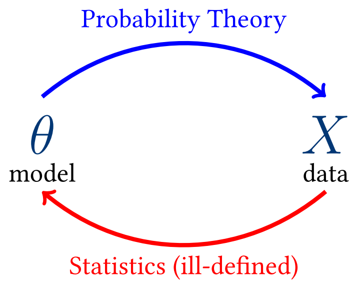
Probability and Statistics

Probability: Given Model, how likely is Data? $\rightarrow$ Well-formed since these are Mathematical questions.
Statistics: Given Data, how likely is Model? $\rightarrow$ Ill-formed since many Models can generate the same data!



There are two interpretations for what the Probability of an Event means:

1. **Frequentists**: *limiting frequency* of the Event
2. **Bayesians**: *subjective belief* that the Event occurs

Probability Space

Probability Space: a triple (Ω, F, P) consisting of:

1. Ω the **Sample Space**
2. $F \subseteq 2^{\Omega}$ a σ-**algebra**[1] on Ω i.e.
   (a) $\Omega \in F$
   (b) $E \in F \Rightarrow E^{\complement} \in F$
   (c) $E_1, E_2, \dots \in F \Rightarrow \bigcup_{i=1}^{\infty} E_i \in F$
3. P: $F \mapsto [0,1]$ a **Probability Measure** i.e.
   (a) $P(E) \geq 0$ for $E \in F$
   (b) $P(\Omega) = 1$
   (c) $P\left(\bigsqcup_{i=1}^{\infty} E_i\right) \Rightarrow \sum_{i=1}^{\infty} P(E_i)$ for $E_i \in F$

Given **Events** $E_i, E \in F$, P also satisfies:

1. **Upward and Downward continuity** of P:
   (a) $E_i \uparrow E \Rightarrow \lim\limits_{n\to\infty} P(E_n) = P(E)$
   (b) $E_i \downarrow E \Rightarrow \lim\limits_{n\to\infty} P(E_n) = P(E)$
2. **Monotonicity** of P:
   (a) $E_i \subseteq E_j \Rightarrow P(E_i) \leq P(E_j)$

Conditional Probability

We can compute Probabilities of Events Conditioned on other Events.
Conditional Probability of event A on event B with $P(B) > 0$ is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

A set of events $\{A_i\}$ are **Mutually Independent** if, for any subset of $\{A_j\}_{j\in k}$:

$$P\left(\bigcap_{j\in k} A_j\right) = \prod_{j\in k} P(A_j)$$

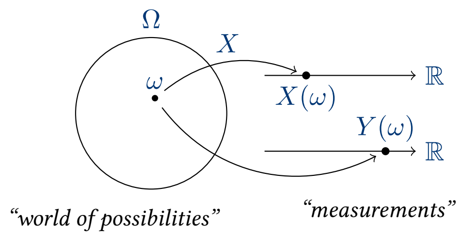
Law of Total Probability: Given Events A and **Partition** $\{B_i\}$ (i.e. where $\bigsqcup_{i=1}^{\infty} B_i = \Omega$)

$$P(A) = \sum_{i=1}^{\infty} P(A \mid B_i)P(B_i)$$

Random Variables

A **Random Variable** is a $\mathbb{B}$-Measurable function X: $(\Omega, F) \mapsto (\mathbb{R}, \mathbb{B})$[2]

1. For $A \in \mathbb{B}$ we can compute $P(X \in A)$
2. $P(X \in A) := P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\})$
3. $P(X^{-1}(\cdot)) := P_X(\cdot)$ which is called the **Push-Forward Measure** of P by X on $\mathbb{R}$
4. Hence X induces a new Probability Space $(\mathbb{R}, \mathbb{B}, P_X)$ from the original (Ω, F, P)



Random Variables can be uniquely determined by their **CDF**:
$F_X(t) := P_X((-\infty, t]) = P(X \leq t)$

1. Right Continuous
2. Non-Negative
3. $\lim\limits_{t\to\infty} F_X(t) = 1$, $\lim\limits_{t\to-\infty} F_X(t) = 0$

Any function satisfying above properties is the CDF for some random variable.

Random Variables studied are usually either **Continuous** or **Discrete** (they can also be **Singular** or **Mixed**).

1. If F_X **Absolutely Continuous** then X is a Continuous RV.[3]
   (a) **Absolutely Continuous**: F differentiable a.e. and $\exists f(x)$ s.t. $F_X(x) = \int_{-\infty}^{x} f(u)du$
   (b) $\Rightarrow \frac{d}{dx}F_X(x) = f(x)$ wherever F is differentiable
   (c) f is called the **PDF**
   (d) f is unique a.e. (may not be everywhere!)
   (e) If X also Non-Negative then **Hazard** of X is $\lambda(t) = \frac{f(t)}{1-F(t)}$
      i. $1 - F(t) = \exp\left(-\int_0^t \lambda(x)\,dx\right)$
      ii. $\lambda(t)$ interpreted as instantaneous survival rate at time t.
      iii. $\lambda(t) = c\ \forall t \iff X \sim Exp(c)$
2. If $X(\Omega)$ is countable then X is Discrete.
   (a) $f(x) := P(\{X = x\})$
   (b) analogously, $F_X(t) = \sum_{i=0}^{t} f(i)$[4]
   (c) f is called the **PMF**

Random Vectors

Joint CDF for $\vec{X} = (X_1, X_2, ..., X_n)$ is
$F(t_1, ..., t_n) = P(X_1 \leq t_1, ..., X_n \leq t_n)$

1. **Marginal PDF** of $\vec{X}_{1:p} = (X_1, \cdots X_p)$ is

$$f_{\vec{X}_{1:p}}(\vec{u}_{1:p}) = \int_{\vec{X}_{(p+1):n}} f_{\vec{X}}\left(\vec{u}_{1:p}, \vec{X}_{(p+1):n}\right) d\vec{X}_{(p+1):n}$$

2. **Conditional PDF** on $\vec{X}_{1:p}$ given $\vec{X}_{(p+1):n}$ is

$$f_{\vec{X}_{1:p}|\vec{X}_{(p+1):n}}\left(\vec{u}_{1:p}, \vec{u}_{(p+1):n}\right) = \frac{f_{\vec{X}}\left(\vec{u}_{1:p}, \vec{u}_{(p+1):n}\right)}{f_{\vec{X}_{(p+1):n}}\left(\vec{u}_{(p+1):n}\right)}$$

3. **kth Order Statistic** $X_{(k)}$ of $\vec{X}$ is the kth smallest value
   (a) $f_{X_{(1)}}(u) = \sum_{i=1}^n f_{X_i}(u) \prod_{j \neq i}(1 - F_{X_j}(u))$
   (b) $f_{X_{(n)}}(u) = \sum_{i=1}^n f_{X_i}(u) \prod_{j \neq i} F_{X_j}(u)$

Moments of a Random Variable

$\mathbb{E}_X(X^r) := \mathbb{E}(X^r)$ is the textbf$r$th Moment of X under the distribution of X

1. $\mathbb{E}(X) = \int_0^\infty 1 - F_X(t)\, dt - \int_{-\infty}^0 F_X(t)\, dt$
2. If X Continuous, $\mathbb{E}(X) = \int_{-\infty}^\infty t \cdot f_X(t)\, dt$
3. **LOTUS**: $\mathbb{E}(g(X)) = \int_{-\infty}^\infty g(t) \cdot f_X(t)\, dt$[5]
4. Moments need not exist! (i.e. $E(|X^r|) = \pm\infty$)

Can generate Moments using the **MGF** of X:
$M_X(t) = \mathbb{E}(\exp(Xt))$, if $\exists \epsilon > 0$ s.t. $\forall |t| < \epsilon$, $M_X(t) < \infty$

1. $\exists \epsilon > 0$ s.t. $\forall |t| < \epsilon$, $M_X(t) = M_Y(t) \Rightarrow X$ and Y have same distribution
2. $\mathbb{E}(|X^r|) = \frac{\partial^r}{\partial^r t} M_X(t)\big|_{t=0}$, if M_X exists.
3. If $\{X_i\}$ independent RVs, then $M_{\sum X_i}(t) = \prod M_{X_i}(t)$

Moments most commonly analyzed are:

1. **Mean** of X: $\mathbb{E}(X) := \mu_X$
2. **Variance** of X: $Var(X) = \mathbb{E}((X - \mu_X)^2) = \sigma_X^2$

For random vectors $\vec{X}$ we have:

1. $\mathbb{E}(\vec{X}) = [\mathbb{E}(X_1), \cdots, \mathbb{E}(X_n)] = \vec{\mu}$
2. $Cov(\vec{X}) = \mathbb{E}[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^\intercal] = \Sigma$

If X and Y are Random Variables on the same Probability Space

1. **Law of Total Expectation**: If $\mathbb{E}(|X|) < \infty$
   $\mathbb{E}(X) = \mathbb{E}_Y(\mathbb{E}_{X|Y}(X \mid Y))$
2. **Law of Total Variance**: If $Var(X) < \infty$
   $Var(X) = \mathbb{E}(Var(X \mid Y)) + Var(\mathbb{E}(X \mid Y))$

Parametric Model

A **Parametric model** is a family of distributions that is defined by a fixed finite number of parameters[6]. Formally,

$$\mathcal{P}_\Theta = \{p_\theta(\cdot; \theta) \mid \theta \in \Theta\}$$

1. $p_\theta(\cdot; \theta)$ is a possible density depending on the **Parameter** θ, and Θ is the **Parameter Space**
2. Most important Parametric family: **Normal Distribution**:
   (a) $X \sim \mathcal{N}_p(\mu, \Sigma)$ with $\mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p \times p}$ Symmetric and Positive Definite iff
   (b) $\forall a \in \mathbb{R}^p$ we have that $a^\intercal x \sim \mathcal{N}_p(a^\intercal \mu, a^\intercal \Sigma a)$
   (c) If Σ non-singular, $f(x) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \mu)^\intercal \Sigma^{-1}(x - \mu)\right\}$
3. Another important family: **Multinoulli Distribution**
   (a) X is a discrete RV over K choices.We encode X as a **one-hot encoding**: a random vector taking values in the unit bases in $\mathbb{R}^K$.
   (b) i.e. $X(\Omega) = \{e_1, e_2, \dots, e_K\}$ where
   $e_j = \begin{pmatrix} 0 & \dots & \underset{\uparrow\; j^{\text{th}} \text{ coordinate}}{1} & \dots & 0 \end{pmatrix}^T \in \mathbb{R}^K$
   (c) $\Theta = \Delta_K$ is the **Probability Simplex** on K choices, and is given by:
   $\Delta_K = \left\{\pi \in \mathbb{R}^K \;;\; \forall j\; \pi_j \geq 0 \text{ and } \sum_{j=1}^K \pi_j = 1\right\}$
   (d) $f(x) = p(x; \pi) = \prod_{j=1}^K \pi_j^{x_j}$ where $x_j \in \{0, 1\}$ is the j^{th} component of x
4. From this we get the: **Multinomial Distribution**
   (a) $X = \sum_{i=1}^n X_i$ where each X_i are IID multinoulli with same parameter π.
   (b) $X(\Omega) = \left\{(n_1, \dots, n_K) \;;\; \forall j\; n_j \in \mathbb{N} \text{ and } \sum_{j=1}^K n_j = n\right\}$

Statistical Decision Theory

A general theory for using Statistics to make decisions under uncertainty. Specifically, given data $D \in \mathcal{D}$, $D \sim P$ for $P \in \mathcal{P}$[7] and set of possible actions $\mathcal{A}$. Note: Θ can be used as $\mathcal{P}$ if using a Parametric family, in which case $P := P_\theta$.

1. Our **Decision Rule** is represented by $\delta : \mathcal{D} \mapsto \mathcal{A}$
2. The **Loss** (cost) of doing an action is given by $L : \mathcal{P} \times \mathcal{A} \mapsto \mathbb{R}$
3. To compare different δ's, can look at the (Frequentist) **Risk** $R(P, \delta) = E_{D \sim P}[L(P, \delta(D))]$. Problem: Risk of any δ changes with P, so must account for this unless δ is Admissible.
   (a) δ_1 **Dominates** δ_2 (for given loss function L) if

$$R(P, \delta_1) \leq R(P, \delta_2) \forall P \in \mathcal{P} \; and$$

$$\exists P \in \mathcal{P},\; R(P, \delta_1) < R(P, \delta_2)$$

   (b) We say that a decision rule δ is **Admissible** if $\nexists \delta_0$ s.t. δ_0 dominates δ.
4. If no δ is Admissible must use a critereon to decide on the optimal one. For Parametric Models:

(a) **Minimax Criteria**: Optimal δ minimizes Risk in worst case scenerio

$$\delta_{minimax} = \min_{\delta} \max_{P \in \mathcal{P}} R(P, \delta)$$

(b) Add a **Weighting** π over Θ (can be interpreted as a Prior)

$$\delta_{weight} = \arg\min_{\delta} \int_{\Theta} R(P_\theta, \delta)\pi(\theta)d\theta$$

(c) **Bayesian Statistical Decision Theory**: Minimize

$$\delta_{bayes}(D) = \arg\min_{\delta} R_B(\delta|D)$$

where $R_B(\delta|D) = \int_\Theta L(P_\theta, \delta)p(\theta|D)d\theta$

i. $p(\theta|D)$ is the posterior for a given prior $\pi(\theta)$.
ii. δ chosen this way is optimal for the given D, since any uncertainty (θ) is integrated out!
iii. $\delta_{bayes} = \delta_{weight}$ if we set π as the prior for Θ.

Maximum Likelihood Estimation

Given some data $x_1, \cdots, x_n$. We want to infer the model which generated the data.
Likelihood Function for some IID observations , coming from a Parametric model is denoted as $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = p(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

Bayesian Statistics

The Bayesian approach is very simple philosophically: it treats all uncertain quantities as random variables.

$$p(\theta \mid X = x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}$$

where,

$p(\theta \mid X = x)$ is the *posterior belief*,

$p(x \mid \theta)$ is the *likelihood* or the observation model,

$p(\theta)$ is the *prior belief* and

$p(x)$ is the *normalization* or ”marginal likelihood”

Notes

1. For some Ω we cannot use 2^{Ω} as a σ-algebra since this may contain sets which do not satisfy all of the axioms. See [1] for an example.

2. Where $\mathbb{B}$ is the **Borel σ-algebra**: the smallest σ-algebra containing all the open intervals. This must contain all intervals of the form $(-\infty, x]$, and since X measurable $\Rightarrow$ F_X guarenteed to exist.

3. Continuity of X as a function on Ω has nothing to do with its continuity as a Random Variable (which depends on the absolute continuity of its CDF) [2]

4. For Discrete RVs, taking P to be the **Counting Measure** and $F = 2^{\Omega}$, it can be shown that Lebesgue integrals are sums. Throughout this cheatsheet whenever we display integrals the reader can replace these with sums as needed.

5. To be explicit, can write $\mathbb{E}_{X\sim f}[g(x)] = \int g(x)f(x)dx$

6. Note: Models with infinite sized Θ are called **Non Parametric**.

7. Often P will describe an IID process, e.g. $D = (X_1, ..., X_n)$ where $X_i \overset{iid}{\sim} P_0$. In this case, the loss is usually written w.r.t P_0 instead of P.

References

[1] J. S. Rosenthal, *A first look at rigorous probability theory.* World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, second ed., 2006.

[2] pidgeot, "On clarifying the relationship between distribution functions in measure theory and probability theory." Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/976739 (version: 2014-10-16).