**Name**: Dat Quoc Ngo

NET-ID: dqn170000

Course#: CS6364.002

Homework 2

## Q1: In at least one page (single space, Font 10), describe in your own words the difference between artificial intelligence and machine learning based on your readings of chapter 1 in the AI textbook and chapter 1 in the PRML textbook.

The study of Artificial Intelligence has been actively carried under different fields, such as philosophy, cognitive science, neuroscience, etc. Even though each field has its separate approach to AI, the early approaches aim at a general agent that is capable of observing (a.k.a sensing), memorizing, and behaving rationally like humans.

Marvin Minsky and his students made early breakthroughs in AI, Logic and argued that AI agents could perform commonsense reasoning based on logical relationships between artifacts. Its implementation is a directed graph in which each node is an artifact carrying information; the directed edge performs the commonsense reasoning. A However, a major drawback is the computational complexity that only 40 artifact nodes may exhaust modern computers. In the testing phase, the commonsense reasoning does not perform well as expected.

Instead of a general AI agent, scientists proposed to decompose AI into multiple sub-tasks which gave birth to different AI-related fields: Computer Vision, Natural Language Processing, Speech Recognition, etc. Early approaches to AI and its subfields relied on rules written by humans that consisted of many assumptions and did not perform well. The first Russian-English program was designed with naïve grammar mappings and word replacement that led to meaningless translation due to lack of contextual understanding. Since 1970s, statistics, mathematics, and the development of Computer Science with new programming languages and tools have played an important role in improving AI. For example, the emergence of neural networks, Hidden Markov Model, and Bayesian Network has led to birth of many statistical modeling algorithms that played important roles in currently state-of-the-art Speech Recognition techniques.

Despite outperforming rule-based approaches, statistics-based algorithms could not achieve best accuracy without accessing to large-scale datasets to eliminate bias assumptions and to fully rely on samples collected from realistic scenarios. Due to the rapid growth of internet from 1990s, large-scale datasets become accessible which led to seemingly impossible applications, such as the Stanley driverless vehicle finishing 132 miles across the Mojave dessert.

Unlike AI research focusing on replicating human behaviors, machine learning is an AI subfield that focuses on recognizing regular patterns based on statistics. Consider the example of recognizing back-and-white handwritten digits. Given a image of the digit target t in size of 28x28, the digit target t could be classified by handcrafted rules (e.g. curve shape) or statistics which considered the distribution of features that are color pixels. Before mapping pixels to distribution, pixels undergo the data preprocessing step to either reduce the feature dimensionality (e.g. eliminating unnecessary pixels) or mapping pixels to the higher dimension for better feature representation. Based on the linear or non-linear relationship among feature samples, linear (e.g. linear regression or logistic regression) or non-linear (e.g. SVM, Trees) models are selected to model the feature distribution and to make prediction. Each machine learning algorithm has a series of weights that should be carefully initialized for robust performance in testing. The weight series could be treated as human memory that given stimulus x, we deliver reaction based on memory of observing similar stimuli in the past.

Speaking of data preprocessing, standardization is usually performed to generalize continuous or discrete input variables into a same range of values. Standardizing is necessary in statistical modeling which assumes that each feature make equal contribution to modeling. Beside standardization, eliminating outlier samples is a common practice. Since machine learning is defined as recognizing regular patterns; hence, outliers are not expected.

Unlike AI with handcrafted rules, machine learning algorithms update themselves through error and gradient. For the sake of simplicity, error is categorized into regression (i.e. Root Mean Square Error) and classification (i.e. Cross Entropy). Given calculated errors, gradients could be found by measuring the rate of change in machine learning algorithms to reach local minimum error. Gradients are then used to update the algorithm weights. The process is repeated for T iterations until errors reach the global minimum point. This repeating process is call training driven by errors and gradients which sets machine learning apart from rule-based and logic-based AI approaches.

## Q2: In at least one page (single space, Font 10), describe in your own words the difference between artificial intelligence and deep learning based on your readings of chapter 1 in the AI textbook and chapter 1 in the DL textbook.

A common characteristic shared by AI and Deep Learning is inspired by neuroscience or how brain’s neurons interact. AI explained by neuroscience demonstrates the brain structure as a network of neurons in which each consists of a nerve cell and synapses for the inter-neuron communication. Similarly, the core technology of Deep Learning is a stack of N layers with M neurons per layer. Each neuron consists of an activation function (e.g. rectified linear unit) to represent the “on” and “off” states by AI cognitive science. Output of neurons in the previous layer are inputs of neurons in the next layer. This setting resembles the inter-connection through synapses. Hence, media may claim that Deep Learning is inspired by neuroscience. However, this claim is not true anymore due to the lack of information about the brain activity to be used as guide for Deep Learning.

As a subfield of Machine Learning, Deep Learning algorithms is scalable for generalization to represent opened world that could not be done by rule-based AI approaches. The 1950s Russian-English translation program is a good example. Since it was designed with handcrafted grammar mappings and naïve word replacement, the translated text is grammatical but nonsensical. The state-of-art Deep Learn algorithms (i.e. sequence-to-sequence models) can translate multi-lingual text meaningfully. Google Translate is a famous product implementing Deep Learning for the translation task.

One benefit of Deep Learning compared with AI is its capability of extracting features from raw data which is so called representation learning. Reconsider the 1950s Russian-English translation program. As mentioned above, the use 2 features, grammar mapping and word replacement in AI led to the poor performance. Instead, Deep Learning algorithms in Google Translate extract “black-box” features directly from raw text without human intervention. Even though “black-box” features are not easily identified, the feature list include more than simply grammar mapping and word replacement.

The core difference between AI and Deep Learning is about how to make predictions. An AI program is algorithmic. In other words, an AI program consists of a list of logic steps which replicate exactly all human actions. Considering the problem of recognizing face. The challenge in the handcrafted AI program is how to describe facial features (e.g. curves of eyebrows, colors) using logical algorithms since the facial feature representation is not discrete, Deep Learning consists of multiple layers that the higher-level layer captures more abstract features using activation functions. In a 3-layer Deep Learning model, the first or bottom layer may extract pixel-level features (e.g. white and black pixels of pupils); the 2nd layer may extract bigger curves (e.g. eyes, noses, lips); the 3rd layer may capture the entire face. Activation functions are unlike logical algorithms since they denote the “on” and “off” states of neurons and determine how much each feature (e.g. pixel or curve or shape) contributes to the face uniqueness. In other words, the multi-layer setting, and activation functions generalize the decision-making process.

Multilayer Perceptron is the fundamental model in any complex Deep Learning architecture. There are Deep Learning architecture variants to improve the representation learning process in different data types. Convolutional Neural Network (CNN) convolves a x-by-x matrix to gather significant features matrix-like data (e.g. tables, image, video). Recurrent Neural Network (RNN) sequentially reads features at each input step in a fixed order. RNN is suitable for sequential data such as text and sound. Similarly, Graph Neural Network (GNN) is suitable for graph-like data such as chemical molecules since it traverses the entire graph before making prediction. the. In general, for an input type, a correspondingly suitable Deep Learning architecture generalizes the decision-making process in which multiple tasks could be performed in the end-to-end setting. This ability is hardly achieved in the early AI approach that leads to the “pipeline error”.

As explained, the early AI program includes a series of programmed actions to simulate a mini environment that any change in the environment requires the human intervention to algorithmically rearrange the action order. Hence, AI is limited in scalability of handling unseen data. Unlike AI, due to the representation learning, Deep Learning is capable of adapting to changes in environments when more and larger datasets are provided for finetuning (aka post-training). The finetuning is a similar to training that no human intervention is required.

## Q3: In at least one page (single space, Font 10), describe in your own words the difference between machine learning and deep learning based on your readings of chapter 1 in the PRML textbook and chapter 1 in the DL textbook.

Since Deep Learning (DL) is a subfield of Machine Learning (ML), both ML and DL are driven by back-propagation and the Stochastic-Gradient-Descent-like optimization. The similarity enables ML and DL to update themselves given any change in features and to maintain their performance rate in the inference setting. However, the major difference between them is in feature extraction.

In Machine Learning, data processing is a common practice to standardize features and select relevant features. The relevant features are selected based on either domain knowledge or a series of trials. Regardless of domain knowledge and trials, the feature selection does not always guarantee yielding relevant features since the study of humans’ brains is not enough to provide information as guide to select features well. The feature selection is also time-consuming that most of the project timeline is for this step.

Instead of manually selecting features, the multi-layer architecture and activation functions in Deep Learning can receive raw data and abstractly extract features. The Deep Learning architecture usually resembles the funnel shape that the bottom layer is bigger than the top layer. In this design, the bottom layer accepts raw data as input and extract simple features; the next layers act as a filter and a compressor that selectively extract important features and combine them into more abstract features. Input to the output layer (i.e. classifier or regressor) is the highly abstract information of the original input. The “filtering and compressing” mechanism shifts the feature selection task from engineers to activation functions which are optimized iteratively by errors and gradients. As a result, less time spent on feature selection. The left tasks for engineers are to design good Deep Learning architectures, good error formulas, and good optimization algorithms.

Both Deep Learning and Machine Learning requires engineers and researchers to design good error formulas and optimization algorithms. However, the progress in designing better Machine Learning algorithms is slow. ML algorithms are mostly designed based on observation of distribution and relationship of features. Consider Logistic Regression and Support-Vector-Machine (SVM) algorithms for classification. Logistic Regression expects features are linear correlated which is not always correct and available. On the other hand, SVM supports the non-linear relationship between features that was proven to be useful the handwritten digit recognition in 1990s by LeCun. As a result, ML algorithms contains bias which come from the algorithm design and assumptions in the hand-designed feature selection.

In Deep Learning, engineers and researchers can simply increase models’ size by stacking more layers or expanding layers’ size. Within 10 years, bigger and more complex models are reported to reduce the classification error of the handwritten digit recognition from 30% to below 5%. The sudden error drop attracts more attention research that leads to completely new Deep Learning architectures such as Transformer and Generative Adversarial Network (GAN). The rising complexity of Deep Learning algorithms requires other special hardware like Graphic Processing Unit (GPU) than general-purpose CPUs. Linear algebra or matrix computation plays an important role in Deep Learning which GPUs can handle since they are initially designed for matrix computation in graphics.

Beside increasing models’ size, increasing the dataset size has been reported to improve both regression and classification performance. However, large-scale datasets are costly and not always available due to the annotation effort. Hence, there exists the research effort in training Deep Learning models with less data for the quick delivery-to-service reason in industry. Since Machine Learning bases on statistical distribution of selected features, Machine Learning is generally considered faster and low latency compared to Deep Learning. The latency of ML algorithms is mostly concerned with the number of selected features.

## Q4: State the condition under which the Binomial distribution arises?

Binomial distribution is used to calculate probability of a discrete random variable which have binary values.

## Q5: If X Bin(12, 0.7) calculate

N (number of trials) = 12

P (success rate) = 0.7

Q (fail rate) = 1 – 0.7 = 0.3

#### P(X=9)

## P(X=9) = 12C9 \* (0.7)^9 \* (0.3)^(12-9) = 220 \* (0.7)^9 \* (0.3)^3 = 0.2397

#### P(X>10)

P(X > 10) = P(X=11) + P(X=12)

= 12C11 \* 0.7^11 \* 0.3^1 + 12C12 \* 0.7^12\*0.3^0

= 12\* 0.7^11 \* 0.3 + 0.7^12 = 0.085

#### P(X<=11)

P(X <= 11) = 1 – P(X = 12)

= 1- 12C12 \* 0.7^12\*0.3^0

= 1 – 0.7^12 = 0.9862

## Q6: A lecturer uses a laptop to give a series of 8 lectures. There is a 5% chance that the laptop causes problems in any given lecture. What is the probability of observing 2 out of 8 lectures in which the laptop exhibits ‘technical difficulties’.

N (number of lectures) = 8

P (chance of laptop to cause problem in any given lecture) = 0.05

Q (chance of laptop not to cause problems) = 1 – 0.05 = 0.95

P(X = 2) = 8C2 \* 0.05^2 \* 0.95^6 = 28 \* 0.05^2 \* 0.95^6 = 0.051

## Q7: Two groups of twelve children are taught two different methods of arith- metic. (Assume that a child in group one is matched in terms of their arith- metic ability with a child in group 2 before the start of the study). What is the probability that at least 9 children from one of the groups will ob- tain higher scores than the other group? What other assumptions have you made?

We assume that the chance of gaining higher scores from students in each group is equal. Hence, p (higher score of group 1) = 0.5 and q (higher score of group 2) = 0.5 Total trials n = 12. Moreover, the probability of students with higher scores from each group is equal.

P(X >= 9 that X are students from group1) = P(Y >= 9 that Y are students from group 2)

= 1 – P(X = 10) – P(X=11) – P(X=12)

= 1 - 12C10 \* 0.5^10 \* 0.5\*2 - 12C11 \* 0.5^11 \* 0.5\*1 - 12C12 \* 0.5^12 \* 0.5\*0

= 1- 66 \* 0.5^12 – 12 \* 0.5^12 – 1\* 0.5^12

= 1 – 79\*0.5^12 = 0.9807

## Q8: Two teams, A and B, play a series of games. If team A has probability 0.4 of winning each game, is to it’s advantage to play the best three out of five games or the best four out of seven? Assume outcomes of succesive games are indepedent.

P (team A’s chance of winning each game) = 0.4

Q (team B’s chance of winning each game) = 0.6

Probability of team A winning more than or equal to 3 out of 5 games

P(X >= 3) = 5C3 \* 0.4^3 \* 0.6^2 + 5C4 \* 0.4^4 \* 0.6 + 5C5 \* 0.4^5 \* 0.6^0

= 10 \* 0.4^3 \* 0.6^2 + 5 \* 0.4^4 \* 0.6 + 0.4^5 = 0.31744

Probability of team A winning more than or equal to 4 out of 7 games

P(X >= 4) = P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7)

= 7C4 \* 0.4^4 \* 0.6^3 + 7C5 \* 0.4^5 \* 0.6^2 + 7C6 \* 0.4^6 \* 0.6 + 7C7 \* 0.4^7 \* 0.6^0

= 35 \* 0.4^4 \* 0.6^3 + 21 \* 0.4^5 \* 0.6^2 + 7 \* 0.4^6 \* 0.6 + 0.4^7

= 0.2897

Comparing the 2 results, it is advantage to play the best 3 out of 4 games.

## Q9: State the conditions under which the Poisson distribution can be used to ap- proximate a Binomial distribution?

When,

* n (number of trials) is large (say > 50)

p is small (say < 0.1)

Then a Bin (n, p) can be approximated with a Po(λ) where λ = np

#### If Bin(200, 0.01), calculate P(X <= 2)

n = 200, p = 0.01

### using a Binomial distribution

P(X <= 2) = P(X = 0) + P(X = 1) + P(X = 2)

= 200C0 \* 0.01^0 \* 0.99^200 + 200C1 \* 0.01^1 \* 0.99^199 + 200C2 \* 0.01^2 \* 0.99^198

= 0.99^200 + 200 \* 0.01^1 \* 0.99^199 + 19900\* 0.01^2 \* 0.99^198

= 0.6767

### using a Poisson approximation to the Binomial

λ = np = 200 \* 0.01 = 2

P(X <= 2) = P(X = 0) + P(X = 1) + P(X = 2)

= e^(-2) \* 2^0 / 0! + e^(-2) \* 2^1 / 1! + e^(-2) \* 2^2 / 2!

= e^(-2) \* (1+ 2 + 4/2) = e^(-2) \* 5 = 0.6767

## Q10: The mean number of bacteria per millimetre of a liquid is known to be 4. Assuming that the number of bacteria follows a Poisson distribution, find the probability that, in 1ml of liquid, there will be

λ = 4

1. no bacteria

P(X = 0) = e^(-4) \* 4^0 / 0! = 0.0183

1. 4 bacteria

P(X = 4) = e^(-4) \* 4^4 / 4! = 0.1954

1. Less than 3 bacteria

P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)

= e^(-4) \* 4^0 / 0! + e^(-4) \* 4^1 / 1! + e^(-4) \* 4^2 / 2!

= e^(-4) \* (1 + 4 + 16 / 2) = e^(-4) \* (1 + 4 + 8) = e^(-4) \* 13

= 0.2381

### Find the probability that

1. in 3ml of liquid there will be less than 2 bacteria

In 1ml, λ = 4

In 3ml, 3λ = 12

P(X < 2) = P(X = 0) + P(X = 1)

= e^(-12) \* 12^0 / 0! + e^(-12) \* 12^1 / 1!

= e^(-12) \* (1 + 12) = e^(-12) \* 13 = 7.987 \* 10^-5

1. in 0.5 ml of liquid there will be more than 2 bacteria

In 1ml, λ = 4

In 0.5ml, λ/2 = 2

P(X > 2) = 1 – P(X = 0) – P(X = 1) – P(X = 2)

= 1 - e^(-2) \* 2^0 / 0! + e^(-2) \* 2^1 / 1! - e^(-2) \* 2^2 / 2!

= 1 - e^(-2) \* (1 + 2 + 4/2) = 1 - e^(-2) \* 5 = 0.3233

## Q11: Stroke patients with aphasic deficits are each given a number of straight- forward tasks in a psychometric test. The number of errors made by 123 patients are shown in the table below. Calculate the mean and variance of the number of errors per patient and comment on these values. Fit a Poisson distribution and comment on how well it fits the observed data.

A picture containing calendar

Description automatically generated

Assume that the largest number of errors is 5.

Total errors = 5 \* 0 + 1 \* 30 + 2 \* 56 + 3 \* 15 + 4 \* 10 = 262

Each patient has their error statistics as following:

* The mean number of errors = 262 / 123 = 2.13
* The variance of the number of errors per patient is

= (5\*(0 – 2.13) ^ 2 + 30\*(1 – 2.13) ^ 2 +56\* (2 – 2.13) ^ 2 + 15\*(3 – 2.13) ^ 2 + 10\*(4 – 2.13) ^ 2 + 7\*(5 – 2.13) ^ 2)/122

= 165.92 / 122 = 1.35

Given the mean and variance, the smallest and largest number of errors is 0.97 and 3.29 which are close to the error counts 1 and 3 with high number of patients.

Fitting to the Poisson distribution,

Each patient has their error statistics as following:

* Mean = 2.13

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of errors | 0 | 1 | 2 | 3 | 4 | 5 or more |
| P(X = x) | 0.12 | 0.25 | 0.27 | 0.19 | 0.102 | 0.068 |
| Number of patients | 5 | 30 | 56 | 15 | 10 | 7 |
| Expected number of patients | 14.76 | 30.76 | 33.21 | 23.37 | 12.55 | 8.36 |

Look at the table, the Poisson distribution reflects the observed data table. However, there is the difference in value that values seem to shrink to the center.

## Q12: In a large town, one person in 80, on average, has blood type X. If 200 blood donors are taken at random, find an approximation to the probability that they include at least five persons having blood type X.

### How many donors must be taken at random in order that the probability of including at least

### one donor of type X shall be 0.9 or more?

Averagely, one person in 80 has blood type X

Then, 2.5 persons in 200 have blood type X

P(Y >= 5) = 1 – P(Y = 0) – P(Y = 1) – P(Y = 2) – P(Y = 4)

= 1 – e^(-2.5)\*2.5^0 / 0! - e^(-2.5)\*2.5^1 / 1! - e^(-2.5)\*2.5^2 / 2! - e^(-2.5)\*2.5^3 / 3! - e^(-2.5)\*2.5^4 / 4!

= 1 - e^(-2.5)(1 + 2.5 + 2.5^2 / 2! + 2.5^3 / 3! + 2.5^4 / 4!)

= 0.108

The probability of including at least one donor of type X shall be 0.9 or more

P(Y >=1) = 1 – P(Y = 0) >= 0.9

P(Y = 0) <= 0.1

e^(-x) \* x^0 / 0! <= 0.1

e^(-x) <= 0.1

-x >= ln(0.1) = -ln(10)

x >= ln(10) = 2.3

Mean = 2.3 donors-> Number of donors = 2.3 \* 80 / 1= 184

## Q13: Telephone calls reach a secretary independently and at random, internal ones at a mean rate of 2 in any 5 minute period, and external ones at a mean rate of 1 in any 5 minute period. Calculate the probability that there will be more than 2 calls in any period of 2 minutes.

### 

Internal calls: averagely 2 in any 5 minutes -> λ1 = 2

External calls: averagely 1 in any 5 minutes -> λ2 = 1

Then, the average of internal and external calls in λ = λ1 + λ2 = 1 + 2 = 3 in 5 minutes.

Then, in 2 minutes, the average of internal and external calls is λ = 3 \* 2 / 5 = 1.2

Probability of having more than 2 calls in any period of 2 minutes

P(X > 2) = 1 – P(X = 0) – P(X = 1) – P(X = 2)

= 1 - e^(-1.2) \* 1.2^0 / 0! - e^(-1.2) \* 1.2^1 / 1! - e^(-1.2) \* 1.2^2 / 2!

= 1 - e^(-1.2) \* (1 + 1.2 + 1.2^2 / 2) = 0.12

## Q14: IfX∼N(200,202), calculate

1. P(X<210)

Standardize to Normal Distribution:

N(0, 20) -> P(X < (210 – 200) / 20) = P(X < 0.5) = 0.6915

#### P(X<195)

Standardize to Normal Distribution:

N(0, 20) -> P(X < (195 – 200) / 20) = P(X < -0.25) = P(X > 0.25) = 1 - 0.5987 = 0.4013

#### P (X > 215)

Standardize to Normal Distribution:

N(0, 20) -> P(X > (215 – 200) / 20) = P(X > 0.75) = 1-P(X < 0.75) = 0.2266

#### P(185 < X < 205)

Standardize to Normal Distribution:

N(0, 20) -> P((185 – 200) / 20 < X < (205 – 200) / 20) = P(-0.75 < X < 0.25)

= P(X < 0.25) – P(X < -0.75)

= P(X < 0.25) - P(X > 0.75) = 0.5987 – (1 – 0.7734) = 0.3721

#### A such that P(X >a)=0.2

P(X > a) = 1 – P(X < a) = 0.2 -> P(X < a) = 0.8

Standardizing we get P((X – 200) / 20 < (a – 200) / 20) = P(X < (a – 200) / 20) = 0.8

-> (a – 200) / 20 = 0.85

-> a – 200 = 0.85 \* 20 = 17 -> a = 217

## Q15: The number of accidents on a certain railway line occur at an average rate of one every 2 months. Find the probability that

We can formulate the railway accidents using Poisson distribution that models the count of railway accidents in a given time interval.

The average accident rate is one every 2 months

1. there are 25 or more accidents in 4 years,

Then, in 4 years, the scaled average accident rate is 24.

Since, the average accident rate is larger than 20, we can use Normal Distribution approximate Poisson Distribution such that

* Mean = 24
* Variance = 24

N (24, 24) -> P(X >= 25)

Standardizing, we get

N(0, 1) -> P(X >= (24.5 – 24) / sqrt24) = P(X >= 0.102) = 1 – P(X <= 0.102)

From tables, we know that P(X <= 0.1) = 0.5398 and P(X <= 0.11) = 0.5438

P(X <= 0.102) = 0.2 \* 0.5438 + 0.80 \* 0.5398 = 0.5406

P(X >= 0.102) = 1 – P(X <= 0.102 = 1 – 0.5406 = 0.4594

1. there are 30 or less accidents in 5 years.

Then, in 5 years, the scaled average accident rate is 30.

Since, the average accident rate is larger than 20, we can use Normal Distribution approximate Poisson Distribution such that

* Mean = 30
* Variance = 39

N (30, 30) -> P(X <= 30)

Standardizing, we get

N(0, 1) -> P(X <= (30.5 – 30) / sqrt30) = P(X <= 0.0913)

From tables, we know that P(X < 0.09) = 0.5359 and P(X < 0.1) = 0.5398

P(X <= 0.0913) = 0.13 \* 0.5398 + 0.87 \* 0.5359 = 0.5364

## Q16: In a certain country the heights of adult males have mean 170cm and stan- dard deviation 10cm, and the heights of adult females have mean 160cm and standard deviation 8cm; for each sex the distribution of heights approx- imates closely to a normal probability model. On the hypothesis that height is not a factor in selecting a mate, calculate the probability that

N\_h(170, 10^2)

N\_w(160, 8^2)

#### a husband and wife selected at random are both taller than 164cm

After standardizing, we get

P(h > 164) = P(h > -0.6) = P(h < 0.6) = 0.7257

P(w > 164) = P(w > 0.5) = 1- P(w < 0.5) = 1 – 0.6915 = 0.3085

Since probabilities of husband and wife are independent, then the probability of selecting a husband and a wife randomly = 0.7257 \* 0.3085 = 0.2238

#### in a randomly selected husband and wife the wife is taller than the husband

Wife is taller than the husband

N(w – h) = N(-10, -36)

After standardizing, we get

P(w) – P(h) = P(w -h < 0) = P(Z < (0 + 10) / (-36**)) =** P(Z < -0.2777) = 1 – P(Z < 0.2777)

From tables, we know that P(Z < 0.27) = 0.6064 and P(Z < 0.28) = 0.6103

P(Z < 0.2777) = 0.77 \* 0.6103 + 0.23 \* 0.6064 = 0.609403

P(Z < -0.2777) = 1 – P(Z < 0.2777) = 1 – 0.609403 = 0.390597

#### (iii)  the average height of a random couple is greater than 168cm.

Average height of the couple

N(avg) = N(170/2 + 160/2, 100/2 + 64/2) = N(165, 82)

After standardizing, we get

P(Z > 168) = P(Z > (168 – 165) / 82)) = P(Z > 0.0366) = 1 – P(Z < 0.0366)

From tables, we know that P(Z < 0.03) = 0.512 and P(Z < 0.04) = 0.516

P(Z < 0.0366) = 0.66 \* 0.516 + 0.34 \* 0.512 = 0.51464

P(Z > 0.0366) = 1- 0.51464 = 0.48536

## Q17: 10% of chocolates produced in a factory are mis-shapes. In a sample of 1000 chocolates find the probability that the number of mis-shapes is

This problem can be formulated by Binomial distribution that the discrete random variable is whether if the chocolate is mis-shapes or not. Since p = 0.1 < 0.5 and n = 1000 > 30, we can use Normal Distribution to approximate the Binomial Distribution

Mean = np = 0.1 \* 1000 = 100

Variance = npq = 0.1 \* 1000 \* 0.9 = 90

-> N(100, 90)

#### less than 80

After standardizing, we get

P(Z < 80) = P(Z < (80.5 – 100) / sqrt90)) = P(Z < -2.0555) = 1 – P(Z < 2.0555)

From tables, we know that P(Z < 2.05) = 0.9798 and P(Z < 2.06) = 0.9803

P(Z < 2.05555) = 0.55 \* 0.9803 + 0.45 \* 0.9798 = 0.980075

P(Z < -2.0555) = 1 – 0.980075 = 0.019925

#### between 90 and 115 inclusive

After standardizing, we get

P(89.5 <= Z <= 115.5) = P(-0.1068 <= Z <= 0.1634)

= P(Z <= 0.1634) – P(Z <= -0.1068)

= P(Z <= 0.1634) – (1 - P(Z <= 0.1068))

From tables, we know that P(Z <= 0.16) = 0.5636, P(Z <= 0.17) = 0.5675, P(Z <= 0.11) = 0.5438, and P(Z <= 0.12) = 0.5478

P(Z <= 0.1634) = 0.34 \* 0.5675 + 0.66 \* 0.5636 = 0.5649

P(Z <= 0.1068) = 0.68 \* 0.5478 + 0.32 \* 0.5438 = 0.5465

P(-0.1068 <= Z <= 0.1634) = 0.5649 – (1 – 0.5465) = 0.1114

#### 120 or more

After standardizing, we get

P(Z >= 119.5) = P(Z >= 2.0555) = 1 – P(Z <= 2.0555)

From (a), P(Z < = 2.0555) = 0.980075

P(Z >= 2.0555) = 1 – 0.980075 = 0.019925

#### Q18: A sample of 100 apples is taken from a load. The apples have the following distribution of sizes

#### A picture containing text, clock Description automatically generated

#### Determine the mean and standard deviation of these diameters.

#### Assuming that the distribution is approximately normal with the estimated mean and standard deviation find the range of size of apples for packing, if 5% are to be rejected as too small and 5% are to be rejected as too large.

Mean = (6 \* 11 + 7 \* 21 + 8 \* 38 + 9 \* 17 + 10 \* 13) / 100 = 8

Variance = (44 + 21 + 0 + 17 + 52) / 100 = 1.3535

Standard Deviation = sqrt(1.3535) = 1.1634

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Diameter | 6 | 7 | 8 | 9 | 10 |
| Frequency | 11 | 21 | 38 | 17 | 13 |
| D \* F | 66 | 147 | 304 | 153 | 130 |
| F(D-mean)^2 | 44 | 21 | 0 | 17 | 52 |

Say the range of size of applies is bounded by a and b.

P(A < Z < B) = P(Z < B) – P(Z < A)

That A = (a – mean) / var; B = (b – mean) / var

* P(Z < B) = 1 – 5% = 0.95 -> B = 1.64 -> (b – 8) / 1.3535 = 1.64 -> b = 10.21974
* P(Z < A) = 0.05 -> (a -8) / 1.3535 = -2.58 -> a = 4.50797

The range of size of applies is between 4.50797 and 10.21974