CO245 - Probability and Statistics

15th January 2020

Probability is a mathematical formalism used to describe and quantify uncertainty.

Sample Spaces and Events

- **sample space** S or Ω

  a set containing the possible outcomes of a random experiment

  for example; sample space of two coin tosses $S = \{(H, H), (H, T), (T, H), (T, T)\}$
- **event** E $(E \subseteq S)$

  any subset of the sample space (collection of some possible events)

  for example; event of the first coin being heads in two tosses $E = \{(H, H), (H, T)\}$

  the extremes are $\varnothing$ (the null event) which will never occur, or S (the universal event) which will always occur - there is only uncertainty when the events are strictly between the events, such that $\varnothing \subset E \subset S$
- **elementary event** singleton subset containing exactly one element from S

When performing a random experiment, the outcome will be a single element $s^* \in S$. Then an event $E \subseteq S$ has **occurred** iff $s^* \in E$. If it has not occurred, then $s^* \notin E \Leftrightarrow s^* \in \bar{E}$ (can be read as not E).

With a set of events $\{E_1, E_2, \dots\}$, we can have the following set operations;

- $\bigcup\limits_i E_i = \{s \in S \mid \exists i.\ [s \in E_i]\}$ will only occur if at least one of the events E_i occurs ("or")
- $\bigcap\limits_i E_i = \{s \in S \mid \forall i.\ [s \in E_i]\}$ will only occur if all of the events E_i occurs ("and")
- $\forall i, j.\ E_i \cap E_j = \varnothing$ $(i \neq j)$ if they are mutually exclusive (at most one can occur)

σ-algebra

In an uncountably infinite set, the event set you are assigning probabilities to cannot be every subset, as the probabilities cannot be made to sum to 1 under reasonable axioms.

We define the σ-algebra as the subset of events which we can assign probabilities to. We want to define a probability function P that corresponds to the subsets of S that we wish to **measure**. This set of subsets is referred to as $\mathfrak{S}$ (the event space), with the following three properties (corresponding to the axioms of probability);

- nonempty $S \in \mathfrak{S}$
- closed under complements $E \in \mathfrak{S} \Rightarrow \bar{E} \in \mathfrak{S}$
- closed under countable union (therefore any countable set is fine) $E_1, E_2, \dots \in \mathfrak{S} \Rightarrow \bigcup\limits_i E_i \in \mathfrak{S}$

A probability measure on the pair $(S, \mathfrak{S})$ is a mapping $P : \mathfrak{S} \to [0, 1]$, satisfying the following three axioms;

- $\forall E \in \mathfrak{S}.\ [0 \leq P(E) \leq 1]$
- $P(S) = 1$
- countably additive, for **disjoint subsets** $E_1, E_2, \dots \in \mathfrak{S}$ $P\left(\bigcup\limits_i E_i\right) = \sum\limits_i P(E_i)$

From these, we can derive the following;

- $P(\bar{E}) = 1 - P(E)$

$$\underbrace{P(E) + P(E)}_{\text{disjoint}} = P(\underbrace{E \cup \bar{E}}_{E \cup \bar{E} = S}) = P(S) = 1$$

- $P(\varnothing) = 0$ special case of the above, when $E = S$
- for any events E and F $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

16th January 2020

Independent Events

It's important to note that independent events are **not** the same as disjoint events. Two events E and F are independent iff $P(E \cap F) = P(E)P(F)$ - sometimes written as $E \perp F$. Generally, a set of events $\{E_1, E_2, \dots\}$ are set to be independent if for any finite subset $\{E_{i_1}, E_{i_2}, \dots, E_{i_n}\}$;

$$P\left(\bigcap_{j=1}^{n} E_{i_j}\right) = \prod_{j=1}^{n} P(E_{i_j})$$

Where we have $\{i_j \mid 1 \leq j \leq n\}$ is any set of distinct positive integers. Note that independence is more than just pairwise independence.

We propose that if events E and F are independent, then $\bar{E}$ and F are also independent. Note that E and $\bar{E}$ form a partition of S (they are disjoint, and union to S). $F = (E \cap F) \cup (\bar{E} \cap F)$ is a disjoint union (and also a partition of F), this gives us $P(F) = P(E \cap F) + P(\bar{E} \cap F) \Rightarrow P(\bar{E} \cap F) = P(F) - P(E \cap F)$;

$$\begin{aligned}
P(\bar{E} \cap F) &= P(F) - P(E \cap F) && E \text{ and } F \text{ are independent}, \Rightarrow \\
&= P(F) - P(E)P(F) && \Rightarrow \\
&= (1 - P(E))P(F) && \text{probability of complement}, \Rightarrow \\
&= P(\bar{E})P(F) && \text{hence independent}, \blacksquare
\end{aligned}$$

Interpretations of Probability

In order to assign meaning to P, we need to have some interpretation of probability, such as the following;

- **classical**

  If S is finite, and the elementary events are "equally likely", then for an event $E \subseteq S$, the probability is the number of outcomes in E out of the total number of possible outcomes (S);

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

  This idea of "equally likely" (uniform) can be extended to infinite spaces. Instead of taking the set cardinality, another standard measure (such as area or volume) can be used instead.

- **frequentist**

  The idea is that if someone were to perform the same experiment (E may or may not occur) in identical random situations many times, then the proportion of times E occurs will tend to some limiting value, which would be $P(E)$.

- **subjective**

  Not assessed. Probability is the degree of belief held by an individual (see *De Finetti*) - suppose a random event $E \subseteq S$ is to be performed, and an individual enters a game regarding this experiment, with two choices;

- gamble — if E occurs they win \$1, otherwise if $\bar{E}$ occurs they win \$0
- stick — regardless of the outcome, the individual receives $\$P(E)$

The critical value of $P(E)$, where the individual is indifferent between the choices, is their probability of E.

Dependent Probabilities and Conditional Probability

For the standard example of flipping a coin and rolling a die (assuming both fair), we have independence - the probability of each elementary event is $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$.

However, consider the case where we have two die, where the first is fair, and the second is a "top", where we only have odd numbers (such that a roll of a 2 is mapped to a 5, 4 to 3, and 6 to 1). When we now flip the coin, if it is heads, we use the normal die, otherwise if it is tails, we use the "top". As expected, this is no longer independent.

For two events E and F in S, where $P(F) \neq 0$, we can define the probability of E occurring, given that we know F has occurred to be;

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Note that this also holds for independence ($P(E)$ doesn't change, as expected);

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)P(F)}{P(F)} = P(E)$$

An example of this is as follows - suppose we roll two normal dice, with one from each hand. The sample space is all the ordered pairs of possible values $S = \{(1,1),(1,2),\dots,(6,6)\}$. Let the event E be defined as the die from the left hand has a higher value than the die from the right hand. Looking at all possible combinations, we have;

$$P(E) = \frac{15}{36}$$

Suppose we now know F, the value of the left die being 5, has occurred. Since we know F has occurred, the only events that could have happened are $F = \{(5,1),(5,2),\dots,(5,6)\}$. Similarly, the only sample space elements in E that could've occurred are $E \cap F = \{(5,1),(5,2),(5,3),(5,4)\}$. Our probability is as follows;

$$\frac{|E \cap F|}{|F|} = \frac{4}{6} = \frac{\frac{4}{36}}{\frac{1}{6}} = \frac{P(E \cap F)}{P(F)} \equiv P(E|F)$$

One way to think about probability conditioning as a shrinking of the sample space, with events being replaced by intersections with the reduced space, and a rescaling of the probabilities. For example, with $F = S$, we have the following;

$$P(E) = \frac{P(E)}{1} = \frac{P(E \cap S)}{P(S)} = P(E|S)$$

Furthermore, we can extend the idea of independence of events with respect to a probability measure P to conditional probabilities. $P(\cdot|F)$ is a valid probability measure which obeys the axioms of probability on the set F. For three events E_1, E_2, F, the event pair E_1 and E_2 are conditionally independent given F (sometimes written as $E_1 \perp E_2 | F$) if and only if;

$$P(E_1 \cap E_2 | F) = P(E_1|F)P(E_2|F)$$

Bayes Theorem

For two events E and F in S, we have $P(E \cap F) = P(F)P(E|F)$, and $P(E \cap F) = P(E)P(F|E)$ (interchanging, and noting commutativity of $\cap$). Hence we have Bayes Theorem;

$$P(E|F) = \frac{P(E)P(F|E)}{P(F)}$$

Partition Rule

Consider a set of events $\{F_1, F_2, \dots\}$, which form a partition of S (they are disjoint, and union together to form S). Then for any event $E \subseteq S$, the partition rule states;

$$P(E) = \sum_i P(E|F_i)P(F_i)$$

The proof is as follows;

$$\begin{aligned}
E &= E \cap S \\
&= E \cap \bigcup_i F_i && \text{by definition of partitions} \\
&= \bigcup_i (E \cap F_i) && \text{by distributivity of intersection} \\
P(E) &= P\left(\bigcup_i (E \cap F_i)\right) \\
&= \sum_i P(E \cap F_i) && \text{disjoint union} \\
&= \sum_i P(E|F_i)P(F_i)
\end{aligned}$$

Note that $\{E \cap F_1, E \cap F_2, \dots\}$ is disjoint if $\{F_1, F_2, \dots\}$ is. Assume there is an element $s \in E \cap F_i$ and $s \in E \cap F_j$ (where $i \neq j$), if it is in both, then $s \in F_i$ and $s \in F_j$, which is not possible.

Note that $\{F, \bar{F}\}$ forms a partition of S, therefore by the Law of Total Probability we have;

$$P(E) = P(E \cap F) + P(E \cap \bar{F}) = P(E|F)P(F) + P(E|\bar{F})P(\bar{F})$$

Terminology

- conditional probabilities $P(E|F)$
- joint probabilities $P(E \cap F)$
- marginal probabilities (margins of a table) $P(E)$

  margins of a table

Likelihood and Posterior Probability

Suppose we have a probability model with parameters θ, that define a model instance (such as μ and σ), and a set of observations (or evidence) X.

- **likelihood function** (probability of the evidence, given the parameters) $P(X|\theta)$

  what is the probability our model will predict that evidence?
- **posterior probability** (probability of the parameters, given the evidence) $P(\theta|X)$

  what is the probability the actual parameters are θ, given our evidence?
- **prior probability** (not taking into account the evidence) $P(\theta)$

This is related by Bayes theorem;

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$\text{posterior probability} \propto \text{likelihood} \times \text{prior probability}$$

This is then divided by the normalising constant;

$$\sum_\theta P(X|\theta)P(\theta) = P(X)$$

22nd January 2020

Example Questions

1. There are 5000 VLSI chips, 1000 from company X (which has a 10% chance of being defective), and 4000 from company Y (which has a 5% chance of being defective). If a chip is defective, what is the probability it came from company X?

   Let E be the event that the randomly selected chip was made by X, and F be the event that the chip is defective.

$$
\begin{aligned}
P(E) &= \frac{1000}{5000} \\
&= 0.2 \\
P(\bar{E}) &= \frac{4000}{5000} \\
&= 0.8 \\
P(F|E) &= 0.1 && \text{given} \\
P(F|\bar{E}) &= 0.05 && \text{given} \\
P(E \cap F) &= P(F|E)P(E) \\
&= 0.02 \\
P(\bar{E} \cap F) &= P(F|\bar{E})P(\bar{E}) \\
&= 0.04
\end{aligned}
$$

   This gives us enough to fill in the table, as well as the missing entries with the law of total probabilities;

	E	$\bar{E}$	
F	0.02	0.04	0.06
$\bar{F}$	0.18	0.76	0.94
	0.2	0.8	

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{0.02}{0.06} = \frac{1}{3}$$

2. A multiple choice question has c available choices. Let p be the probability the student knows the right answer. When he doesn't know, he chooses an answer at random. Given that the answer the student chooses is correct, what is the probability that the student knew the correct answer?

   Let A be the event that the question is answered correctly, and K be the event that the student knew the correct answer. We therefore want to find $P(K|A)$.

$$P(K|A) = \frac{P(A|K)P(K)}{P(A)}$$

   We know $P(A|K) = 1$ (given that they don't purposely choose a wrong answer), and $P(K) = p$. By the partition rule, we have $P(A) = P(A|K)P(K) + P(A|\bar{K})P(\bar{K})$. Substituting values we get;

$$P(A) = 1 \cdot p + \frac{1}{c} \cdot (1 - p) = p + \frac{1-p}{c}$$

   Therefore,

$$P(K|A) = \frac{p}{p + \frac{1-p}{c}} = \frac{cp}{cp + 1 - p}$$

3. A new HIV test is claimed to correctly identify 95% of people who are really HIV positive and 98% of people who are really HIV negative.

(a) If only 1 in a 1000 of the population are HIV positive, what is the probability that someone who tests positive actually has HIV?

Let H be the event that someone has the virus ($P(H) = 0.001$), and T be the event that someone tests positive. Similar to above, we want to find the following, and can use the partition rule again.

$$P(H|T) = \frac{P(T|H)P(H)}{P(T)} = \frac{P(T|H)P(H)}{P(T|H)P(H) + P(T|\bar{H})P(\bar{H})} \approx 0.045$$

Therefore, less than 5% of those who test positive really have HIV.

(b) Is this acceptable? no

(c) Would a repeat test be appropriate for someone who tests positive?

Let T_i denote the event that the i^{th} test is positive. Suppose that the correctness of the test stays the same, and the test results are conditionally independent.

$$\begin{aligned}
P(H|T_1 \cap T_2) &= \frac{P(T_1 \cap T_2|H)P(H)}{P(T_1 \cap T_2)} \\
&= \frac{P(T_1 \cap T_2|H)P(H)}{P(T_1 \cap T_2|H)P(H) + P(T_1 \cap T_2|\bar{H})P(\bar{H})} \\
&= \frac{P(T_1|H)P(T_2|H)P(H)}{P(T_1|H)P(T_2|H)P(H) + P(T_1|\bar{H})P(T_2|\bar{H})P(\bar{H})} \\
&\approx 0.693
\end{aligned}$$

23rd January 2020

Simple Random Variables

Suppose we have identified a sample space S and a probability measure $P(E)$ on (measurable subsets) $E \subseteq S$. A random variable is a mapping from the sample space to the real numbers, such that a random variable $X : S \to \mathbb{R}$. Each element $s \in S$ is assigned a numerical value $X(s)$ (not always unique). We denote the outcome of the random experiment as s^*, the corresponding unknown outcome of the random variable $X(s^*)$ will be referred to as X.

The probability measure P defined on S induces a **probability distribution function**, P_X, on the random variable $X \in \mathbb{R}$. For each $x \in \mathbb{R}$, let $S_x \subseteq S$ be the set containing the elements of S which are mapped by X to numbers no greater than x, precisely $S_x = X^{-1}((-\infty, x])$.

$$P_X(X \leq x) = P(S_x)$$

We define the image of S under X as the range of the random variable X;

$$\text{range}(X) \equiv X(S) = \{x \in \mathbb{R} \mid \exists s \in S.\ [X(s) = x]\}$$

Consider this applied to the experiment of a fair coin toss, with $S = \{H, T\}$, probability measure $P(\{H\}) = P(\{T\}) = \frac{1}{2}$, and a random variable $X : \{H, T\} \to \mathbb{R}$ (such that $X(T) = 0$ and $X(H) = 1$);

$$X^{-1}((-\infty, x]) = \begin{cases} \varnothing & x < 0 \\ \{T\} & 0 < x < 1 \\ \{H, T\} & x \geq 1 \end{cases}$$

$$\begin{aligned}
P_X(X \leq x) &= \begin{cases} P(\varnothing) & x < 0 \\ P(\{T\}) & 0 < x < 1 \\ P(\{H, T\}) & x \geq 1 \end{cases} \\
&= \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}
\end{aligned}$$

The **cumulative distribution function** of a random variable X, $F_X(x)$ is the probability that X takes a value less than or equal to x;

$$F_X(x) = P_X(X \leq x)$$

To verify a function $F_X(x)$ is a valid cdf, we need to verify the following properties;

- $0 \leq F_X(x) \leq 1$, $\forall x \in \mathbb{R}$
- $\forall x_1, x_2 \in \mathbb{R}.\ [x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)]$ monotonicity
- $F_X(-\infty) = 0$, and $F_X(\infty) = 1$

Note that for finite intervals $(a, b] \subseteq \mathbb{R}$; $P_X(a < X \leq b) = F_X(b) - F_X(a)$. Unless there is ambiguity, we can generally omit the subscript of P_X, to just write P - thus we just consider the random variable from the start, letting the range of X be the sample space.

We define a random variable as simple if it can only take a finite number of possible values. Suppose X is simple, and can take m values $\mathcal{X} = \{x_1, x_2, \dots, x_m\}$, ordered $x_1 < x_2 < \dots < x_m$. Each $s \in S$ is mapped to one of these values by X. The sample space S can then be partitioned into m disjoint subsets, $\{E_1, E_2, \dots, E_m\}$, such that $s \in E_i \Leftrightarrow X(s) = x_i$. Therefore we have $P_X(X = x_i) = P(E_i)$, and $P_X(X = x_i) = F_X(x_i) - F_X(x_{i-1})$, with $x_0 = -\infty$.

A random variable is simply a numeric relabelling of our underlying sample space.

Discrete Random Variables

A random variable is discrete if it can take only a **countable** number of possible values (the range is countable). Therefore a simple random variable is a special case of a discrete random variable. Similar to above, we can partition S into a countable collection of disjoint subsets. For a discrete random variable X, F_X is a monotonic increasing step function with jumps at points in $\mathcal{X} = \{x_1, x_2, \dots\}$, where $x_1 < x_2 < \dots$, continuous on the right.

For a discrete random variable X and $x \in \mathbb{R}$, we define the **probability mass function**, $p_X(x)$ or just $p(x)$ as;

$$p_X(x) = P_X(X = x)$$

Given that X can take the values $\mathcal{X} = \{x_1, x_2, \dots\}$, then the following must hold;

- $0 \leq p_X(x) \leq 1$, $\forall x \in \mathbb{R}$
- $\sum\limits_{x \in \mathcal{X}} p_X(x) = 1$

Either the probability mass function (pmf) or the cumulative distribution function (cdf) of a random variable fully characterises its distribution, as we can work one out from the other;

- $p(x_i) = F(x_i) - F(x_{i-1})$
- $F(X_i) = \sum\limits_{j=1}^{i} p(x_j)$

Link to Statistics

Consider the set of data $(x_1, x_2, \dots, x_n)$ as n realisations of a random variable X. The frequency counts in the histogram for that set of data can be seen as an estimate for the probability mass function. Similarly, a cumulative histogram is an estimate of the cumulative distribution function.

Expectation

We define the **expectation** (also written as $E(X)$ or μ_X) of a discrete random variable X as

$$E_X(X) = \sum_x x p_X(x)$$

This gives a weighted average of the possible values, with the weights coming from the probability of a particular outcome. Occasionally referred to as the mean of the distribution.

The expectation of a function of a random variable is denoted $E\{g(X)\}$, where $g : \mathbb{R} \to \mathbb{R}$. We notice that $g(X)(s) = (g \circ X)(s)$ is also a random variable, therefore the expectation is;

$$E_X\{g(X)\} = \sum_x g(x) p_X(x)$$

Note that for a linear function $g(X) = aX + b$, where $a, b \in \mathbb{R}$, we have $E_X(aX + b) = aE_X(X) + b$. Similarly, for two linear functions g, h, $E_X(g(x)+h(x)) = E_X(g(x))+E_X(h(x))$. Therefore expectation is a linear operator.

The variance is the expectation of X, with $g(X) = (X - E(X))^2$. This is denoted $\text{Var}_X(X)$, or sometimes σ_X^2.

$$\text{Var}_X(X) = E_X((X - E_X(X))^2) = E(X^2) - E(X)^2$$

The variance of a linear function of a random variable is as follows;

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

The standard deviation of a random variable, $\text{sd}_X(X)$ (also σ_X) is the square root of the variance.

$$\text{sd}_X(X) = \sqrt{\text{Var}_X(X)}$$

The skewness γ_1 of a discrete random variable X is defined;

$$\gamma_1 = \frac{E_X((X - E_X(X))^3)}{\text{sd}_X(X)^3} = \frac{E_X((X - \mu)^3)}{\sigma^3}$$

The part in violet is when $\mu = E(X), \sigma = \text{sd}(X)$.

Example Questions

1. If X is a random variable taking the integer value scored with a single roll of a fair die, what is

   (a) the expected value

$$\begin{aligned} E(X) &= \sum_{x=1}^{6} x p(x) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{21}{6} \qquad (= 3.5) \end{aligned}$$

   (b) the variance

$$\begin{aligned} \text{Var}(X) &= \sum_{x=1}^{6} x^2 p(x) - 3.5^2 \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} - \left(\frac{7}{2}\right)^2 \\ &= \frac{35}{12} \end{aligned}$$

2. A student gets X marks answering a single multiple choice question with four options, where 3 marks are awarded for a correct answer, and -1 for a wrong answer - what is

   (a) the expected value

$$\begin{aligned} E(X) &= 3 \cdot P(\text{correct}) + -1 \cdot P(\text{incorrect}) \\ &= 3 \cdot \frac{1}{4} + -1 \cdot \frac{3}{4} \\ &= 0 \end{aligned}$$

   (b) the standard deviation

$$\begin{aligned} E(X^2) &= 3^2 \cdot P(\text{correct}) + (-1)^2 \cdot P(\text{incorrect}) \\ &= 9 \cdot \frac{1}{4} + 1 \cdot \frac{3}{4} \\ &= 4 & \Rightarrow \\ \text{sd}(X) &= \sqrt{3 - 0^2} \\ &= \sqrt{3} \end{aligned}$$

29th January 2020

Probability Generating Function

The **probability generating function**, $G_X(z)$ or just $G(z)$, is defined as;

$$G_X(z) = E_X(z^X) = \sum_x p_X(x) z^x$$

Moments of a random variable X, M_n and M_n^f, are defined as follows;

$$\begin{aligned} M_n &= E(X^n) & n^{\text{th}} \text{ moment} \\ M_n^f &= E(X(X-1)\dots(X-n+1)) & n^{\text{th}} \text{ factorial moment} \end{aligned}$$

It's also important to note that the first moment, M_1 is the mean, and the second moment $M_2 = \text{Var}(X) + E(X)^2$. Generally, we can use the factorial moments, as they will also contain the polynomial term - but can be obtained from taking derivatives of G;

$$\begin{aligned} G^n(z) &= E(X(X-1)\dots(X-n+1)z^{X-n}) & \Rightarrow \\ M_n^f &= G^n(1) \\ M_0 &= M_0^f \\ &= G(1) \\ &= 1 \\ M_1 &= M_1^f \\ &= G'(1) \\ M_2 &= M_2^f + M_1^f \\ &= G''(1) + G'(1) \end{aligned}$$

This gives us the variance as follows;

$$\text{Var}(X) = M_2 - M_1^2 = G''(1) + G'(1) - G'(1)^2$$

Sum of Random Variables

If we let $X_1, X_2, \ldots, X_n$ be n random variables (not necessarily with the same distribution, nor necessarily independent).

$$S_n = \sum_{i=1}^{n} X_i \qquad \text{sum of those variables}$$

$$E(S_n) = \sum_{i=1}^{n} E(X_i)$$

$$E\left(\frac{S_n}{n}\right) = \frac{\sum\limits_{i=1}^{n} E(X_i)}{n} \qquad \text{expectation of the average}$$

if they are **independent**

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\sum\limits_{i=1}^{n} \mathrm{Var}(X_i)}{n^2}$$

if they are **independent** and **identically distributed**, with $E(X_i) = \mu_X$ and $\mathrm{Var}(X_i) = \sigma_X^2$

$$E\left(\frac{S_n}{n}\right) = \mu_X$$

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma_X^2}{n} \qquad \text{variance decreases with more samples}$$

Note that we can prove the independent variance result as follows, "inductively";

$$\begin{aligned}\mathrm{Var}(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\ &= \underbrace{E(X^2) - E(X)^2}_{\mathrm{Var}(X)} + \underbrace{E(Y^2) - E(Y)^2}_{\mathrm{Var}(Y)} + \underbrace{2E(XY) - 2E(X)E(Y)}_{=0 \text{ if independent}}\end{aligned}$$

We're able to obtain the probability generating function of a sum of **independent** random variables as;

$$G_{S_n}(z) = \prod_{i=1}^{n} G_{X_i}(z)$$

This is due to the following;

$$\begin{aligned}G_{S_n}(z) &= E(z^{\sum_{i=1}^{n} X_i}) \\ &= E\left(\prod_{i=1}^{n} z^{X_i}\right) \\ &= \prod_{i=1}^{n} E(z^{X_i}) && \text{only if independent} \\ &= \prod_{i=1}^{n} G_{X_i}(z)\end{aligned}$$

Additionally, if they are also **identically distributed**, $G_{S_n}(z) = G_{X_i}(z)^n$

Discrete Distributions

Some examples of commonly encountered distributions are as follows;

- **bernoulli** $X \sim \text{Bernoulli}(p)$

  Consider an experiment that has two possible outcomes; with a random variable X taking 1 with probability p, and 0 with probability $1 - p$. The probability mass function is (note that since there are only two cases, it can be written out per case);

  $$p(x) = p^x(1-p)^{1-x}, \text{ for } x = 0, 1$$

  The standard formulae for can then be used to obtain the following results;

  $$\begin{aligned}
  \mu &= E(X) \\
  &= 0 \cdot (1-p) + 1 \cdot p \\
  &= p \\
  \sigma^2 &= E(X^2) - E(X)^2 \\
  &= 0^2 \cdot (1-p) + 1^2 \cdot p - p^2 \\
  &= p(1-p) \\
  G(z) &= (1-p)z^0 + pz^1 \\
  &= 1 - p + pz \\
  &= 1 - p(1-z)
  \end{aligned}$$

- **binomial** $X \sim \text{Binomial}(n, p)$

  Consider n identical, independent Bernoulli(p) trials $X_1, \dots, X_n$. Let X be the total number of 1s observed in the n trials;

  $$X = \sum_{i=1}^{n} X_i$$

  Therefore, X is a random variable taking values in $\{0, 1, 2, \dots, n\}$. The probability mass function is (for $0 \leq x \leq n$);

  $$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

  The following values can be obtained with the standard formulae, or by taking the sums of random variables;

  $$\begin{aligned}
  \mu &= np \\
  \sigma^2 &= np(1-p) \\
  \gamma_1 &= \frac{1-2p}{\sqrt{np(1-p)}}
  \end{aligned}$$

30th January 2020

Derivation of Binomial PMF with PGF

$$\begin{aligned}
G_{\text{bern}}(z) &= (1-p) + pz \\
G_{\text{bin}}(z) &= (G_{\text{bern}}(z))^n && \text{sum of independent, identically distributed RVs} \\
&= ((1-p) + pz)^n \\
\text{coeff. of } z^x &= \binom{n}{x} p^x (1-p)^{n-x} && 0 \leq x \leq n
\end{aligned}$$

Distributions (Continued)

Continuing from the last lecture;

- **geometric** $X \sim \text{Geometric}(p)$

  Consider a potentially infinite sequence of independent Bernoulli(p) random variables $X_1, X_2, \dots$ We can define a quantity X as the index of the first Bernoulli trial to result in a 1.

  $$X = \min\{i \mid X_i = 1\}$$

  Therefore X is a random variable, with values $X \in \mathbb{Z}^+ = \{1, 2, \dots\}$. The probability mass function can be deduced intuitively; for the x^{th} trial to be the first resulting in a 1, we must've had $x - 1$ trials resulting in a 0, and the last one resulting in a 1, therefore, for $x \in \mathbb{Z}^+$;

  $$p(x) = p(1-p)^{x-1}$$

  The mean, variance, and skewness are;

  $$\mu = \frac{1}{p}$$
  $$\sigma^2 = \frac{1-p}{p^2}$$
  $$\gamma_1 = \frac{2-p}{\sqrt{1-p}} \qquad \text{note it is always positive}$$

- **poisson** (rate parameter λ) $X \sim \text{Poi}(\lambda)$

  The previous three distributions were concerned with the success or failure of a trial. Let X be a random variable on $\mathbb{N}$, with the probability mass function;

  $$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

  To verify this is a valid pmf, we need to check the following;

  - positive for all x yes, all components are positive (assuming $\lambda > 0$)
  - sums to 1 yes, $\sum\limits_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$ (Taylor expansion)

  These random variables are concerned with the number of random events per unit of time or space, when there is a constant underlying "rate" of events occurring across this unit. Some examples of this are the number of particles emitted by a radioactive substance in a given time, the number of minor car crashes per day, the number of jobs arriving at a server per hour, or the number of potholes in each mile of road.

  Note that it has equal mean and variance (fine as it is dimensionless), and that the skewness is always positive, but decreases as λ increases;

  $$\mu = \lambda$$
  $$\sigma^2 = \lambda$$
  $$\gamma_1 = \frac{1}{\sqrt{\lambda}}$$

The probability generating function can be derived as follows;

$$\begin{aligned}
G_{\text{Poi}(\lambda)}(z) &= \sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}z^x \\
&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda z)^x}{x!} \\
&= e^{-\lambda}e^{\lambda z} \\
&= e^{-\lambda(1-z)}
\end{aligned}$$

An example of fitting the distribution to data goes back to the idea of counting radioactive particles. For 2608 time intervals, each of length 7.5 seconds, the number of particles emitted was measured as follows (such that there were 57 time intervals with 0 emissions, 203 time intervals with 1 emission, and so on);

x	0	1	2	3	4	5	6	7	8	9	≥ 10
$O(n_x) = n_x$	57	203	383	525	532	408	273	139	45	27	16
$E(n_x)$	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1

Let the average number per interval to be the total number of particles divided by the total number of intervals;

$$\frac{\sum_x xn_x}{\sum_x n_x} = \frac{10094}{2608} \approx 3.87$$

Let us set the mean of the Poisson distribution to the observed (sample) mean, therefore $\lambda = 3.87$. We can then say our expectation of the number of 0 counts is $n \cdot p(0) \approx 54.4$. The expected values are added in violet in the table.

As the two sets of numbers are sufficiently close - it suggests the Poisson approximation is good.

- **discrete uniform** $X \sim U(\{1, 2, \dots, n\})$

  Let X be a random variable on $\{1, 2, \dots, n\}$, with the following probability mass function ($x = 1, 2, \dots, n$);

  $$p(x) = \frac{1}{n}$$

  We have the following values for the mean and variance (note that the skewness is clearly 0, as intuitively it looks at the tilt of the distribution, which doesn't really apply here);

  $$\begin{aligned}
  \mu &= \frac{n+1}{2} \\
  \sigma^2 &= \frac{n^2-1}{12} \\
  \gamma_1 &= 0
  \end{aligned}$$

Approximation of Binomial Distribution with Poisson Distribution

The Poisson distribution with rate parameter np, $\text{Poi}(np)$, can approximate $\text{Binomial}(n, p)$, when p is small ($p < 0.1$), and n is large, which is often the case. This is useful as the generating function for the Poisson distribution is easy to deal with, and tabulating a single $\text{Poi}(\lambda)$ encompasses all the possible corresponding $\text{Binomial}(n, \frac{\lambda}{n})$.

An example of this is as follows - suppose a manufacturer produces chips, with 1% being defective. Find the probability that in a box of 100 chips, none are defective. This can be a binomial distribution, Binomial$(100, 0.01)$ - however as n is large, and p is small, we can approximate this distribution with Poi(1).

$$p(0) \approx \frac{e^{-1}\lambda^0}{0!} \approx 0.3679$$

Note that the actual value, with the Binomial distribution, ≈ 0.366.

A sketch of this proof is as follows;

$$\begin{aligned}
&\text{let } p = \frac{\lambda}{n} && \text{and let } n \to \infty \text{, hence } p \to 0 \\
&G(z) = (1 - p(1-z))^n && \text{pgf of Binomial random variable} \\
&\quad = \left(1 - \frac{\lambda(1-z)}{n}\right)^n && \text{substituting our value for } p \\
&\quad \to e^{-\lambda(1-z)} && \text{as } n \to \infty
\end{aligned}$$

This proof shows that the pgf of a Binomial random variable approaches that of a Poisson random variable.

Continuous Random Variables

Recall that a random variable is defined as a mapping $X : S \to \mathbb{R}$, from the sample space S to the real numbers. This induces a probability measure $P_X(B) = P(\{X^{-1}(B)\})$, where $B \subseteq \mathbb{R}$. Note that the inverse set is the set of all items in the sample space that would result in elements of B. We define X to be continuous if there exists a function $f_X : \mathbb{R} \to \mathbb{R}$ (the **probability density function** of X);

$$P_X(B) = \int_{x \in B} f_X(x) \, \mathrm{d}x$$

Note the following consequences;

- the probability assigned to a singleton subset $B = \{x\}$, where $x \in \mathbb{R}$ is zero for a continuous random variable;

$$P_X(X = x) = P_X(\{x\}) = 0$$

- the probability of a countable set $B = \{x_1, x_2, \dots\} \subseteq \mathbb{R}$ will also have zero probability measure;

$$P_X(X \in B) = P_X(X = x_1) + P_X(X = x_2) + \cdots = 0 + 0 + \cdots = 0$$

- the range of a continuous random variable must therefore be uncountable (otherwise would not sum to 1)

Example Questions

1. Suppose that 10 users are authorised to use a particular computer system, but that the system crashes if 7 or more users attempt to be logged on simultaneously. Suppose that each user has the same probability $p = 0.2$ of wishing to log in during the each hour. What's the probability that the system will crash in a given hour?

   The probability of exactly x users wanting to log in can be represented as a binomial distribution $X \sim$ Binomial$(10, 0.2)$ (assuming independence). Therefore we want to find $P(7 \leq X \leq 10)$;

$$p(7) + p(8) + p(9) + p(10) = \binom{10}{7} 0.2^7 0.8^3 + \cdots + \binom{10}{10} 0.2^1 00.8^0 \approx 0.00086$$

   Using this, we can calculate the mean time to a crash as approximately 1163 hours, with a geometric distribution.

2. Suppose people have problems logging on to a particular website once every 5 attempts, on average.

   (a) Assuming that the attempts are independent, what is the probability that an individual will not succeed until the 4^{th}?

   Let $p = \frac{4}{5} = 0.8$. This is a geometric distribution, hence;

$$p(4) = (1-p)^3 p = 0.2^3 0.8 = 0.0064$$

   (b) On average, how many trials must one make until succeeding?

   The mean is $\mu = \frac{1}{p} = 1.25$.

   (c) What's the probability the first successful attempt is the 7^{th} or later?

   Due to the nature of a geometric series, it will eventually have a success. For it to be on attempt 7, or later, there must have already been 6 failures. Therefore $P(X \geq 7) = (1-p)^6 = 0.2^6 = 0.000064$.