# Contents

**1 Background** **3**
1.1 Probability vs Statistics . . . . . . . . . . . . . . . . . . . . . 3
1.2 Double Integrals . . . . . . . . . . . . . . . . . . . . . . . . . 3

**2 Elementary (Discrete) Probability** **5**
2.1 Foundations . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
2.2 Conditional Probabilities . . . . . . . . . . . . . . . . . . . . 8
2.3 Independence . . . . . . . . . . . . . . . . . . . . . . . . . . 10

**3 Discrete Random Variables** **13**
3.1 Random Variables . . . . . . . . . . . . . . . . . . . . . . . . 13
3.2 Discrete Random Variables . . . . . . . . . . . . . . . . . . . 14
3.3 Discrete Distributions . . . . . . . . . . . . . . . . . . . . . . 16
3.4 Multiple Variables . . . . . . . . . . . . . . . . . . . . . . . . 17
3.5 Moments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

3.6 Tails . . . . . . . . . . 23

**4 Continuous Random Variables** **24**

4.1 Single Variables . . . . . . . . . . 24

4.2 Multiple Variables . . . . . . . . . . 25

4.3 Poisson Processes . . . . . . . . . . 25

**5 (Discrete) Markov Chains** **26**

**6 Estimators** **27**

6.1 Evaluation . . . . . . . . . . 27

6.2 Maximum Likelihood Estimation (MLE) . . . . . . . . . . 27

# 1 Background

## 1.1 Probability vs Statistics

- Probabilistic reasoning works forwards from a given mathematical model to the single possible prediction about given events under that model
- Statistical reasoning works backwards from given data about events to one of many possible mathematical models that could explain that data

## 1.2 Double Integrals

- Naive approach: Blindly compute inner then outer e.g. $\int_{y=0}^{y=\infty} \int_{x=0}^{x=y} e^{-y} dx dy = \int_{y=0}^{y=\infty} [xe^{-y}]_{x=0}^{x=y} dy = \int_{y=0}^{y=\infty} ye^{-y} - 0e^{-y} = ... = 1$

- *Fubini's Theorem: Suppose* $f(x,y)$ *is continuous throughout* $R = \{(x,y) \in \mathbb{R}^2 : a \leq x \leq b$ *and* $c \leq y \leq d\}$. *Then,* $\iint\limits_R f(x,y)dR = \int_c^d \int_a^b f(x,y)dxdy = \int_a^b \int_c^d f(x,y)dydx$

- Can be easier to reverse the order as per Fubini e.g. $\int_{y=0}^{y=\infty} \int_{x=0}^{x=y} e^{-y}dxdy = \int_{x=0}^{x=\infty} \int_{y=x}^{y=\infty} e^{-y}dydx = \int_{x=0}^{x=\infty} [-e^{-y}]_{y=x}^{y=\infty}dx = \int_{x=0}^{x=\infty} e^{-x}dx = [-e^{-x}]_{x=0}^{x=\infty} = 1$

# 2 Elementary (Discrete) Probability

## 2.1 Foundations

- **All probabilities are defined based on a given experiment**
- **The sample space** $\Omega =$ **the set of possible outcomes of the experiment**
- **The sample space** $\Omega$ **is discrete iff** $\Omega$ **is a countable set** (recall this means it **may be infinite**)
- **The sample space** $\Omega$ **is continuous** iff it is not discrete **iff** $\Omega$ **is an uncountably infinite set**
- **An event E is** $\subseteq \Omega$ i.e. zero or more outcomes
- **An event is elementary iff it is singleton** i.e. a single outcome

- Deduce that the set of events is closed under the standard set operations
- **Events $E_1$ and $E_2$ are mutually exclusive iff $E_1 \cap E_2 = \emptyset$**
- Recall that a family of sets (e.g events) $E_1, ..., E_n$ partition a set $F$ iff $\bigcup_{i \in [n]} E_i = F$ and $\forall i \in [n]. \forall j \in [n] \setminus \{i\}. E_i \cap E_j = \emptyset$
- **The 3 axioms of probability:**

  1. **Non-negativity: Let $E$ be an event. Then, $\mathbf{P}(E) \geq 0$**
  2. **Additivity: Let $E_i$ and $E_j$ be mutually exclusive events. Then, $\mathbf{P}(E_i \cup E_j) = \mathbf{P}(E_i) + \mathbf{P}(E_j)$**
  3. **Normalization: $\mathbf{P}(\Omega) = 1$ or equivalently $\mathbf{P}(\emptyset) = 0$**

- Note that additivity means that (by induction) the probability of any event can be broken down into a sum of probabilities of elementary events that partition it

- Proposition: $\mathrm{P}(E \cap F) = \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cup F)$

  Proof:

  By additivity, $\mathrm{P}(A) = \sum_{\omega \in A} \mathrm{P}(\omega)$ for each $A$. Recall that the members of each $A$ (the $\omega$s) are elementary events.

  Pick arbitrary $\omega \in E \cup F$. Deduce that $\omega \in E \vee \omega \in F$.

  Case $\omega \in E$ but $\omega \notin F$: Then, $\mathrm{P}(\omega)$ occurs in $\mathrm{P}(E)$ and $\mathrm{P}(E \cup F)$ but not in $\mathrm{P}(F)$ and so simplifies down to $\mathrm{P}(\omega) - \mathrm{P}(\omega) = 0$ in $\mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cup F)$

  Case $\omega \in F$ but $\omega \notin E$: Then, $\mathrm{P}(\omega)$ occurs in $\mathrm{P}(F)$ and $\mathrm{P}(E \cup F)$ but not in $\mathrm{P}(E)$ and so simplifies down to $\mathrm{P}(\omega) - \mathrm{P}(\omega) = 0$ in $\mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cup F)$

  Case $\omega \in E$ and $\omega \in F$: Then, $\mathrm{P}(\omega)$ occurs in $\mathrm{P}(F)$ and $\mathrm{P}(E \cup F)$ but not in $\mathrm{P}(E)$ and so simplifies down to $\mathrm{P}(\omega) + \mathrm{P}(\omega) - \mathrm{P}(\omega) = \mathrm{P}(\omega)$ in $\mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cup F)$. Moreover, $\omega \in E \cap F$.

  Deduce that we have covered each term of $\mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cup F)$ exactly once. Thus, we have shown that $\mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cup F) = \sum_{\omega \in E \cap F} \mathrm{P}(\omega)$ and we know that this equals $\mathrm{P}(E \cap F)$ as required.

Corollary **(Union Bound):** $\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F) - \mathbf{P}(E \cap F) \leq \mathbf{P}(E) + \mathbf{P}(F)$ **[as** $\mathbf{P}(E \cap F) \geq 0$ **by non-negativity axiom]**

## 2.2 Conditional Probabilities

- **Probability that E occurred given that F occurred** $= \mathbf{P}(E|F) = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(F)}$
  Corollary: $\mathrm{P}(E \cap F) = \mathrm{P}(F)\mathrm{P}(E|F) = \mathrm{P}(E)\mathrm{P}(F|E)$

- **Chain Rule: Let** $\mathbf{P}(\bigcap_{i=1}^{n} E_i) > 0$**. Then,** $\mathbf{P}(\bigcap_{i=1}^{n} E_i) = \prod_{i=1}^{n} \mathbf{P}(E_i | \cap_{j=1}^{i-1} E_j)$.
  For example, $\mathrm{P}(E_1)\mathrm{P}(E_2|E_1)\mathrm{P}(E_3|E_1 \cap E_2)\mathrm{P}(E_4|E_1 \cap E_2 \cap E_3) = \mathrm{P}(E_1 \cap E_2)\mathrm{P}(E_3|E_1 \cap E_2)\mathrm{P}(E_4|E_1 \cap E_2 \cap E_3) = \mathrm{P}(E_1 \cap E_2 \cap E_3)\mathrm{P}(E_4|E_1 \cap E_2 \cap E_3) = \mathrm{P}(E_1 \cap E_2 \cap E_3 \cap E_4)$ as claimed.

- **Theorem (Law of total probability): Let** $E$ **be an event and** $F_1, ..., F_n$ **be a partition of** $\Omega$**. Then,** $\mathbf{P}(E) = \sum_{i \in [n]} \mathbf{P}(E \cap F_i) =$

$\sum_{i\in[n]} \mathbf{P}(E|F_i)\mathbf{P}(F_i)$. For example, $\mathrm{P}(A) = \mathrm{P}(A|B) + \mathrm{P}(A|\neg B)$.
Proof: Recall that $E \subseteq \Omega$ and thus deduce that $E = E \cap \Omega$. Hence, $E = E \cap (F_1 \cup ... \cup F_n) = E \cap F_1 \cup ... \cup E \cap F_n$. Recall that $F_1, ..., F_n$ are pairwise disjoint and thus deduce that $E \cap F_1, ..., E \cap F_n$ are pairwise disjoint. Hence, by additivity, $\mathrm{P}(A) = \mathrm{P}(A \cap B_1) + ... + \mathrm{P}(A \cap B_n)$ as required.

- **Proposition: Let $E$ and $F$ be events and $G_1, ..., G_n$ be a partition of $\Omega$. Then, $\mathbf{P}(E|F) = \sum_{i\in[n]} \mathbf{P}(E|F_i \cap G_i)\mathbf{P}(F_i|G_i)$.**

  Proof: $\mathrm{P}(E|F) = \frac{\mathrm{P}(E\cap F)}{\mathrm{P}(F)}$. Applying law of total probability: $\mathrm{P}(E \cap F) = \sum_{i\in[n]} \mathrm{P}(E \cap F \cap G_i) = \sum_{i\in[n]} \mathrm{P}(E|F \cap G_i)\mathrm{P}(F \cap G_i) = \sum_{i\in[n]} \mathrm{P}(E|F \cap G_i)\mathrm{P}(G_i|F)\mathrm{P}(F)$. Thus, $\mathrm{P}(E|F) = \frac{\mathrm{P}(E\cap F)}{\mathrm{P}(F)} = \sum_{i\in[n]} \frac{\mathrm{P}(E|F\cap G_i)\mathrm{P}(G_i|F)\mathrm{P}(F)}{\mathrm{P}(F)} = \sum_{i\in[n]} \mathrm{P}(E|F \cap G_i)\mathrm{P}(G_i|F)$ as required.
  Corollary: If $\forall i \in [n].F \perp\!\!\!\perp G_i$, then $\mathrm{P}(E|F) = \sum_{i\in[n]} \mathrm{P}(E|F_i \cap G_i)\mathrm{P}(G_i)$; a closer symmetry with the law of total probability.

- **Theorem (Bayes' Theorem):** $\mathbf{P}(F|E) = \frac{\mathbf{P}(E|F)\mathbf{P}(F)}{\mathbf{P}(E)} = \frac{\mathbf{P}(E|F)\mathbf{P}(F)}{\mathbf{P}(E|F)\mathbf{P}(F)+\mathbf{P}(E|\neg F)\mathbf{P}(\neg F)}$
  Proof: Recall that $\mathrm{P}(E \cap F) = \mathrm{P}(E|F)\mathrm{P}(F) = \mathrm{P}(F|E)\mathrm{P}(E)$. Thus, $\mathrm{P}(F|E) = \mathrm{P}(F|E) = \frac{\mathrm{P}(E|F)\mathrm{P}(F)}{\mathrm{P}(E)}$.
  By the definition of $\neg$, $F \cap \neg F = \emptyset$ and $F \cup \neg F = \Omega$. Hence, we can apply the law of total probability: $\mathrm{P}(E) = \mathrm{P}(E|F)\mathrm{P}(F) + \mathrm{P}(E|\neg F)\mathrm{P}(\neg F)$ completing the proof.

## 2.3 Independence

- **Events $A_1$, ..., $A_n$ are independent iff $\forall S \in 2^{\{A_1,\ldots,A_n\}}.\mathbf{P}(S) = \prod_{X\in S}\mathbf{P}(X)$.** For example, $A_1, A_2, A_3, A_4$ are independent iff $\mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1)\mathrm{P}(A_2)$ and $\mathrm{P}(A_1 \cap A_3) = \mathrm{P}(A_1)\mathrm{P}(A_3)$ and $\mathrm{P}(A_1 \cap A_4) = \mathrm{P}(A_1)\mathrm{P}(A_4)$ and $\mathrm{P}(A_2 \cap A_3) = \mathrm{P}(A_2)\mathrm{P}(A_3)$ and $\mathrm{P}(A_2 \cap A_4) = \mathrm{P}(A_2)\mathrm{P}(A_4)$ and $\mathrm{P}(A_3 \cap A_4) = \mathrm{P}(A_3)\mathrm{P}(A_4)$ and $\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3)$

and $P(A_1 \cap A_2 \cap A_4) = P(A_1)P(A_2)P(A_4)$ and $P(A_2 \cap A_3 \cap A_4) = P(A_2)P(A_3)P(A_4)$ and $P(A_1 \cap A_2 \cap A_3 \cap A_4) = P(A_1)P(A_2)P(A_3)P(A_4)$
**Corollary: Events $E$ and $F$ are independent ($E \perp\!\!\!\perp F$) iff $\mathbf{P}(E \cap F) = \mathbf{P}(E)\mathbf{P}(F)$**

- Proposition: Assuming that $P(E) > 0$ and $P(F) > 0$ respectively, $E \perp\!\!\!\perp F$ iff $P(E|F) = P(E)$ iff $P(F|E) = P(F)$
  Proof: Trivial from Bayes' theorem

- Proposition: If $E_1$ and $E_2$ are mutually exclusive events, then $E_1$ and $E_2$ are independent only if at least one of $E_1$ and $E_2$ has probability 0
  Proof: $E_1$ and $E_2$ are mutually exclusive $\Rightarrow P(E_1 \cap E_2) = 0$.
  $E_1$ and $E_2$ are independent $\Rightarrow P(E_1)P(E_2) = P(E_1 \cap E_2) = 0 \Rightarrow$ $(P(E_1) = 0 \ \vee \ P(E_2) = 0)$.

- Events $A_1$, ..., $A_n$ are pairwise independent iff $\forall i \in [n].\forall j \in [n]\setminus\{i\}.P(A_i \cap A_j) = P(A_i)P(A_j)$

- Proposition: There exists a family of events that are pairwise independent but not fully independent. Proof: Exercise

- Events $E$ and $F$ are conditionally independent given an event G iff $\mathrm{P}(E \cap F|G) = \mathrm{P}(E|G)\mathrm{P}(F|G)$

- Proposition: There exists a family of events that are conditionally independent given some additional event but not fully independent. Proof: Exercise

- Proposition: There exists a family of events that are fully independent but for which there exists an event they are not conditionally independent given. Proof: Exercise

# 3 Discrete Random Variables

## 3.1 Random Variables

- In CS130 we only considered events: an abstraction over elementary outcomes (members of sets (or singleton sets)). In this module we focus on random variables: an abstraction over events (sets).

- A random variable $X$ is a function $\Omega \mapsto \mathbb{R}$. Note that, unlike $\mathrm{P} : 2^{\Omega} \mapsto \mathbb{R}$ which assigns a number to each set (event (collection of outcomes of experiments)), **a random variable assigns a number to each member of the set $\Omega$ (an outcome of an experiment** (an elementary event)**)**.

- As $X^{-1}$ assigns to each number a set of outcomes, **considering a random variable being equal to a certain value exactly corresponds to considering a certain event** $((X = i) = \{\omega : X(\omega) = i\} = \{\omega : \omega \in X^{-1}(i)\})$

- Proposition: $\{(X = i) : i \in \mathbb{R}\}$ is a partition of $\Omega$.
  Proof: Pick arbitrary $i \in \mathbb{R}$.
  $\cup_{i\in\mathbb{R}}(X = i) = \Omega$ as $\mathbb{R} \supseteq$ codomain($X$) and so this union must be the entire domain of $X$. (Many of the sets being unioned will be $\emptyset$ but this is fine).
  Pick arbitrary $j \in \mathbb{R}$ such that $i \neq j$. Then, $(X = i) \cap (X = j) = \emptyset$ as any element in this intersection would be an elementary outcome associated with more than one number by $X$ and so would contradict $X$ being a function.

## 3.2 Discrete Random Variables

- **A discrete random variable has countable range (e.g. $\mathbb{N}$) whereas a continuous random variable has uncountable range (e.g. $\mathbb{R}$)**. For the rest of this chapter all random variables are discrete

- $\mathbf{p}_X(i) =$ **probability mass function (pmf) of** $X = \mathbf{P}(X = i)$

- $\mathbf{p}_{X|A}(i) =$ **probability mass function of** $X$ **given** $A = \frac{\mathbf{P}(\{X=i\}\cap A)}{\mathbf{P}(A)}$. $X|A$ can be viewed as an event in its own right, with this pmf.

- $\mathbf{F}_X(i) =$ **cumulative distribution function of** $X = \mathbf{P}(X \leq i) = \mathrm{P}(\cup_{j\leq i} X = j)$ by definition $= \sum_{j\leq i} \mathbf{P}(X = j)$ **by additivity** (as the $X = j$ are a partition of $X \leq i$)

- $\neg\mathbf{F}_X(i) =$ **tail function of** $X = \mathbf{P}(X > i) = 1 - \mathbf{F}_X(i)$. Note this is not $\mathrm{P}(X \geq i)$

- Proposition: Let $X$ be a random variable. Then, $\sum_{i\in\mathbb{R}} \mathrm{P}(X = i) = 1$
  Proof:
  i) Pick arbitrary $i \in \mathbb{R}$. Pick arbitrary $j \in \mathbb{R}$ such that $i \neq j$. Then, $(X = i) \cap (X = j) = \emptyset$ as any element in this intersection would be an elementary outcome associated with more than one number by $X$ and so would contradict $X$ being a function.

ii) $\cup_{i\in\mathbb{R}}(X = i) = \Omega$ as $\mathbb{R} \supseteq \text{range}(X)$ and so this union must be the entire domain of $X$. (Many of the sets being unioned will be $\emptyset$ but this is fine). Combining i) and ii) $\{(X = i) : i \in \mathbb{R}\}$ is a partition of $\Omega$. Thus, as each $X = i$ is an event, we can apply the law of total probability to obtain $\sum_{i\in\mathbb{R}} \mathrm{P}(X = i) = \mathrm{P}(\Omega)$. Finally, by the normalization axiom, $\mathrm{P}(\Omega) = 1$.

## 3.3 Discrete Distributions

- **If $X \sim$ Bernoulli$(p)$, then range$(X) = \{0, 1\}$ and $\mathbf{p}_X(1) = p$ and $\mathbf{p}_X(0) = 1 - p$.** Interpretation: Single trial of a binary experiment; 1 iff success; 0 iff failure; $p =$ probability of success

- **If $X \sim$ Binomial$(n, p)$, then range$(X) = \{\underline{0}, ..., n\}$ and $\mathbf{p}_X(i) = \binom{n}{i}(p)^i(1-p)^{n-i}$.** Interpretation: Number of successes out of $n$ independent Bernoulli trials each with probability of success $p$

- **If $X \sim$ Geometric$(p)$, then range$(X) = \mathbb{N}_{>0}$ and $\mathbf{p}_X(i) = (1 - p)^{i-1}p$ and $\neg\mathbf{F}_X(i) = (1 - p)^i$.** Interpretation: Number of independent Bernoulli trials (each with probability of success $p$) until the first success (hence the 1 indexing as at least one trial must be required)

- **If $X \sim$ Poisson$(\lambda)$, then range$(X) = \mathbb{N}$ and $\mathbf{p}_X(i) = e^{-\lambda}\frac{\lambda^i}{i!}$.** Interpretation: Number of events per unit time where the average number of events per unit time is $\lambda$

## 3.4 Multiple Variables

- **Joint probability mass function of random variables $X$ and $Y$ = $\mathbf{p}_{X,Y}(x, y) = \mathbf{P}(X = x \cap Y = y)$**

- Proposition: $\mathrm{p}_X(x) = \sum_y \mathrm{p}_{X,Y}(x, y)$
  Proof: Recall that $\{Y = y\}$ is a partition of $\Omega$.
  $\mathrm{p}_X(x) = P(X = x) = P(X = x \cap \cup_y Y = y) = P(\cup_y(X = x \cap Y = y)) = \sum_y P(X = x \cap Y = y)$ as the sets being unioned are mutually disjoint $= \sum_y \mathrm{p}_{X,Y}(x, y)$ by definition

- $X \perp\!\!\!\perp Y$ **iff** $\forall x, y. p_{X,Y}(x, y) = p_X(x)p_Y(y)$

## 3.5 Moments

- $k^{\text{th}}$ **moment of** $X = E[X^k] = \sum_x (x^k \mathbf{P}(X = x))$
- **Expectation of** $X = E[X] = 1^{\text{st}}$ **moment of** $X$
- If $X \sim$ Bernoulli$(p)$, then $E[X] = 1p + 0(1 - p) = p$
- Proposition: If $X \sim$ Poisson$(\lambda)$, then $E[X] = \lambda$
  Proof: Recall that the Maclaurin series for $e^x$ is $\sum_{i=0}^{\infty} \frac{x^i}{i!}$.
  $E[X] = \sum_{i=0}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = 0 + e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} = e^{-\lambda}\lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = e^{-\lambda}\lambda \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$.
- By the definition of expectation, $E[g(X)] = \sum_x (g(x)\mathrm{P}(X = x))$ and so in general $E[g(X)] \neq g(E[X])$

- **Theorem: If $X \perp\!\!\!\perp$Y, then** $E[g(X)f(Y)] = E[g(X)]E[f(Y)]$. Note that here $g(X)f(Y)$ denotes a random variable of the arithmetic product of $g(X)$ and $f(Y)$ not the Cartesian product of their distributions whereas it will be the Cartesian product in $P_{g(X)f(Y)}$.
  Proof:
  Lemma: $X \perp\!\!\!\perp Y \Rightarrow g(X) \perp\!\!\!\perp f(Y) \Rightarrow \forall x, y. P_{g(X)f(Y)}(x, y) = P_{g(X)}(x)P_{f(Y)}(y)$
  Proof: Exercise
  $E[g(X)f(Y)] = \sum_{x \in g(X)} \sum_{y \in f(Y)} (g(X)f(Y)\mathrm{P}_{g(X)f(Y)}(x, y)) =$
  $\sum_{x \in g(X)} \sum_{y \in f(Y)} (g(X)f(Y)\mathrm{P}_{g(X)}(x)P_{f(Y)}(y)) =$
  $\sum_{x \in g(X)} (g(x)\mathrm{P}_{g(X)}(x)E[f(Y)]) = E[g(X)]E[f(Y)])$.
  Corollary: $E[XY] = E[X]E[Y]$

- **Theorem (Linearity of expectation):** $E[g(X) + f(Y)] = E[g(X)] + E[f(Y)]$ irrespective of whether $X \perp\!\!\!\perp$Y.
  Proof: $E[g(X)+f(Y)] = \sum_{x \in g(X)} \sum_{y \in f(Y)} ((g(X)+f(Y))\mathrm{P}_{g(X)f(Y)}(x, y)) =$
  $\sum_{x \in g(X)} \sum_{y \in f(Y)} (g(X)\mathrm{P}_{g(X)f(Y)}(x, y)) +$
  $\sum_{y \in f(Y)} \sum_{x \in g(X)} (f(Y)\mathrm{P}_{g(X)f(Y)}(x, y)) =$

$\sum_{x \in X}(g(X)\mathrm{P}_X(x)) + \sum_{y \in Y}(f(Y)\mathrm{P}_Y(y)) = E[g(X)] + E[f(Y)]$
Corollary: $E[X + Y] = E[X] + E[Y]$

- **Proposition: If $X \sim$ Binomial$(n, p)$, then $E[X] = np$**
  Proof: Consider a family of i.i.d. random variables $Y_i \sim \text{Bernoulli}(p)\ \forall i \in [n]$. Deduce that then $X = \sum_{i \in [n]} Y_i$. By linearity of expectation, $E[X] = \sum_{i \in [n]} E[Y_i] = np$.

- **Theorem: Let $F_i$s partition $\Omega$. Then, $E[X] = \sum_{i \in [n]} E[X|F_i]\mathbf{P}(F_i)$**
  Proof: $E[X] = \sum_{x \in X} x\mathrm{P}_X(x) = \sum_{x \in X} x(\sum_{i \in [n]} \mathrm{P}_{X|F_i}(x)\mathrm{P}(F_i)) = \sum_{i \in [n]} \sum_{x \in X} x\mathrm{P}_{X|F_i}(x)\mathrm{P}(F_i) = \sum_{i \in [n]} \mathrm{P}(F_i) \sum_{x \in X} x\mathrm{P}_{X|F_i}(x) = \sum_{i \in [n]} \mathrm{P}(F_i)E[X|F_i]$
  Corollary: $E[X] = E[X|A]\mathrm{P}(A) + E[X|\neg A]\mathrm{P}(\neg A)$

- **Proposition: If $X \sim$ Geometric$(p)$, then $E[X] = \frac{1}{p}$**
  Proof: Consider $A =$ first trial is a failure.
  Then, $E[X] = E[X|A]P(A) + E[X|\neg A]P(\neg A)$. If $\neg A$ has occurred, it has taken exactly one trial to get a success; thus $E[X|\neg A] = 1$. As trials

are independent the expected number of remaining trials after a failure is $E[X]$; thus, $E[X|A] = E[X] + 1$ as we have already had 1 trial.
Thus, $E[X] = (E[X]+1)(1-p) + (1)(p) = (1-p)E[X] + 1 - p + p = E[X] - pE[X] + 1$. Thus, $pE[X] = 1$ as required.
Alternative proof: $E[X] = \sum_{i=1}^{\infty} i(1-p)^{i-1}p = p\sum_{i=1}^{\infty} i(1-p)^{i-1} = p\sum_{i=1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}p}((1-p)^i) = p\frac{\mathrm{d}}{\mathrm{d}p}(\sum_{i=1}^{\infty}(1-p)^i) = p\frac{\mathrm{d}}{\mathrm{d}p}(\frac{1}{1-(1-p)}) = p\frac{\mathrm{d}}{\mathrm{d}p}(\frac{1}{p}) = p\frac{1}{p^2} = \frac{1}{p}$.

- Proposition: If $X \sim$ Geometric($p$), then $E[X^2] = \frac{2-p}{p^2}$
  Proof: Consider $A =$ first trial is a failure.
  Then, $E[X^2] = E[X^2|A]P(A) + E[X^2|\neg A]P(\neg A) =$
  $E[(X+1)^2](1-p) + 1^2 p = (1-p)[E[X^2] + 2E[X] + 1] + p =$
  $(1-p)[E[X^2] + \frac{2+p}{p}] + p$
  $pE[X^2] = \frac{2+p}{p} - \frac{p(2+p)}{p} + \frac{p(p)}{p} = \frac{2+p-2p-p^2+p^2}{p} = \frac{2-p}{p}$ as required

- $Var(X) =$ **variance of** $X = E[(X - E[X])^2]$

- **Proposition:** $Var(X) = E[X^2] - E[X]^2$
  Proof: $Var(X) = E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2] = E[X^2 + (-2E[X])X + (E[X]^2)] = E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - 2E[X]^2 + E[X]^2 = E[X^2] - E[X]^2$

- **Proposition: If** $X \sim$ **Geometric**$(p)$**, then** $Var(X) = \frac{1-p}{p^2}$
  Proof: $Var(X) = E[X^2] - E[X]^2 = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$

- **Theorem: $\underline{\textbf{If } X \perp\!\!\!\perp Y\textbf{,}}$ then** $Var(X+Y) = Var(X) + Var(Y)$
  Proof: $Var(X+Y) = (E[(X+Y)^2]) - (E[X+Y]^2) = (E[X^2] + 2E[XY] + E[Y^2]) - (E[X]^2 + 2E[X]E[Y] + E[Y]^2)$.
  As $X \perp\!\!\!\perp Y$, E[XY] = E[X]E[Y].
  Thus, $Var(X+Y) = E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 = Var(X) + Var(Y)$ as required
  Corollary: If $k$ is a constant, then $Var(X+k) = Var(X)$

- **Proposition: If $k$ is a constant, then $Var(kX) = k^2Var(X)$**
  $Var(kX) = E[k^2X^2] - E[kX]^2 = k^2E[X^2] - k^2E[X]^2 = k^2Var(X)$
- **Proposition: If $X \sim$ Binomial$(n, p)$, then $Var(X) = np(1-p)$**
  Proof: $X = \sum_{i\in[n]} X_i$ where $X_i \sim$ Bernoulli$(p)$.
  $Var(X_i) = E[(X_i)^2] - E[X_i]^2 = p - p^2 = p(1-p)$.
  As $X_i$s are independent, $Var(X) = \sum_{i\in[n]} Var(X_i) = np(1-p)$

## 3.6 Tails

# 4 Continuous Random Variables

## 4.1 Single Variables

- $\mathbf{p}_X(i)$ **= probability density function (pdf) of** $X \neq \mathrm{P}(X = i)$**)**
- **If** $X \sim$ **Exp**$(\lambda)$**, then range**$(X) = \mathbb{N}$ **and** $\mathbf{p}_X(i) = \lambda e^{-\lambda i}$**.** Interpretation: Waiting interval between events (distributed Poisson($\lambda$)) per unit time

## 4.2 Multiple Variables

## 4.3 Poisson Processes

# 5 (Discrete) Markov Chains

# 6 Estimators

## 6.1 Evaluation

## 6.2 Maximum Likelihood Estimation (MLE)