothesis is true or because the test has low power.

4 The **Wald test** is based on the normal approximation. Consider a null hypothesis $\theta = 0$ and the alternative hypothesis $\theta \neq 0$, and suppose that $\widehat{\theta}$ is approximately normally distributed. The Wald test rejects the null hypothesis at the 5% significance level if $|\widehat{\theta}| > 1.96\operatorname{se}(\widehat{\theta})$.