Applied Statistics

Beatriz Seoane Bartolomé

September 14, 2023

CONTENTS

1

STATISTICAL EVENTS AND FUNDAMENTALS OF PROBABILITY

- **Probability:** Probability is a mathematical description of uncertain or random events.
- **Statistics:** Statistics involves the collection, analysis, and interpretation of data.

1.1 RANDOM VARIABLES

In probability theory we deal with variables that have no fixed value and are subject to chance. These variables are called *random* variables, denoted by X, and are characterized by a probability distribution P. We will say that $X \sim P$. This means that we cannot predict the exact value that X will take, but we can determine the probability of the various outcomes. We will denote a particular value or *realization* of the random variable as x and the set of all possible values that X can take (with non-zero probability) will be represented by Ω and called *event space*.

These random variables can be either discrete or continuous. For discrete random variables, the list of possible different realizations can be enumerated. For example, a six-sided die has an outcome space $\Omega = \{1,2,3,4,5,6\}$, or the 7 days of the week. In this case, the probability of observing x_n is $P(X = x_n) = p_n$. The sum of all probabilities over all possible events is equal to 1, i.e. $\sum_n p_n = 1$.

For continuous random variables, the random variables can take values from a continuous range of real numbers represented by an interval in $\mathbb{R}$. We describe the probabilities by the *probability distribution function* $p(x)$. Now the sum of all outcomes is replaced by an integral $\int_{x\in\Omega} p(x)\,dx = 1$. Also

$$P(a < X < b) = \int_a^b p(x)dx.$$

1.2 EVENT REALIZATIONS

Given an event space Ω, an *event* A is defined as a subset of Ω. The occurrence of an event A is determined by a specific realization x such that $x \in A$. This subset might represent outcomes like obtaining an even number on a dice throw or x existing within an interval $a < x < b$ for continuous variables. The probability of event A is represented by $P(A)$. For discrete variables, it's calculated as the sum of the probabilities of all outcomes in A, whereas for continuous variables, it's calculated as the integral of the probability density function over the interval A.

Example 1.1. Some probabilistic models.

1. Tossing a fair coin. There are two possible outcomes, both having the same probability

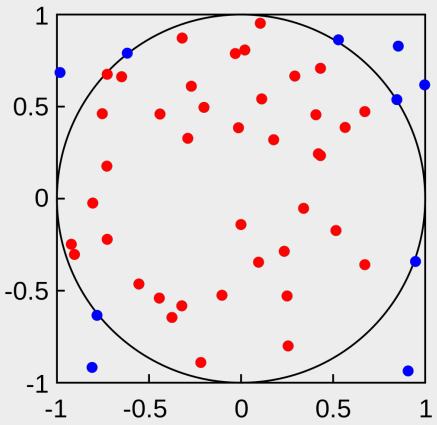
$$\Omega = \{\text{heads,tails}\},$$

$$P(\text{heads}) = P(\text{tails}) = \frac{1}{2},$$

2. Paradigmatic example of a Monte Carlo integration is the estimation of π.

$$\Omega = \{\text{inside circle,outside}\},$$

$$P(\text{inside circle}) = \frac{\text{Area circle}}{\text{Area square}}.$$



Example 1.2. Consider a random variable X representing the outcome of a 6-faced dice. The set of possible outcomes is $\Omega = \{1,2,3,4,5,6\}$. We define three events:

- Event A is the occurrence of an even number ($A = \{2,4,6\}$).
- Event B is the occurrence of an odd number ($B = \{1,3,5\}$).
- Event C is the occurrence of a number less than or equal to 3 ($C = \{1,2,3\}$).

We assume that all numbers have an equal probability of $1/6$. Then

$$P(A) = P(B) = P(C) = 1/2$$

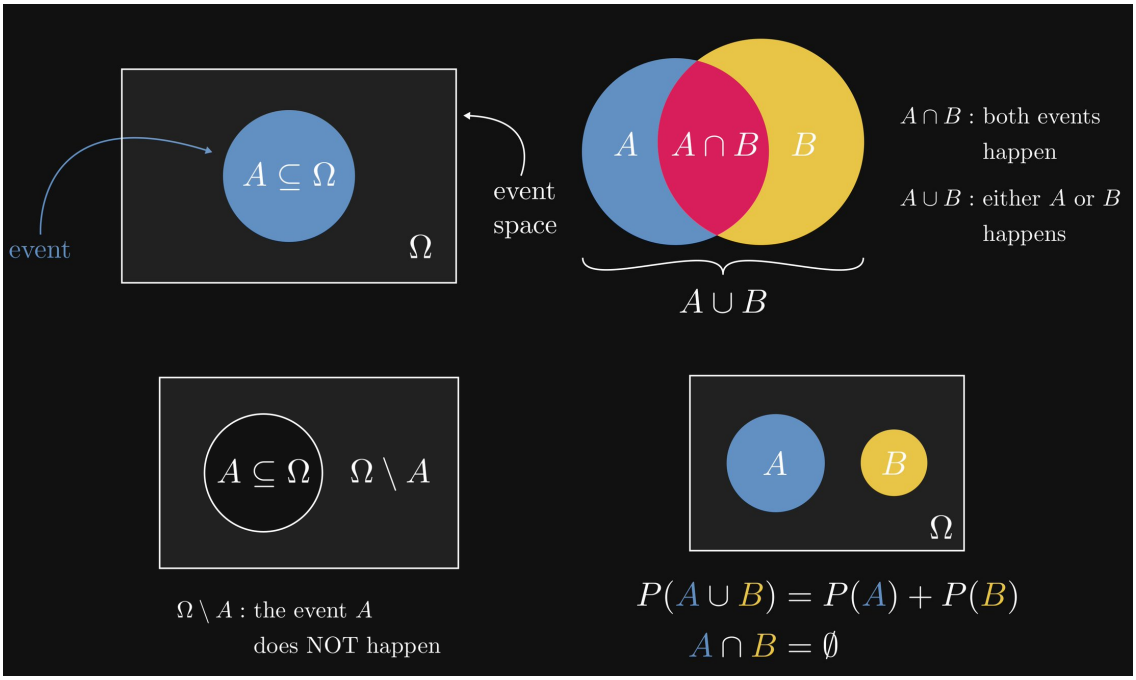
Some inherent properties of probabilities include:

- The probability $P(A)$ always lies in the range [0, 1], i.e., $0 \leq P(A) \leq 1$.
- An event with no outcomes, represented as the empty set $\emptyset$, has a probability of 0, i.e., $P(\emptyset) = 0$.

Furthermore, events allow for various operations:

- **Union:** $A \cup B$ represents an event that encompasses all outcomes belonging to either A or B.

- **Intersection:** $A \cap B$ denotes an event comprising outcomes that are common to both A and B.
- **Complementary:** $\bar{A}$ is the event representing all outcomes not in A. Consequently, $\bar{A} \cap A = \emptyset$ and $\bar{A} \cup A = \Omega$.



1.2.1 *Mutually exclusive events*

A set of N events, denoted as $\{A_k\}_{k=1,\ldots,N}$, is termed *mutually exclusive events* if:

- Any two distinct events in the set don't share outcomes, i.e., $A_j \cap A_k = \emptyset$ for all $j \neq k$.
- The union of all events encapsulates the complete outcome space Ω, i.e., $\bigcup_{j=1}^{N} A_j = \Omega$.

Then, we have $P\left(\bigcup_{j=1}^{N} A_j\right) = \sum_{j=1}^{N} P(A_j)$.

In fact, any measure that satisfy these two *Kolmogorov's axioms*,

- $P(\Omega) = 1$,
- $P\left(\bigcup_{j=1}^{N} A_j\right) = \sum_{j=1}^{N} P(A_j)$,

is a probability measure, and all results from probability measure follows from these two axioms.

1.3 CONDITIONAL PROBABILITY AND BAYES' THEOREM

Using the properties of probability, we can derive relationships for conditional probabilities.

Considering two events A and B, the union's probability $P(A \cup B)$ is represented as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.1}$$

From here, we can express the conditional probability $P(A|B)$ - the probability of event A occurring given that B has already occurred. It is derived as:

$$P(A \cap B) = P(B) \cdot P(A|B). \tag{1.2}$$

Reformulating this, we get Bayes' theorem:

$$P(A|B) = P(A) \times \frac{P(B|A)}{P(B)} \tag{1.3}$$

1.3.1 *Incompatible and independent events*

If the events A and B are *mutually exclusive* or *incompatible*, then their intersection is zero, i.e. $P(A \cap B) = 0$. This condition states that if one event occurs, then the other cannot occur. Symbolically, this is captured by:

$$P(A|B) = P(B|A) = 0. \tag{1.4}$$

Events A and B are termed *independent* if the occurrence of one does not influence the probability of the other. This is mathematically represented as:

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B). \tag{1.5}$$

Another equivalent condition for independence is:

$$P(A \cap B) = P(A) \times P(B). \tag{1.6}$$

This means that the joint probability of both events occurring is simply the product of their individual probabilities.

Example 1.3. In the example above with the 6-faded dice

- $P(A \cap C) = 1/6$, $P(B \cap C) = 2/6$.
- $P(A \cup C) = 5/6$, $P(B \cup C) = 4/6$, $P(A \cup B) = 1$.
- $P(A|C) = 1/3$, $P(B|C) = 2/3$, $P(C|A) = 1/3$, and $P(C|B) = 2/3$.

For example, $P(A|C) = 1/3$, which can be computed as the probability of event A occurring given that event C has occurred. Alternatively, we can compute it as $\frac{P(A \cap C)}{P(C)} = \frac{1/6}{1/2} = 1/3$.

1.3.2 *Multiplication of the conditional probability*

Now, let's consider N non-excluding events $\{B_k\}$. The joint probability of all these events can be expressed recursively as:

$$\begin{aligned} P(\cap_{k=1}^{N} B_k) &= P(B_N|\cap_{k=1}^{N-1} B_k)P(\cap_{k=1}^{N-1} B_k) \\ &= P(B_N|\cap_{k=1}^{N-1} B_k)P(B_{N-1}|\cap_{k=1}^{N-2} B_k)P(\cap_{k=1}^{N-2} B_k) \\ &= \dots \\ &= P(B_N|\cap_{k=1}^{N-1} B_k)P(B_{N-1}|\cap_{k=1}^{N-2} B_k)\cdot P(B_3|B_2\cap B_2)P(B_2|B_1)P(B_1) \end{aligned}$$

1.1 Exercise

Estimate the probability that at least 2 students share the same birthday in a class of 35.

Let it be $A = \{\text{at least two students share the same birthday}\}$.

Then $\bar{A} = \{\text{there are no two students that share the same birthday}\}$ and $P(A) = 1 - P(\bar{A})$. Let also be

$$B_1 = \{\text{The second student does not have the same birthday than the first}\},$$

$$B_2 = \{\text{The third student does not have the same birthday than the first and the second}\},$$

$$\dots$$

$$B_k = \{\text{The } k+1\text{th student does not have the same birthday than the } k \text{ precedent}\},$$

We can use the expression above to compute

$$P(\bar{A}) = P(\cap_{j=1}^{n-1} B_j) = P(B_{n-1}|\cap_{j=1}^{n-2} B_j)\cdots P(B_3|B_2\cap B_1)P(B_2|B_1)P(B_1) \tag{1.7}$$

, with

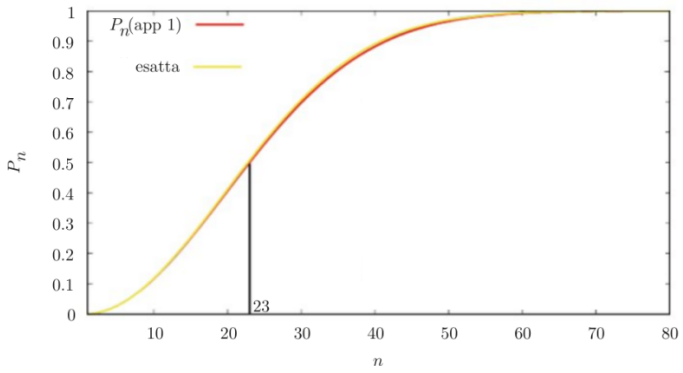
$$P(B_1) = \frac{364}{365}, P(B_2|B_1) = \frac{363}{365} \qquad \cdots P(B_k|\cap_{j=1}^{k-1} B_j) = \frac{365-k}{365}.$$

Then, the probability we wanted to compute is

$$\begin{aligned} P(A) = 1 - P(\bar{A}) &= 1 - \prod_{k=1}^{n-1}\left(1 - \frac{k}{365}\right) = \frac{365!}{365^n(365-n)!} \\ &= 1 - \exp\left[\sum_{k=1}^{n-1}\log\left(1-\frac{k}{365}\right)\right] \approx 1 - \exp\left[\sum_{k=1}^{n-1}\left(-\frac{k}{365}\right)\right] = 1 - \exp\left(-\frac{n(n-1)}{2\cdot 365}\right) \end{aligned}$$

We approximated $k/365 \ll 1$, and $\log(1-k/365) \approx -k/365$, and we used the arithmetic sum

$$\begin{aligned} S = \sum_{k=1}^{n-1} k &= 1+2+3+\cdots+(n-2)+(n-1) \\ S &= (n-1)+(n-2)+(n-3)+\cdots+2+1 \\ \Rightarrow 2S &= (n-1)n \end{aligned}$$



1.4 BAYES FORMULA AND STATISTICAL TESTS

Suppose that $\{A_k\}_{k=1}^N$ is a set of mutually exclusive events. In such a scenario, we can express $P(B)$ as a sum of probabilities with respect to all possible causal events:

$$P(B) = \sum_j P(A_j \cap B) = \sum_j P(B|A_j)P(A_j). \tag{1.8}$$

Given the above, Bayes' theorem can be written as follows:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)} = \frac{P(B|A_k)P(A_k)}{\sum_j P(B|A_j)P(A_j)}. \tag{1.9}$$

We can interpret the above formula in the following way: Let each event A_k represent a possible cause of the event B. Since $\{A_k\}_{k=1}^N$ is defined as a complete set of mutually exclusive events, this means that at least one of these events A_k must be the true cause of B. Therefore, the expression $P(A_k|B)$ denotes the probability that a given event A_k is the cause of the observed event B.

Example 1.4.

Imagine you are trying to find the cause of a certain event (e.g., a noise in a house). There could be several reasons: a creaking door, a blowing wind, or a radio playing.

In this analogy:

- Each possible cause (door, wind, radio) corresponds to events $\{A_k\}_{k=1}^N$.
- The noise you are investigating corresponds to event B.

The probability $P(B)$ of hearing this sound can be represented as a composite of all these possible reasons. Each reason contributes a part to the total probability of hearing this sound.

The Bayes' theorem helps to estimate the probability that a particular cause A_k (like wind) is behind the noise, i.e., to estimate $P(A_k|B)$.

After hearing the noise, the theorem essentially evaluates, “What is the probability that this specific cause (e.g., wind) caused the noise?"

Example 1.5. **The Unicorn Glitter Test** Background: In a fictional world there are unicorns. Only 1% of the horse-like animals in this world are unicorns; the rest are ordinary horses. The problem is that unicorns tend to magically hide their horn so as not to be recognized. So it is not easy to tell whether a horse is an ordinary horse or a magical unicorn.

However, we know that unicorns especially like to roll around in glittery meadows, which means that we can use the presence of glitter in the animals' manes as a test to detect unicorns. We know that

- If an animal is a unicorn, the test will correctly identify the glitter in its mane 90% of the time. This is because unicorns love rolling in glittery meadows.
- If an animal is a common horse, at 10% of the time it will mistakenly have glitter in its mane. Perhaps some of these horses had a wild party or accidentally got into the glitter meadow.

So the question to answer is: If an animal's mane has glitter, what is the probability that it is actually a unicorn?

Let:

- U be the event that an animal is a unicorn.
- G be the event that an animal's mane tests positive for glitter.

We want to find the probability $P(U|G)$, i.e., the probability that an animal is a unicorn given that its mane tested positive for glitter. Using Bayes' theorem:

$$P(U|G) = \frac{P(G|U) \times P(U)}{P(G)}$$

Where:

- Likelihood: $P(G|U) = 0.90$ (Probability of glitter given that it's a unicorn)
- Prior: $P(U) = 0.01$ (Probability that a randomly selected animal is a unicorn)
- Marginal: $P(G)$ the general probability to find glitter on a horse-like animal.

We do not know the marginal probability of finding glitter, but we can obtain it:

$$P(G) = P(G|U) \times P(U) + P(G|not\ U) \times P(not\ U)$$

$$P(G) = 0.90 \times 0.01 + 0.10 \times 0.99$$

$$P(G) = 0.108$$

This means that in about a 11% of cases when we test a mane, we find glitter.

Plugging these values into the Bayes' formula:

$$P(U|G) = \frac{0.90 \times 0.01}{0.109}$$

$$P(U|G) \approx 0.0826$$

So, if an animal's mane has glitter, there's only about an 8.26% chance that it's actually a unicorn. The rest are just party-hard horses!

Conclusion: Even in a world sprinkled with whimsy and magic, Bayes' theorem helps us discern reality from fantasy. Always remember: glitter doesn't always mean it's a unicorn!

This was just a funny example, but this formula is also highly valuable for evaluating the performance of predictors, such as image classifiers in Machine Learning or diagnostic tests for viral infections. In general, we are interested in assessing the precision of conditional probabilities. For instance, we want to determine how often we can correctly classify a given label, or how likely it is that someone is infected by COVID if they had a positive test.

Example 1.6. **Spam predictor**

Consider a machine learning predictor designed to classify emails as "spam" or "not spam". We use Bayes' theorem to understand its precision.

Definitions. Let:

- S: Event that an email is spam.
- P: Event that our predictor classifies an email as spam.

We want to find: $P(S|P)$: Probability that an email is truly spam given our predictor classifies it as spam.

Given Probabilities

$P(S) = 0.6$	(Prevalence of spam emails)
$P(P) = 0.7$	(Probability predictor classifies as spam)
$P(P|S) = 0.9$	(True Positive Rate)

Using Bayes' theorem:

$$P(S|P) = \frac{P(P|S) \times P(S)}{P(P)}$$

Substituting in our given probabilities:

$$P(S|P) = \frac{0.9 \times 0.6}{0.7} = 0.7714$$

Thus, if our predictor classifies an email as spam, there's a 77.14% chance it is truly spam.

Remember, the real power of this analysis is in its application. The above example highlights that even with a relatively high true positive rate (90%), the predictor's actual reliability in a real-world scenario (when it says an email is spam) is slightly lower at 77.14% because of the prevalence of spam emails and the overall rate at which the predictor classifies emails as spam.

Example 1.7. Imagine considering a test set containing images of only three categories: dog, cat, and bird. Our dataset does not need to be balanced between categories, but we know the proportion of each class. Let us assume that the probability of images being dogs is denoted as $P(\text{'dog'})$, for cats as $P(\text{'cat'})$, and for birds as $P(\text{'bird'})$.

Our classifier assigns a label (DOG, CAT, or BIRD) to each image it sees. We use capital letters to distinguish the predicted labels from the true ones.

To test the reliability of my classifier, we aim to evaluate the probability that images classified as 'DOG' are indeed dogs, denoted as $P(\text{dog}|\text{DOG})$, as well as the same probabilities for the rest of the labels. Computing these conditional probabilities directly is not as straightforward as the reverse one $P(\text{DOG}|\text{dog})$.

Indeed, one can compute $P(\text{DOG}|\text{dog})$, by feeding our classifier with images of dogs and counting how often they are classified as DOG, CAT, or BIRD. It is important to note that both conditional probabilities are not the same. By employing Bayes' formula, we get

$$P(\text{dog}|\text{DOG}) = \frac{P(\text{DOG}|\text{dog})P(\text{dog})}{P(\text{D}|\text{dog})P(\text{dog}) + P(\text{D}|\text{cat})P(\text{cat}) + P(\text{D}|\text{bird})P(\text{bird})}. \tag{1.10}$$

Clearly, in the special case when the dataset is perfectly balanced, i.e., $P(\text{'dog'}) = P(\text{'cat'}) = P(\text{'bird'})$, $P(\text{dog}|\text{DOG})$ is simply the accuracy at which the label 'DOG' is properly assigned to a 'dog'. However, this is not true if the dataset is unbalanced.

For instance, let's assume $P(\text{'dog'}) = 0.1$, $P(\text{'cat'}) = 0.5$, and $P(\text{'bird'}) = 0.4$, and that the predictor predicts $P(\text{DOG}|\text{dog}) = 0.99$, $P(\text{DOG}|\text{cat}) = 0.1$, and $P(\text{DOG}|\text{bird}) = 0.01$, then the accuracy of the 'DOG' prediction is

$$P(\text{dog}|\text{DOG}) = \frac{0.99 \cdot 0.1}{0.99 \cdot 0.1 + 0.1 \cdot 0.5 + 0.01 \cdot 0.4} = 0.65$$

This result strongly differs from $P(\text{DOG}|\text{dog}) = 0.99$. Normally, for a predictor to be good, we are not only worried about predicting properly a given label, the accuracy, but also that the prediction is specific enough, in

our case, I want that the probability of getting 'DOG' when I show a pic of a 'cat' is small

$$P(\text{cat}|\text{DOG}) = \frac{P(\text{cat})P(\text{DOG}|\text{cat})}{P(\text{DOG})} = \frac{0.5 \cdot 0.1}{0.99 \cdot 0.1 + 0.1 \cdot 0.5 + 0.01 \cdot 0.4} = 0.33,$$

which is much higher than $P(\text{DOG}|\text{cat}) = 0.1$. Hence, one must be very cautious when evaluating accuracies in unbalanced datasets.

The probabilities discussed above are normally empirically obtained using the number of entries in the datasets with a given classification. We will discuss in the next chapter to which extent this is true, and the error we commit. For now, let me assume that this is true.

Let it be $\{A_k\}$ the true labels and $\{B_k\}$ the predicted labels events. Let it be $N(A_k)$ the number of samples with the k-th label, and $N(B_k)$ the number of datapoints with the k-th predicted label, and N_{T} the dataset size. Then

$$P(A_k) \approx \frac{N(A_k)}{N_{\text{T}}}. \tag{1.11}$$

In the same way

$$P(A_k|B_j) \approx \frac{N(A_k \cap B_j)}{N(B_j)}, \tag{1.12}$$

which allows us to infer directly all the conditional probabilities just counting the number of samples with a given label and a given predicted label. In the previous example, it is sufficient to compute

$$P(\text{dog}|\text{DOG}) \approx \frac{N(\text{DOG} \cap \text{dog})}{N(\text{DOG})}. \tag{1.13}$$

Now note that in order to have a good predictor, we want not only that the it predicts a given label with high accuracy, but also that it is specific, that is that also identifies well that a cat or a bird is not a dog.

Let me discuss different ways of evaluating the quality of a prediction in the following simpler case.

1.5 QUALITY ESTIMATORS FOR PREDICTOR EVALUATION

We simplify our scenario by considering two labels: healthy (H) and sick (S), with test results limited to positive (p) or negative (n). This dichotomy establishes the basic terminology:

- **True Positives (TP)**: $N(S \cap p)$, representing correctly identified infections.
- **False Positives (FP)**: $N(H \cap p)$, indicating healthy individuals who were incorrectly diagnosed as infected.
- **True Negatives (TN)**: $N(H \cap n)$, indicating correctly identified healthy individuals.
- **False Negatives (FN)**: $N(S \cap n)$, i.e., infected individuals misdiagnosed as healthy.

Using these definitions, various quality estimators can be formulated:

1. **Sensitivity (Sen) or Recall or True Positive Rate (TPR):** represents the proportion of actual positive cases that were correctly identified. It is formulated as follows:

$$\text{Sen} = \text{Recall} = P(p|S) \approx \frac{N(S \cap p)}{N(S)} = \frac{TP}{TP + FN} \tag{1.14}$$

Sensitivity indicates how reliably the test detects the disease when it is actually present.

2. **Specificity (Spe) or True Negative Rate (TNR):** Indicates the proportion of actual negatives correctly identified. It is given by:

$$\text{Spe} = P(n|H) \approx \frac{N(H \cap n)}{N(H)} = \frac{TN}{TN + FP} \tag{1.15}$$

The high specificity underscores the ability of the test to correctly identify non-diseased individuals.

3. **Precision (Pre) or Positive Predictive Value (PPV):** This metric provides information about the accuracy of positive predictions. More specifically:

$$\text{Pre} = \text{PPV} \approx P(S|p) \approx \frac{N(S \cap p)}{N(p)} = \frac{TP}{TP + FP} \tag{1.16}$$

Precision indicates the probability that individuals with a positive test result truly have the disease. While high precision indicates a low false positive rate, it does not account for false negatives.

4. **Negative Predictive Value (NPV)**: This value complements precision by focusing on negatives:

$$P(H|n) \approx \frac{N(H \cap n)}{N(n)} = \frac{TN}{TN + FN} \tag{1.17}$$

It indicates the probability that a person with a negative test result is truly disease-free.

5. **Accuracy (Acc)**: Represents the overall percentage of correct predictions relative to all predictions:

$$\text{Acc} = P(S \cup p) + P(H \cup n) \approx \frac{TP + TN}{N_T} \tag{1.18}$$

However, accuracy can be misleading, especially for unbalanced data sets where one class may be significantly underrepresented. A test that naively places all instances in the majority class will still achieve high accuracy, making it unreliable in such cases.

6. $\mathbf{F_1}$ **score:** is a useful score to balance between Sensibility and Precision, (it is the harmonic mean between both)

$$\text{F}_1 = 2\frac{\text{Pre} \times \text{Sen}}{\text{Pre} + \text{Sen}} \approx 2\frac{N(S \cap p)}{N(p) + N(S)}. \tag{1.19}$$

Note that $N(S \cap p) \leq \min\{N(p), N(S)\}$, and thus $\text{F}_1 \leq 1$, where the value 1 is only obtained if both p and S is the same event ($N(S \cap p) = N(S) = N(p)$).

1.5.1 *Random Guess Scenario: A Baseline Evaluation*

The effectiveness of a predictor can be gaged not only by its pure performance measures, but also by its comparative superiority over a baseline. One such baseline is the **random guess scenario** - a hypothetical case in which the predictor produces purely random results.

Under the random scenario, the event of being sick (S) and the event of testing positive for the disease (p) are considered independent. This means that the conditional probabilities of being sick when there is a

test result and vice versa are just the marginal probabilities of these events. Mathematically, this can be expressed as follows:

$$\begin{aligned} P(S|p) = P(S|n) = P(S) &\approx \frac{N(S)}{N_T} \\ P(H|p) = P(H|n) = P(H) = 1 - P(S) &\approx \frac{N(H)}{N_T} \\ P(S \cap p) &= P(S)P(p) \\ P(H \cap p) &= P(H)P(p) \end{aligned}$$

Consequently, the expected values of the quality metrics in this scenario are:

$$\begin{aligned} \text{Sen}^r = \text{Recall}^r &= P(p) \\ \text{Spe}^r &= P(n) = 1 - P(p) \\ \text{Pre}^r = \text{PPV}^r &= P(S) \\ \text{NPV}^r &= P(H) = 1 - P(S) \\ \text{Acc}^r &= P(S)P(p) + P(H)P(n) \end{aligned}$$

An important observation here is the inherent symmetry of these results for an uncorrelated predictor:

$$\text{Sen}^r = \text{Recall}^r = 1 - \text{Spe}^r \quad \text{and} \quad \text{Pre}^r = \text{PPV}^r = 1 - \text{NPV}^r \tag{1.20}$$

or that the measures of Precision and Recall (Sensitivity) are be completely decorrelated in the random baseline, each related with a different kind of statistics, the first is linked with the proportion of sick individuals in the dataset, and the second to the proportion of positive results raised by the test. The implications are noteworthy. Even when the performance of a predictor is equivalent to a random estimate, certain metrics can appear deceptively promising, especially when the dataset or predictor is heavily biased toward one class. This phenomenon underscores the need to evaluate all metrics together rather than in isolation to truly assess the quality of a predictor.

For this reason, there are several common ways to combine these measures to assess and compare the quality of different tests (or different hyperparameters)

- The ROC curve and the area under the curve (AUC)
- The Precision-Recall Curves
- The F_1 score

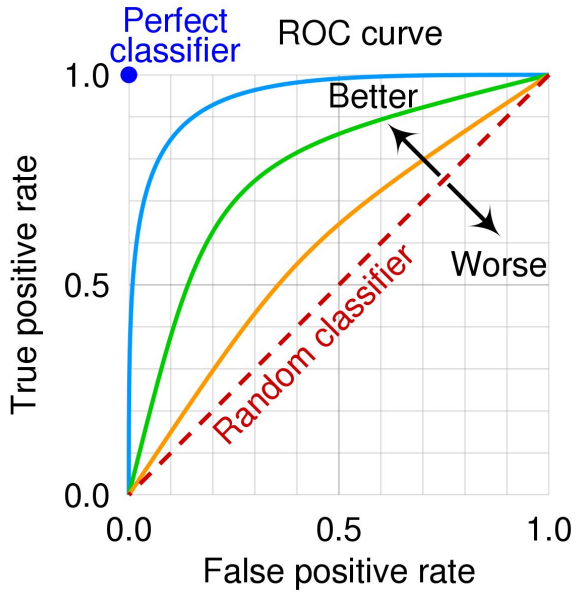
and these tests are particularly important when dealing with unbalanced data sets where one class clearly outnumbers the other, due to two challenges

- **Biased Model**: a model can have high accuracy by predicting only the majority class.
- **Overlooking Minority Class**: The importance of the minority class may be overlooked.

1.5.2 *The ROC Curve*

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the trade-off between sensitivity and specificity for every possible cutoff. Another possibility, but less common, is to represent the Precision against the 1-NPV

- The x-axis represents $1 -$ Specificity (False Positive Rate).
- The y-axis represents Sensitivity.
- A perfect predictor would have an area under the ROC curve (AUC) of 1, while a random predictor would have an AUC of 0.5. Remember that the baseline for a random test is $\text{Sen}^r = 1 - \text{Spe}^r$.



1.6 FREQUENTISTS VS. BAYESIAN INTERPRETATIONS

The fundamental disagreement between the frequentist and Bayesian approaches arises when we try to define the probability of an event based on observed data.

Frequentist Approach:

1. Probability is viewed in terms of the frequency of an event when an experiment is repeated many times.
2. For instance, consider flipping a coin multiple times. The proportion of times we observe a tail is given by:
$$P(\text{tails}) \approx \frac{\text{number of tails observed}}{\text{total number of flips}}$$
3. As we flip the coin more and more, this ratio will get closer and closer to the actual probability of getting a tail. This phenomenon is backed by the law of large numbers, a topic we'll delve into later.

Bayesian Approach:

1. Here, probability is seen as a measure of our belief or confidence about an event.
2. Our beliefs can change based on new observations. For instance, the likelihood of rain varies based on whether the sky is clear or cloudy. If the sky is clear, our belief in the chance of rain drops, whereas a cloudy sky boosts our belief that it might rain.

To summarize, while frequentists see probability as a result of long-term frequency, Bayesians see it as a measure of current belief, which can be updated with new information. The beliefs based on observations can be expressed in terms of conditional probabilities, $P(A|B)$, with A the event, and B the observation. I.e. $P(x|k\text{heads}, n-k\text{tails})$, in the coin toss example.

Conditional probabilities allow us to progressively refine our probabilistic models as we obtain new information. This principle is embodied in the Bayes formula. We will return to "Bayesian statistics" later in this course.

1.7 INTRODUCTION BAYESIAN STATISTICS

Conditional probabilities allow us to refine our probabilistic models as we obtain new information. This principle is embodied in the Bayes formula, from which we derive the term "Bayesian statistics". It is important to clarify that the Bayes formula is a mathematically proven principle, not just a subjective interpretation. Let us take apart the components of Bayes' theorem:

$$\underbrace{P(A|B)}_{\text{posterior}} = \underbrace{P(A)}_{\text{prior}} \times \frac{\overbrace{P(B|A)}^{\text{likelihood}}}{\underbrace{P(B)}_{\text{marginal}}}$$

- $P(A|B)$: The so-called *posterior* probability indicates our revised belief about the event A after considering the evidence B.
- $P(A)$: The so-called *prior* probability reflects our initial belief or assumptions about A before considering evidence.
- $P(B)$: This is called the *marginal* probability, which represents the general probability of the evidence B occurring in any context.
- $P(B|A)$: The *likelihood*, illustrates the probability that the evidence B occurs under the assumption that our belief about A is correct.

In principle one can update the posterior to include all the

Example 1.8. ***Pregnancy Test.*** Let us consider a pregnancy test with a sensitivity of $P(+|\text{preg}) = 90\%$ and a false-positive rate of $P(+|\text{not preg}) = 50\%$. Given that the average probability of being pregnant after a sexual encounter is 15%, we aim to determine the probability of truly being pregnant after receiving one, two, or five positive test results.

To find the posterior probability of being pregnant given a positive test result, we can use Bayes' theorem, given by

$$P(\text{preg}|+) = \frac{P(+|\text{preg})P(\text{preg})}{P(+|\text{preg})P(\text{preg}) + P(+|\text{not preg})P(\text{not preg})},$$

where $P(\text{preg})$ is our prior belief of the probability of being pregnant, which is 15%.

After the first test, we find

$$P(\text{preg}|+) = \frac{0.9 \cdot 0.15}{(0.9 \cdot 0.15) + (0.5 \cdot 0.85)} = \frac{0.135}{0.135 + 0.425} \approx 0.241.$$

This means that after a single positive test, the probability of being pregnant is only about 24.1%.

For subsequent tests, we will update our prior with the posterior probability obtained from the previous test. For instance, for the second test, we use $P(\text{preg}) = 0.241$:

$$P(\text{preg}|++) = \frac{0.9 \cdot 0.241}{(0.9 \cdot 0.241) + (0.5 \cdot (1 - 0.241))} \approx 0.364.$$

Repeating this procedure for a series of tests, we find that after five positive tests,

$$P(\text{preg}|5+) = 0.769.$$