1 Statistics in Data Science

What statistics do Data Scientists need?
- understanding of probability and distributions
- statistical significance
- hypothesis testing

What tools do Data Scientists use that rely on statistics?
- almost everything
- exploratory analysis (e.g. mean, posterior)
- models (e.g. Regression, regularization, Bayesian Graphical Models)
- parameter estimation (e.g. Gibbs sampling, Expectation Maximization)
- and it’s hidden in machine learning too (e.g. SVM, neural networks)

2 Defining Probability

Definition 1. Let $S$ be a sample space. We begin by defining a probability measure on $S$ as
\[ P : \mathcal{P}(S) \mapsto [0, 1], \]  
where $\mathcal{P}$ is the power set of $S$.

Definition 2. We define power sets to be the set of all subsets of $S$ (or any combination of events in $S$). Formally,
\[ \mathcal{P}(S) = \{ A | A \subseteq S \} \]

Example 1. So if $S$ is the set \{H, T\} then $\mathcal{P}(S) = \{ \emptyset, \{H\}, \{T\}, \{H, T\} \}$

Effectively, power sets are the sets of all combinations of elements in $S$. We just defined a mapping that relates each element in the power set to a value between 0 and 1. We also implicitly created a way to measure the probability of a particular event or set of events.

What additional properties do we want our probability measure to have?

Definition 3. Given a sample space $S$, a probability function $P$ that satisfies

1. $P(S) = 1$.
2. Let $A$ be a set in $\mathcal{P}$, then $0 \leq P(A) \leq 1$ for all $A \in \mathcal{P}(S)$.
3. If $A_1, A_2, \ldots \in \mathcal{P}(S)$ and $A_i \cap A_j = \emptyset, i \neq j$ (pairwise disjoint), then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

\[^1\text{This is a slightly more restrictive definition of probability than is necessary but it is sufficient for the content of this course.}\]
So we began defining a function that maps any subset of $S$ to a value between 0 and 1. This buys us a way of **measuring** the chance of any combinations of events in $S$. Also, we’ve made no assumptions about $S$ other than that it is a set of events. This means we can use the basic set operations.

Why is a probability function so important?

We want to define a function $f$ that takes a set of predictors $X$ and maps it to an outcome $Y$. $f(X) + \epsilon = Y$ or $f(X) \mapsto Y$. But how do we define how “good” our function is? We need a measure. The probability function we defined above is an agreed upon measure.

## 3 Probability distributions

We don’t usually have access to the underlying probability structure. So we need a way to go from our sample space to the real numbers. To do that we introduce random variables.

**Definition 4.** A **random variable** is a function from a sample space $S$ into the real numbers.

**Example 2.** For example, if we’re flipping a coin, the final outcome is related to the flip of the coin, the coin’s weight, the density of the air etc. We cannot know for certain whether a given flip will be a heads or a tail (hence the word random) but once we observe the outcome it is recorded as a 0 or a 1. This is what we mean when we say it takes events and maps them to real numbers (Casella, Berger) Suppose we have a sample space $S = \{s_1, \ldots, s_n\}$ with a probability function $P$ and we define a random variable $X$ with the range $\mathcal{X} = \{x_1, \ldots, x_m\}$. We can define a probability function $P_X$ on $\mathcal{X}$. We will observe $X = x_i$ if and only if the outcome of the random experiment is an $s_j \in S$ such that $X(s_j) = x_i$. So

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}).$$

### 3.1 Cumulative Distribution Function

**Definition 5.** With every random variable $X$, we associate a **cumulative distribution function** of $X$:

$$F_X(x) = P_X(X \leq x), \text{ for all } x$$

**Definition 6.** The function $F(x)$ is a cumulative distribution function or cdf if and only if the following three conditions hold:

a. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$

b. $F_X(x)$ is a non-decreasing function of $x$

c. $F_X(x)$ is right continuous; that is, for every number $x_0, \lim_{x \uparrow x_0} F(x) = F(x_0)$

Why do we need cdfs?

To prove things about cdfs we need to be able to evaluate limits (does the cdf go to 1) and to differentiate the function $F(x)$ to confirm it is non-decreasing. But even without proofs we can do
a lot of things. We can compute the probability of seeing events within a particular range and It allows us to define how likely a more extreme event is. It allows us to define identically distributed—the cdfs of two random variables are identical. It constrains our measurement space and allows us to map our probability function to an interpretable and measurable quantity.

Some useful facts that are fairly easy to prove:

\[ P(a < X \leq b) = F(b) - F(a) \]  
\[ P(X > x) = 1 - F(x) \]