Elementary Probability

Dom Hutchinson

October 5, 2020

Probability is the study of predicting the likelihood of future events. While *Statistics* is the analysis of data from past events.

1 Definition

Definition 1.1 - *Axioms of Probability*
Let Ω be a sample space and $\mathbb{P} : \mathcal{F} \to [0,1]$ be a probability measure. The *Axioms of Probability* state

- $\mathbb{P}(A) \in [0,1]\ \forall\ A \subset \Omega$.
- $\mathbb{P}(\emptyset) = 0$.
- $\mathbb{P}(\Omega) = 1$.
- $\mathbb{P}\left(\bigcup\limits_{i=1}^{n} A_i\right) = \sum\limits_{i=1}^{n} \mathbb{P}(A_i)$ when $A_1, \dots, A_n$ are all pairwise disjoint.

Definition 1.2 - *Permutation*
A *Permutation* is when selecting r objects from a set of n and the order of selection does matter.
There are ${}^nP_r := \frac{n!}{(n-r)!}$ possible ways to do this.

Definition 1.3 - *Combination*
A *Permutation* is when selecting r objects from a set of n and the order of selection does not matter.
There are ${}^nC_r := \binom{n}{r} = \frac{n!}{(n-r)!r!}$ possible ways to do this.

1.1 Probability Space

Definition 1.4 - *Sample Space, Omega*
A *Sample Space* Ω is the set of all possible events

Definition 1.5 - *Sigmafield, $\mathcal{F}$*
A *Sigmafield*, $\mathcal{F}$, is a set of subsets of a *Sample Space* which fulfil the *Axioms of Probability.*

- $\emptyset \in \mathcal{F}$
- $\forall \{A_1, \dots, A_n\} \subset \mathcal{F}, \quad \left(\bigcup\limits_{i=1}^{n} A_i\right) \in \mathcal{F}.$
- $\forall A \in \mathcal{F}, \quad A^c \in \mathcal{F}$

The events in $\mathcal{F}$ are said to be *$\mathcal{F}$-Measurable.* If $\mathcal{F}_1, \mathcal{F}_2$ are *Sigmafields* then $\mathcal{F}_1 \subset \mathcal{F}_2$ is a *Sigmafield.*

Definition 1.6 - *Probability Measure,* $\mathbb{P}$
A *Probability Measure* is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ which satisfies

- $\mathbb{P}(\emptyset) = 0$.
- $\mathbb{P}(\Omega) = 1$.
- $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$ when $A_1, \dots, A_n$ are all pairwise disjoint.

Definition 1.7 - *Probability Space,* $(\Omega, \mathcal{F}, \mathbb{P})$
A *Probability Space* is a triple of: a sample space Ω; a sigmafield $\mathcal{F}$; and, a probability measure $\mathbb{P}$.

Definition 1.8 - *Random Variable,* X
A *Random Variable* $X : \Omega \to \mathbb{R}$ is a function which maps events to real-values. *Random Variables* represent the possible outcomes of random phenomenon. *Random Variables* have a *Probability Distribution* which specifies the likelihood of events occuring (the distribution is often unknown). *Random Variables* can take either discrete or continuous values.

Definition 1.9 - *Random Vector,* $\boldsymbol{X}$
A *Random Vector* is a vector whose values depend on random events. Each element can be assigned a different distribution (dependent or independet). Often IID variables are considered as a *Random Vector* to compress notation.

1.2 Probability Mass Functions

Definition 1.10 - *Probability Distribution*
Probability Distributions are functions which return the probability of a specific outcome of a *Random Variable* occuring. See `ProbabilityDistributions.pdf` for some common & well defined distributions.

Definition 1.11 - *Probability Function,* $f_X(\cdot)$
A *Probability Mass Function* is the probability distribution for a discrete random variable.
A *Probability Density Function* is the probability distribution for a continuous random variable.

$$\begin{aligned} f_X(x) &:= \mathbb{P}(X = x) \\ &= \int f_{X,Y}(x, y)dy \end{aligned}$$

Definition 1.12 - *Cummulative Probability Function,* $F_X(\cdot)$
A *Cummulative Probability Function* gives the probability of observing a value less than, or equal to, the value specified.

$$\begin{aligned} F_X(x) &:= \mathbb{P}(X \leq x) \\ &= \int_{-\infty}^{x} f_X(y)dy \\ &= \sum_{i=-\infty}^{x} f_X(i) \end{aligned}$$

Definition 1.13 - *Joint Probability Function,* $f_{X,Y}(\cdot,\cdot)$
A *Joint Probability Function* is the probability distribution for multiple random variables and returns the probability of the random variables having specifed values at the same time.

$$f_{X,Y}(x,y)=\mathbb{P}(X=x,Y=y)$$

Definition 1.14 - *Conditional Probability Function,* $f_{X|Y}(\cdot|\cdot)$
A *Conditional Probability Function* is defined for multiple random variables (say X & Y) and defines the probability of X having a specific value given that Y has a specific value.

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

1.3 Describing Random Variables

Definition 1.15 - *Expected Value,* $\mathbb{E}(\cdot)$
The *Expected Value* of a random variable is the weighted average value and equivalent to the arithmetic mean. Let $X\sim f_X(\cdot)$

$$\begin{aligned}\mathbb{E}(X) &:= \int xf_X(x)dx && \text{[Continuous RV]}\\ &:= \sum_x xf_X(x) && \text{[Continuous RV]}\end{aligned}$$

The *Conditional Expected Value* of a random variable is its expected value, given another random variable has a specified value.

$$\begin{aligned}\mathbb{E}(X|Y=y) &= \int xf_{X|Y}(x|y)dx\\ &= \sum_x xf_{X|Y}(x|y)\end{aligned}$$

Definition 1.16 - *Variance, Var*$(\cdot)$
The *Variance* of a random variable measures the expected spread of values around the expected value.

$$\begin{aligned}\text{Var}(X) &:= \mathbb{E}[(X-\mathbb{E}(X))^2]\\ &= \mathbb{E}(X^2)-\mathbb{E}(X)^2\end{aligned}$$

The *Conditional Variance* of a random variable is the variance of the random variable given the value of another random variable(s).

$$\text{Var}(X|Y)=\mathbb{E}\left([X-\mathbb{E}(X|Y)]^2|Y\right)$$

Definition 1.17 - *Percentage Points,* x_α
A *Percentage Point* $x_\alpha\in\mathbb{R}$ of a random variable X is the value such that $\alpha\in[0,1]$ of the distributions mass is less than x_α.

$$\mathbb{P}(X<x_\alpha)=\alpha \quad \int_{-\infty}^{x_\alpha} xf_X(x)dx=\alpha \quad \sum_{x=-\infty}^{x_\alpha} xf_X(x)=\alpha$$

1.4 Dependence & Correlation

Definition 1.18 - *Covariance, Cov($\cdot$)*
Covariance is a measure of the relationship between two random variables. A value close to zero indicates no relationship; a negative value indicates a negative correlation; and, a positive value indicates a positive correlation.

$$\begin{array}{rcl}\text{Cov}(X,Y) & := & \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] \\ & = & \mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)\end{array}$$

Definition 1.19 - *Pearson's Correlation Coefficient, Corr($\cdot,\cdot$)*
Pearson's Correlation Coefficient is a measure of the relationship between two random variables. Similar to *Covariance* except it is valued in $[-1,1]$.

$$\text{Corr}(X,Y)=\frac{\text{Cov}(X,Y)}{\text{Var}(X)\text{Var}(Y)}$$

Definition 1.20 - *Independent Random Variables*
A set of random variables $X_1,\dots,X_n$ are *Mututally Independent* iff

$$\forall\{x_1,\dots,x_n\}\quad f_{X_1,\dots,X_n}(x_1,\dots,x_n)=\prod_{i=1}^n f_{X_i}(x_i)$$

If $X_1,\dots,X_n$ are mutually independent and have the same distribution then they are said to be *Independent Indetically Distributed* (IID) random variables.
If $X_1,\dots,X_n$ are mutually independent then $\mathbb{E}(X_1)\dots\mathbb{E}(X_n)$ and $\text{Cov}(X_i,X_j)=0\ \forall\ i\neq j$.

1.5 Moments

Definition 1.21 - *Moments*
The n^{th} *Moment* of a random variable X is $\mathbb{E}(X^n)$.

Definition 1.22 - *Moment Generating Function*
A *Moment Generating Function* is an alternative specification of a *Probability Distribution*. *MGF*s are unique for each distribution and thus is two distributions have the same *MGF* then they are the same distribution.

$$\begin{array}{rcll}\mathcal{M}_X(t) & := & \mathbb{E}(e^{tX}) & \text{for } t\in\mathbb{R} \\ & = & \displaystyle\int e^{tx}f_X(x)dx & \\ & = & \displaystyle\sum_x e^{tx}f_X(x) & \end{array}$$

1.6 Intervals

Definition 1.23 - *Random Interval, $\mathcal{I}(\cdot)$*
A *Random Interval* is an interval whose bounds depend on random variable(s).

Definition 1.24 - *Wald's Confidence Interval*
A *Wald Confidence Interval* is a *Random Interval* $\mathcal{I}(\boldsymbol{X}) := [L(\boldsymbol{X}), U(\boldsymbol{X})]$ used to give a continuous range of possible values for an unknown parameter, dependent on observed data.
The *Coverage* of a *Confidence Interval* is the probability of the true parameter value being in the interval.
A $1-\alpha$ *Confidence Interval* is a *Confidence Interval* with coverage of at least $1-\alpha$.

$$\mathbb{P}(\theta^* \in \mathcal{I}(\boldsymbol{X})) \geq 1-\alpha$$

Consider a bijective, continuously differentiable transformation of a parameter $\tau := g(\theta)$. A *Confidence Interval* for $\tau(\theta^*)$ can be derived as

- $[g(L(\boldsymbol{X})), g(U(\boldsymbol{X}))]$ if τ is *increasing.*
- $[g(U(\boldsymbol{X})), g(L(\boldsymbol{X}))]$ if τ is *decreasing.*

Definition 1.25 - *Wilks' Confidence Set*
A *Wilks Confidence Set* is the set of estimators which are sufficient close in likelihood to the *Maximum Likelihood Estimate.* See `StatisticalModels.tex` for more.

Theorem 1.1 - *Convergence of Confidence Intervals*

1.7 Convergence

Definition 1.26 - *Convergence in Probability, $Z_n \to_{\mathbb{P}} Z$*
A sequence of random variables $\{Z_n\}_{n\in\mathbb{N}}$ *Converges in Probability* to random variable Z if

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}(|Z_n - Z| > \varepsilon) = 0$$

Definition 1.27 - *Convergence in Distribution, $Z_n \to_D Z$*
A sequence of random variables $\{Z_n\}_{n\in\mathbb{N}}$ *Converges in Distribution* to random variable Z if

$$\forall z \in Z \text{ where } F_Z(z) \text{ is continuous } \lim_{n\to\infty} F_{Z_n}(z) = F_Z(z)$$

Definition 1.28 - *Convergence in Quadratic Mean, $Z_n \to_{qm} Z$*
A sequence of random variables $\{Z_n\}_{n\in\mathbb{N}}$ *Converges in Quadratic Mean* to random variable Z if

$$\lim_{n\to\infty} \mathbb{E}\left[(Z_n - Z)^2\right] = 0$$

Remark 1.1 - *Hierarchy of Convergences*

- $Z_n \to_{qm} Z \implies Z_n \to_{\mathbb{P}} Z$
- $Z_n \to_{\mathbb{P}} Z \implies Z_n \to_D Z$
- $\forall\ a \in \mathbb{R} \quad Z_n \to_{\mathbb{P}} a \iff Z_n \to_D a$

Theorem 1.2 - *Continuous Mapping Theorem*

1. $Z_n \to_{\mathbb{P}} Z \implies g(Z_n) \to_{\mathbb{P}} g(Z)$

2. $Z_n \to_D Z \implies g(Z_n) \to_D g(Z)$

Theorem 1.3 - *Slutsky's Theorem*
If $Y_n \to_D Y$ and $Z_n \to_D c$ for $c \in \mathbb{R}$. Then

1. $Y_n + Z_n \to_D Y + c$
2. $Y_n Z_n \to Yc$
3. $\dfrac{Y_n}{Z_n} \to_D \dfrac{Y}{c}$

2 Theorems

Theorem 2.1 - *Bayes Theorem*
Consider $X \sim f_X(\cdot, \theta)$

$$\underbrace{\mathbb{P}(\theta|X)}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}(X|\theta)}^{\text{Likelihood}}\ \overbrace{\mathbb{P}(\theta)}^{\text{Prior}}}{\underbrace{\mathbb{P}(X)}_{\text{Evidence}}}$$

Theorem 2.2 - *Binomial Theorem*
Let $a, b \in \mathbb{R}$ then $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$.

Theorem 2.3 - *Boole's Inequality*
Let $A_1, \dots, A_n$ be a set of events.

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$$

The probability of all events occuring is never greater than the sum of the probability of the events occuring independently.

Theorem 2.4 - *Central Limit Theorem*
Let $X_1, \dots, X_m$ be iid random variables with $\mathbb{E}(X) = \mu$, $\text{Var}(X) = \sigma^2$ and $\bar{X}_n$ be the sample mean of the first n. Then for large n

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

Theorem 2.5 - *Chain Rule*
Let $A_1, \dots, A_n$ be a set of events.

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \prod_{i=1}^{n} \mathbb{P}\left(A_i \middle| \bigcap_{j=1}^{i-1} A_j\right)$$

Theorem 2.6 - *Chebyshev's Inequality*

Let X be a random variable, define $\mu := \mathbb{E}(X),\ \sigma^2 := \text{Var}(X)$ and let $c \in \mathbb{R}$.

$$\mathbb{P}(|X-\mu| > c) \leq \frac{\sigma^2}{c^2}$$

Theorem 2.7 - *de Moivre-Laplace Theorem*
Let $X_n \sim \text{Binomial}(n,p)$ be a sequence of binomial random variables with fixed probability p and $a < b$.

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{X_n - np}{\sqrt{np(1-p)}} \leq b\right) = \Phi(b) - \Phi(a)$$

Theorem 2.8 - *Inclusion-Exclusion Principle*
Let $A_1, \dots, A_n$ be a set of events.

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n (-1)^{i+1}\left(\sum_{1\leq j_1<\cdots<j_i\leq n} |A_{j_1} \cap \cdots \cap A_{j_i}\right)$$

Theorem 2.9 - *Lack of Memory Property*

$$\mathbb{P}(X = x+n | X > n) = \mathbb{P}(X = x)$$

Theorem 2.10 - *Law of Total Expectation*

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$$

Theorem 2.11 - *Markov's Inequality*
Let X be a non-negative random variable and $a > 0$ then $\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}(X)$.

Theorem 2.12 - *Partition Theorem*
Let $B_1, \dots, B_n$ be a disjoint partition of the sample space with $\mathbb{P}(B_i) > 0\ \forall\ i$.

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Theorem 2.13 - *Weak Law of Large Numbers*
Let $X_1, \dots, X_n$ be a set of IID RVs each with mean μ.

$$\forall\ \epsilon > 0 \quad \mathbb{P}\left(|\bar{X} - \mu| > c\right) \underset{n\to\infty}{\longrightarrow} 0$$

3 Identities

$$\begin{array}{rcll}
(A\cup B)^c &=& A^c\cap B^c & \\
(A\cap B)^c &=& A^c\cup B^c & \\
1-\mathbb{P}(A\cup B) &=& \mathbb{P}(A^c\cap B^c) & \text{[de Morgan's Law]}\\
1-\mathbb{P}(A\cap B) &=& \mathbb{P}(A^c\cup B^c) & \text{[de Morgan's Law]}\\
\binom{n}{k} &=& \binom{n}{n-k} & \\
\binom{n}{k} &=& \binom{n-1}{k-1}+\binom{n-1}{k} & \text{[Pascal's Identity]}\\
p_X(x) &=& \int p_{X,Y}(x,y)dy & \\
\mathbb{P}(A^c|B) &=& 1-\mathbb{P}(A|B) & \\
\mathbb{P}(\emptyset|B) &=& 0 & \\
\mathbb{P}(A\cup C|B) &=& \mathbb{P}(A|B)+\mathbb{P}(C|B)-\mathbb{P}(A\cap C|B) & \\
\mathbb{E}(aX+b) &=& a\mathbb{E}(X)+b & \\
\mathbb{E}(X+Y) &=& \mathbb{E}(X)+\mathbb{E}(Y) & \\
\mathbb{P}(X=x) &=& \sum_{y\in Y}\mathbb{P}(X=x|Y=y) & \\
\mathrm{Var}(X) &=& \mathbb{E}(X^2)-[\mathbb{E}(X)]^2 & \\
\mathrm{Var}(aX+b) &=& a^2\mathrm{Var}(X) & \\
\mathrm{Var}(X+Y) &=& \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y) & \\
\mathrm{Cov}(X,Y) &=& \mathbb{E}(XY)-\mathbb{E}(X)\cdot\mathbb{E}(Y) & \\
\mathrm{Cov}(aX,bY) &=& ab\mathrm{Cov}(X,Y) & \\
\mathrm{Cov}(X,Y+Z) &=& \mathrm{Cov}(X,Y)+\mathrm{Cov}(X,Z) & \\
\mathcal{M}_{aX+b}(t) &=& e^{tb}\mathcal{M}_X(ta) & \\
\mathcal{M}_{aX+bY}(t) &=& \mathcal{M}_X(at)\cdot\mathcal{M}_Y(bt) &
\end{array}$$