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;

- $\bigcup_i E_i = \{s \in S \mid \exists i.\ [s \in E_i]\}$ will only occur if at least one of the events E_i occurs ("or")
- $\bigcap_i E_i = \{s \in S \mid \forall i.\ [s \in E_i]\}$ 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) $E_1, E_2, \dots \in \mathfrak{S} \Rightarrow \bigcup_i E_i \in \mathfrak{S}$

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** $E_1, E_2, \dots \in \mathfrak{S}$ $P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$

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 $S = \{(1,1),(1,2),\ldots,(6,6)\}$. 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 $F = \{(5,1),(5,2),\ldots,(5,6)\}$. 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 $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$, ordered $x_1 < x_2 < \cdots < x_m$. 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, $\{E_1, E_2, \ldots, E_m\}$, 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 $\mathcal{X} = \{x_1, x_2, \ldots\}$, where $x_1 < x_2 < \ldots$, 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 $\mathcal{X} = \{x_1, x_2, \ldots\}$, then the following must hold;

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

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;

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

Link to Statistics

Consider the set of data $(x_1, x_2, \ldots, x_n)$ 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 $\mathrm{Var}_X(X)$, or sometimes σ_X^2.

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

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

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

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

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

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

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

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

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

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

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

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

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;

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

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).

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

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

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

if they are **independent**

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

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

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;

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

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 $X_1, \ldots, X_n$. 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 $\{0, 1, 2, \ldots, n\}$. 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, $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$ (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;

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

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda z)^x}{x!}$$
$$= e^{-\lambda} e^{\lambda z}$$
$$= e^{-\lambda(1-z)}$$

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;

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

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** $X \sim U(\{1, 2, \ldots, n\})$

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

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

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);

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

Approximation of Binomial Distribution with Poisson Distribution

The Poisson distribution with rate parameter np, Poi(np), can approximate 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 Poi(λ) encompasses all the possible corresponding 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 $X \sim \text{Binomial}(10, 0.2)$ (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 $P(X \geq 7) = (1-p)^6 = 0.2^6 = 0.000064$.

5th February 2020

Continuous Cumulative Distribution Function

The **cumulative distribution function**, $F_X(x)$, where the continuous random variable takes a value less than or equal to x is as follows (for all real x);

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

Therefore, if we are given the cumulative distribution function of a random variable X, we can define the probability density function as;

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = F_X'(x)$$

The pdf cannot be negative, as it is the derivative of the cdf (which is always non-decreasing); therefore we have the following properties for a pdf $f_X(x)$;

- $f_X(x) \geq 0$, $\forall x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$ $\qquad = F_X(\infty)$

The probability that a continuous random variable X lies in an interval $(a, b]$ is as follows (note that the inclusive / exclusive bounds don't really matter, as the probability of it being exactly a or b is zero);

$$\begin{aligned} P_X(a < X \leq b) &= P_X(X \leq b) - P_X(X \leq a) \\ &= F_X(b) - F_X(a) \\ &= \int_a^b f_X(x)\,\mathrm{d}x \qquad \text{the area under the pdf curve between } a \text{ and } b \end{aligned}$$

The following observations can also be made;

- the cumulative distribution function F_X of a continuous random variable X is a non-decreasing function, and is also absolutely continuous
- since $P(X = x) = 0$, $F_X(x) = P(X \leq x) \equiv P(X < x)$
- small h, $f_X(x)h$ is approximately the probability that X takes a value in $[x, x+h)$
- we don't require $f_X(x) \leq 1$, since the probability distribution function $f_X(x)$ is not itself a probability, unlike the probability mass function of a discrete random variable
- the probability density function f_X, when it exists, completely characterises its distribution, so we often just specify f_X

Transformation of Random Variables

Given a continuous random variable X, let Y be the **transformed** random variable $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ is continuous and strictly monotonic (such that g^{-1} exists). Consider the two cases for g;

- Suppose g is monotonic **increasing**, then we have, for $y \in \mathbb{R}$;

$$Y \leq y \Leftrightarrow X \leq g^{-1}(y)$$

  Then we have the following result, for the cumulative distribution function of Y;

$$F_Y(y) = P_Y(Y \leq y) = P_X(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

  Via differentiation (with the chain rule), we have the following result for the probability density function f_Y;

$$f_Y(y) = F'_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) g^{-1'}(y)$$

  We can also be sure that the derivative of $g^{-1}(y)$, with respect to y, is positive as we assumed it was increasing.

- Suppose g is monotonic **decreasing**, then we have, for $y \in \mathbb{R}$;

$$Y \leq y \Leftrightarrow X \geq g^{-1}(y)$$

  Then we have the following result, for the cumulative distribution function of Y;

$$F_Y(y) = P_Y(Y \leq y) = P_X(X \geq g^{-1}(y)) = 1 - P_X(X \leq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

  Via differentiation (with the chain rule), we have the following result for the probability density function f_Y;

$$f_Y(y) = F'_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}(1 - F_X(g^{-1}(y))) = -f_X(g^{-1}(y)) g^{-1'}(y)$$

  We can also be sure that the derivative of $g^{-1}(y)$, with respect to y, is negative as we assumed it was decreasing.

In either case, when $Y = g(X)$, we have;

$$f_Y(y) = f_X(g^{-1}(y)) |g^{-1'}(y)|$$

Expectation and Variance of a Continuous Random Variable

Similar to before, we can define the mean, or expectation, as follows;

$$\mu_X = E_X(X) = \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x$$

And generally, for a function of interest $g : \mathbb{R} \to \mathbb{R}$;

$$E_X(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) \, \mathrm{d}x$$

Similarly, we can define the variance as follows;

$$\sigma_X^2 = \mathrm{Var}_X(X) = E((X - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) \, \mathrm{d}x$$

It's also possible to use the following form (which can be derived similar to the discrete case, but with linearity of integration);

$$\mathrm{Var}_X(X) = \int_{-\infty}^{\infty} x^2 f_X(x) \, \mathrm{d}x - \mu_X^2 = E(X^2) - E(X)^2$$

Again, for a linear transformation, $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$

Moment Generating Function and Characteristic Function

The moment generating function $M_X(t)$ for a continuous random variable X is;

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,\mathrm{d}x$$

As such, the n^{th} moment is;

$$M_n = \left.\frac{\mathrm{d}^n}{\mathrm{d}t^n} M_X(t)\right|_{t=0}$$

However, this integral may not exist, such as for $f_X(x) > e^{-tX}$ and $t > 0$.

We also define the **characteristic function** as the Fourier transform of the probability density function (which always exists);

$$\phi_X(t) = E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\,\mathrm{d}x$$

The n^{th} moment is then;

$$M_n = (-i)^n \left.\frac{\mathrm{d}^n}{\mathrm{d}t^n} \phi_X(t)\right|_{t=0}$$

Note that these are both instances of a probability generating function, where $M(t) = G(e^t)$ and $\phi(t) = G(e^{it})$.

Sums of Independent Random Variables

As the moment generating function and characteristic functions are both instances of a probability generating function, the same properties hold with respect to the sums of independent random variables;

$$S_n = \sum_{j=1}^{n} X_j$$

Therefore, if $X_1, X_2, \ldots, X_n$ are independent, then;

$$\phi_{S_n}(t) = \prod_{j=1}^{n} \phi_{X_j}(t) \text{ and } M_{S_n}(t) = \prod_{j=1}^{n} M_{X_j}(t)$$

This can be proven as follows, for the characteristic function;

$$\begin{aligned}
\phi_{S_n}(t) &= E(e^{it\sum_{j=1}^{n} X_j}) \\
&= E\left(\prod_{j=1}^{n} e^{itX_j}\right) \\
&= \prod_{j=1}^{n} E\left(e^{itX_j}\right) && \text{by independence} \\
&= \prod_{j=1}^{n} \phi_{X_j}(t)
\end{aligned}$$

Quantiles

We define the lower quartile as a point $\frac{1}{4}$ of the way through the ordered sample, and similarly $\frac{1}{2}$ for the median, $\frac{3}{4}$ for the upper quartile. Generally, for a random variable X, we define the α-quantile $Q_X(\alpha)$, for $0 \leq \alpha \leq 1$, as the least number satisfying;

$$P(X \leq Q_X(\alpha)) = \alpha = F_X(Q_X(\alpha)) \Rightarrow Q_X(\alpha) = F_X^{-1}(\alpha)$$

For example, the median of a random variable X is;

$$F_X^{-1}\left(\frac{1}{2}\right)$$

Example Questions

1. Suppose we have a continuous random variable X with probability density function (with some unknown constant c), given by;

$$f(x) = \begin{cases} cx^2 & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

   (a) determine c

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x &= \int_{-\infty}^{0} f(x)\,\mathrm{d}x + \int_{0}^{3} f(x)\,\mathrm{d}x + \int_{3}^{\infty} f(x)\,\mathrm{d}x \\ &= \left[\frac{1}{3}cx^3\right]_0^3 \\ &= 9c \\ &= 1 \qquad \text{property of pdf,} \Rightarrow \\ c &= \frac{1}{9} \end{aligned}$$

   (b) find the cumulative distribution function of X

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{27}x^3 & 0 \le x \le 3 \\ 1 & x > 3 \end{cases}$$

   (c) calculate $P(1 < X < 2)$

$$\begin{aligned} P(1 < X < 2) &= F(2) - F(1) \\ &= \frac{8}{27} - \frac{1}{27} \\ &= \frac{7}{27} \\ &\approx 0.2593 \end{aligned}$$

   (d) calculate $E(X)$

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x \\ &= \int_{0}^{3} \frac{1}{9}x^3\,\mathrm{d}x \qquad \text{bounds changed} \\ &= \left[\frac{1}{36}x^4\right]_0^3 \\ &= 2.25 \end{aligned}$$

   (e) calculate $\mathrm{Var}(X)$

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - 9 \end{aligned}$$

$$
\begin{aligned}
&= \int_0^3 \frac{1}{9}x^4 \,\mathrm{d}x - 9 \\
&= \left[\frac{1}{45}x^5\right]_0^3 - 9 \\
&= \frac{3^5}{45} - 2.25 \\
&= 0.3375
\end{aligned}
$$

(f) calculate the median of X

$$
\begin{aligned}
F(x) &= \frac{1}{2} && \Rightarrow \\
\frac{1}{27}x^3 &= \frac{1}{2} && \Rightarrow \\
x^3 &= \frac{27}{2} && \Rightarrow \\
x &= \frac{3}{2^{\frac{1}{3}}} \\
&\approx 2.3811
\end{aligned}
$$

6th February 2020

Continuous Distributions

Some encountered distributions are as follows;

- **uniform** $X \sim \mathrm{U}(a, b)$

  A continuous random variable X with a range (a, b) has a uniform distribution if its probability density function (or cumulative distribution function) is;

$$
f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}
$$

$$
F(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}
$$

  Note that $\mathrm{U}(0, 1)$ is considered to be the "standard" uniform distribution. Let $X \sim \mathrm{U}(0, 1)$, therefore $F_X(x) = x$, for $0 \leq x \leq 1$. To map this to a general interval (a, b), where $a < b \in \mathbb{R}$, we define a new random variable $Y = a + (b - a)X$, and observe the following;

$$
Y \leq y \Leftrightarrow a + (b-a)X \leq y \Leftrightarrow X \leq \frac{y-a}{b-a}
$$

  Therefore $Y \sim \mathrm{U}(a, b)$;

$$
F_Y(y) = P(Y \leq y) = P\left(X \leq \frac{y-a}{b-a}\right) = F_X\left(\frac{y-a}{b-a}\right) = \frac{y-a}{b-a}
$$

  With the standard methods, we can compute the mean and variance to be the following;

$$
\begin{aligned}
\mu &= \frac{a+b}{2} \\
\sigma^2 &= \frac{(a+b)^2}{12}
\end{aligned}
$$

- **(negative) exponential** $X \sim \text{Exp}(\lambda)$

  Suppose X is a random variable with range $\mathbb{R}^+ = [0, \infty)$, and a probability density function (or cumulative distribution function) of;

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$
$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

  Then X is a (negative) exponential random variable with rate parameter λ. Note that this is a **memoryless property** (see below).

  These are used to model time until occurrence of a random event when there is an assumed constant risk, or rate, of the event happening over time. Hence it is frequently used as a simplest model. Some examples of this include the time to failure of a component in a system, the time until the next request arrives at a webserver, or the distance along a road between potholes.

  Note these examples have a duality with the ones we saw for a Poisson distribution. However, instead of the number of events, we are now considering the time between events. Thus we claim if events in a random process occur according to a Poisson distribution with rate λ, then the time between consecutive events has an exponential distribution with parameter λ. The proof of this as follows;

  - suppose the number of events occurring in the "time" interval $[0, x]$ $(\forall x > 0)$, denoted as N_x, follows a Poisson distribution with rate parameter λ, such that $N_x \sim \text{Poi}(\lambda x)$ - this process is known as a **(homogeneous) Poisson process** (homogeneous because the rate stays the same)

    note that we multiply by x as there would be more events, intuitively, for a longer period
  - let X be the time until the first event of this process occurs - therefore if $X > x$ means that no event has occurred in the time interval

$$\begin{aligned} P(X > x) &\equiv P(N_x = 0) \\ &= \frac{(\lambda x)^0 e^{-\lambda x}}{0!} \\ &= e^{-\lambda x} \qquad \Rightarrow \\ P(X \leq x) &= 1 - e^{-\lambda x} \end{aligned}$$

  - therefore $X \sim \text{Exp}(\lambda)$

- **normal (Gaussian)** $X \sim N(\mu, \sigma^2)$

  A normal (or Gaussian) random variable with a range across all real numbers has the following probability density function (and corresponding cumulative distribution function);

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, \mathrm{d}t$$

  It can be shown that X has mean μ and variance σ^2. Note that when $\mu = 0$, and $\sigma = 1$, we have the standard normal distribution, $Z \sim N(0, 1)$, which has the following simplified probability density function, and cumulative distribution function;

$$f(z) \equiv \phi(z)$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$F(z) \equiv \Phi(z)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} \, \mathrm{d}t$$

As the cdf is not analytically available, numerical integration procedures are used to approximate probabilities to an arbitrary degree of accuracy. Statistical tables contain values of the standard normal cdf $\Phi(z)$, for a range of z values, and the quantiles $\Phi^{-1}(\alpha)$, for a range of values $\alpha \in (0, 1)$. Since it is only tabulated for $z \geq 0$ (due to the symmetry), we can do the following;

$$\Phi(z) = 1 - \Phi(-z)$$

$$P(Z > z) = 1 - \Phi(z)$$

Suppose $X \sim N(\mu, \sigma^2)$, then applying a linear transformation with constants $a, b \in \mathbb{R}$, $aX + b$ also has a normal distribution (follows from the general results for expectation and variance);

$$X \sim N(\mu, \sigma^2) \Rightarrow aX + b \sim N(a\mu + b, a^2\sigma^2)$$

This allows us to standardise any normal random variable;

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X - \mu}{\sigma} \sim N(0, 1)$$

This allows us to write the cdf of X in terms of Φ;

$$\text{let } Z = \frac{X - \mu}{\sigma} \text{ then } X \leq x \Leftrightarrow \frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \Leftrightarrow Z \leq \frac{x - \mu}{\sigma}$$

Therefore

$$F_X(x) = P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

See **central limit theorem** below, for more applications.

Memoryless Property

Note that the complementary cumulative distribution function (shortened to ccdf) is $P(X > x)$, for a random variable X.

Note that for an exponential distribution $X \sim \text{Exp}(\lambda)$, the cdf is $1 - e^{-\lambda x}$, and therefore the ccdf is $e^{-\lambda x}$.

Consider the probability that X is greater than $x + s$, given that we know it is already greater than s;

$$P(X > x + s | X > s) = \frac{P((X > x + s) \cap (X > s))}{P(X > s)} = \frac{P(X > x + s)}{P(X > s)}$$

The third equality can be established, since it is obvious that if it is greater than $x + s$, it must already be greater than s. Taking X as an exponential distribution, with parameters λ;

$$P(X > x + s | X > s) = \frac{e^{-\lambda x + s}}{e^{-\lambda s}} = e^{-\lambda x}$$

This is an exponential ccdf with parameter λ. If we think of this (exponential) random variable as the time to an event, the knowledge that we have already waited s for the event doesn't give us any information about how much longer we will have to wait - hence it has **no memory**.

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed (iid) random variables from any probability distribution, with mean μ and variance σ^2. We know the following;

$$E\left(\sum_{i=1}^{n} X_i\right) = n\mu$$
$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2$$

subtracting $n\mu$

$$E\left(\sum_{i=1}^{n} X_i - n\mu\right) = 0$$
$$\text{Var}\left(\sum_{i=1}^{n} X_i - n\mu\right) = n\sigma^2$$

dividing by $\sqrt{n}\sigma$

$$E\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}\right) = 0$$
$$\text{Var}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}\right) = 1$$

This gives the result

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$

Which can also be written as

$$\lim_{n\to\infty} \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \Phi \text{ where } \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The CLT implies for large n;

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ or } \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

The results are exact when $X_i \sim N(\mu, \sigma^2)$, the results are exact, as the sum of independent normally distributed random variables is also randomly distributed. A sketch proof of this is as follows - given n iid random variables $X_1, X_2, \ldots, X_n$ let;

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}} \text{ where } Y_i = \frac{X_i - \mu}{\sigma}$$

Note that we've normalised the random variable X_i, but we could've taken X_i with mean 0 and variance 1 without loss of generality. The characteristic function of Z_n is therefore (using a property of a probability generating function of a sum of iid random variables);

$$\phi_{Z_n}(t) = \phi_Y\left(\frac{t}{\sqrt{n}}\right)^n$$

As a side note, we can apply Taylor's theorem to the characteristic function as follows;

$$\phi_X(t) = \phi_X(0) + \phi'_X(0)t + \frac{\phi''_X(0)t^2}{2!} + \cdots$$
$$\phi_X^{(n)}(0) = i^n M_n \qquad \text{using the definition for the } n^{\text{th}} \text{ moment of } X$$

Note that because we've normalised Y, the first moment is 0, and the second moment is 1. As such, we can write the characteristic function of Y as follows;

$$\phi_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{i^2t^2}{2n} + M_3\frac{i^3t^3}{6n^{\frac{3}{2}}} + \cdots$$

Hence as $n \to \infty$;

$$\Phi_{Z_n}(t) = \left(1 + \frac{i^2t^2}{2n} + M_3\frac{i^3t^3}{6n^{\frac{3}{2}}} + \cdots\right)^n \to e^{-\frac{t^2}{2}} \text{ (characteristic function of N}(0,1))$$

Note that as n goes to infinity, the any term after t^2 go towards zero, much faster than the t^2 term. Therefore, by Levy's continuity theorem; $Z_n \sim N(0,1)$ as $n \to \infty$.

The simplest application of this would be a set of iid Bernoulli(p) random variables $X_1, X_2, \ldots$ with $\mu = p$ and $\sigma^2 = p(1-p)$. Then for any n (has $\mu = np$ and $\sigma^2 = np(1-p)$);

$$\sum_{i=1}^{n} X_i \sim \text{Binomial}(n,p)$$

By CLT, for large n;

$$\sum_{i=1}^{n} X_i \sim N(np, np(1-p))$$

This then shows that, for a large n, Binomial$(n,p) \approx N(np, np(1-p))$. However, the LHS is a discrete distribution, whereas the RHS is a continuous distribution.

Example Questions

1. An analogue signal received at a detector (measured in microvolts) may be modelled as a Gaussian random variable $X \sim N(200, 256)$.

   (a) What is the probability that the signal will exceed 240μV?

$$\begin{aligned} P(X > 240) &= 1 - P(X \leq 240) \\ &= 1 - \Phi\left(\frac{240-200}{16}\right) \\ &\approx 1 - 0.994 \\ &= 0.006 \end{aligned}$$

   (b) What is the probability that the signal is larger than 240μV, given that it is greater than 210μV?

$$\begin{aligned} P(X > 240 | X > 210) &= \frac{P(X > 240)}{P(X > 210)} \\ &= \frac{1 - P(X \leq 240)}{1 - P(X \leq 210)} \\ &= \frac{1 - \Phi\left(\frac{240-200}{16}\right)}{1 - \Phi\left(\frac{210-200}{16}\right)} \\ &\approx \frac{1 - 0.994}{1 - 0.734} \\ &= 0.023 \end{aligned}$$

12th February 2020

Joint Random Variables

Suppose we have two random variables X, Y, which are mappings from the sample space to the real numbers. The sample space has a probability measure $P(E)$, where $E \subseteq S$. S could be the set of outcomes from two experiments, with X relating to the first experiment, and Y to the second. We can then define marginal probability distributions P_X and P_Y for $B \subseteq \mathbb{R}$;

$$P_X(B) = P(X^{-1}(B)) \text{ and } P_Y(B) = P(Y^{-1}(B))$$

The **joint probability distribution** P_{XY} is defined (for $B_X, B_Y \subseteq \mathbb{R}$) as;

$$P_{XY}(P_X, B_Y) = P(\{X^{-1}(B_X) \cap Y^{-1}(B_Y)\})$$

Therefore, it is the probability measure of the set of all points in the sample space that get mapped into B_X by X, and also into B_Y by Y. In practice, it is the probability that $X \in B_X$ and $Y \in B_Y$.

We can specifically define a region $B_{XY} \subseteq \mathbb{R}^2$, and therefore the corresponding sample space points;

$$S_{XY} = \{s \in S \mid (X(s), Y(s)) \in B_{XY}\} \text{ and define } P_{XY}(B_{XY}) = P(S_{XY})$$

We can then define the **joint cumulative distribution function** for $x, y \in \mathbb{R}$ as;

$$F_{XY}(x, y) = P_{XY}((-\infty, x], (-\infty, y])$$

We can also recover the marginal cumulative distribution functions for X and Y as;

- $F_X(x) = F_{XY}(x, \infty)$ $\qquad x \in \mathbb{R}$
- $F_Y(y) = F_{XY}(\infty, y)$ $\qquad y \in \mathbb{R}$

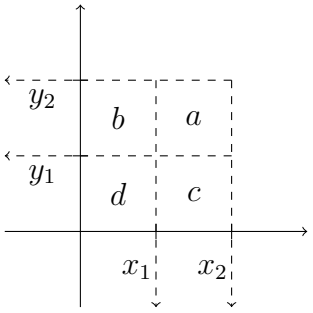
To verify that F_{XY} is a valid joint cumulative distribution function, the following conditions must hold;

- $0 \leq F_{XY}(x, y) \leq 1$, $\forall x, y \in \mathbb{R}$
- $\forall x_1, x_2, y_1, y_2 \in \mathbb{R}$: monotonicity
  $x_1 < x_2 \Rightarrow F_{XY}(x_1, y_1) \leq F_{XY}(x_2, y_1)$ and
  $y_1 < y_2 \Rightarrow F_{XY}(x_1, y_1) \leq F_{XY}(x_1, y_2)$
- $F_{XY}(x, -\infty) = F_{XY}(-\infty) = 0$ and $F_{XY}(\infty, \infty) = 1$, $\forall x, y \in \mathbb{R}$

Interval probabilities are done as follows;

$$P_{XY}(x_1 < X \leq x_2, Y \leq y) = F_{XY}(x_2, y) - F_{XY}(x_1, y)$$
$$P_{XY}(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \underbrace{F_{XY}(x_2, y_2)}_{a+b+c+d} - \underbrace{F_{XY}(x_1, y_2)}_{b+d} - \underbrace{F_{XY}(x_2, y_1)}_{c+d} + \underbrace{F_{XY}(x_1, y_1)}_{d}$$

With an interval on both, we can visualise it as follows (which is why we add the last region);



Joint Probability Mass Function

If X and Y are both discrete random variables, we can define the joint probability mass function (for $x, y \in \mathbb{R}$) as;

$$p_{XY}(x, y) = P_{XY}(X = x, Y = y)$$

And recover the marginal probability mass functions p_X and p_Y by the law of total probability;

- $p_X(x) = \sum_y p_{XY}(x, y)$ $\forall x \in \mathbb{R}$
- $p_Y(y) = \sum_x p_{XY}(x, y)$ $\forall y \in \mathbb{R}$

Similarly, for p_{XY} to be valid, the following conditions must hold;

- $0 \leq p_{XY}(x, y) \leq 1$, $\forall x, y \in \mathbb{R}$
- $\sum_y \sum_x p_{XY}(x, y) = 1$

Joint Probability Density Function

On the other hand, we say X and Y are **jointly continuous** if there exists a **joint probability density function** $f_{XY} : \mathbb{R} \times \boldsymbol{R} \to \mathbb{R}$, such that the following holds ($B_{XY} \subseteq \mathbb{R} \times \mathbb{R}$);

$$P_{XY}(B_{XY}) = \int_{(x,y) \in B_{XY}} f_{XY}(x, y) \, \mathrm{d}x\mathrm{d}y$$

We can then define the **joint cumulative distribution function** F_{XY}, for $x, y \in \mathbb{R}$ as;

$$F_{XY}(x, y) = \int_{t=-\infty}^{t=y} \int_{s=-\infty}^{s=x} f_{XY}(s, t) \, \mathrm{d}s \, \mathrm{d}t$$

The joint probability density function can be identified as follows;

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y)$$

The marginal densities f_X and f_Y can be obtained as follows;

$$\begin{aligned}
f_X(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) \\
&= \frac{\mathrm{d}}{\mathrm{d}x} F_{XY}(x, \infty) \\
&= \frac{\mathrm{d}}{\mathrm{d}x} \int_{y=-\infty}^{y=\infty} \int_{s=-\infty}^{s=x} f_{XY}(s, y) \, \mathrm{d}s \, \mathrm{d}y \\
&= \int_{y=-\infty}^{y=\infty} f_{XY}(x, y) \, \mathrm{d}y \\
f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} \int_{x=-\infty}^{x=\infty} \int_{s=-\infty}^{s=y} f_{XY}(x, s) \, \mathrm{d}s \, \mathrm{d}x \\
&= \int_{x=-\infty}^{x=\infty} f_{XY}(x, y) \, \mathrm{d}x
\end{aligned}$$

Similarly, for f_{XY} to be valid, the following conditions must hold;

- $f_{XY}(x, y) \geq 0$, $\forall x, y \in \mathbb{R}$
- $\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} f_{XY}(x, y) \, \mathrm{d}x \, \mathrm{d}y = 1$

Independence of Random Variables and Conditional Distributions

Two random variables X, Y are independent if, and only if, $\forall B_X, B_Y \subseteq \mathbb{R}$;

$$P_{XY}(B_X, B_Y) = P_X(B_X)P_Y(B_Y)$$

Or more specifically, $\forall x, y \in \mathbb{R}$;

- **discrete** $p_{XY}(x, y) = p_X(x)p_Y(y)$
- **continuous** $f_{XY}(x, y) = f_X(x)f_Y(y)$

We define the conditional probability distribution $P_{Y|X}$, for $B_X, B_Y \subseteq \mathbb{R}$ as follows;

$$P_{Y|X}(B_Y|B_X) = \frac{P_{XY}(B_X, B_Y)}{P_X(B_X)}$$

Which is the revised probability of Y falling in B_Y, given we know $X \in B_X$. Therefore, X and Y are independent if, and only if, $P_{Y|X}(B_Y|B_X) = P_Y(B_Y)$, $\forall B_X, B_Y \subseteq \mathbb{R}$. Or more specifically, $\forall x, y \in \mathbb{R}$;

- **discrete** $p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$
- **continuous** $f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$

Similarly, X and Y are independent if, and only, if $p_{Y|X}(y|x) = p_Y(y)$ or $f_{Y|X}(y|x) = f_Y(y)$, $\forall x, y \in \mathbb{R}$.

We can also define the partition rule as follows;

- **discrete**

$$p_X(x) = \sum_y p_{X|Y}(x|y)p_Y(y)$$

- **continuous**

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)\,\mathrm{d}y \textbf{ and also } F_X(x) = \int_{y=-\infty}^{\infty} F_{X|Y}(x|y)f_Y(y)\,\mathrm{d}y$$

13th February 2020

Expectation of a Function of Random Variables

Let $g : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a bivariate function of the random variables X and Y. We define the expectations as follows;

- **discrete**

$$E_{XY}(g(X, Y)) = \sum_y \sum_x g(x, y)p_{XY}(x, y)$$

- **continuous**

$$E_{XY}(g(X, Y)) = \int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} g(x, y)f_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

From these definitions, we have the following;

- if $g(X, Y) = g_1(X) + g_2(Y)$; $E_{XY}(g_1(X) + g_2(Y)) = E_X(g_1(X)) + E_Y(g_2(Y))$
- if $g(X, Y) = g_1(X)g_2(Y)$ and X, Y independent $E_{XY}(g_1(X)g_2(Y)) = E_X(g_1(X))E_Y(g_2(Y))$
- if $g(X, Y) = XY$ and X, Y independent $E_{XY}(XY) = E_X(X)E_Y(Y)$

The conditional expectation of a random variable Y, given that a random variable $X = x$ is;

- **discrete**

$$E_{Y|X}(Y|x) = \sum_y y p_{Y|X}(y|x)$$

- **continuous**

$$E_{Y|X}(Y|x) = \int_{y=-\infty}^{\infty} y f_{Y|X}(y|x)\,\mathrm{d}y$$

In either case, the conditional expectation is a function of x, and not of the random variable Y.

Note that we can also define a random variable as follows;

$$Z = E_{Y|X}(Y|X) \text{ where } Z(s) = E_{Y|X}(Y|X(s))$$

This then gives us;

$$E_Y(Y) = E_X(E_{Y|X}(Y|X))$$

I hope the tower rule is not examinable.

Covariance and Correlation

For a single random variable X, we considered the expectation of $g(X) = (X - \mu_X)(X - \mu_X)$ as the variance, and denoted it σ_X^2. The extension of this, to two variables, is the expectation of $g(X, Y) = (X - \mu_X)(Y - \mu_Y)$, which we denote the **covariance**, σ_{XY}.

$$\begin{aligned}\sigma_{XY} &= \mathrm{Cov}(X, Y) \\ &= E_{XY}((X - \mu_X)(Y - \mu_Y)) \\ &= E_{XY}(XY) - \mu_X\mu_Y\end{aligned}$$

Note that for **independent** random variables, $E_{XY}(XY) = E_X(X)E_Y(Y) = \mu_X\mu_Y$, therefore $\sigma_{XY} = 0$. This measures how two random variables change in tandem with one another.

This is closely related to the idea of correlation, and therefore the correlation of X and Y can be defined as;

$$\rho_{XY} = \mathrm{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

This is invariant to the scale of the random variables (and therefore lies between -1 and 1). Additionally, if X and Y are independent, then $\sigma_{XY} = \rho_{XY} = 0$ (does not necessarily apply the other way around).

Example Questions

1. a Let X, Y be independent exponential random variables with parameters λ, μ respectively. What is the probability $X < Y$?

   s1: direct

$$\begin{aligned}P(X < Y) &= \int_{x<y} f_{XY}(x, y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=y} f_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=y} f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y && \text{by independence} \\ &= \int_{y=-\infty}^{y=\infty} f_Y(y)\int_{x=-\infty}^{x=y} f_X(x)\,\mathrm{d}x\,\mathrm{d}y && f_Y(y) \text{ is constant w.r.t } x \\ &= \int_{y=-\infty}^{y=\infty} F_X(y)f_Y(y)\,\mathrm{d}y && \text{by definition of cdf}\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty (1 - e^{-\lambda y})\mu e^{-\mu y}\,\mathrm{d}y \\
&= \left[-e^{-\mu y}\right]_0^\infty - \left[-\frac{\mu}{\mu+\lambda}e^{-(\mu+\lambda)y}\right]_0^\infty \\
&= 1 - \frac{\mu}{\mu+\lambda} \\
&= \frac{\lambda}{\mu+\lambda}
\end{aligned}
$$

s2: intuition (with partition rule)

$$
\begin{aligned}
P(X<Y) &= \int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=y} f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=y} f_{X|Y}(x|y)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{y=-\infty}^{y=\infty} F_{X|Y}(y|y)f_Y(y)\,\mathrm{d}y && \text{intuition here} \\
&= \int_{y=-\infty}^{y=\infty} F_X(y)f_Y(y)\,\mathrm{d}y && \text{by independence} \\
&= \cdots
\end{aligned}
$$

2. Suppose that the lifetime, X, and brightness, Y, of a light bulb are modelled as continuous random variables. Let their joint probability distribution function be given by;

$$f(x,y) = \lambda_1\lambda_2 e^{-\lambda_1 x - \lambda_2 y} \text{ for } x, y > 0$$

Are lifetime and brightness independent?

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \lambda_1\lambda_2 e^{-\lambda_1 x - \lambda_2 y}\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \lambda_1 e^{-\lambda_1 x}\lambda_2 e^{-\lambda_2 y}\,\mathrm{d}y \\
&= \lambda_1 e^{-\lambda_1 x}\int_{-\infty}^{\infty} \lambda_2 e^{-\lambda_2 y}\,\mathrm{d}y \\
&= \lambda_1 e^{-\lambda_1 x} && \text{note it is a pdf, hence integrates to 1} \\
f_Y(y) &= \lambda_2 e^{-\lambda_2 y} && \text{same as above} \\
f_{XY}(x,y) &= f_X(x)f_Y(y)
\end{aligned}
$$

Hence X and Y are independent.

3. Suppose continuous random variables $(X, Y) \in \mathbb{R}^2$ have joint probability density function

$$f(x,y) = \begin{cases} 1 & |x| + |y| < \frac{1}{\sqrt{2}} \\ 0 & \text{otherwise} \end{cases}$$

Determine the marginal probability density functions for X and Y.

First determine the range of values for y;

$$|x| + |y| < \frac{1}{\sqrt{2}} \Leftrightarrow |y| < \frac{1}{\sqrt{2}} - |x| \Leftrightarrow -\left(\frac{1}{\sqrt{2}} - |x|\right) < y < \frac{1}{\sqrt{2}} - |x|$$

$$
\begin{aligned}
f_X(x) &= \int_{y=-\infty}^{y=\infty} f(x,y)\,\mathrm{d}y \\
&= \int_{y=-\left(\frac{1}{\sqrt{2}}-|x|\right)}^{y=\frac{1}{\sqrt{2}}-|x|} 1\,\mathrm{d}y \\
&= \sqrt{2}-2|x| \\
f_Y(y) &= \sqrt{2}-2|y| \qquad \text{same as above}
\end{aligned}
$$

Hence $f_X(x)f_Y(y) \neq f_{XY}(x,y)$, therefore X and Y are not independent.

19th February 2020

Statistical Modelling

Statistics is the science of developing knowledge through the use of empirical data. The goal is to use statistical methods to relate the measurements of a sample to the characteristics of the entire population. Modern statistical methods are largely driven by the notion of a model, which is an approximation to a structure, which could've led to the data (commonly used models are parametric). This reduces the problem of learning about the underlying population to learning about a finite set of parameters.

Estimation Theory

For a sample of data $\underline{x} = (x_1, \ldots x_n)$, we consider the these observed values as realisations of corresponding random variables $\underline{X} = (X_1, \ldots, X_n)$. Note that the underlining denotes a collection.

If the distribution of a single random draw X has probability distribution (probability mass function for discrete, and probability density for continuous) $P_{X|\theta}(\cdot|\theta)$, where θ is a generic parameter, we then assume our n data point random variables $\underline{X}$ are iid, with $P_{X|\theta}(\cdot|\theta)$

Example Models

(I) Suppose we want the ages of people in this course - we could ask everyone in this population their age, and calculate the population mean μ, and the population variance σ^2. This may not always be feasible - we can collect a **sample** of 20 observations from the population, and calculate a sample mean $\bar{x}$, and sample variance s^2;

$$\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i \text{ and } s^2 = \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2$$

If this is done with replacement (for iid model), then we know (by CLT) the sample mean $\bar{x}$ is a random variable whose distribution is approximated by;

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{20}\right)$$

(II) Suppose we have a random variable X representing a random draw from our underlying population;

$$X \sim \text{Binomial}(10, p) \text{ with } P_X(x) = \binom{10}{x} p^x (1-p)^{10-x}$$

Where p is unknown. Suppose we draw a sample size of 100 from the population (100 independent binomial random variables), and observe the following;

x	0	1	2	3	4	5	6	7	8	9	10
frequency	2	16	35	22	21	3	1	0	0	0	0

How might we estimate p?

Estimators

Consider a sequence of random variables $\underline{X} = (X_1, \ldots, X_n)$, corresponding to n iid data samples from a population with distribution P_X. Let $\underline{x} = (x_1, \ldots, x_n)$ be the values we observe for these random variables. We define a **statistic** (which is a random variable) as;

$$T = T(X_1, \ldots, X_n) = T(\underline{X})$$

The realised value of a statistic is $t = t(\underline{x})$. For example;

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is a statistic, and } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ is a realised value}$$

If a statistic $T(\underline{X})$ is to be used to approximate parameters of the distribution $P_{X|\theta}(\cdot|\theta)$, we say T is an **estimator** for those parameters. The realised value of the estimator for a data sample is an **estimate**.

20th February 2020

Estimators (Continued)

Consider sample data $\underline{x} = (x_1, \ldots, x_n)$, which comes from an exponentially distributed population;

$$X_i \sim \text{Exp}(\lambda) \text{ with pdf } f_X(x) = \lambda e^{-\lambda x}$$

We can then attempt to estimate the mean ($= \frac{1}{\lambda}$), or attempt to estimate the variance ($= \frac{1}{\lambda^2}$). Some candidates for estimators are as follows - but we need to quantify which is better;

1) first data point $T_1(\underline{X}) = X_1$

2) sample mean $T_2(\underline{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$

3) median (suppose data is ordered) $T_3(\underline{X}) = X_{\frac{n+1}{2}}$

Suppose we knew the parameter values θ for our population distribution $P_{X|\theta}(\cdot|\theta)$. Since we considered our sampled data as iid realisations from this distribution, any statistic $T = T(\underline{X})$ is also a random variable with some distribution, depending on the same parameters. If we can identify this **sampling distribution of our statistic** $P_{T|\theta}$, we can then find the expectation, variance, etc. of our statistic. Provided n is large, CLT gives an approximation for $P_{T|\theta}$, if

$$T = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is the sample mean}$$

Regardless of the form of $P_{X|\theta}$, we know that approximately;

$$\bar{X} \sim N\left(E(X), \frac{\text{Var}(X)}{n}\right)$$

Considering the initial example, where we have an exponential distribution, it can be shown (somehow) our statistic $T = \bar{X}$ is a continuous random variable with probability density function;

$$f_{T|\lambda}(t|\lambda) = \frac{(n\lambda)^n t^{n-1} e^{-n\lambda t}}{(n-1)!}$$

As an aside, the probability density function for Gamma(α, β) is

$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{(\alpha-1)!}$$

Hence, we have $\alpha = n$ and $\beta = n\lambda$;

$$T \sim \text{Gamma}(n, n\lambda)$$

Note that $\text{Gamma}(\alpha, \beta)$ has expectation $\frac{\alpha}{\beta}$, we can then say;

$$E(\bar{X}) = E_{T|\lambda}(T|\lambda) = \frac{n}{n\lambda} = \frac{1}{\lambda}$$

Note that this is the same as the mean of our population distribution, $E(X)$. This suggests that the sample mean $\bar{X}$ is a good statistic for estimating the mean of an exponential distribution.

Bias of Estimators

The **bias** of an estimator T for a parameter θ is;

$$\text{bias}(T) = E(T|\theta) - \theta$$

If this value is 0, then the estimator is unbiased. Note that in the previous example, the sample mean gave an unbiased estimate of the mean of an exponential distribution (which is $\frac{1}{\lambda}$). This is true for any distribution - the sample mean $\bar{x}$ is an unbiased estimate for the population mean μ;

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{n\mu}{n} = \mu$$

Similarly, the **bias-corrected sample variance** is as follows;

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Note that this is not the same as the sample variance, which has a denominator of n instead of $n-1$ as it has one too many degrees of freedom (it would be valid if we knew the population mean μ, instead of using $\bar{X}$). The proof for the bias-correct sample variance is as follows;

$$\begin{aligned}
E\left(S_{n-1}^2\right) &= \frac{1}{n-1}\sum_{i=1}^{n} E\left((X_i - \bar{X})^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} E\left(\left(X_i - \frac{1}{n}\sum_{k=1}^{n} X_j\right)^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} E\left(X_i^2 - \frac{2}{n}\sum_{j=1}^{n} X_i X_j + \frac{1}{n^2}\sum_{j,k=1}^{n} X_j X_k\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} E\left(X_i^2 - \frac{2}{n}X_i^2 - \frac{2}{n}\sum_{j\neq i} X_i X_j + \frac{1}{n^2}\sum_{j=1}^{n} X_j^2 + \frac{1}{n^2}\sum_{j\neq k} X_j X_k\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{n-2}{n}E\left(X_i^2\right) - \frac{2}{n}\sum_{j\neq i} E(X_i X_j) + \frac{1}{n^2}\sum_{j=1}^{n} E\left(X_j^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j} E(X_j X_k)\right)
\end{aligned}$$

we can use the following;

$$\begin{aligned}
E(X_i X_j) &= E(X_i)E(X_j) \\
&= \mu^2 \\
E(X_i^2) &= \sigma^2 + \mu^2
\end{aligned}$$

continuing on;

$$
\begin{aligned}
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{n-2}{n}(\sigma^2+\mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(\sigma^2+\mu^2) + \frac{1}{n^2}n(n-1)\mu^2\right) \\
&= \frac{1}{n-1}n\left(\frac{n-2}{n}(\sigma^2+\mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(\sigma^2+\mu^2) + \frac{1}{n^2}n(n-1)\mu^2\right) \\
&= \frac{1}{n-1}\left((n-2)(\sigma^2+\mu^2) - 2(n-1)\mu^2 + (\sigma^2+\mu^2) + (n-1)\mu^2\right) \\
&= \cdots \\
&= \sigma^2
\end{aligned}
$$

Efficiency of Estimators

Let $\hat{\Theta}(\underline{X})$ and $\tilde{\Theta}(\underline{X})$ be two unbiased estimators for a parameter θ. Suppose we have the sampling distributions for these estimators $P_{\hat{\Theta}|\theta}$, and $P_{\tilde{\Theta}|\theta}$ respectively (and so can calculate their means and variances). We say $\hat{\Theta}$ is more efficient than $\tilde{\Theta}$ if;

$$\forall\theta \ \mathrm{Var}_{\hat{\Theta}|\theta}(\hat{\Theta}|\theta) \leq \mathrm{Var}_{\tilde{\Theta}|\theta}(\tilde{\Theta}|\theta)$$

Additionally, we can say it is strictly more efficient if there is a θ such that it is a strict inequality.

If $\hat{\Theta}$ is more efficient than any other possible estimator, then we can say it is efficient.

Example of Bias and Efficiency

Suppose we have a population with mean μ, and variance σ^2, with n samples $\underline{X}$. Two estimators of μ are as follows (both of which have expectation μ, and are therefore unbiased);

- the sample mean $\hat{M} = \bar{X}$

  $E(\bar{X}) = \mu$ (proven)

  $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$
- the first observation $\tilde{M} = X_1$

  $E(X_1) = \mu$ (trivial)

  $\mathrm{Var}(X_1) = \sigma^2$

Looking at the variances, for $n \geq 2$, $\hat{M}$ is more efficient than $\tilde{M}$ as an estimator for μ.

Consistency of Estimators

The worst aspect of taking the first observation is that it does not change with a larger sample, whereas the variance the sample mean decreases with larger n. We say the $\hat{\Theta}$ is a **consistent** estimator for θ if;

$$\forall\epsilon > 0 \ P(|\hat{\Theta} - \theta) > \epsilon) \to 0 \text{ as } n \to \infty$$

Additionally, if the estimator is unbiased;

$$\lim_{n\to\infty}\mathrm{Var}(\hat{\Theta}) = 0 \Rightarrow \hat{\Theta} \text{ is consistent}$$

Therefore $\bar{X}$ is a consistent estimator of μ for any underlying population.

26th February 2020

Maximum Likelihood Estimator

The general setting is as follows; we have n iid samples with probability distribution (density in the case of continuous, mass in the case of discrete) function $P_{X|\theta}(x_i)$. We denote the set of observed data $\underline{x} = (x_1, \ldots, x_n)$. The joint probability of getting the data $\underline{x}$ is the **likelihood function**, which is a function of θ for fixed data - "the probability to observe this data";

$$L(\theta|\underline{x}) \triangleq \prod_{i=1}^{n} P_{X|\theta}(x_i)$$

Our guess for $\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\, L(\theta|\underline{x})$, which is the value of θ that maximises the likelihood function.
However, we can maximise the log-likelihood function;

$$\ell(\theta|\underline{x}) = \log L(\theta|\underline{x}) = \sum_{i=1}^{n} \log P_{X|\theta}(x_i)$$

Therefore our $\hat{\theta}$ must satisfy the following;

$$\underbrace{\frac{\partial}{\partial\theta}\ell(\theta|\underline{x}) = 0\Big|_{\theta=\hat{\theta}}}_{\text{stationary point}} \text{ and } \underbrace{\frac{\partial^2}{\partial\theta^2}\ell(\theta|\underline{x}) < 0\Big|_{\theta=\hat{\theta}}}_{\text{maximum point}}$$

For this to work, our probability distribution function must be differentiable.

Applying this to our example, based on the binomial distribution, where we have 100 samples $\underline{x}$, and

$$P_{X|p}(x_i) = \binom{10}{x_i} p^{x_i}(1-p)^{10-x_i}$$

$$\begin{aligned}
L(p|\underline{x}) &= L(p) \\
&= \prod_{i=1}^{100} \binom{10}{x_i} p^{x_i}(1-p)^{10-x_i} && \Rightarrow \\
\ell(p) &= \sum_{i=1}^{100} \ln\left(\binom{10}{x_i} p^{x_i}(1-p)^{10-x_i}\right) \\
&= \sum_{i=1}^{100} \ln\binom{10}{x_i} + \ln(p)\sum_{i=1}^{100} x_i + \ln(1-p)\left(10\cdot 100 - \sum_{i=1}^{100} x_i\right) && \Rightarrow \\
\frac{\partial}{\partial p}\ell(p) &= \frac{\sum_{i=1}^{100} x_i}{p} - \frac{10\cdot 100 - \sum_{i=1}^{100} x_i}{1-p} && \text{set} = 0, \text{ solve for } p, \Rightarrow \\
\hat{p}_{\text{MLE}} &= \frac{\sum_{i=1}^{100} x_i}{10\cdot 100} \\
&= \frac{\bar{x}}{10} \\
\frac{\partial^2}{\partial p^2}\ell(p) &= \cdots \\
&= -100\left(\frac{\bar{x}}{p^2} + \frac{10-\bar{x}}{(1-p)^2}\right) && \text{check maximum}
\end{aligned}$$

Plugging in the values from the the example, we get $p \approx 0.257$.

Another example is to work out the MLE for μ for random variables following the normal distribution;

$$X \sim N(\mu, \sigma^2) \text{ and } f_X(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned}
L(\mu|\underline{x}) &= L(\mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} && \Rightarrow \\
\ell(\mu) &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2 \\
&= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{1}{2\sigma^2} 2\mu \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} n\mu^2 && \Rightarrow \\
\frac{\partial}{\partial\mu}\ell(\mu) &= \frac{\sum_{i=1}^{n} x_i}{\sigma^2} - \frac{n\mu}{\sigma^2} && \text{set} = 0, \text{ solve for } \mu, \Rightarrow \\
\hat{\mu}_{\text{MLE}} &= \bar{x}
\end{aligned}$$

However, is the MLE a good estimator of θ?

- not necessarily unbiased
- for large n, the MLE is approximately normally distributed with mean θ
- it is consistent
- always asymptotically efficient, and if there exists an efficient estimator, it's the MLE

We can quantify the degree of uncertainty with confidence intervals.

27th February 2020

Uncertainty of Estimates

It is often insufficient to report an estimate $\hat{\theta}$ for an unknown θ - we usually want to quantify the uncertainty of this estimate. If we had the true sampling distribution of our statistic, $P_{T|\theta}$, then the **variance of this distribution** would give this measure. We can plug in our estimated value ($\hat{\theta}$), and hence use the approximated sampling distribution $P_{T|\theta}$.

Confidence Intervals

By the central limit theorem, we have the following (for any X_i with mean μ, and variance σ^2);

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Using the sample mean as the estimator for the population mean, we can further approximate this as;

$$\bar{X} \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

If we knew the true population variance σ^2, then we'd be able to say if the true mean parameters μ was $\bar{x}$, then for a large n, with 95% probability we would've observed $\bar{X}$ within the 95% **confidence interval**;

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

More generally, we can define the $100(1-\alpha)\%$ confidence interval for μ as;

$$\left(\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Note that z_α is the α-quantile of the standard normal (same as using $\Phi^{-1}(\alpha)$).

An example of this is as follows. Suppose there was a survey to investigate the proportion of employees who thought the board was doing a good job has 1000 employees selected at random, with 732 saying they did. Find a 99% confidence interval for the proportion in the population who thought the board was doing a good job.

We can model this as 1000 Bernoulli(p) random variables, with probability p saying good, and probability $1-p$ saying bad. Note that we also know $\mu = p$ for a Bernoulli distribution, hence our estimate for the population mean is $\hat{p} = \bar{x} = 0.732$. Since we've assumed a Bernoulli(p) distribution, we also have the variance as;

$$\begin{aligned}\sigma^2 &= p(1-p)\\ &= \bar{x}(1-\bar{x})\\ &\approx 0.196\end{aligned}$$

We can then construct a 99% confidence interval, taking $\Phi^{-1}(0.995) \approx 2.576$, for $\hat{p} = \bar{x} = 0.732$;

$$\left(\bar{x} - 2.576\frac{\sqrt{0.196}}{\sqrt{1000}}, \bar{x} + 2.576\frac{\sqrt{0.196}}{\sqrt{1000}}\right) \approx (0.696, 0.768)$$

However, it's important to note that this was only an approximate confidence interval, as we relied on CLT, as well as assumed the population variance.

If $\underline{X} = (X_1, \ldots, X_n)$ is an iid sample from $N(\mu, \sigma^2)$, then we do not need the central limit theorem, and the following is exact;

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Then the confidence interval is exact, given we know σ^2.

Normal Distribution with Unknown Variance

In practice, when we aim to fit a normal distribution to real data, both μ and σ^2 are unknown. Similarly, if $\underline{X}$ is an iid sample from $N(\mu, \sigma^2)$, we have exactly (without proof);

$$\underbrace{\frac{\bar{X}-\mu}{\frac{s_{n-1}}{\sqrt{n}}} \sim t_{n-1}}_{\sigma^2 \text{ unknown}} \text{ versus } \underbrace{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)}_{\sigma^2 \text{ known}}$$

Where s_{n-1} is the bias-correct sample standard deviation, and t_ν is the student's t distribution with ν degrees of freedom. Therefore an **exact** $100(1-\alpha)\%$ confidence interval for μ is;

$$\left(\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{s_{n-1}}{\sqrt{x}}, \bar{x} + t_{n-1,1-\frac{\alpha}{2}}\frac{s_{n-1}}{\sqrt{x}}\right)$$

Where $t_{\nu,\alpha}$ is the α-quantile of t_ν (tabulated). We can make the following comments about t_ν;

- heavier tailed than the normal distribution

  this means that if we know σ^2, then we should use the normal distribution as it will give us tighter (less wide) confidence intervals
- $\lim_{\nu\to\infty} t_\nu = N(0,1)$ - it approaches the standard normal distribution, hence $t_{\infty,\alpha} = z_\alpha$
- for $\nu > 40$, $t_\nu \approx N(0,1)$

For example, if we were to construct a 95% confidence interval we'd do the following. First note that it is a two-sided test, so the quantile we're interested in is 0.975. Since $\nu = n - 1 = 9$, we will look at $t_{9,0.975}$, hence we have 2.26.

Hypothesis Testing

Suppose we want to know if exposure to asbestos is associated with lung disease. We take a group of rats, and divide them into two groups - one which is exposed, and one which is kept unexposed. We then compare the rates of lung disease. Consider the following two hypotheses;

- null hypothesis — disease rate is the same
- alternative hypothesis — disease rate is different (not same)

If the exposed group has a much higher rate of lung disease, then we reject the null hypothesis and conclude that our evidence favours the alternative hypothesis.

Suppose we have a random iid sample, $\underline{X} = (X_1, \ldots, X_n)$ of random variable from an unknown distribution P_X. We then hypothesise probability models for X. specifically, we fix upon a parametric family $P_{X|\theta}$, and test whether our hypothesised parameter values are plausible - that we can assume $\theta = \theta_0$ for some particular θ_0.

For example, let $X \sim N(\mu, \sigma^2)$, and we may wish to test whether $\mu = 0$ is plausible.

Formally, we partition the parameter space Θ into Θ_0 and Θ_1 (disjoint) and test;

$$\underbrace{H_0 : \theta \in \Theta_0}_{\text{null hypothesis}} \text{ versus } \underbrace{H_1 : \theta \in \Theta_1}_{\text{alt. hypothesis}}$$

- **simple** hypothesis — form $\theta = \theta_0$

  $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ — two-sided test (most often)
- **composite** hypothesis — form $\theta > \theta_0$ or $\theta < \theta_0$

  $H_0 : \theta > \theta_0$ versus $H_1 : \theta < \theta_0$ — one-sided test

4th March 2020

Literally just a tutorial.

5th March 2019 (old Panopto)

Rejection Region

To test the validity of H_0, we choose a test statistic $T(\underline{X})$, with the distribution P_T, under H_0. Then, we wish to identify a region $R \subset \mathbb{R}$ containing low probability values of T, under the assumption H_0 is true. Let the low probability be α (such as 5%).

$$P(T \in R | H_0) = \alpha$$

We then calculate the observed test statistic $t(\underline{x})$ for our observed data $\underline{x}$.

- if $t \in R$ — "**reject** the null hypothesis at the 100α% level"

- if $t \notin R$ "**retain** the null hypothesis at the $100\alpha\%$ level"

Changing the significance level, $\alpha \in (0, 1)$, will result in the following;

- as $\alpha \to 0$, we are less likely to reject our null hypothesis (rejection region becoming smaller)
- as $\alpha \to 1$, we are more likely to reject our null hypothesis (rejection region becoming larger)

we define the p-value of the data as the threshold between us rejecting and not rejecting the null hypothesis. A smaller p-value suggests stronger evidence against H_0.

Note the duality with confidence intervals. If we have constructed a $100(1-\alpha)\%$ confidence interval for some parameter θ - this is the set of values for for which there **wouldn't** be sufficient evidence to reject the null hypothesis $H_0 : \theta = \theta_0$ at the $100\alpha\%$ level.

Errors

There are two types of errors in the outcome of a hypothesis test;

type I: Rejecting H_0 when H_0 is true, which has the probability;

$$\alpha = P(T \in R|H_0)$$

Hence the significance level of a hypothesis test is also the type I error rate.

type II: Not rejecting H_0 when H_1 is true, which has the probability;

$$\beta = P(T \notin R|H_1)$$

Some real scenarios of the following are as follows;

- H_0 says a person is innocent

type I: an innocent person is convicted
type II: a guilty person is free

- H_0 says a patient's symptoms after treatment A are indistinguishable from a placebo

type I: falsely indicates the treatment is more effective than a placebo
type II: fails to demonstrate the treatment is more effective than a placebo (even though it is)

The **power** of a hypothesis test is defined as

$$1 - \beta = 1 - P(T \notin R|H_1) = P(T \in R|H_1)$$

For a fixed significance level $\alpha = P(T \in R|H_0)$, a well chosen statistic T and rejection region R will have high power (thus maximising the probability of rejecting the null hypothesis when the alternative hypothesis is true).

Population Mean with Known Variance

Suppose $\underline{X}$ has n samples, iid $N(\mu, \sigma^2)$ (whether exactly, or by CLT), with known σ^2, and unknown μ. We wish to test for $\mu = \mu_0$ (for some specific value). Our hypotheses are as follows;

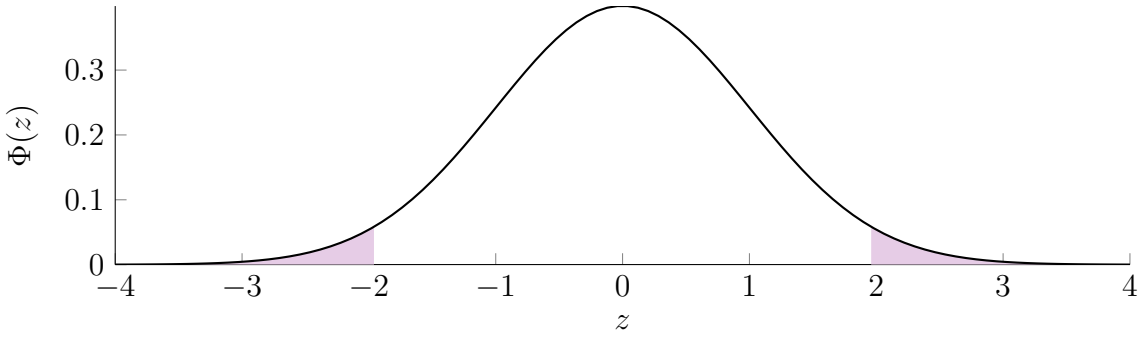
$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

Under H_0 (assuming H_0 is true), we have know both the mean and variance. We have a known distribution for the test statistic ($\bar{X}$);

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

We define our rejection region to be the $100\alpha\%$ **tails** of the standard normal distribution;

$$R = (-\infty, -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty)$$



The rejection region is shaded, for $\alpha = 5\%$.

Thus $P(Z \in R|H_0) = \alpha$, therefore, we reject our the null hypothesis (at the $100\alpha\%$ significance level) when our observed test statistic $\bar{x}$;

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in R$$

The p-value is given by;

$$2(1 - \Phi(|z|)) = 2\left(1 - \Phi\left|\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right|\right)$$

Example of Hypothesis Testing (Know Variance)

A company makes packets of snack foods, with the weight on the bags being 454 grams. Suppose we know the variance is 70. We then have the mass, in grams) of 50 randomly sampled packets - not writing them out.

- $H_0 : \mu = 454$ null hypothesis
- $H_1 : \mu \neq 454$ alternative hypothesis

While the weight distribution of each packet is unknown, the test statistic (sample mean) is distributed as follows (by CLT);

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

The realised values are;

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{451.22 - 454}{\frac{\sqrt{70}}{\sqrt{50}}} \approx -2.35$$

- Let us choose a significance level of $100\alpha\%$, with $\alpha = 0.05$, thus our rejection region is;

  $$R = (-\infty, -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty) \approx (-\infty, -1.96) \cup (1.96, \infty)$$

  We have $z \in R$, therefore we reject the null hypothesis at the 5% level (hence we are 95% sure that the mean weight is incorrect).
- Let us choose another significance level of $100\alpha\%$, with $\alpha = 0.01$, thus our rejection region is;

  $$R = (-\infty, -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty) \approx (-\infty, -2.576) \cup (2.576, \infty)$$

  We have $z \notin R$, therefore we retain the null hypothesis at the 1% level.

6th March 2019 (old Panopto)

Proof of p-value

We claim that the p-value, $\alpha_0 = 2(1 - \Phi(|z|))$. By the definition of the p-value, we want the critical point, that is when $z_{1-\frac{\alpha}{2}} = |z|$ Therefore, we have;

$$\Phi^{-1}(1 - \tfrac{\alpha_0}{2}) = |z| \Rightarrow \alpha_0 = 2(1 - \Phi(|z|))$$

Population Mean with Unknown Variance

This is done similarly to the case with a known variance, but instead we use the bias corrected sample standard deviation s_{n-1}, and the student t distribution, with $n - 1$ degrees of freedom.

$$T = \frac{\bar{X} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \sim t_{n-1}$$

Therefore, for a test

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

The rejection region of our observed test statistic

$$t = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}}$$

Is given by

$$R = (-\infty, -t_{n-1,1-\frac{\alpha}{2}}) \cup (t_{n-1,1-\frac{\alpha}{2}}, \infty)$$

Again $P(T \in R|H_0) = \alpha$.

Example of Hypothesis Testing (Unknown Variance)

A piece of computer code takes a random time to run on a computer, with an average time of 6 seconds. The code is optimised, and the programmer wishes to know whether the mean run time has changed. The re-optimised code is run 16 times, with a sample mean run time of 5.8 seconds, and a bias-corrected sample standard deviation of 1.2 seconds. Is the code any faster?

- $H_0 : \mu = 6$ (code isn't faster) null hypothesis
- $H_1 : \mu \neq 6$ (code is faster) alternative hypothesis

We can invoke the central limit theorem (as n is fairly large), thus (under H_0);

$$T = \frac{\bar{X} - \mu}{\frac{s_{n-1}}{\sqrt{n}}} \sim t_{n-1}$$

Plugging in values, we have our realised test statistic as;

$$t = \frac{5.8 - 6}{\frac{1.2}{\sqrt{16}}} = -\frac{2}{3}$$

Looking at the value of $t_{15,0.975}$ (at a 5% significance level), we have 2.13. Therefore we retain the null hypothesis at a 5% significance level, thus the code isn't faster.

Samples from Two Populations

Suppose we have a random sample $\underline{X} = (X_1, \ldots, X_{n_1})$ from an unknown population distribution P_X, and another sample $\underline{Y} = (Y_1, \ldots, Y_{n_2})$, with another unknown population distribution P_Y. We are often interested in testing whether P_X and P_Y have the same means;

$$H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X \neq \mu_Y$$

A **special** case of this is when $\underline{X}$ and $\underline{Y}$ are paired, if $n_1 = n_2 = n$, and the data are collected as pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, such that for each i, X_i and Y_i are possibly dependent. For a test of equal means, we can then consider the sample of differences $Z_1 = X_1 - Y_1, \ldots Z_n = X_n - Y_n$, and then test for $H_0 : \mu_Z = 0$ (with the previous single sample methods).

An example of this is a sample of n individuals, with X_i representing the heart rate before light exercise, and Y_i representing the heart rate after, for the i^{th} person.

Consider the following cases (excluding the special case with paired data);

- **known variance**

  Suppose we have $\underline{X} = (X_1, \ldots, X_{n_1})$ iid with $N(\mu_X, \sigma_X^2)$, with μ_X unknown, and $\underline{Y} = (Y_1, \ldots, Y_{n_2})$ iid with $N(\mu_Y, \sigma_Y^2)$, also with μ_Y unknown - also the two samples are independent. Then we still have the following;

  $$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_1}\right) \text{ and } \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right)$$

  From here, it follows that the difference in sample means is distributed as;

  $$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right) \Rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim N(0, 1)$$

  Therefore, under $H_0 : \mu_X = \mu_Y$, obviously $\mu_X - \mu_Y = 0$, we have;

  $$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim N(0, 1)$$

  Given that we know σ_X^2 and σ_Y^2, we have the test statistic as;

  $$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}$$

  This can be compared against the quantiles of a standard normal distribution, in the same way as before.

- **unknown variance**

  Suppose we have $\underline{X}$ and $\underline{Y}$ the same as before, with the means unknown, but also the variances unknown. If we know $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, with unknown σ^2, we can still proceed as follows;

  $$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

  And similarly we can take a test statistic under H_0 for the same reason as;

  $$\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

We however cannot combine the samples into one large sample - we don't know whether the means are the same (since that's what we're attempting to work out), and thus we would over-estimate the variance. The **bias-corrected pooled sample variance** is defined as;

$$S^2_{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_1}(Y_i - \bar{Y})^2}{n_1 + n_2 - 2} = \underbrace{\frac{n_1 - 1}{n_1 + n_2 - 2} S^2_{n_1-1} + \frac{n_2 - 1}{n_1 + n_2 - 2} S^2_{n_2-1}}_{\text{see below}}$$

The part above shows that $s^2_{n_1+n_2-2}$ is an unbiased estimate of σ^2, as it is a weighted average of the bias-corrected sample variances for each of the individual samples (both of which are unbiased estimates).

By substituting $S_{n_1+n_2-2}$ in for σ, we have;

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_{n_1+n_2-2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

And similarly, assuming $H_0 : \mu_X = \mu_Y$, we have;

$$T = \frac{\bar{X} - \bar{Y}}{S_{n_1+n_2-2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Therefore, our rejection region for a $100\alpha\%$ significance level is;

$$R = (-\infty, -t_{n_1+n_2-2,1-\frac{\alpha}{2}}) \cup (t_{n_1+n_2-2,1-\frac{\alpha}{2}}, \infty)$$

Example of Hypothesis Testing (Two Populations - Known Variance)

Suppose the same piece of C code was repeatedly run after compilation under two different compilers. The sample mean and bias-corrected sample variance for compiler 1 were 114s, and 310s respectively, and the corresponding figures for compiler 2 were 94s and 290s. Both sets of data were based on 15 runs each. Suppose compiler 2 is a refined version of compiler 1 - hence if μ_1 and μ_2 were the run times of code under compilers 1 and 2 respectively, assume $\mu_2 \leq \mu_1$. At the 5% level, conduct the hypothesis;

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 > \mu_2$$

Let the realised value of the test statistic be;

$$t = \frac{\bar{x} - \bar{y}}{s_{n_1+n_2-2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{114 - 94}{\sqrt{\frac{310+290}{2}}\sqrt{\frac{1}{15} + \frac{1}{15}}} \approx 3.162$$

Note that this is a one sided test. Therefore we compare it with;

$$t_{n_1+n_2-2,1-\alpha} = t_{28,0.95} \approx 1.71$$

As it falls in the rejection region, we reject the null hypothesis at a 5% significance level (therefore the second compiler produced faster code).

11th March 2020

Goodness of Fit

The previous sections relied on the data following a normal distribution, or having the sample mean be approximated to be normally distributed (via CLT). Another problem concerns model checking, which can be addressed through a more general consideration of count data for simple (discrete and finite) distributions.

Let X be a simple random variable, taking values in the range $\{x_1, \ldots, x_k\}$ with probability mass function (for $1 \leq j \leq k$) depending on an unknown parameter θ;

$$p_j = P(X = x_j|\theta)$$

We can then summarise a random sample size n from the distribution of X with the **observed frequency counts**;

$$\underline{O} = (O_1, \ldots, O_k) \text{ at the points } x_1, \ldots, x_k \text{ thus } \sum_{j=1}^{k} O_j = n$$

The observed frequency counts are the number of times we observe each value in n samples.

Suppose our null hypothesis is $H_0 : \theta = \theta_0$. Therefore, under H_0, we know the probability mass function p_j. The **expected frequency counts** can then be calculated as;

$$\underline{E} = (E_1, \ldots, E_k) \text{ where } E_j = np_j \text{ thus } \sum_{j=1}^{k} E_j = n$$

Therefore, to test for goodness of fit, we want to quantitatively compare the observed frequencies $\underline{O}$ with the expected frequencies $\underline{E}$.

Chi-Square Test

To test the hypotheses;

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

We used the chi-square statistic;

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

If H_0 were true, then X^2 would follow a chi-square distribution χ^2_ν, with $\nu = k-p-1$ degrees of freedom (k is the number of values X can take, and p is the number of parameters being estimated). For this approximation to be valid, we need $\forall j.\ E_j \geq 5$ - if not, we can merge values ($E_a + E_b = F_a = n(p_a + p_b)$).

Clearly, from our definition of the chi-square statistic, larger values of X^2 correspond to larger deviations from the model proposed by the null hypothesis, and if $X^2 = 0$, the observed counts match exactly to what our model expects.

We always perform a **one-sided goodness of fit** test using the χ^2 statistic, looking at the upper tail of the distribution. The rejection region, at the $100\alpha\%$ level, is given by;

$$R = \{x^2 \mid x^2 > \chi^2_{k-p-1,1-\alpha}\}$$

Examples

- A study in the Journal of the American Medical Association gave the following causes of 95 ADEs (adverse drug effects);

cause	number of ADEs
lack of knowledge of drug	29
rule violation	17
faulty dose checking	13
slips	9
other	27

Test whether the true percentages of ADEs differ across the 5 causes.

Our null hypothesis, H_0, is that they all share the same proportion. Therefore, our expected value $E_j = \frac{95}{5} = 19$, for $1 \leq j \leq 5$.

$$\begin{aligned} X^2 &= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \\ &= \sum_{i=1}^{5} \frac{(O_i - 19)^2}{19} \\ &= \frac{(29-19)^2}{19} + \frac{(17-19)^2}{19} + \frac{(13-19)^2}{19} + \frac{(9-19)^2}{19} + \frac{(27-19)^2}{19} \\ &= 16 \end{aligned}$$

We then compare this to the χ^2_ν distribution, with $\nu = 5 - 0 - 1 = 4$ (5 possible values, and 0 parameters). Let $\alpha = 0.05$, hence at a 5% level, we have $\chi^2_{4,0.95} = 9.49$, hence we reject the null hypothesis at a 5% level (since 16 ¿ 9.49) - the true percentages do differ (not all the same proportion).

- Recall the example from discrete random variables concerning the number of particles emitted by a radioactive substance. For 2608 time intervals, each of length 7.5 seconds, these are the number of particles emitted.

x	0	1	2	3	4	5	6	7	8	9	≥ 10
O	57	203	383	525	532	408	273	139	45	27	16
E	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1
$\frac{(O-E)^2}{E}$	0.124	0.267	1.461	0	1.096	0.534	1.452	0.012	7.723	0.166	0.071

In total, we observe 10094 counts, thus the mean number per interval is 3.87. Since the mean of a Poisson(λ) random variable is λ, we can set $\lambda = 3.87$.

Then we have;

$$X^2 = \sum_{i=1}^{11} \frac{(O_i - E_i)^2}{E_i} \approx 12.906$$

We compare the statistic with a $\chi^2_{11-1-1} = \chi^2_9$ distribution (11 possible values, and 1 parameter). At the 5% level, let $\alpha = 0.05$, $\chi^2_{k-p-1,1-\alpha} = \chi^2_{9,0.95} = 16.91$. Since $12.906 < 16.91$, we retain the null hypothesis at the 5% level, hence the model Poisson(3.87) is good.

Independence Testing

Suppose two random variables X (with range $\{x_1, \ldots, x_k\}$) and Y (with range $\{y_1, \ldots, y_l\}$) are jointly distributed with unknown probability mass function p_{XY}. We often want to determine whether the two random variables are independent, which is the same as determining whether

$$p_{XY}(x, y) = p_X(x) p_Y(y)$$

Then n iid samples from the joint distribution of (X, Y) can be represented as a list of counts $n_{i,j}$ of the number of times we observe the pair (x_i, y_i), obviously with $1 \leq i \leq k$ and $1 \leq j \leq l$. This can be tabulated as follows - a $k \times l$ contingency table;

	y_1	y_2	$\cdots$	y_l	
x_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,l}$	$n_{1,\times}$
x_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,l}$	$n_{2,\times}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	$n_{k,1}$	$n_{k,2}$	$\cdots$	$n_{k,l}$	$n_{k,\times}$
	$n_{\times,1}$	$n_{\times,2}$	$\cdots$	$n_{\times,l}$	n

Under our null hypothesis, H_0, we take X and Y to be independent. Therefore, the expected values are given from the product of the row and column sums (for $1 \leq i \leq k, 1 \leq j \leq l$);

$$\hat{n}_{i,j} = \frac{n_{i,\times} \cdot n_{\times,j}}{n}$$

This is because we can approximate $p_X(x_i)$ by $\hat{p}_{i,\times} = \frac{n_{i,\times}}{n}$, and similarly $p_Y(y_j)$ by $\hat{p}_{\times,j} = \frac{n_{\times,j}}{n}$. Under independence, we can estimate it as;

$$\hat{p}_{i,j} = \hat{p}_{i,\times} \cdot \hat{p}_{\times,j} = \frac{n_{i,\times} \cdot n_{\times,j}}{n^2}$$

We then perform the standard test, comparing it to χ^2_ν, where $\nu = (k-1)(l-1)$ degrees of freedom.