STAT 230 Course Notes

Muhammad Talha, taught by Diana K. Skrzydlo

Fall 2017

Contents

1 Chapter 1

There are three definitions of probability.

1.1 Classical

If you have a bunch of possible outcomes that are *equally* likely, then the probability of an event is:

$$p = \frac{\text{total \# of ways that event can occur}}{\text{total \# of outcomes}}$$

E.g. Find the probability of rolling a 1 with a six sided die.

$$p = \frac{1}{6}$$

1.2 Relative Frequency

If you have the ability to repeat an event under the same conditions, the probability of an event is the proportion of time it occurs in infinite repetitions.

1.3 Subjective

The subjective probability of an event is how confident a person is that it will occur.

2 Chapter 2

An experiment is a process that has multiple possible results and can be repeated. For example, rolling a dice is an experiment. There are 6 possible results and it can be repeated.

A trial is one iteration of an experiment. For example, rolling one die. An outcome is one of the possible results from one trial of an experiment.

A sample space is a set of all possible outcomes from one trial of an experiment. For example, the sample space for rolling one die is $\{1, 2, 3, 4, 5, 6\}$.

An event is a subset of a sample space. For example, the event for rolling a 1 is $\{1\} \subseteq \{1, 2, 3, 4, 5, 6\}$. If an outcome is inside an event, then the event occurred. An event that contains one element is called simple, otherwise it is called compound.

2.1 Probability of an event

The probability of an event A is $p(A)$ and it conforms to the following axioms:

- $0 \leq p(A) \leq 1$
- $\sum_{a \in A} p(a) = 1$

Notice that the probability of an event is the sum of the probabilities of its simple events.

E.g. Let A be the event for rolling an odd number, A = $\{2, 4, 6\}$.

$$p(A) = p(\{2\}) + p(\{4\}) + p(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

An easier alternative is using the definition of classical probability, which results in the same answer.

3 Chapter 3

3.1 Counting Techniques

Counting deals with the problem of counting how many ways there are to do something. There are two major techniques that help with counting problems.

3.2 Addition Rule

Suppose we want to count the number of ways to do task 1 *or* task 2. If task 1 can be done in p ways and task 2 can be done with q ways, then the number of ways to do task 1 *or* task 2, provided they don't overlap, is $p + q$.

E.g. Recall that a character is either a digit (0-9) or a letter (a-z). The number of possible characters are $10 + 26 = 36$.

Note that the addition rule can only be used in situations where the two tasks don't overlap, i.e. cannot happen at the same time. For example, counting the number of ways to roll 1 dice or roll 1 dice is $\frac{1}{6}$, not $\frac{2}{6}$.

3.3 Multiplication Rule

Suppose we want to count the number of ways to do task 1 *and* task 2. The number of ways to do this, provided they don't depend on each other, is $p \times q$.

E.g. How many ways are there to flip a coin (H, T) and roll a dice (1-6)? There are $2 \times 6 = 12$ ways.

Note that the multiplication rule can only be used in situations where the two tasks don't depend on each other, i.e. whether or not task 1 occurs does not affect whether or not task 2 will occur. For example, the number of ways to flip a head and a tail (at the same time) is 0.

3.4 Sampling

Sampling means picking items from a set. There are two ways to sample items from a set: with or without replacement.

With replacement means its possible to get the same item twice. For example, its possible to flip a coin and get a head, and then flip a coin again to get

another head.

Without replacement means its *not* possible to get the same item twice. For example, if we draw a card from a deck of 52 cards we will never draw that same card (with same rank and suite) again.

3.4.1 Sampling with replacement

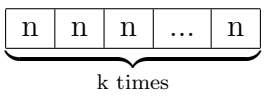
Suppose we have n items in a set and we want a sample of size k. How many ways are there to do this?

We can think of this problem as a bunch of tasks where each task is picking one of the k items from the set. So then we have k tasks.

How many ways are there to do task 1 (picking first of k items)?
n

How many ways are there to do task 2 (picking second of k items)?
Note that we are sampling *with* replacement, so it is possible to get the same result twice. That is, its possible to get the first item we picked in task 1.
So there are n ways to do task 2.

In general, there are n ways to pick the ith item ($1 \leq i \leq k$).



Now, we want the number of ways to do task 1 *and* task 2 *and* .. *and* task k. By the multiplication rule, there are

$$n^k \tag{1}$$

ways to sample k items from a set containing n items with replacement.

E.g. Suppose 3 students go to either one of 4 schools and students can go to the same school. How many ways are there for these 3 students to go to school?
There are 4^3 ways.

3.4.2 Sampling without replacement

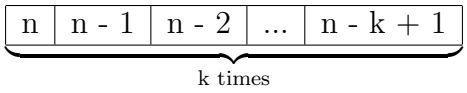
Suppose we have n items in a set and we want a sample of size k. How many ways are there to do this?

We can split the problem up into k tasks just like we did in the previous section.

How many ways are there to do task 1 (picking first of k items)?
n

How many ways are there to do task 2 (picking second of k items)?
Note that we are sampling *without* replacement, so it is not possible to get the same result twice. That is, it is not possible to get the first item we picked in task 1.
So there are $n - 1$ ways to do task 2.

In general, there are $n - (i - 1)$ ways to pick the ith item $(1 \leq i \leq k)$.



Now, we want the number of ways to do task 1 *and* task 2 *and* .. *and* task k. By the multiplication rule, there are

$$n(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!} = n^{(k)} \tag{2}$$

ways to sample k items from a set containing n items without replacement. The abbreviated notation $n^{(k)}$ is often referred to as n to k factors.

E.g. Suppose 3 students go to either one of 4 schools and students cannot go to the same school. How many ways are there for these 3 students to go to school?
There are $4^{(3)}$ ways.

E.g. Suppose a PIN is a 4-digit number where each digit cannot be repeated. How many possible PINs are there?
There are $10^{(4)}$ possible PINs.

3.4.3 Examples of with/without replacement

E.g. An IP address is 4 numbers, each between 0-255. How many possible IP addresses are there?

Note that an IP address can repeat numbers (E.g. 192.168.1.1) so we are sampling with replacement. There are 256^4 possible IP addresses.

E.g. What is the probability a random IP address contains only odd numbers?

Let A be the event that contains all IP addresses that contain only odd numbers.

$$p(A) = \frac{|A|}{|S|} = \frac{128^4}{256^4} = \frac{1}{16}$$

E.g. What is the probability a random IP address has at least one 0? [1]

We could try solving this problem by considering all possible sub cases, in which case we would need to deal with having one 0's, two 0's, three 0's, etc. Lets instead use the complement.

$$p(A) = 1 - p\left(\bar{A}\right) = 1 - p(\text{IP address contains no 0's}) = 1 - \frac{255^4}{256^4}$$

E.g. What is the probability an IP address has all distinct numbers?
Note that if an IP address has all distinct numbers, that means numbers cannot be repeated. Therefore, we are sampling without replacement.

$$p(A) = \frac{256^{(4)}}{256^4}$$

E.g. Suppose five people (A, B, C, D, E) apply to four jobs (1, 2, 3, 4). What is the probability A gets the job?
How many possible ways can the jobs be filled up? $5^{(4)}$.
This can be solved with or without using the complement.

[1] In general, if a problem contains the wording "at least", consider using the complement.

1. Lets first solve it without using the complement. Lets consider all the sub cases where A gets a job.

$$\begin{array}{cccc} \text{A} & \_ & \_ & \_ & 4^{(3)} \\ \_ & \text{A} & \_ & \_ & 4^{(3)} \\ \_ & \_ & \text{A} & \_ & 4^{(3)} \\ \_ & \_ & \_ & \text{A} & 4^{(3)} \end{array}$$

   There are 4 sub cases and the number of ways each sub case can happen is $4^{(3)}$. Therefore:

$$p\left(A\right) = \frac{4 \times 4^{(3)}}{5^{(4)}}$$

2. Now, using the complement:

$$p\left(\text{A gets a job}\right) = 1 - p\left(\text{A does not get a job}\right) = 1 - \frac{4^{(4)}}{5^{(4)}}$$

3.5 Permutations

A permutation is an ordering of k objects selected from n objects without replacement.

How many permutations are there? $n^{(k)}$

Note that a permutation *is* a ordering. There is, order matters when dealing with a permutation.

3.6 Combinations

A combination is a subset of k objects selected from n objects without replacement.

Note that a combination *is* a subset. That is, order does not matter when dealing with a combination.

How many combinations are there?

Lets first find out how many possibilities there are when order matters. $n^{(k)}$

Now, how can we get rid of the duplicates?

E.g. Consider the set $1, 2, 3$. How many ways can this set be rearranged?

$$\{1,2,3\}, \{1,3,2\}, \{2,1,3\}, \{2,3,1\}, \{3,1,2\}, \{3,2,1\}$$

We see there are 6, or 3! ways to arrange three numbers. In general, there are $k!$ ways to rearrange k objects. For each ordering of k objects in our permutation, we need to divide by the number of rearrangements of k objects in order to remove the duplicates.

There are

$$\frac{n^{(k)}}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k} \tag{3}$$

ways to select k objects from n objects without replacement and order doesn't matter. The abbreviated notation $\binom{n}{k}$ is often referred to as n choose k.

There are several properties of $\binom{n}{k}$ that you will study in MATH 239 but we will just discuss two of them.

1. $$\binom{n}{k} = \binom{n}{n-k}$$

   This result can be proved using arithmetic but the logical proof is more interesting and easier to remember. The number of ways to include k objects is to equal to the number of ways to exclude $n-k$ objects. After excluding $n-k$ objects, we will be left with n objects.

2. $$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

Suppose one of the n objects is special. Then $\binom{n}{k}$ contains all the ways to choose k objects from n objects (whether or not we include the special object in our k objects). However, this should be equal to the number of ways to choose k objects including the special object plus the number of ways to choose k objects excluding the special object.

Note that the two sub cases don't overlap (cannot include and exclude the special object at the same time) so we can use the addition rule.

E.g. Suppose you select 5 cards from a deck of 52 cards. Recall a standard deck of 52 cards has 4 suites and 13 ranks. What is the probability of getting a full house (3 cards of one rank and 2 cards of another rank)?

First, how many possible ways are there to select 5 cards out of 52 cards? $\binom{52}{5}$

We want 3 cards of one rank *and* 2 cards of another rank.
How many ways are there to select 3 cards of one rank? We need to select a rank: $\binom{13}{1}$. Choosing a rank will get us 4 cards (one for each suite) *and* we need to pick 3 cards of out these 4 cards: $\binom{4}{3}$. Similarly, to get 2 cards of another (i.e. different) rank: $\binom{12}{1}$ *and* out of those 4 cards we need to pick 2: $\binom{4}{2}$. So we get:

$$\frac{\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}}{\binom{52}{5}}$$

3.7 Arrangements with alike objects

How many arrangements are there of the word: S T A T I S T I C S (10 letters, 3 S's, 3 T's, 1 A, 2 I's, 1 C)?
It is not 10!.

The problem is that there exists alike objects inside the word. For example, the letter S is repeated and two S letters cannot be distinguished between each other. As a result, there are no longer 10 possible choices for the first "slot". The first letter can now only be S, T, A, I, C (5 choices). How can we deal with this case?

What we can do is first treat all letters as distinct. Then there are 10! ways to rearrange the word. Now, for each arrangement of the word, say:

$$S_1\, T_1\, A_1\, T_2\, I_1\, S_2\, T_3\, I_2\, C_1\, S_3$$

there are multiple ways to rearrange that word so that it will look exactly the same:

$$\begin{gathered} S_1 ... S_2 ... S_3 \\ S_1 ... S_3 ... S_2 \\ S_2 ... S_1 ... S_3 \\ S_2 ... S_3 ... S_1 \\ S_3 ... S_1 ... S_2 \\ S_3 ... S_2 ... S_1 \\ ... \\ T_i ... T_j ... T_k \\ ... \\ I_i ... I_j \\ ... \end{gathered}$$

So in order to get rid of the over-counting we did by treating all letters as distinct, we need to divide by all possible ways to rearrange the alike objects. The number of ways to rearrange the S's inside the word is 3!, the number of ways to rearrange the T's inside the word is 3! and similarly 1!, 2!, 1! for A, I, C respectively.

So there are

$$\frac{10!}{3! \times 3! \times 1! \times 2! \times 1!}$$

ways to rearrange the word S T A T I S T I C S.

In general, if you have n objects with n_1 type 1's, n_2 type 2's, ..., n_k type k's, there are:

$$\frac{n!}{n_1! \times n_2! \times ... \times n_k!}$$

ways to rearrange the n objects.

4 Chapter 4

In this chapter we will formalize a lot of the concepts we learned in chapter 3, mathematically.

4.1 Set theory for probability

Recall that an event is a subset of a sample space. Therefore, all concepts from set theory learned in MATH 135 extends to events. Let A, B be events of a sample space.

$$
\begin{gathered}
p\left(\text{A } \textit{or} \text{ B occurs}\right) = p\left(A \cup B\right) \\
p\left(\text{A } \textit{and} \text{ B occurs}\right) = p\left(A \cap B\right) \\
p\left(A\right) = 1 - p\left(\bar{A}\right) \\
p\left(\overline{A \cup B}\right) = p\left(\bar{A} \cap \bar{B}\right) \\
p\left(\overline{A \cap B}\right) = p\left(\bar{A} \cup \bar{B}\right) \\
p\left(A \cup B\right) = p\left(A\right) + p\left(B\right) - p\left(A \cap B\right) \\
p\left(A \cup B \cup C\right) = p\left(A\right) + p\left(B\right) + p\left(C\right) - p\left(A \cap B\right) - p\left(A \cap C\right) \\
-p\left(B \cap C\right) + p\left(A \cap B \cap C\right)
\end{gathered}
$$

Note that $p\left(A \cap B\right)$ is commonly abbreviated as $p\left(AB\right)$.

Two events, A and B, are called mutually exclusive if $AB = \emptyset$.
Two events are independent if $p\left(AB\right) = p\left(A\right)p\left(B\right)$.

4.1.1 Addition Rule Revised

Recall that if task 1 can be done in p ways and task 2 can be done in q ways, then the number of ways to do task 1 *or* task 2, assuming they don't overlap, is $p + q$. Lets extend this concept to events. How many ways are there to do A *or* B, that is, $A \cup B$? There are $|A| + |B| - |AB|$ ways. However, if A and B don't overlap (i.e. cannot happen at the same time), then $AB = \emptyset, |AB| = 0$. So there would only be $|A| + |B|$ ways.

In general, we can only use the old addition rule when two events are mutually exclusive, that is, cannot happen at the same time. Otherwise, we need to subtract the number of ways the two events can happen at the same time.

4.1.2 Multiplication Rule Revised

Recall that if task 1 can be done in p ways and task 2 can be done in q ways, then the number of ways to do task 1 *and* task 2, assuming they don't depend on each other, is $p \times q$. Lets extend this concept to events. How many ways are there to do A *and* B, that is, AB? There are $|AB|$ ways. However, if A and B don't depend on each other, then $AB = A \times B$[2], $|AB| = |A||B|$. So there would only be $|A||B|$ ways.

In general, we can only use the old multiplication rule when two events are independent, that is, whether or not one event occurs does not increase/decrease the probability that the other event occurs.

Note that the reasoning for why $AB = A \times B$ when events A and B don't depend on each other is tricky, and a bit unintuitive at first. Consider two events: A = flipping a coin and getting head, B = rolling a dice and getting an odd number. Clearly, $A = \{H\}, B = \{2, 4, 6\}$. We know that flipping a head does not affect whether or not we get an odd number, so the total number of ways we can flip a head *and* roll an odd number should just be all pairs of the two events, $(H, 2), (H, 4), (H, 6)$, that is, $1 \times 3 = 3$ ways. In general, if two events are not independent, then we cannot do this and we need to apply another technique.

4.1.3 Examples using revised rules

E.g. Suppose we roll 2 dice. What is the probability that both dice rolls give an odd number?

We want $p\,(\text{first dice is odd } \textit{and} \text{ second dice is odd})$. Are the two events independent? Whether or not we roll an odd number on the first dice should not increase/decrease the probability of rolling an odd number on the second dice. So they are independent and by the multiplication rule, the total number of ways are: $\frac{3}{6} \times \frac{3}{6} = \frac{1}{4}$.

E.g. Let A = first dice rolls a 3. Let B = the sum of two dice rolls is 7. Are A and B independent?

In this case it is tricky to tell from intuition. We can try to figure out if $p\,(AB) = p\,(A)\,p\,(B)$ and if it does, then A and B are independent. Clearly,

[2] $A \times B$ is the cartesian product of A and B.

$p(A) = \frac{1}{6}$ and $p(B) = \frac{|\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}|}{6^2} = \frac{6}{6^2} = \frac{1}{6}$.
$p(AB) = \frac{|\{(3,4)\}|}{6^2} = \frac{1}{6^2}$.

Clearly, $p(AB) = p(A)\,p(B)$ so A and B are independent.

E.g. Let A = first dice rolls a 3. Let C = the sum of two dice rolls is 8. Are A and C independent?

Lets try to figure out if $p(AC) = p(A)\,p(C)$ and if it does, then A and C are independent.

$p(A) = \frac{1}{6}$ and $p(C) = \frac{|\{(2,6),(3,5),(4,4),(5,3),(6,2)\}|}{6^2} = \frac{5}{6^2}$.
$p(AC) = \frac{|\{(3,5)\}|}{6^2} = \frac{1}{6^2}$.

We see that $p(AC) \neq p(A)\,p(C)$ so A and C are dependent. How could we have arrived to the same conclusion by using intuition? It seems that whether or not the first dice roll is 3 increases/decreases the probability that the total is equal to 8. Suppose the first dice roll is not 3, and instead is 1. Then that makes C impossible since the highest possible total is 7. So then A and C are dependent.

4.2 Conditional Probability

The concept of conditional probability is related to dependence between two events. The probability of an event A, given another event B has happened is:

$$p(A|B) = \frac{p(AB)}{p(B)} \tag{4}$$

Note that if A and B are independent, then this simplifies to $p(A|B) = \frac{p(A)p(B)}{p(B)} = p(A)$. That is, whether or not B occurs does not affect the probability that A occurs.

Suppose that A and B are dependent. Why does the conditional probability of A given B divide by the probability of B? The reason why is because we limit the sample space to B. We are essentially asking the question: what is the probability of event A happening in all those cases where B happens.

E.g. Let A = first dice rolls a 3. Let C = the sum of two dice rolls is 8. What is $p(A|C), p(C|A)$?

Recall $p(A) = \frac{1}{6}$, $p(C) = \frac{5}{6^2}$ and $p(AC) = \frac{1}{6^2}$ where A and C are dependent. Then:

$$p(A|C) = \frac{p(AC)}{p(C)} = \frac{\frac{1}{6^2}}{\frac{5}{6^2}} = \frac{1}{5}$$
$$p(C|A) = \frac{p(CA)}{p(A)} = \frac{\frac{1}{6^2}}{\frac{1}{6}} = \frac{1}{6}$$

Notice that the probability of A increased from $\frac{1}{6}$ to $\frac{1}{5}$. The reason why is because if C occurs, that is, the total is 8, then the first dice roll could not have been a 1 (otherwise the total would be 7). There are only 5 possibilities the first dice roll could be, which is 1 less than the number of possibilities if the total was not 8. Therefore, A is more likely. Also, notice that C is more likely.

In general, dependence is a two way street. If A makes B more likely, then B makes A more likely. Similarly if A makes B less likely, then B makes A less likely.

E.g. Suppose we flip a coin 4 times. Let A = event where we get at least 3 heads. Let B = the event where the first flip is tails. What is $p(A|B)$ and $p(B|A)$?

We could use the definition of conditional probability but instead we will solve this problem much more intuitively. This involves figuring out how many times A and B can occur together and dividing it by how many times B can occur.

How many ways can B occur? The first flip must be a tails and the other 3 flips can be head or tails. So there are 2^3 ways.
Of those 2^3 cases, how many include event A? There is only one case that satisfies both A and B, $\{T, H, H, H\}$. So $p(A|B) = \frac{1}{2^3}$.

Similarly for $p(B|A)$, how many ways can A occur? There are 2 sub cases: getting exactly 3 heads or getting exactly 4 heads. $4+1 = 5$. Of those 5 cases, how many satisfy B? There is only 1, as mentioned above. So $p(B|A) = \frac{1}{5}$.

4.3 Product Rule

Recall for two events A and B, the conditional probability of A given B is: $p(A|B) = \frac{p(AB)}{p(B)}$ and B given A is: $p(B|A) = \frac{p(AB)}{p(A)}$. The product rule is just a rearrangement of conditional probability. The probability of A *and* B is:

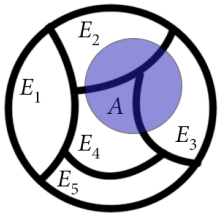
$$p(AB) = p(B)\,p(A|B) = p(A)\,p(B|A) \tag{5}$$

Note that the product rule does make logical sense. The probability of A *and* B is the probability that B occurs times the probability that A occurs given B has occurred, or the probability that A occurs times the probability that B occurs given A has occurred. This intuitive rule is also extendable:

$$p(ABC) = p(A)\,p(B|A)\,p(C|AB)$$

4.4 Partition Rule

Suppose you have a sample space S that can be partitioned into a bunch of subsets that combine to make up S. That is, $S = E_1 \cup E_2 ... \cup E_k$. Any event could have parts of E_i and so we can split it up into those parts.



Then for any event A: $A = AE_1 \cup AE_2 ... \cup AE_k$.

$$p(A) = p(AE_1) + p(AE_2) + ... + p(AE_k) = \sum_{i=1}^{k} (AE_i)$$

$$p(A) = \sum_{i=1}^{k} p(E_i)\,p(A|E_i) \tag{6}$$

Note that a often used partition of S is simply an event B and its complement $\bar{B}$ since $S = B \cup \bar{B}$.

4.5 Bayes Rule

Bayes rule is used to reverse the sign of conditional probability. For example, if we want to find $p(A|B)$ but only have $p(B|A)$, we can use bayes rule.

$$p(A|B) = \frac{p(AB)}{p(B)} = \frac{p(A)\,p(B|A)}{p(B)} \tag{7}$$

4.6 Example using all rules

E.g Suppose 2% of a population have a disease. A blood test for the disease gives a 5% false positive rate and a 1% false negative rate. What is the probability of a random person tests positive for the disease?

The first thing to note is that our sample space will contain the entire population of people. The second thing to note that is we have a partition of the sample space: all the people who have the disease and all the people who don't.

Let D = person has the disease. Let T = person tests positive for the disease. We know $p(D) = 0.02, p(T|\bar{D}) = 0.05, p\left(\bar{T}|D\right) = 0.01$.

$S = D \cup \bar{D}$. So by the partition rule and product rule:

$$\begin{aligned} p(T) &= p(D)\,p(T|D) + p\left(\bar{D}\right)p\left(T|\bar{D}\right) \\ &= 0.02(1-0.01) + 0.98(0.05) = 0.0688 \end{aligned}$$

Suppose that someone tests positive. What is the probability they have the disease?

We want $p(D|T)$. By Bayes rule:

$$\begin{aligned} p(D|T) &= \frac{p(D)\,p(T|D)}{p(T)} \\ &= \frac{0.02(1-0.01)}{0.0688} \\ &= 0.288 \end{aligned}$$

5 Chapter 5

We will now introduce the concept of random variables. As the name implies, random variables are random (i.e. they can take multiple values). The values that a random variable can take is called the range of the random variable.

Convention: X, Y, Z are often abbreviated as random variables and x, y, z are values they can take, respectively. Random variable is often abbreviated as r.v.

Random variables can be discrete or continuous. Discrete random variables are random variables who have a discrete range (finite or countably infinite). Continuous random variables are random variables who have a continuous range (uncountable infinite, like $\mathbb{R}$). For example a random variable that can take a value anywhere between 0 and 1.0 is continuous as there are uncountably infinite values in that range. This chapter will discuss discrete random variables.

The probability that X takes a value x is called the probability function of X:

$$f(x) = P(X = x) \quad \forall x \in X \tag{8}$$

The probability function of a random variable conforms to following axioms:

- $0 \leq f(x) \leq 1$
- $\sum_{x \in X} f(x) = 1$

The probability that X takes a value $\leq$ x is called the cumulative distribution function of X:

$$F(x) = P(X \leq x) \quad \forall x \in \mathbb{R} \tag{9}$$

Note that cumulative distribution function is often abbreviated as c.d.f. Also note that it is defined for all real numbers.

The cumulative distribution function of a random variable conforms to following axioms:

- $0 \leq F(x) \leq 1$

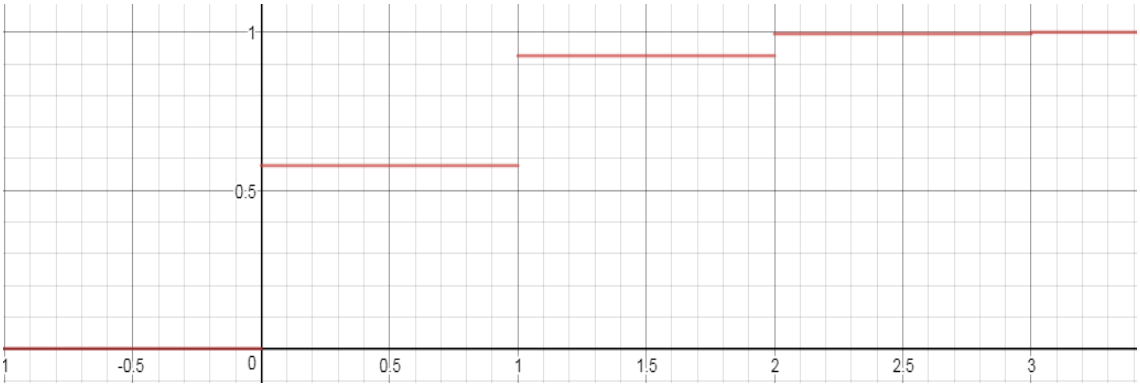
- $\lim_{F(x)\to\infty} F(x) = 1$
- $\lim_{F(x)\to-\infty} F(x) = 0$
- $F(x)$ is non-decreasing

If x_o is the minimum then $F(x_o) = f(x_o)$.
If x_o is the maximum then $F(x_o) = 1$.
$F(x)$ is non-decreasing because it is the sum of the values before it, that is, it is at least $F(x-1)$.

E.g. Graph the c.d.f for the following random variable X = number of 1's in 3 dice rolls:

x	0	1	2	3
$f(x)$	$5^3/6^3$	$4 \cdot (5^2)/6^3$	$15/6^3$	$1/6^3$

$$F(0) = P(X \leq 0) = f(0) = \frac{125}{216}$$
$$F(1) = P(X \leq 1) = f(0) + f(1) = \frac{5^3}{6^3} + \frac{4 \cdot 5^2}{6^3} = \frac{200}{216}$$
$$F(2) = P(X \leq 2) = f(0) + f(1) + f(2) = \frac{200}{216} + \frac{15}{6^3} = \frac{215}{216}$$
$$F(3) = P(X \leq 3) = 1$$



So we see that $F(x)$ is a right continuous step function with discontinuities at each value of x in X's range.

Also note that $f(x) = F(x) - F(x-1)$, that is, $P(X = x) = P(X \leq x) - P(X \leq x - 1)$.

5.0.1 Discrete Uniform Random Variable

Let X be a random variable that takes values in the range $\{a, a+1, ..., b\}$ such that each value is equally likely. Then we say X *is a* discrete uniform random variable:

$$X \sim DU[a, b] \tag{10}$$

E.g. Let X = value on die. Then $X \sim DU[1, 6]$.

What is the probability function of X?

$$f(x) = P(X = x) = \frac{1}{b - a + 1} \tag{11}$$

What is the cumulative distribution function of X?

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < a \\ 1 & x > b \\ \sum_{i=a}^{\lfloor x \rfloor} \frac{1}{b-a+1} = \frac{\lfloor x \rfloor - a + 1}{b-a+1} & a \leq x \leq b \end{cases} \tag{12}$$

Note that, in the discrete uniform case, the cumulative distribution function has "jumps" of equal heights.

5.0.2 Hypergeometric Random Variable

Suppose you have N objects, r of which are successes and $N - r$ are fails. Suppose we draw n objects without replacement and order doesn't matter. Let X be a random variable that counts the number of successes in those n objects. Then we say X *is a* hypergeometric random variable:

$$X \sim Hyper(N, r, n) \tag{13}$$

E.g. In lotto 6/49 six numbers are drawn from $\{1, 2, ..., 49\}$ to be winning numbers. Players draw 6 numbers[3]. Let X = number of winning numbers a player draws. Then $X \sim Hyper(49, 6, 6)$.

What is the probability function of X?

$$f(x) = p(\text{we draw } x \text{ successes } and \ n - x \text{ failures}) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \tag{14}$$

What is the range of X?
The largest number of successes we can draw is either limited by how many successes there are, r, or how many objects were picking, n. The smallest number of successes we draw is either limited by 0 or the maximum number of fails we can draw, $n - (N - r)$.

For example, suppose N = 49, r = 6, n = 3. The highest number of successes we can draw is 3. Now suppose N = 49, r = 6, n = 9. The highest number of successes we can draw is 6.

Similarly, suppose N = 49, r = 6, n = 6. The lowest number of successes we can draw 0. Now suppose N = 10, r = 6, n = 6. The lowest number of successes we can get is 2.

In general, the range is $\text{Max}(0, n - (N - r)) \leq x \leq \text{Min}(n, r)$.
Unfortunately there is no closed form of the cumulative distribution function for hypergeometric random variables.

5.0.3 Binomial Random Variable

The binomial random variable relies on a concept known as bernoulli trials. Bernoulli trials are trials of an experiment that conform under the following axioms:

- Trials are independent.
- Each trial is either a success or a failure.
- The probability of success, p, is constant.

[3]"Draw" often implies without replacement and order doesn't matter.

Suppose there are n trials and the probability of success is p. Let X = number of successes in those n trials. Then we say X is a binomial random variable:

$$X \sim Bin(n, p) \tag{15}$$

E.g. Flipping 10 coins. Let X = number of heads in those 10 flips. Then $X \sim Bin(10, 0.5)$.

What is the probability function of X?

$$f(x) = p(\text{we get } x \text{ successes } \textit{and } n - x \text{ failures}) = \binom{n}{x} p^x (1-p)^{n-x} \tag{16}$$

Note that $p^x(1-p)^{n-x}$ is the probability of getting x successes and $n-x$ fails in a row. That is, $\underbrace{pp...p}_{\text{x times}} \underbrace{(1-p)(1-p)...(1-p)}_{\text{n - x times}}$. In reality, those successes and failures can occur in any order, as long as they add up to x successes and $n-x$ failures. So we need to account for all the ways to place x successes in n trials (with the rest of the trials being failures) $\binom{n}{x}$.

What is the range of X?
The smallest number of successes we get can get in n trials is 0. The largest number of successes we can get is n. So the range is $\{0, 1, 2, 3, ..., n\}$.

Unfortunately there is no closed form of the cumulative distribution function for binomial random variables.

5.0.4 Approximation of Hypergeometric using Binomial

Suppose we have N objects, r of which are successes and $N-r$ are fails. Suppose we draw n objects without replacement and order doesn't matter and $n << N$. That is, the number of objects that were drawing is much smaller than the total number of objects. Let X = number of successes in n objects.

Then, whether we sample with or without replacement doesn't really matter because it is extremely unlikely to get the same object twice.

If we treat the problem without replacement then $X \sim Hyper(N, r, n)$.

If we treat the problem with replacement then $X \sim Bin(n, \frac{r}{N})$.

5.0.5 Negative Binomial Random Variable

Suppose there are n trials and the probability of success is p. Let X = number of failures before the kth success. Then we say X is a negative binomial random variable:

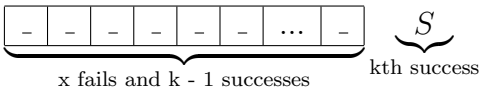
$$X \sim NB(k, p) \tag{17}$$

E.g. Roll a dice until you roll 6 1's. Let X = number of non 1's before the 6th 1, $X \sim NB(6, \frac{1}{6})$.

E.g. Suppose the probability of passing a course for a student is p and courses are independent. Let X = number of failed courses before passing 40 courses, $X \sim NB(40, p)$.

What is the probability function of X?

$$\begin{aligned} f(x) &= p(x \text{ fails before kth success}) \\ &= p(x \text{ fails } \textit{and} \text{ k - 1 successes before kth success}) \end{aligned}$$



$$f(x) = \binom{x+k-1}{x} p^k (1-p)^x = \binom{x+k-1}{k-1} p^k (1-p)^x \tag{18}$$

Note that $p^k(1-p)^x$ is the probability of getting k successes and x fails in a row. That is, $\underbrace{pp...p}_{\text{k times}} \underbrace{(1-p)(1-p)...(1-p)}_{\text{x times}}$. In reality, those successes and failures can occur in any order, as long as they add up to x fails and $k-1$ successes. So we need to account for all the ways to place x fails in $x+k-1$ trials (with the rest of the trials being successes) $\binom{x+k-1}{x}$.

What is the range of X?
The smallest number of fails before the kth success is 0. However, in theory, there is no maximum number of fails before the kth success. So the range is $\{0, 1, 2, 3, ...\}$.

E.g. Suppose a startup is looking for 5 investors. Each investor independently agrees with probability 0.2. Founders ask one investor at a time until 5 investors say yes. How many investors do they ask?

Let X = number of investors that say no before the 5th investor says yes. Then $X \sim NB(5, 0.2)$. Let Y = total number of investors they ask. Then $Y = X + 5$.

$$\begin{aligned} f(x) = P(X = x) &= P(Y - 5 = x) \\ &= P(Y = x + 5) = \binom{x + 5 + 5 - 1}{x + 5} p^5 (1-p)^{x+5} \\ &= \binom{x - 9}{x + 5} p^5 (1-p)^{x+5} \end{aligned}$$

E.g. Suppose someone sends a bit stream of 0's and 1's over a noisy connection. The probability that a bit gets flipped is 0.01. Find the probability it takes 50 bits to observer 5 errors.

Let X = number of non-flipped bits before the 5th error. Then $X \sim NB(5, 0.01)$.

$$\begin{aligned} p(\text{takes 50 bits to observe 5 errors}) &= P(X = 45) \\ &= \binom{49}{4} 0.01^5 (1 - 0.01)^{45} \end{aligned}$$

5.0.6 Geometric Random Variable

The geometric random variable is just a special case of the negative binomial random variable. Suppose there are n trials and the probability of success is p. Let X = number of fails before the first success. Then we say X is a geometric random variable:

$$X \sim Geo(p) \tag{19}$$

What is the probability function of X?

$$f(x) = p(x \text{ fails before first success}) = p(1-p)^x \tag{20}$$

What is the range of X?
The smallest number of fails before a success is 0. However, just like the negative binomial random variable, in theory, there is no maximum number of fails before a success. So the range is $\{0, 1, 2, 3, ...\}$.

What is cumulative distribution function of X?

$$\begin{aligned} F(x) = P(X \leq x) &= 1 - P(X \geq x+1) \\ &= 1 - (f(x+1) + f(x+2) + ...) \\ &= 1 - (p(1-p)^{x+1} + p(1-p)^{x+2} + ...) \\ &= 1 - \frac{p(1-p)^{x+1}}{1-(1-p)} \text{by geometric series} \\ &= 1 - (1-p)^{x+1} \end{aligned}$$

E.g. Suppose a safe has a 10 digit code. You randomly try keys with replacement (don't keep track of keys you already tried). What is the probability of cracking the safe in 11 tries or fewer?

Let X = number of failed attempts before unlocking the safe. Then $X \sim Geo(\frac{1}{10^{10}})$.

$$F(11) = 1 - (1 - \frac{1}{10^{10}})^{11}$$

5.0.7 Poisson Random Variable

The poisson random variable relies on a concept known as a poisson process. A poisson process is a process where events occur randomly with continuous time and:

- The number of events in non-overlapping time intervals are independent
- Events occur one at a time (probability of two or more events occurring at the same time is 0)
- Events occur at a constant average rate, λ

Suppose you have a poisson process and rate λ and Let X = number of events that occur in t units of time. Then we say X is a poisson random variable:

$$X \sim Poi(\mu = \lambda t) \tag{21}$$

Note: Make sure λ and t are measured in the same time units.

The poisson process is often used to model:

- Births in a large population
- Number of requests to a server
- Number of collisions in a nuclear reaction
- Traffic accidents

The probability function of X is:

$$f(x) = \frac{e^{-\mu}\mu^x}{x!} \tag{22}$$

What is the range of X?
The smallest number of events that occur in t units of time is 0. There is no maximum so the range is $\{0, 1, 2, 3, ...\}$.

E.g. Server requests come according to a poisson process with rate 100 $\frac{requests}{minute}$. Find the probability of 1 request and 1 second.

Let X = number of requests that come in one second. Then $X \sim Poi(\frac{100 \text{ requests}}{60 \text{ secs}} *$ 1 sec) $= Poi(\frac{100}{60})$.

$$P(X = 1) = \frac{e^{-\frac{100}{60}} \frac{100}{60}^1}{1!}$$

5.0.8 Approximation of Binomial using Poisson

Suppose you have n trials and the probability of success is p. Let X = number of successes in those n trials. Suppose also that $n \to \infty$ and $p \to 0$. That is, the number of trials goes to infinity and the probability of success goes to 0.

Then X can be modeled by a poisson random variable: $X \sim Poi(\mu = np)$.

Lets prove this by showing that if we impose the conditions of $n \to \infty$ and $p \to 0$ on the binomial random variable, it becomes the poisson random variable:

$$\begin{aligned}
f(x) &= \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{n(n-1)(n-2)...(n-x+1)}{x!} \frac{\mu^x}{n^x} \left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^x}{x!} \lim_{n\to\infty} \frac{n(n-1)(n-2)...(n-x+1)}{n^x} \left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^x}{x!} \lim_{n\to\infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} ... \frac{n-x+1}{n} \left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^x}{x!} \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n-x}
\end{aligned}$$

Recall from calculus: $\lim_{n\to\infty} \left(1+\frac{a}{n}\right) = e^a$.

$$f(x) = \frac{e^{-\mu}\mu^x}{x!}$$

5.0.9 Important Series

Memorize the following:

Geometric Series: If $|r| < 1$:

$$\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}$$

$$\sum_{k=0}^{n} ar^k = \frac{a(1-r)^{n+1}}{1-r}$$

Binomial Theorm:

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$$

Hypergeometric Identity:

$$\binom{a+b}{n} = \sum_{k=0}^{n} \binom{a}{k} \binom{b}{n-k}$$

Exponential Series:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

6 Chapter 6

Not covered in STAT 230.

7 Chapter 7

Knowing the entire distribution is nice but sometimes we want a few numbers that summarize the distribution.

7.0.1 Summarizing Data

E.g. Let X = number of kids in a family.
Suppose a group was surveyed and it was found that 5 families had 1 kid, 12 families had 2 kids, 7 families had 3 kids and 2 families had 4 kids.

Sample mean (denoted $\overline{X}$): The sample mean of a random variable is the average of the random variable. It is called the sample mean because it is often different for every sample. The sample mean of X is the average number of kids in a family.

$$\begin{aligned}\overline{X} &= \frac{\text{Total number of kids}}{\text{Total number of families}} \\ &= \frac{1 \cdot 5 + 2 \cdot 12 + 3 \cdot 7 + 4 \cdot 2}{5 + 12 + 7 + 2} \\ &= 2.23\end{aligned}$$

Median: The median is also called the middle value. The median of a random variable is the middle sorted value. The median of X is the middle value if we sorted the number of kids from lowest to highest. Notice that X has two middle values since it is an even number.

$$Median(X) = \frac{2+2}{2} = 2$$

Mode: The mode is the most common value. The mode of X is the most common number of kids for a family to have.

$$Mode(X) = 2$$

7.0.2 Expected Value

The expected value of a random variable is its mean, that is, the average value of the random variable.

Recall from the above example that the sample mean of X, the number of kids in a family is:

$$\begin{aligned}\overline{X} &= \frac{\text{Total number of kids}}{\text{Total number of families}} \\ &= \frac{1 \cdot 5 + 2 \cdot 12 + 3 \cdot 7 + 4 \cdot 2}{5 + 12 + 7 + 2} \\ &= \frac{1 \cdot 5 + 2 \cdot 12 + 3 \cdot 7 + 4 \cdot 2}{26} \\ &= 1.8\end{aligned}$$

We can rearrange this expression to get a much simpler form:

$$\begin{aligned}\overline{X} &= 1 \cdot \frac{1}{26} + 2 \cdot \frac{12}{26} + 3 \cdot \frac{7}{26} + 4 \cdot \frac{2}{26} \\ &= \sum_{x=1}^{4} x f(x)\end{aligned}$$

So we can think of the the sample mean (and thus the mean) as the weighted average of x with the weights being the probabilities.

In general, for a random variable X, the expected value of X is:

$$E[X] = \sum_{\text{all x}} x f(x) = \mu \tag{23}$$

Just like the sample mean, we can think of the expected value of X as the weighted average of x with the weights being the probabilities that x occurs.

E.g. Let X = number of kids in a family with the probability function:

X	1	2	3	4	5
$f(x)$	0.43	0.4	0.12	0.04	0.01

Suppose we have another function, $g(X)$, that determines the tax credit for a family with x amount of kids where $g(X) = 1000 + 250X$. What is the

average tax credit for a family, $E[g(X)]$?

Just like $E[X]$ is the weighted average of x with the weights being the probabilities that x occurs, $E[g(X)]$ is the weighted average of g(x) with the weights being the probabilities that x occurs.

$$\begin{aligned} E[g(X)] = &(\text{tax credit for 1 kid})(\text{probability of having 1 kid})+ \\ &(\text{tax credit for 2 kid})(\text{probability of having 2 kid})+ \\ &(\text{tax credit for 3 kid})(\text{probability of having 3 kid})+ \\ &(\text{tax credit for 4 kid})(\text{probability of having 4 kid}) \\ = 1250 \cdot 0.43 + 1500 \cdot 0.4 + 1750 \cdot 0.12 + 2000 \cdot 0.04 + 2250 \cdot 0.01 = 1450 \end{aligned}$$

However, notice that we could have gotten $E[g(X)]$ in a much easier way:

$$\begin{aligned} E[g(X)] &= E[1000 + 250X] \\ &= 1000 + 250E[X] \\ &= 1000 + 250 \cdot 1.8 = 1450 \end{aligned}$$

E.g. Now suppose $g(X) = \frac{2000}{X}$. What is the average tax credit for a family, $E[g(X)]$?

$$E[g(X)] = \frac{2000}{1} \cdot 0.43 + \frac{2000}{2} \cdot 0.4 + \frac{2000}{3} \cdot 0.12 + \frac{2000}{4} \cdot 0.04 + \frac{2000}{5} \cdot 0.01 = 1364$$

Could we have gotten $E[g(X)]$ in a much easier way?

$$\begin{aligned} E[g(X)] &= E\left[\frac{2000}{X}\right] \\ &= \frac{2000}{E[X]} \\ &= \frac{2000}{1.8} = 1111 \neq 1364 \end{aligned}$$

The reason why it worked in the first example is because expectation is a linear operator. That is, $E[aX + b] = aE[X] + b$.

Lets prove this using the definition of expectation:

$$
\begin{aligned}
E[aX+b] &= \sum_{\text{all x}} (ax+b) f(x) \\
&= \sum_{\text{all x}} axf(x) + \sum_{\text{all x}} bf(x) \\
&= a \sum_{\text{all x}} xf(x) + b \sum_{\text{all x}} f(x) \\
&= aE[X] + b
\end{aligned}
$$

In general:

$$E[aX+b] = aE[X] + b \tag{24}$$

$$E[g(X)] = g(E[X]) \text{ if } g(X) \text{ linear} \tag{25}$$

7.0.3 Expectations of Named Distributions

Suppose $X \sim Bin(n, p)$. What is $E[X]$?

$$
\begin{aligned}
E[X] &= \sum_{x=0}^{n} x f(x) \\
&= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{xn!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!((n-1)-(x-1))!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!((n-1)-(x-1))!} p p^{x-1} (1-p)^{(n-1)-(x-1)} \\
&= np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} \text{ Let } y = x-1 \\
&= np \underbrace{\sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{(n-1)-y}}_{\text{This is the probability function of } Y \sim Bin(n-1,p) \text{ summed over its range}} \\
&= np
\end{aligned}
$$

Note that the result also makes logical sense. If we have n trials and the probability of success is p. The expected number of successes should be np. It should be proportional to n and p. As the number of trials increases so does the expected number of successes. As the probability of success increases so does the expected number number of successes.

Suppose $X \sim Poi(\mu)$. What is $E[X]$?

$$\begin{aligned}
E[X] &= \sum_{x=0}^{\infty} x f(x) \\
&= \sum_{x=0}^{\infty} x \frac{e^{-\mu}\mu^x}{x!} \\
&= \sum_{x=1}^{\infty} \frac{e^{-\mu}\mu^x}{(x-1)!} \text{ Let } y = x-1 \\
&= \mu \underbrace{\sum_{y=0}^{\infty} \frac{e^{\mu}\mu^y}{y!}}_{\text{This is the probability function of } Y \sim Poi(\mu) \text{ summed over its entire range}} \\
&= \mu
\end{aligned}$$

Note that the result also makes logical sense. If we had a poisson process with rate λ and were counting the number of events in t units of time, it makes sense that the expected number of events is λt. It should be proportional to λ and t. As the average rate that events occur increases so does the expected number of events. As we watch for longer periods of time, the expected number of events increases.

The expected value for other named distributions can be found on the formula sheet.

7.0.4 Variance

We want a way to measure how spread out a distribution is around its mean value. That is, we want to know how much it *varies* around its mean value. The variance of a random variable X is defined to be:

$$Var(X) = E[(X - E[X])^2] = \sigma^2 \tag{26}$$

The variance of a random variable X is the average squared distance from the mean. The reason why we define the variance to be the squared distance from the mean is because we want to work with the absolute value. However, we cannot define variance to be the absolute value because the function is undefined at x = 0, which is troublesome to deal with. Also note that the variance

is always non-negative.

An alternate definition of variance can be derived if we simplify:

$$\begin{aligned} Var\,(X) &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - 2E[X]^2 + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

So the variance can also be defined as follows:

$$Var\,(X) = E[X^2] - E[X]^2 \tag{27}$$

Recall to find $E[X^2]$:

$$E[X^2] = \sum_{\text{all x}} x^2 f\,(x)$$

E.g. Let X = time to complete coding with the probability function:

X	10	12	20	22
$f\,(x)$	0.72	0.08	0.18	0.02

What is the $Var\,(X)$?

Recall $Var\,(X) = E[X^2] - E[X]^2$.

$$\begin{aligned} E[X^2] &= \sum_{\text{all x}} x^2 f\,(x) \\ &= 10^2 \cdot 0.72 + 12^2 \cdot 0.08 + 20^2 \cdot 0.18 + 22^2 \cdot 0.02 \\ &= 165.2 \text{ minute}^2 \end{aligned}$$

$$\begin{aligned} E[X] &= \sum_{\text{all x}} x f\,(x) \\ &= 10 \cdot 0.72 + 12 \cdot 0.08 + 20 \cdot 0.18 + 22 \cdot 0.02 \\ &= 12.2 \end{aligned}$$

$$Var\,(X) = 165.5 - 12.2^2 = 16.36 \text{ minute}^2$$

Notice that the variance is always in squared units of X. This can be difficult to interpret so we have another unit of measurement, called the standard deviation. The standard deviation of a random variable X is:

$$SD\,(X) = \sqrt{Var\,(X)} = \sigma \tag{28}$$

The standard deviation of X in the previous example is:

$$Var\,(X) = \sqrt{16.36} = 4.04 \text{ minutes}$$

Question: What is $Var\,(Y = aX + b)$?

$$\begin{aligned} Var\,(Y = aX + b) &= E[(Y - E[Y])^2] \\ &= E[(aX + b - E[aX + b])^2 \\ &= E[(aX + b - aE[X] - b)^2] \\ &= E[(aX - aE[X])^2] \\ &= E[a^2(X - E[X])^2] \\ &= a^2 E[(X - E[X])^2] \\ &= a^2 Var\,(X) \end{aligned}$$

In general:

$$Var\,(aX + b) = a^2 Var\,(X) \tag{29}$$

Notice that b does not affect the variance. The reason why is because shifting the distribution by a constant doesn't affect the spread of the data at the mean.

7.0.5 Variances of Named Distributions

Suppose $X \sim Bin(n, p)$. What is $Var\,(X)$?

$$\begin{aligned} Var\,(X) &= E[X^2] - E[X]^2 \\ &= E[X^2 - X + X] - E[X]^2 \\ &= E[X(X-1) + X] - E[X]^2 \\ &= E[X(X-1)] + E[X] - E[X]^2 \end{aligned}$$

$$\begin{aligned} E[X(X-1)] &= \sum_{x=0}^{n} x(x-1) f\,(x) \\ &= \sum_{x=2}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \sum_{x=2}^{n} \frac{n(n-1)(n-2)!}{(x-2)!(n-x)!} p^x (1-p)^{n-x} \\ &= \sum_{x=2}^{n} \frac{n(n-1)(n-2)!}{(x-2)!((n-2)-(x-2))!} p^2 p^{x-2} (1-p)^{(n-2)-(x-2)} \\ &= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!((n-2)-(x-2))!} p^{x-2} (1-p)^{(n-2)-(x-2)} \\ &= n(n-1)p^2 \sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} (1-p)^{(n-2)-(x-2)} \text{ Let } y = x-2 \\ &= n(n-1)p^2 \underbrace{\sum_{y=0}^{n-2} \binom{n-2}{y} p^y (1-p)^{(n-2)-y}}_{\text{This is the probability function of } Y \sim Bin(n-2,p) \text{ summed over its entire range}} \\ &= n(n-1)p^2 \end{aligned}$$

$$\begin{aligned} Var\,(X) &= n(n-1)p^2 + np - (np)^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= -np^2 + np \\ &= np(1-p) \end{aligned}$$

Suppose $X \sim Poi(\mu)$. We find that the variance of X is: $Var\,(X) = \mu$.

The variance for other named distributions can be found on the formula sheet.

8 Chapter 8

We will now introduce the concept of continuous random variables. Continuous random variables are random variables with a uncountably infinite range.

E.g. Let X = temperature on a given day. Given infinite precision, there are uncountably infinite possibilities.

E.g. Let Y = heights of a population.

In general, if the thing that are you are modeling can be measured with infinite precision, it can be modeled by a continuous random variable.

What is the probability function for a continuous random variable, X?
Since the range is uncountably infinite, $f(x) = P(X = x) = 0, \forall x$. Therefore the probability function is useless when dealing with continuous random variables.

The cumulative distribution function for a continuous random variable is the same, $F(x) = P(X \leq x)$. However, $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = F(b) - F(a)$, since $P(X = a) = P(X = b) = 0$. That is, the end points don't matter when dealing with continuous random variables.

Also note that the cumulative distribution function for a continuous random variable is actually continuous. Recall that the c.d.f for discrete random variables had discontinuities at each value of x where the heights would correspond to $f(x)$. However, since $f(x) = 0, \forall x$, each of those jump heights are now equal to 0 and the c.d.f is a continuous function.

To describe continuous random variables, we define another function called the probability density function that is the derivative of the cumulative distribution function, $f(x) = F'(x)$. The properties of $f(x)$ is as follows:

- $f(x) \geq 0$ since $F(x)$ is a non-decreasing function. Note that the p.d.f can be bigger than 1 so it should not be interpreted as a probability
- $$F(x) = \int_{-\infty}^{x} f(x)\, dx$$

- $$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)\,dx$$
- $$\int_{\infty}^{\infty} f(x) = 1$$

E.g. Suppose you have a continuous random variable on range 0 - 1:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 & x > 1 \\ x^2 & 0 \leq x \leq 1 \end{cases}$$

Find $f(x)$, $P(X = 0.25)$, $P(X \leq 0.25)$:

$$P(X = 0.25) = 0$$
$$f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & elsewhere \end{cases}$$
$$P(X \leq 0.25) = F(0.25) = 0.25^2$$
$$P(X \leq 0.25) = \int_0^{0.25} 2x dx = x^2|_0^{0.25} = 0.25^2$$

8.0.1 Expected Value and Variance

Rule of thumb: when working with continuous random variables, replace $\sum$ with $\int_{-\infty}^{\infty}$ and use the probability density function instead of the probability function.

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx \tag{30}$$
$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,dx \tag{31}$$
$$Var(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \left(\int_{-\infty}^{\infty} x f(x)\,dx\right)^2 \tag{32}$$

8.0.2 Percentiles

The pth percentile of a random variable in the value x_p such that $P(X \leq x_p) = p$.

E.g. Suppose X has the p.d.f function: $f(x) = cx^2, 0 < x < 2$ and 0 otherwise. Find c and $E[X]$.

$$\int_0^2 cx^2 dx = 1 = c\left[\frac{x^3}{3}\right]_0^2 = \frac{8c}{3} = 1 \rightarrow c = \frac{3}{8}$$

$$E[X] = \int_0^2 x\frac{3}{8}x^2 dx = \frac{3}{8}\left[\frac{x^4}{4}\right]_0^2 = 1.5$$

8.0.3 Uniform Random Variable

Let X be a continuous random variable that takes real values between (a, b) such that all sub-intervals of a fixed length are equally likely. Then we say X is a uniform random variable:

$$X \sim U(a, b) \tag{33}$$

Note that the uniform random variable is not defined as: every value in (a, b) is equally likely because $P(X = x) = 0, \forall x$. So, if we did define it that way, every continuous random variable would be a uniform random variable.

What is the probability density function?
Since every sub-interval of fixed length is equally likely, $F(x)$ grows linearly. So $f(x) = F'(x) = c$ (constant).

$$\int_a^b cdx = 1 = c(b-a) \rightarrow f(x) = c = \frac{1}{b-a}$$

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

What is the cumulative distribution function?

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{a}^{x} \frac{1}{b-a}dx = \frac{1}{b-a}x|_a^x = \frac{x-a}{b-a}$$
$$= \begin{cases} \frac{x-a}{b-a} & a < x < b \\ 0 & x < a \\ 1 & x > b \end{cases}$$

What is the expected value and variance?

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{a}^{b} x\frac{1}{b-a}dx$$
$$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)}$$
$$= \frac{b+a}{2}$$
$$Var(X) = E[X^2] - E[X]^2 = ... = \frac{(b-a)^2}{12}$$

E.g. Suppose there is a spinner that can land on a value been 0 and 4. Let X = value spinner lands on. Then $X \sim U(0,4)$, $f_x(x) = \frac{1}{4}$, 0 otherwise, $F_x(x) = \frac{x}{4}$, $0 < x < 4$. Let Y = $\frac{1}{X}$. Find the p.d.f of Y.

What is the range of Y?
We know the range of X is (0, 4). So the range of Y is $(-\infty, \frac{1}{4})$.

$$F_Y(y) = P(Y \le y) = P\left(\frac{1}{X} \le y\right)$$
$$= P\left(X \ge \frac{1}{y}\right) = 1 - P\left(X < \frac{1}{y}\right)$$
$$= 1 - P\left(X \le \frac{1}{y}\right) = 1 - F_x\left(\frac{1}{y}\right)$$

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}\left(1 - F_x\left(\frac{1}{y}\right)\right)$$
$$= -F_x\left(\frac{1}{y}\right) * -\frac{1}{y^2} = \frac{1}{4y}\frac{1}{y^2} = \frac{1}{4y^2}, 0 < \frac{1}{y} < 4$$
$$= \frac{1}{4y^2}, 0 < 1 < 4y \rightarrow y > \frac{1}{4}$$

$$f_Y(y) = \begin{cases} \frac{1}{4y^2} & y > \frac{1}{4} \\ 0 & otherwise \end{cases}$$

8.0.4 Exponential Random Variable

Suppose you have a poisson process with rate λ. Instead of counting how many times an event occurs in a fixed amount of time (which is discrete), we count the waiting time until the next event happens. Let X be the waiting time until the next event. Then we say X has a exponential distribution:

$$X \sim Exp(\theta = \frac{1}{\lambda}) \tag{34}$$

What is the cumulative distribution function of X?

$$F(x) = P(X \leq x) = 1 - P(X \geq x)$$
$$= 1 - p(\text{time until the next event occurs is} > \text{x})$$
$$= 1 - p(\text{no events between 0 and x})$$

Let Y = number of events between 0 and x. Then Y is a discrete random variable: $Y \sim Poi(\lambda x)$.

$$F(x) = 1 - p(\text{no events between 0 and x})$$
$$= 1 - \frac{e^{-\lambda x}(\lambda x)^0}{0!}$$
$$= 1 - e^{\lambda x} = 1 - e^{-\frac{x}{\theta}}$$
$$= \begin{cases} 1 - e^{-\frac{x}{\theta}} & x > 0 \\ 0 & otherwise \end{cases}$$

What is the probability density function of X?

$$f(x) = F'(x) = \begin{cases} \frac{1}{\theta}e^{-\frac{x}{\theta}} & x > 0 \\ 0 & otherwise \end{cases}$$

What is the expected value of X?

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)\,dx = \int_{0}^{\infty} xf(x)\,dx \\ &= \int_{0}^{\infty} \frac{x}{\theta}e^{-\frac{x}{\theta}}dx \end{aligned}$$

We could use integration by parts. However, lets instead use a gamma function. The gamma function is used when you are integrating a polynomial * exponential.

$$\Gamma(\alpha) = \int_{0}^{\infty} x^{\alpha-1}e^{-x}dx \tag{35}$$

There are several useful properties of the gamma function that helps avoid doing integration by parts.

- $\Gamma(\alpha) = (\alpha-1)!$ if α is an integer and $\alpha \geq 1$
- $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ if α is not an integer
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

$$\begin{aligned} E[X] &= \int_{0}^{\infty} \frac{x}{\theta}e^{-\frac{x}{\theta}}dx \text{ Let } y = \frac{x}{\theta} \\ &= \int_{0}^{\infty} ye^{-y}\theta dy \\ &= \theta\int_{0}^{\infty} y^{2-1}e^{-y}dy = \theta(2-1)! = \theta \end{aligned}$$

So we see that $E[X]$ is proportional to $\frac{1}{\lambda}$. Notice that this result makes logical sense. As the rate that events occur increases, the expected waiting time

decreases. As the rate that events occur decreases, the expected waiting time increases.

What is the variance and standard deviation of X?

$$\begin{aligned} E[X^2] &= \int_0^\infty \frac{x^2}{\theta} e^{-\frac{x}{\theta}} dx \text{ Let } y = \frac{x}{\theta} \\ &= \int_0^\infty \theta^2 y^2 e^{-y} dy \\ &= \int_0^\infty \theta^2 y^{(3-1)} e^{-y} dy \\ &= \theta^2 (3-1)! = 2\theta^2 \end{aligned}$$

$$\begin{aligned} Var\,(X) &= E[X^2] - E[X]^2 \\ &= 2\theta^2 - \theta^2 = \theta^2 \\ SD\,(X) &= \theta \end{aligned}$$

E.g. Suppose buses arrive to a stop according to a poisson process with rate $\lambda = 5\frac{buses}{hour} = \frac{1}{12}\frac{buses}{minutes}$. What is the probability you wait at least 15 minutes for a bus to arrive? What is the probability you wait at least 21 minutes given you've already wait 6 minutes?

$$\begin{aligned} P\,(X \geq 15) &= 1 - P\,(X < 15) \\ &= 1 - (1 - e^{-\frac{15}{12}}) = 0.2865 \end{aligned}$$

$$\begin{aligned} P\,(X \geq 21 | X \geq 6) &= \frac{P\,(X \geq 21 \text{ and } X \geq 6)}{P\,(X \geq 6)} = \frac{P\,(X \geq 21)}{P\,(X \geq 6)} \\ &= \frac{1 - F\,(21)}{1 - F\,(6)} \\ &= \frac{1 - (1 - e^{-\frac{21}{12}})}{1 - (1 - e^{-\frac{6}{12}})} = 0.2865 \end{aligned}$$

It seems all the time we waited was irrelevant in getting the next bus to come sooner. In general, for the exponential distribution, $P\,(X > s + t | X > s) =$

$P(X > t)$. This is called the memory less property. The memory less property means the time already waited is irrelevant in determining the future waiting time. This means that the exponential distribution is not useful to model any process that has a "wearing down" effect. For example, if the exponential distribution was used to model human life, the probability that a person would live 1 additional year would be the same for any age.

8.0.5 Normal Random Variable

The normal random variable is extremely useful. Many real life phenomenons follow the normal distribution such as heights of a population, weights of a population, grades etc. We say X follows the normal distribution, $X \sim N(\mu, \sigma^2)$, if it has probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \forall x - \infty < x < \infty \tag{36}$$

There are a few properties of the normal random variable:

- The expected value is μ
- The variance is σ^2
- $f(x)$ is symmetric around μ
- $f(x) \to 0$ as $x \to \pm\infty$

When the probability density function for a normal random variable is graphed it looks like a bell curve shape, centered around μ. Approximately 68% of the area is within 1 standard deviation of the mean, $x \pm \sigma$. Approximately 95% of the area is within 2 standard deviations of the mean, $x \pm 2\sigma$.

There is no closed form of the cumulative distribution function so we use a table to evaluate the c.d.f. The table can be found on the formula sheet. To calculate $F(2.15) = 0.98422$, look at row 2.1 and column 0.05.

There is also a special case of the normal distribution, called the standard normal distribution. We say Z follows the standard normal distribution if it follows the normal distribution and has $\mu = 0$ and $\sigma^2 = 1$, $Z \sim N(0, 1)$. The probability density function of a standard normal random variable is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}z^2} \tag{37}$$

The table given on the formula sheet shows the c.d.f of $Z \sim N(0,1)$.

One of the most useful properties of the normal distribution is its symmetry. For example, since the standard normal distribution is symmetric around 0, $P(Z \geq x) = P(Z \leq -x)$.

E.g. $P(Z \leq -0.51) = P(Z \geq 0.51) = 1 - P(Z \leq 0.51)$

E.g. Find the 87th percentile of the standard normal distribution.

We want a value c such that $P(Z \leq c) = 0.87$. The formula sheet also provides the inverse function for the c.d.f which we can use to answer this question. The regular c.d.f takes an x and gives a probability p. The inverse function of the c.d.f takes a probability p and gives the value x. So we find $c = 1.1264$.

E.g. Let Z = noise on a connection where $Z \sim N(0,1)$. Find a value z such that 95% of the time a random z is within $(-z, z)$.

We want $P(-z \leq Z \leq z) = 0.95$. Notice that if the range (-z, z) takes 95% of the area, the remaining area is 5%. That means $P(Z \leq z) = 0.975 \rightarrow z = 1.96$.

Suppose we have a normal random variable and we want $F(x)$. How can we do this with just the c.d.f of a standard normal random variable?

It turns out there is a linear transformation we can apply to any normal random variable to make a standard normal random variable. If $X \sim N(\mu, \sigma^2)$ then:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1) \tag{38}$$

Lets prove this using the definition of a normal random variable:

$$\begin{aligned} F_Z(z) = P(Z \leq z) &= P\left(\frac{X - \mu}{\sigma} \leq z\right) \\ &= P(X \leq z\sigma + \mu) \end{aligned}$$

$$
\begin{aligned}
f_z(z) &= F_z'(z) \text{ w.r.t. z} \\
&= F_X'(z\sigma+\mu)\cdot\sigma = f_X(z\sigma+\mu)\cdot\sigma \\
&= \frac{1}{\sqrt{2\pi}\cancel{\sigma}} e^{-\frac{1}{2}\left(\frac{z\cancel{\sigma}+\mu-\mu}{\cancel{\sigma}}\right)^2}\cdot\cancel{\sigma} \\
&= \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}z^2}
\end{aligned}
$$

E.g. Heights of adult men follow a normal distribution, $X \sim N(69, 2.4^2)$. Find the probability a random man is taller than 75 inches. Find the range (a, b) such that 80% of men are within that range.

$$
\begin{aligned}
P(X \geq 75) &= P\left(\frac{X-69}{2.4} \geq \frac{75-69}{2.4}\right) \\
&= P(Z \geq 2.5) \\
&= 1 - P(Z < 2.5) = 1 - 0.99379 = 0.00621
\end{aligned}
$$

We want $P(a \leq X \leq b) = 0.8$. Notice that if the range (a, b) takes 80% of the area, the remaining area is 20%. That means $P(X \leq b) = 0.9$.

$$
\begin{aligned}
P(X \leq b) &= 0.9 \\
P\left(Z \leq \frac{b-69}{2.4}\right) &= 0.9 \\
\frac{b-69}{2.4} &= 1.2816 \rightarrow b = 72.07
\end{aligned}
$$

$$
\begin{aligned}
P(X \leq a) &= 0.1 \\
P\left(Z \leq \frac{a-69}{2.4}\right) &= 0.1 \\
P\left(Z \geq \frac{69-a}{2.4}\right) &= 0.1 \\
P\left(Z < \frac{69-a}{2.4}\right) &= 0.9 \\
\frac{69-a}{2.4} &= 1.2816 \rightarrow b = 65.92
\end{aligned}
$$

9 Chapter 9

We have models for single random variables whether they are discrete or continuous random variables. However, now we want to model several random variables together. In this chapter we will discuss discrete multivariate random variables.

For two random variables X and Y, the joint probability function (j.p.f) is :

$$f(x,y) = P(X = x, Y = y), \forall (x,y). \tag{39}$$

There are some properties of $f(x,y)$:

- $0 \leq f(x,y) \leq 1$
- $\sum\limits_{\text{all x}} \sum\limits_{\text{all y}} f(x,y) = 1$

E.g. Suppose a coin is tossed 3 times.

Let X = number of heads. Let Y = $\begin{cases} 1 & \text{Got H on first toss} \\ 0 & \text{otherwise} \end{cases}$

$f(x,y)$	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

What is $P(Y = 0)$? We can sum over all x. $P(Y = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 0) + P(X = 2, Y = 0) + P(X = 3, Y = 0) = \frac{1}{2}$.

For two random variables X and Y, the marginal probability function (m.p.f) of X and Y is:

$$f_X(x) = P(X = x) = \sum_{\text{all y}} f(x,y) \tag{40}$$

$$f_Y(y) = P(Y = y) = \sum_{\text{all x}} f(x,y) \tag{41}$$

The marginal probability function of a random variable is the probability function of that random variable.

Two random variables X and Y are independent iff: $f(x,y) = f_X(x) f_Y(y)$.

The conditional probability of X given Y is:

$$f(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x,y)}{f_Y(y)} \tag{42}$$

The above definitions can be easily extended to more than two random variables and it is fairly intuitive to do.

9.0.1 Functions of multiple random variables

E.g. Let X = number of courses taken by a full time UW math student. Let Y = $\begin{cases} 1 & \text{if coop} \\ 0 & \text{otherwise} \end{cases}$. Suppose X and Y have the following j.p.f:

$f(x,y)$	3	4	5	6
0	0.09	0.17	0.22	0.01
1	0.05	0.1	0.32	0.04

Let G = $\frac{X}{2} + \frac{Y}{2}$. How do we find the probability function of G?

There is a process we can apply to find the p.f. of any function $G(X,Y)$:

1. Find the range of G
2. To find the probability function of G: for each value of g inside G's range, find pairs (x, y) that result in that value g and sum their probabilities

1. G's range:

g	3	4	5	6
0	1.5	2	2.5	3
1	2	2.5	3	3.5

2. G's p.f:

$f_G(g)$	1.5	2	2.5	3	3.5
	0.09	0.22	0.32	0.33	0.04

In general, for $G = g(X, Y)$, the probability function is:

$$f_G(g) = \sum_{\text{all (x, y) s.t. g(x, y) = g}} f(x, y) \tag{43}$$

One common choice of a function of multiple random variables is $T = X + Y$, called a convolution. The p.f. of T is:

$$\begin{aligned} f_T(t) &= \sum_{\text{all (x, y) s.t. x + y = t}} f(x, y) \\ &= \sum_{\text{all x}} \sum_{\text{all y}} f(x, y) \text{ s.t. } x + y = t \\ &= \sum_{x=0}^{t} f(x, t - x) \end{aligned}$$

Furthermore, if X and Y are independent, then the p.f. becomes: $\sum_{x=0}^{t} f_X(x) f_Y(t - x)$.

Prove that if $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$ and X, Y are independent then $T = X + Y \sim Bin(n + m, p)$.

$$\begin{aligned} f_T(t) &= P(T = t) \\ &= \sum_{x=0}^{t} f_X(x) f_Y(t - x) \\ &= \sum_{x=0}^{t} \binom{n}{x} p^x (1 - p)^{n-x} \binom{m}{t - x} p^{t-x} (1 - p)^{m-t+x} \\ &= p^t (1 - p)^{n+m-t} \sum_{x=0}^{t} \binom{n}{x} \binom{m}{t - x} \\ &= \binom{n + m}{t} p^t (1 - p)^{n+m-t} \text{ by hypergeometric identity} \end{aligned}$$

Therefore, $T \sim Bin(n + m, p)$.

Prove that if $X \sim Poi(\mu_1)$ and $Y \sim Poi(\mu_2)$ and X, Y are independent then $T = X + Y \sim Poi(\mu_1 + \mu_2)$.

$$
\begin{aligned}
f_T(t) &= P(T=t) \\
&= \sum_{x=0}^{t} f_X(x) f_Y(t-x) \\
&= \sum_{x=0}^{t} \frac{e^{-\mu_1}\mu_1^x}{x!}\frac{e^{-\mu_2}\mu_2^{t-x}}{(t-x)!} \\
&= \frac{e^{-(\mu_1+\mu_2)}\mu_2^t}{t!}\sum_{x=0}^{t}\frac{t!}{x!(t-x)!}\left(\frac{\mu_1}{\mu_2}\right)^x \\
&= \frac{e^{-(\mu_1+\mu_2)}\mu_2^t}{t!}\sum_{x=0}^{t}\binom{t}{x}\left(\frac{\mu_1}{\mu_2}\right)^x \\
&= \frac{e^{-(\mu_1+\mu_2)}\mu_2^t}{t!}\left(1+\frac{\mu_1}{\mu_2}\right)^t \text{ by binomial theorm} \\
&= \frac{e^{-(\mu_1+\mu_2)}(\mu_1+\mu_2)^t}{t!}
\end{aligned}
$$

Therefore, $T \sim Poi(\mu_1 + \mu_2)$.

What is the expectation of $g(X,Y)$ is:

$$E[g(X,Y)] = \sum_{\text{all x}}\sum_{\text{all y}} g(x,y) f(x,y)$$

Recall that expectation is a linear operator. In general, $E[g(X,Y)] \neq g(E[X], E[Y])$. However, if $g(X,Y)$ is a linear function:

$$E[aX + bY + c] = aE[X] = bE[Y] + c$$

9.0.2 Multinomial Random Variable

Suppose we have n independent trials where each trial has k outcomes. Let p_i be the probability a trial has outcome i. Let X_i be the number of outcome i's in those n trials. Then we say $(X_1, X_2, ..., X_k)$ follows the multinomial distribution:

$$(X_1, X_2, ..., X_k) \sim Multi(n, p_1, p_2, ..., p_k) \tag{44}$$

What is the probability function of $(X_1, X_2, ..., X_k)$?

$$\begin{aligned} f(x_1, x_2, ..., x_k) &= p(x_1 \text{ outcome 1's}, x_2 \text{ outcome 2's}, ..., x_k \text{ outcome k's}) \\ &= \frac{n!}{x_1! x_2! ... x_k!} p_1^{x_1} p_2^{x_1} ... p_k^{x_k} \end{aligned}$$

Note that $p_1^{x_1} p_2^{x_1} ... p_k^{x_k}$ is the probability of getting x_1 outcome 1's, x_2 outcome 2's, ..., x_k outcome k's in a row. In reality, these outcomes can occur in any order. We need to account for all the ways to rearrange the outcomes: $\frac{n!}{x_1! x_2! ... x_k!}$.

What is the marginal probability function of X_i?
Notice that we have n bernoulli trials with the probability of success being p_i. Therefore:

$$X_i \sim Bin(n, p_i) \tag{45}$$

Question: Are X_i and X_j $(i \neq j)$ independent?
No. As X_i increases (i.e. more trials get outcome i's), X_j decreases (i.e. less trials get outcome j's).

What is the conditional probability of X_i given X_j?

$$\begin{aligned} P(X_i = x_i | X_j = x_j) &= \frac{P(X_i = x_i \text{ and } X_j = x_j)}{P(X_j = x_j)} \\ &= \frac{P(X_i = x_i \text{ and } X_j = x_j \text{ and } X_{\neq i,j} = n - x_i - x_j)}{P(X_j = x_j)} \end{aligned}$$

Notice that $(X_i, X_j, X_{\neq i,j}) \sim Multi(n, p_i, p_j, 1 - p_i - p_j)$.

$$
\begin{aligned}
P\left(X_i = x_i | X_j = x_j\right) &= \frac{\frac{n!}{x_i! x_j! (n-x_i-x_j)!} p_i^{x_i} p_j^{x_j} (1-p_i-p_j)^{n-x_i-x_j}}{\frac{n!}{x_j!(n-x_j)!} p_j^{x_j} (1-p_j)^{n-x_j}} \\
&= \frac{(n-x_j)!}{x_i!(n-x_i-x_j)!} p_i^{x_i} \frac{(1-p_i-p_j)^{n-x_i-x_j}}{(1-p_j)^{n-x_j}} \\
&= \frac{(n-x_j)!}{x_i!(n-x_i-x_j)!} p_i^{x_i} \frac{(1-p_i-p_j)^{n-x_i-x_j}(1-p_j)^{-x_i}}{(1-p_j)^{n-x_j-x_i}} \\
&= \binom{n-x_j}{x_i} \left(\frac{p_i}{1-p_j}\right)^{x_i} \left(\frac{1-p_i-p_j}{1-p_j}\right)^{n-x_j-x_i} \\
&= \binom{n-x_j}{x_i} \left(\frac{p_i}{1-p_j}\right)^{x_i} \left(1-\frac{p_i}{1-p_j}\right)^{n-x_j-x_i}
\end{aligned}
$$

Therefore, $P\left(X_i = x_i | X_j = x_j\right) \sim Bin(n - x_j, \frac{p_i}{1-p_j})$. Note that this result makes logical sense. If the first x_j trials have outcome j then the remaining $n - x_j$ trials cannot have been outcome j. Then the probability of getting outcome i in those $n - x_j$ trials is:

$$
\begin{aligned}
p\left(\text{trial has outcome i} \mid \text{trial cannot be outcome j}\right) &= \frac{\text{trial has outcome i and not j}}{\text{trial cannot be outcome j}} \\
&= \frac{\text{trial has outcome i}}{\text{trial cannot be outcome j}} \\
&= \frac{p_i}{1-p_j}
\end{aligned}
$$

What is the expected value of $E[X_i X_j]$?

$$
E[X_i X_j] = ... = n(n-1) p_i p_j
$$

E.g. Suppose we independently drink a coke, sprite or pepsi for 6 days, one on each day. The probability that a coke, sprite or pepsi is drank on a day is 0.2, 0.3, 0.5 respectively. What is the probability we drink 1 coke, 2 sprite and 3 pepsi?

The first thing to notice is that we have 6 independent trials (days) where each trial has 3 outcomes (drink coke, sprite or pepsi). Let C, S, P be the number of coke, sprite and pepsi we drink in 6 days. Then:

$$(C, S, P) \sim Multi(6, 0.2, 0.3, 0.5)$$
$$P(C = 1, S = 2, P = 3) = \frac{6!}{3!2!1!} 0.2^1 0.3^2 0.5^3$$

9.0.3 Covariance and Corelation

We want a way to measure how two random variables are related. The covariance of two random variables X and Y is the average distance from X's mean times Y's mean.

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] \tag{46}$$

An alternate definition of covariance can be derived if we simplify:

$$\begin{aligned} Cov(X, Y) &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

So the covariance can also be defined as follows:

$$Cov(X, Y) = E[XY] - E[X]E[Y] \tag{47}$$

Note that the covariance looks very similar to the variance of a random variable ($Var(X) = E[X^2] - E[X]^2$). Also notice that the covariance is not defined to be the squared distance. That means it can be negative (unlike the variance).

If the relationship between two random variables, X and Y, is positive (i.e. as X increases, Y increases and as Y increases, X increases) then their covariance will be positive. If the relationship between the two variables is negative (i.e. as X increases, Y decreases and as Y increases, X decreases) then their covariance will be negative.

If X and Y are independent, $Cov(X, Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0$.

The problem with covariance is that its not bounded, $-\infty < Cov\,(X,Y) < \infty$. That means there is no way to interpret the magnitude of the covariance. That is why we have another measurement.

The correlation of two random variables, X and Y, is defined as:

$$Corr\,(X,Y) = \frac{Cov\,(X,Y)}{SD(X)SD(Y)} \tag{48}$$

The correlation is bounded, $-1 \leq Corr\,(X,Y) \leq 1$. A correlation of 0.99 means the two random variables have a strong positive relationship. A correlation of -0.99 means the two random variables have a strong negative relationship.

E.g. In the last example, what is the covariance and correlation of C and S?

$$\begin{aligned} Cov\,(C,S) &= E[CS] - E[C]E[S] \\ &= 6(6-1)(0.2)(0.3) - 6(0.2)(6)(0.3) \\ &= -0.36 \end{aligned}$$

$$\begin{aligned} Corr\,(C,S) &= \frac{Cov\,(C,S)}{SD(C)SD(S)} \\ &= \frac{-0.36}{\sqrt{6(0.2)(1-0.2)}\sqrt{6(0.3)(1-0.3)}} \\ &= -0.327 \end{aligned}$$

Notice that C and S have a medium weak negative relationship. It makes sense that it is negative because, as we drink more cokes, we drink less sprites.

Note that even if the data shows a very strong correlation between two random variables, it does not mean causation. That is, it does not mean one is causing the other. For example, the correlation between global temperature and the number of pirates has a very strong correlation of 0.98. This does not mean the number of pirates in the world is causing global temperature.

In general, for X and Y, X could cause Y, Y could cause X, or X and Y could both be caused by another random variable Z, or it could simply be a coincidence. The random variables need to be studied further in order to determine causation.

9.0.4 Linear Combinations of Random Variables

The next few sections rely on linear combinations of random variables. There are a number of properties that will be used:

- $E[aX + bY + c] = aE[X] + bE[Y] + c$
- $$Y = \sum_{i=1}^{n} a_i X_i$$
$$E[Y] = \sum_{i=1}^{n} a_i E[X_i]$$
- Suppose we have n random variables where each random variable X_i has the same mean μ. Then the sample mean of those random variables is:
$$\overline{X} = \sum_{i=1}^{n} \frac{1}{n} X_i$$
$$E[\overline{X}] = \sum_{i=1}^{n} \frac{1}{n} E[X_i]$$
$$= \frac{\mu n}{n} = \mu$$
- $Var\left(aX + bY + c\right) = a^2 Var\left(X\right) + b^2 Var\left(Y\right) + 2abCov\left(X, Y\right)$
- $$Y = \sum_{i=1}^{n} a_i X_i$$
$$Var\left(Y\right) = \sum_{i=1}^{n} a_i^2 Var\left(X_i\right) + \sum_{i,i\neq j}^{n} \sum_{j,i\neq j}^{n} a_i a_j Cov\left(X_i, X_j\right)$$
$$= \sum_{i=1}^{n} a_i^2 Var\left(X_i\right) + 2 \sum_{i,i<j}^{n} \sum_{j,i<j}^{n} a_i a_j Cov\left(X_i, X_j\right)$$

  Note that in the last simplification we used the fact that each set of indices has a matching result. That is, for example, if $i = 1, j = 3$ then the term is $a_1 a_3 Cov\left(X_1, X_3\right)$ and when $i = 3, j = 1$ the term is $a_3 a_1 Cov\left(X_3, X_1\right)$. These two terms are exactly the same so we can group the indices together.

- Suppose we have n random variables where each random variable X_i has the same variance σ^2 and is independent. Then the sample variance of those random variables is:

$$\begin{aligned}\overline{X} &= \sum_{i=1}^{n} \frac{1}{n} X_i \\ Var\left(\overline{X}\right) &= \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 Var\left(X_i\right) \\ &= \left(\frac{1}{n}\right)^2 n\sigma^2 \\ &= \frac{\sigma^2}{n}\end{aligned}$$

- $Cov\left(aX+b, CY+d\right) = acCov\left(X,Y\right)$
- $Cov\left(X,X\right) = Var\left(X\right)$

9.0.5 Linear Combinations of Independent Normal RVs

Suppose you have a bunch of independent normal random variables, $X_i \sim N(\mu_i, \sigma_i^2)$. Let Y be the sum of those random variables, $Y = \sum_{i=1}^{n} a_i X_i$. Then Y is a normal random variable:

$$Y \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right) \tag{49}$$

That is, the sum of independent normal random variables is a normal random variable. This is difficult to prove so the proof is in the course notes.

E.g. Suppose the weight of a cat follows a normal distribution, $C \sim N(4.1, 1.6^2)$, and the weight of a dog also follows a normal distribution, $D \sim N(9.4, 3.6^2)$. Find the probability a random cat weighs more than a random dog. Suppose C and D are independent.

We want: $P\left(C > D\right) = P\left(C - D > 0\right)$. Notice that $C - D$ is a sum of

independent normal random variables.

$$\begin{aligned} E[C-D] &= E[C]-E[D] = 4.1-9.4 = -5.3 \\ Var\,(C-D) &= Var\,(C)+(-1)^2 Var\,(D) \\ &= 1.6^2+3.6^2 = 15.52 \end{aligned}$$

Then $C-D \sim N(-5.3, 15.52)$.

$$\begin{aligned} P\,(C-D>0) &= 1-P\,(C-D\leq 0) \\ &= 1-P\left(Z\leq \frac{0-(-5.3)}{\sqrt{15.52}}\right) \\ &= 1-P\,(Z\leq 1.35) \\ &= 1-0.91149 \\ &= 0.08851 \end{aligned}$$

E.g. Suppose the height of a cat follows a normal distribution, $X \sim N(24, 1.5^2)$. Find the probability a random cat is within 1 cm of the average height.

$$\begin{aligned} P\,(24-1<X<24+1) &= P\,(23<X<25) \\ &= P\left(\frac{23-24}{1.5}<Z<\frac{25-24}{1.5}\right) \\ &= P\,(-0.67<Z<0.67) \\ &= F(0.67)-F(-0.67) \\ &= F(0.67)-P\,(Z>0.67) \\ &= 2F(0.67)-1 \\ &= 0.49714 \end{aligned}$$

E.g. For the above example, find the probability 5 cats are within 1 cm of the average height.

Let X = height of 5 cats. $\overline{X} = \sum_{i=1}^{5} \frac{1}{5} X_i$. Notice that $\overline{X}$ is the sum of independent

normal random variables.

$$E[\overline{X}] = \sum_{i=1}^{5} \frac{1}{5} * 24 = 24$$
$$Var\left(\overline{X}\right) = \sum_{i=1}^{5} \left(\frac{1}{5}\right)^2 * 1.5^2 = \frac{1.5^2}{5}$$

Then $\overline{X} \sim N(24, \frac{1.5^2}{5})$.

$$\begin{aligned} P\left(24-1 < \overline{X} < 24+1\right) &= P\left(23 < \overline{X} < 25\right) \\ &= P\left(\frac{23-24}{\frac{1.5}{\sqrt{5}}} < Z < \frac{25-24}{\frac{1.5}{\sqrt{5}}}\right) \\ &= P\left(-1.49 < Z < 1.49\right) \\ &= 2F(1.49) - 1 \\ &= 0.86387 \end{aligned}$$

9.0.6 Indicator Variables

An indicator variable I_A is attached to an event A and defined as:

$$I_A = \begin{cases} 1 & \text{if A occurs} \\ 0 & \text{if A does not occur} \end{cases} \tag{50}$$

There are several properties of indicator variables:

- $$\begin{aligned} E[I_A] &= 1 \cdot p\left(A\right) + 0 \cdot p\left(A\right) \\ E[I_A] &= p\left(A\right) \end{aligned}$$

- $$\begin{aligned} Var\left(I_A\right) &= E[I_A^2] - E[I_A]^2 \\ E[I_A^2] &= 1^2 \cdot p\left(A\right) + 0^2 \cdot p\left(A\right) = p\left(A\right) \\ Var\left(I_A\right) &= p\left(A\right)\left(1 - p\left(A\right)\right) \end{aligned}$$

•

$$\begin{aligned} Cov\left(I_A, I_B\right) &= E[I_A I_B] - E[I_A]E[I_B] \\ E[I_A I_B] &= p\left(AB\right) \\ Cov\left(I_A, I_B\right) &= p\left(AB\right) - p\left(A\right)p\left(B\right) \end{aligned}$$

E.g. Use indicator variables to find the mean and variance of a $X \sim Bin(n, p)$.

Let $X_i = \begin{cases} 1 & \text{if ith trial is success} \\ 0 & \text{otherwise} \end{cases}$

Then $X = X_1 + X_2 + ... + X_n = \sum_{i=1}^{n} X_i$.

$$\begin{aligned} E[X_i] &= p\left(\text{trial is success}\right) = p \\ Var\left(X_i\right) &= p(1-p) \end{aligned}$$

Recall that the binomial random variable has n independent trials. That means each $X_i, X_j (i \neq j)$ is independent.

$$\begin{aligned} E[X] &= \sum_{i=1}^{n} E[X_i] = np \\ Var\left(X\right) &= \sum_{i=1}^{n} Var\left(X_i\right) = np(1-p) \end{aligned}$$

E.g. Use indicator variables to find the mean and variance of $X \sim Hyper(N, r, n)$.

Let $X_i = \begin{cases} 1 & \text{if object i is success} \\ 0 & \text{otherwise} \end{cases}$

Then $X = X_1 + X_2 + ... + X_n = \sum_{i=1}^{n} X_i$.

$$\begin{aligned} E[X_i] &= p\left(\text{object i is success}\right) = \frac{r}{N} \\ Var\left(X_i\right) &= \frac{r}{N}\left(1 - \frac{r}{N}\right) \end{aligned}$$

In a hypergeometric random variable, n objects are selected from N objects without replacement where order doesn't matter. If the first object we select

is a success, then the probability of getting a success again goes down. So each X_i is not independent.

$$\begin{aligned}
E[X_iX_j] &= p\,(\text{object i } \textit{and} \text{ j is a sucess}) = \frac{r}{N}\cdot\frac{r-1}{N-1} \\
Cov\,(X_i,X_j) &= E[X_iX_j] - E[X_i]E[X_j] \\
&= \frac{r(r-1)}{N(N-1)} - \left(\frac{r}{N}\right)^2 \\
&= -\frac{r(N-r)}{N^2(N-1)} \\
E[X] &= \sum_{i=1}^{n} E[X_i] = \frac{rn}{N} \\
Var\,(X) &= \sum_{i=1}^{n} Var\,(X_i) + 2\sum_{i,i<j}^{n}\sum_{j,i<j}^{n} Cov\,(X_i,X_j) \\
&= \sum_{i=1}^{n} \frac{r}{N}(1-\frac{r}{N}) + 2\sum_{i,i<j}^{n}\sum_{j,i<j}^{n} -\frac{r(N-r)}{N^2(N-1)} \\
&= \frac{rn}{N}(1-\frac{r}{N}) + 2\cdot\binom{n}{2}\left(-\frac{r(N-r)}{N^2(N-1)}\right) \\
&= \frac{rn}{N}(1-\frac{r}{N}) + n(n-1)\left(-\frac{r(N-r)}{N^2(N-1)}\right) \\
&= \frac{nr}{N}\left(1-\frac{r}{N}\right)\left(\frac{N-n}{N-r}\right)
\end{aligned}$$

E.g. Suppose you have N messages coming to an email server which get randomly distributed among N recipients. Let X = number of correctly delivered messages. Find the mean and variance of X.

Let $X_i = \begin{cases} 1 & \text{ith message is correct} \\ 0 & \text{otherwise} \end{cases}$

Then $X = X_1 + X_2 + ... + X_n = \sum_{i=1}^{N} X_i$.

$$
\begin{aligned}
E[X_i] &= \frac{1}{N} \\
Var\left(X_i\right) &= \frac{1}{N}\left(1-\frac{1}{N}\right) \\
E[X_i X_j] &= p\left(\text{message i } \textit{and} \text{ j are both correctly delivered}\right) \\
&= \frac{1}{N}\cdot\frac{1}{N-1} = \frac{1}{N(N-1)} \\
Cov\left(X_i, X_j\right) &= E[X_i X_j] - E[X_i]E[X_j] \\
&= \frac{1}{N(N-1)} - \left(\frac{1}{N}\right)^2 \\
&= \frac{1}{N^2(N-1)} \\
E[X] &= \sum_{i=1}^{N} E[X_i] = \frac{N}{N} = 1 \\
Var\left(X\right) &= \sum_{i=1}^{N} Var\left(X_i\right) + 2\sum_{i,i<j}^{N}\sum_{j,i<j}^{N} Cov\left(X_i, X_j\right) \\
&= N\left(\frac{1}{N}\right)\left(1-\frac{1}{N}\right) + 2\cdot\frac{N(N-1)}{2}\cdot\frac{1}{N^2(N-1)} \\
&= 1 - \frac{1}{N} + \frac{1}{N} = 1
\end{aligned}
$$

10 Chapter 10

The central limit theorem (C.L.T) says the sum of infinite independent and identically distributed random variables (i.i.d) can be modeled by the normal distribution.

E.g. Suppose you are finding Pokemon according to a poisson process with $\lambda = 30\frac{pokemon}{hour}$. Find the probability that you find your 500th Pokemon between the 17th - 18th hour of play.

If we used a poisson random variable to do this we would have to deal with a few cases: 0 Pokemon between 0 - 17th hour, 500 Pokemon between 17 - 18th hour, 1 Pokemon between 0 - 17th hour, 499 Pokemon between 17 - 18th hour and so on. We would have to deal with 500 sub cases. Lets instead find a way to express X as a sum of random variables and use the central limit theorem.

Let X_i = additional waiting time to catch the ith Pokemon for $i = 1, ..., 500$. Then $X_i \sim Exp(\theta = \frac{1}{30})$.

Then $X = X_1 + X_2 + ... + X_{500} = \sum_{i=1}^{500} X_i$ where X is the waiting time to find the 500th Pokemon. We want $P\left(17 < X < 18\right)$. Since X is the sum of independent and identically distributed random variables, X can be approximated by a normal random variable by C.L.T.

$$E[X] = \sum_{i=1}^{500} E[X_i] = \frac{500}{30}$$

$$Var\left(X\right) = \sum_{i=1}^{500} Var\left(X_i\right) = \frac{500}{30^2} = \frac{500}{900}$$

Then $X \sim N(\frac{500}{300}, \frac{500}{900})$.

$$\begin{aligned} P\left(17 < X < 18\right) &= P\left(\frac{17 - \frac{500}{30}}{\sqrt{\frac{500}{900}}} < Z < \frac{18 - \frac{500}{30}}{\sqrt{\frac{500}{900}}}\right) \\ &= P\left(0.45 < Z < 1.79\right) = 0.28963 \end{aligned}$$

Notice that we applied the C.L.T to approximate a continuous random variable (i.e. exponential) using a continuous random variable (i.e. normal).

E.g. Suppose a file can have corrupt bits. Find the probability that 250 files have between 70 and 85 corrupt bits (inclusive). Let X = number of corrupt bits in a file with probability function:

$f(x)$	0	1	2
	0.8	0.1	0.1

$$E[X] = 1 \cdot 0.1 + 2 \cdot 0.2 = 0.3$$
$$Var(X) = ... = 0.41$$

We could find all the cases but instead we will express X as a sum of random variables and use the central limit theorem.

Let X_i = corrupt bits on ith file for $i = 1, ..., 250$.

Then $X = X_1 + X_2 + ... + X_{250} = \sum_{i=1}^{250} X_i$. Since X is the sum of independent and identically distributed random variables, X can be approximated by a normal random variable by C.L.T.

$$E[X] = 250 \cdot 0.3 = 75$$
$$Var(X) = 250 \cdot 0.41 = 102.5$$

Then $X \sim N(75, 102.5)$. We want $P(70 \leq X \leq 85)$. Note, however, that X is actually a discrete random variable. We are approximating X with a continuous random variable.

The problem with this is that the endpoints don't matter when dealing with continuous random variables. So if we calculated $P(70 \leq X \leq 85)$ it would not include the endpoints 70 and 85. In order to deal with this we will apply a continuity correction. A continuity correction is applied when approximating a discrete random variable using a continuous random variable. In order to apply a continuity correction: replace $x \leq 85$ with $x \leq 85.5$ and $x \geq 70$ with $x \geq 69.5$.

$$P\left(\frac{69.5 - 75}{\sqrt{102.5}} \leq Z \leq \frac{85.5 - 75}{\sqrt{102.5}}\right) = P(-0.54 \leq Z \leq 1.04) = 0.55623$$

10.0.1 Approximating Poisson and Binomial using Normal

Suppose we have a random variable X where $X \sim Poi(\mu)$ where μ is large.

Recall that if you have two random variables X and Y where $X \sim Poi(\mu_1)$ and $Y \sim Poi(\mu_2)$ and X and Y are independent, then $T = X + Y \sim Poi(\mu_1 + \mu_2)$.

We can think of a poisson random variable as a sum of μ independent $Poi(1)$ random variables.

$$X_i \sim Poi(1)$$
$$X = \sum_{i=1}^{\mu} X_i$$

Then $X \sim Poi(\underbrace{1 + 1 + ... + 1}_{\mu \text{ times}}) = Poi(\mu)$.

X is a sum of independent and identically distributed random variables and so, by C.L.T, it can be approximated by a normal random variable.

$$X \sim N(\mu, \mu)$$

Note that approximating a poisson using normal needs a continuity correction as a poisson random variable is discrete.

Suppose we have a random variable X where $X \sim Bin(n, p)$ where n is large.

Let $X_i = \begin{cases} 1 & \text{if ith trial is success} \\ 0 & \text{otherwise} \end{cases}$

Then $X = X_1 + X_2 + ... + X_n = \sum_{i=1}^{n} X_i$. X is a sum of independent and identically distributed random variables so, by C.L.T, it can be approximated by a normal random variable.

$$X \sim N(np, np(1-p))$$

Note that approximating a binomial using normal needs a continuity correction as a binomial random variable is discrete.

E.g. Suppose a large number of users play an online game and each user is independently playing with probability p. Suppose we check 400 users and

find the proportion of them that are currently online. Find the probability that the proportion observed is within ± 0.02 of the true p.

$$P\left(0.18 \leq \frac{X}{400} \leq 0.22\right) = P\left(72 \leq X \leq 88\right)$$

Recall that the binomial random variable does not have a closed form of the cumulative distribution function. Lets instead approximate it using a normal random variable.

By C.L.T, $X \sim N(400 \cdot 0.2, 400 \cdot 0.2 \cdot 0.8) = N(80, 64)$.

$$\begin{aligned} P\left(72 \leq X \leq 88\right) &= P\left(71.5 \leq X \leq 88.5\right) \\ &= P\left(\frac{71.5 - 80}{\sqrt{64}} \leq Z \leq \frac{88.5 - 80}{\sqrt{64}}\right) \\ &= 0.71086 \end{aligned}$$

E.g. For the above example, how many users would need to be asked for the proportion observed to be within ± 0.02 of the true p.

$$\begin{aligned} P\left(0.18 \leq \frac{X}{n} \leq 0.22\right) &= 0.95 \\ P\left(0.18n \leq X \leq 0.22n\right) &= 0.95 \end{aligned}$$

Then, by C.L.T, $X \sim N(0.2n, 0.16n)$. In this case we don't apply a continuity correction since n is unknown.

$$\begin{aligned} P\left(\frac{0.18n - 0.2n}{\sqrt{0.16n}} \leq Z \leq \frac{0.22n - 0.2n}{\sqrt{0.16n}}\right) &= 0.95 \\ P\left(Z \leq \frac{0.22n - 0.2n}{\sqrt{0.16n}}\right) &= 0.975 \\ n &= 1537 \end{aligned}$$