A random experiment $E$ is an experiment such that
1. all possible distinct outcomes of the experiment are known in advance;
2. the actual outcome of the experiment is not known in advance with certainty;
3. the experiment can be repeated under identical conditions.

The sample space, $\Omega$, is the set of all possible outcomes of a random experiment.

A subset $A \subseteq \Omega$ of the sample space is referred to as an event.

The empty set $\emptyset \subseteq \Omega$ is referred to as the impossible event.

The sample space itself, $\Omega \subseteq \Omega$, is referred to as the certain event.
Example of a random experiment

- The random experiment $E$ consists in a *single toss* of an *unbiased coin*.
- The possible outcomes of this experiment are:
  - $\omega_1 = \text{“heads”}$,
  - $\omega_2 = \text{“tails”}$.
- The sample space is thus $\Omega = \{\omega_1 = \text{“heads”}, \omega_2 = \text{“tails”}\}$.
- There are exactly four events — $2^{|\Omega|} = 2^2 = 4$ subsets of $\Omega$:
  - $H = \{\omega_1\} = \text{“heads (obverse) comes up”}$;
  - $T = \{\omega_2\} = \text{“tails (reverse) comes up”}$;
  - $\emptyset = \{\} = \text{“nothing comes up” — if we *do* perform the experiment $E$, this will never occur, so this is indeed the impossible event}$;
  - $\Omega = \{\omega_1, \omega_2\} = \text{“either heads or tails comes up” — if we *do* perform the experiment $E$, this is guaranteed to occur, so this is indeed the certain event (we disregard the possibility of the coin landing on its edge — the third side of the coin; otherwise we’d need a separate outcome in $\Omega$ to model this possibility)}$. 


The classical interpretation of probability

- Let $A$ be an event associated with an experiment $\mathcal{E}$ so that $A$ either occurs or does not occur when $\mathcal{E}$ is performed.
- Assume that $\Omega$ is finite.
- Furthermore, assume that all outcomes in $\Omega$ are equally likely.
- Denote by $M(\cdot)$ the number of outcomes in an event; thus $M(A)$ is the number of outcomes in $A$, $M(\Omega)$ the number of outcomes in $\Omega$.
- Then the probability of $A$ is given by

$$P(A) = \frac{M(A)}{M(\Omega)}.$$
The classical interpretation of probability: an example

- Let us continue our example where the random experiment $\mathcal{E}$ consists in a single toss of an unbiased coin.
- For the event $H = \{\omega_1\}$, according to the classical interpretation of probability,

$$P(H) = \frac{M(H)}{M(\Omega)} = \frac{1}{2}.$$  

- But what if $\Omega$ is not finite?
- And what if the coin is biased?
The frequentist interpretation of probability

- Let $A$ be an event associated with an experiment $\mathcal{E}$ so that $A$ either occurs or does not occur when $\mathcal{E}$ is performed.
- Consider a superexperiment $\mathcal{E}^\infty$ consisting in an infinite number of independent performances of $\mathcal{E}$.
- Let $N(A, n)$ be the number of occurrences of $A$ in the first $n$ performances of $\mathcal{E}$ within $\mathcal{E}^\infty$.
- Then the probability of $A$ is given by

$$P[A] = \lim_{n \to \infty} \frac{N(A, n)}{n}.$$ 

- This interpretation of probability is known as the long-term relative frequency (LTRF) (or frequentist, or objectivist) [Wil01, page 5].
- The claim is that, in the long term, as the number of trials approaches infinity, the relative frequency will converge exactly to the true probability.
- It requires that the probabilities be estimated from samples.
- Unknown quantities, such as means, variances, etc., are considered to be fixed but unknown.
Interpretations of probability

Question

Can you use the frequentist interpretation of probability to compute the probability of the existence of extraterrestrial life?
Interpretations of probability

Bayesian interpretation of probability

- In Bayesian (subjectivist, epistemic, evidential) interpretation, the probability of an event is the degree of belief that that event will occur.
- This degree of belief can be determined on the basis of
  - empirical data,
  - past experience, or
  - subjective plausibility.
- Bayesian probability can be assigned to any statement, whether or not a random experiment is performed.
- Unknown quantities, such as means, variances, etc., are regarded to follow a probability distribution, which expresses our degree of belief about that quantity at a particular time.
- On arrival of new information, the degree of belief can be updated.
Interpretations of probability

The axiomatic interpretation of probability

- Andrey Nikolaevich Kolmogorov (1903–1987): “The theory of probability as a mathematical discipline can and should be developed from axioms in exactly the same way as Geometry and Algebra.” [Kol33]

- **Kolmogorov’s axioms** of probability:
  - **First axiom**: For any event $E$, $\mathbb{P} [E] \in \mathbb{R}$, $\mathbb{P} [E] \geq 0$. (The assumption of finite measure.)
  - **Second axiom**: $\mathbb{P} [\Omega] = 1$. (The assumption of unit measure.)
  - **Third axiom**: For any countable collection of disjoint events $E_1, E_2, \ldots$, $\mathbb{P} [\bigcup_{i=1}^{\infty} E_i] = \sum_{i=1}^{\infty} \mathbb{P} [E_i]$. (The assumption of $\sigma$-additivity.)

- **Consistency**:
  - The LTRF and Bayesian interpretations motivated Kolmogorov’s axioms and are consistent with them.
  - The LTRF interpretation reappears in the axiomatic interpretation as a theorem — the **Strong Law of Large Numbers**.
  - The axioms describe how probability behaves, not what probability is... Or is Kolmogorov saying that what probability is is defined by the way it behaves? (“When I see a bird that walks like a duck and swims like a duck and quacks like a duck, I call that bird a duck.” — Indiana poet James Whitcombe Riley, around 1916.)
Consequences of the axioms

- **Null empty set**: \( \mathbb{P}[\emptyset] = 0 \).
- **Complement rule**: for any event \( A \), \( \mathbb{P}[A^c] = 1 - \mathbb{P}[A] \).
- **Difference rule**: for any events \( A, B \), \( \mathbb{P}[B \setminus A] = \mathbb{P}[B] - \mathbb{P}[A] \).
- **Monotonicity rule**: if \( A \subseteq B \), then \( \mathbb{P}[A] \leq \mathbb{P}[B] \).
- **The upper bound on probability is 1**: for all \( A \), \( \mathbb{P}[A] \leq 1 \).
- **Inclusion-exclusion rule**: for any events \( A, B \),
  \[ \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \.
- **Bonferroni inequality**: for any events \( \mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B] \).
- **Continuity property**: If the events \( A_1, A_2, \ldots \) satisfy \( A_1 \subseteq A_2 \subseteq \ldots \) and \( A = \bigcup_{i=1}^{\infty} A_i \), then \( \mathbb{P}[A_i] \) is increasing and \( \mathbb{P}[A] = \lim_{i \to \infty} \mathbb{P}[A_i] \). If the events \( B_1, B_2, \ldots \) satisfy \( B_1 \supseteq B_2 \supseteq \ldots \) and \( B = \bigcap_{i=1}^{\infty} B_i \), then \( \mathbb{P}[B_i] \) is decreasing and \( \mathbb{P}[B] = \lim_{i \to \infty} \mathbb{P}[B_i] \).
- **Borel–Cantelli Lemma**: For any events \( A_1, A_2, \ldots \), if \( \sum_{i=1}^{\infty} \mathbb{P}[A_i] < \infty \), then
  \( \mathbb{P}[\bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j] = 0 \).

\(^1\) The event \( \mathbb{P}[\bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j] \) is sometimes referred to as “\( A_i \) infinitely often” or as the limit superior of the \( A_i \), \( \lim_{i \to \infty} \sup_{j \geq i} A_j \).
Frequentist vs Bayesian interpretation of probability

- The frequentist approach is (arguably) *objective*.
- The Bayesian approach is (arguably) *subjective*.
- The frequentist approach uses only *new data* to draw conclusions.
- The Bayesian approach uses *both new and past data, and belief*, to draw conclusions.
Conditional probability

- The probability that the event $A$ occurs may be influenced by the information concerning the occurrence of the event $B$.

- The probability of event $A$ given that $B$ has occurred or will occur, called the **conditional probability** of $A$ given $B$ is given by

\[
P(A | B) = \frac{P(A \cap B)}{P(B)}
\]

for $P(B) > 0$.

- We say that the events $A$ and $B$ are **independent** if the information concerning the occurrence of the event $B$ does not influence the probability of the occurrence of the event $A$ (and one can show that, as a consequence, the information concerning the occurrence of the event $A$ does not influence the probability of the occurrence of the event $B$), i.e.

\[
P(A | B) = P(A) , \quad P(B | A) = P(B) .
\]

- This can be equivalently and more concisely expressed by the product rule

\[
P(A \cap B) = P(A) P(B) .
\]
The Law of Total Probability

- When the (unconditional) probability of an event is not given directly, it may be obtainable as a weighted average of various conditional probabilities.

- If $B_1, B_2, \ldots, B_s$ are events such that they form a partition of $\Omega$, i.e. $\Omega = \bigcup_{i=1}^{s} B_i$, and are pairwise disjoint, i.e. for all $1 \leq i \leq s$, $1 \leq j \leq s$, $B_i \cap B_j = \emptyset$, the axioms of probability imply that

\[
P[A] = \sum_{i=1}^{s} P[A | B_i] P[B_i].
\]

- This result is known as the **Law of Total Probability**.
Bayes’s theorem

Let $H$ and $E$ be events ($H$ stands for “hypothesis”, $E$ stands for “evidence”). Bayes’s theorem establishes the relationship between the probability of $H$, $\Pr[H]$, the probability of $E$, $\Pr[E]$, the conditional probability of $H$ given $E$, $\Pr[H \mid E]$, and the conditional probability of $E$ given $H$, $\Pr[E \mid H]$:

$$
\Pr[H \mid E] = \frac{\Pr[E \mid H] \Pr[H]}{\Pr[E]}.
$$

The proof follows immediately from the definition of conditional probability:

$$
\Pr[H \mid E] = \frac{\Pr[H \cap E]}{\Pr[E]} = \frac{\Pr[E \cap H]}{\Pr[E]} = \frac{\Pr[E \mid H] \Pr[H]}{\Pr[E]}.
$$

It is useful to reformulate it for the case when there are multiple alternative hypotheses, $H_1, H_2, \ldots H_s$, $s \in \mathbb{N}^*$. Then, for $1 \leq i \leq s$,

$$
\Pr[H_i \mid E] = \frac{\Pr[E \mid H_i] \Pr[H_i]}{\Pr[E]} = \frac{\Pr[E \mid H_i] \Pr[H_i]}{\sum_{j=1}^{s} \Pr[E \mid H_j] \Pr[H_j]},
$$

where the second equality follows from the Law of Total Probability.
Bayes’s theorem: frequentist interpretation

▷ How does a frequentist interpret

\[ P[H | E] = \frac{P[E | H] P[H]}{P[E]} \]

▷ For a frequentist, probability is a *long-term relative frequency* of outcomes in the hyperexperiment \( E^\infty \).

▷ Let \( N(A, n) \) be the number of occurrences of the event \( A \) in the first \( n \) performances of \( E \) within \( E^\infty \). Then \( P[A] := \lim_{n \to \infty} \frac{N(A, n)}{n} \).

▷ In particular,

\[ P[H] := \lim_{n \to \infty} \frac{N(H, n)}{n}, \quad P[E] := \lim_{n \to \infty} \frac{N(E, n)}{n}. \]

▷ Therefore,

\[ P[H | E] = \frac{P[H \cap E]}{P[E]} = \lim_{n \to \infty} \frac{N(H \cap E, n)}{N(E, n)}, \quad P[E \cap H] = \frac{P[H \cap E]}{P[E]} = \lim_{n \to \infty} \frac{N(H \cap E, n)}{N(H, n)}. \]

▷ The conditional probabilities are, then, *relative frequencies*: \( P[H | E] \) is the proportion of outcomes with property \( H \) out of those with property \( E \).
Bayes’s theorem: Bayesian interpretation

- How does a Bayesian interpret

\[ \Pr[H | E] = \frac{\Pr[E | H] \Pr[H]}{\Pr[E]} ? \]

- For a Bayesian, probability is a degree of belief.
- Before (prior to) observing any new evidence \( E \) — the degree of belief in a certain hypothesis \( H \).
- After (posterior to) — the degree of belief in \( H \) after taking into account that piece of evidence \( E \).
- Thus,
  - \( \Pr[H] \) is the prior, the initial degree of belief in \( H \).
  - \( \Pr[H | E] \) is the posterior, the degree of belief in \( H \) after accounting for \( E \).
  - \( \frac{\Pr[E | H]}{\Pr[E]} \) is the support that the evidence \( E \) provides for the hypothesis \( H \).
  - \( \Pr[E | H] \) is the likelihood, the compatibility of the evidence with the hypothesis.
  - \( \Pr[E] \) is the marginal likelihood of the evidence, irrespective of the hypothesis.
- Using this Bayesian terminology, Bayes’s theorem can be stated as:

\[ \text{posterior} = \text{support} \cdot \text{prior} \propto \text{likelihood} \cdot \text{prior}. \]
Recap: expectations and variances

- Let $X, X_1, X_2$ be $n$-dimensional random variables, $Y$ be an $m$-dimensional random variable, $A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p, C \in \mathbb{R}^{q \times m}, d \in \mathbb{R}^q$.

- Recall that the expectation (or expected value) of the random variable $X$ is denoted by $\mathbb{E}[X]$.

- Recall that
  \[ \text{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top]. \]

- Often the notation $\text{Cov}[X]$ is used instead of $\text{Var}[X]$. $\text{Var}[X]$ is called the variance of $X$, the covariance of $X$, or, if $n > 1$, the covariance matrix of $X$ or the variance-covariance matrix of $X$.

- Recall also that the cross-covariance between $X$ and $Y$ is defined as
  \[ \text{Cov}[X, Y] := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top]. \]

- Somewhat confusingly, the cross-covariance is also sometimes referred to as covariance.

- Finally, the correlation matrix, the matrix of Pearson product-moment correlation coefficients between each of the random variables in the random vector $X$, is given by
  \[ \text{Cor}[X] = \text{diag}(\Sigma)^{-\frac{1}{2}} \Sigma \text{diag}(\Sigma)^{-\frac{1}{2}}, \]
  where $\Sigma = \text{Var}[X]$ and $\text{diag}(\Sigma)$ is the diagonal matrix whose diagonal elements are equal to the corresponding diagonal elements of $\Sigma$. 
Properties of expectations

- **Expectation is a linear operator:**
  - $\mathbb{E}[AX + b] = A\mathbb{E}[X] + b$;
  - $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$.

- If $X_1$ and $X_2$ are independent, then $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \mathbb{E}[X_2]$.

- **Jensen’s inequality:** If $\varphi$ is a convex function, then
  $$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

  if $\varphi$ is a concave function, then
  $$\varphi(\mathbb{E}[X]) \geq \mathbb{E}[\varphi(X)]$$
Properties of variances, cross-covariances and correlation matrices

- \( \text{Var} [X] = \mathbb{E} [XX^\top] - \mathbb{E} [X] \mathbb{E} [X]^\top; \)
- \( \Sigma := \text{Var} [X] \) is an \( n \times n \) symmetric positive-semidefinite matrix;
- \( \text{Var} [AX + b] = A \text{Var} [X] A^\top; \)
- \( \text{Cov} [X, Y] = \text{Cov} [Y, X]^\top; \)
- \( \text{Cov} [X_1 + X_2, Y] = \text{Cov} [X_1, Y] + \text{Cov} [X_2, Y]; \)
- \( \text{Cov} [AX + a, CY + d] = A \text{Cov} [X, Y] C^\top; \)
- \( \text{Var} [X_1 + X_2] = \text{Var} [X_1] + \text{Var} [X_2] + \text{Cov} [X_1, X_2] + \text{Cov} [X_2, X_1]; \)
- if \( X \) and \( Y \) are independent, then \( \text{Cov} [X, Y] = 0; \)
- \( \text{Cor} [X] = \text{Cov} [Z], \) where \( Z \) is an \( n \)-dimensional random vector such that \( (Z)_i = \frac{(X)_i}{\text{Var}(X)_i} \) for \( 1 \leq i \leq n. \)
Interpretations of probability

Bradley Efron.  
Statistics as a unified discipline.  

Andrey Nikolaevich Kolmogorov.  
Grundbegriffe der Wahrscheinlichkeitsrechnung.  

David Ruppert.  
*Statistics and Finance: An Introduction*.  

David Williams.  
*Weighing the Odds: A Course in Probability and Statistics*.  