Réalisé par : SDIKI Yousra
ELARCH Malak
ICHI Zakia

Année universitaire : 2024/2025
Encadré par :M. AMAL Youssouf

Rapport projet statistiques

1.Analyse descriptive :

❖ Analyse du Temps de Détection des Attaques

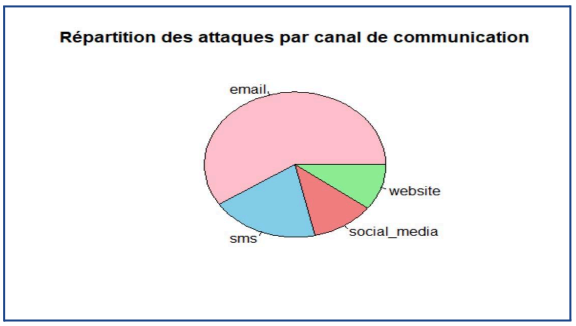
Le temps de détection des attaques de phishing constitue un indicateur clé pour évaluer l'efficacité des mesures de surveillance. Voici les principaux résultats de cette analyse :

- **Temps Moyen de Détection :** Le temps moyen pour détecter une attaque de phishing est de 37.43643 heures. Ce résultat permet d'évaluer le temps typique requis pour détecter une attaque, constituant une référence pour juger de l'efficacité des mécanismes actuels de détection.
- **Médiane du Temps de Détection :** La médiane du temps de détection est de 26 .59666 heures, ce qui indique que 50 % des attaques sont détectées en 26.5966 6 heures ou moins. Cette donnée est cruciale pour évaluer la rapidité générale d es équipes de surveillance face aux menaces
- **Mode du Temps de Détection :** La valeur modale, représentant le temps de détection le plus fréquemment observé, est de 23.42272 heures. Cela met en lumière le délai de détection le plus commun et peut indiquer un temps de réponse typique dans des circonstances standard.

❖ Répartition des Attaques par Canal de Communication

L'analyse des canaux de communication exploités pour les attaques de phishing révèle des tendances intéressantes quant aux méthodes préférées par les attaquants :

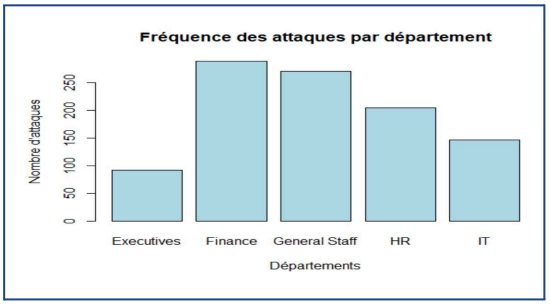
➔ **Graphe :**



➢ **Interprétation :** Un **diagramme circulaire** montre que les attaques de phishing sont principalement réalisées par **email** et **SMS**, avec des parts moindres via les réseaux sociaux et les sites web. Les efforts de protection devraient donc se concentrer en priorité sur les canaux email et SMS.

❖ Fréquence des Attaques par Département :

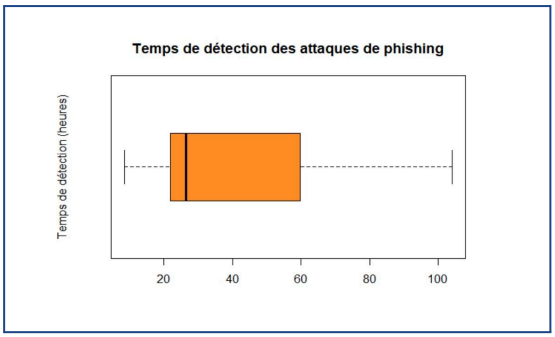
➔ **Graphe**



➢ **Interprétation :** Un **diagramme en barres** indique que les départements **Finance** et **General Staff** sont plus fréquemment visés que d'autres, ce qui peut être lié aux données sensibles qu'ils traitent ou aux interactions externes fréquentes.

❖ **Temps de Détection : Diagramme en Boîte**

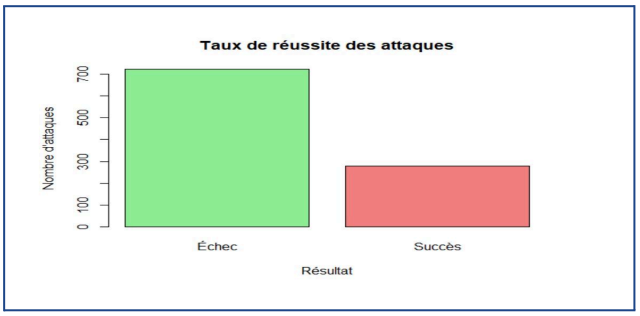
➔ **Graphe :**



➢ **Interprétation** : Un **diagramme en boîte** permet de visualiser la dispersion du temps de détection des attaques de phishing. Cette représentation permet de repérer des variations importantes dans le temps de détection.

❖ **Taux de Réussite des Attaques de Phishing :**

➔ **Graphe** :



➢ **Interprétation :** Un **diagramme en barres** montre la répartition entre les attaques réussies et celles ayant échoué. D'après le graphe , la majorité des attaques ont échoué (plus de 700 attaques) .

Conclusion

Cette analyse descriptive fournit un panorama des caractéristiques des attaques de phishing en termes de temps de détection, de répartition par canal, de cible par département et de taux de réussite. Les statistiques clés telles que la moyenne, la médiane et le mode du temps de détection mettent en évidence les délais typiques pour détecter une attaque, tandis que les graphiques montrent les départements les plus exposés et les canaux de communication les plus à risque. Ces informations sont essentielles pour orienter les efforts de prévention et d'amélioration des mécanismes de détection et de défense contre les attaques de phishing.

2. Analyse bivariée des attaques de phishing : Canal et Succès

Dans le cadre de ce projet, nous avons mené une étude bivariée pour analyser les interactions entre deux variables clés dans le contexte des attaques de phishing : le type de canal utilisé (channel) et le résultat des attaques(success). Objectifs et méthodologie : L'objectif principal de cette étude est d'explorer la relation entre le type de canal et le résultat des attaques. Pour ce faire, une table de contingence a été créée afin d'observer la distribution des résultats en fonction des canaux. Des fréquences relatives ont ensuite été calculées pour analyser la répartition des attaques en pourcentages. Les distributions conditionnelles ont été examinées, notamment X|Y (répartition des canaux conditionnée par le résultat des attaques) et Y|X (répartition des résultats conditionnée par le canal utilisé). Cette analyse nous a permis de répondre à des questions stratégiques, telles que :

- Quels canaux sont les plus utilisés ?
- Quels sont les plus efficaces ?

❖ **Observations et Interprétation du tableau de contingence : L'email** est le canal de phishing dominant, représentant **59,2 %** des attaques avec un total de **592 attaques (424 échecs, 168 succès)**. Cette prédominance peut s'expliquer par la popularité et l'accessibilité des emails. Les canaux les moins utilisés sont **les réseaux sociaux** (112 attaques, 11,2 %) et **les sites web** (102 attaques, 10,2 %). En termes de réussite globale, **278 attaques ont réussi sur**

1000, soit un **taux de réussite de 27,8** %, tandis que les échecs dominent avec 722 attaques (72,2 %). La répartition des succès par canal montre que les attaques par **email** ont un **taux de succès de 28,4** %, tandis que celles par **SMS** ont un **taux de réussite de 27,3 %.** Les attaques via **les réseaux sociaux** affichent le taux de succès le plus élevé, avec **31,3** %, malgré leur fréquence plus faible, tandis que celles via les sites web ont le taux de succès le plus bas, à 21,6 %, suggérant une meilleure sécurisation de cette voie par l'entreprise.

- **Observations et Interprétation du la table de fréquence :** Ce tableau exprime les données en pourcentages pour faciliter les comparaisons et interprétations. Globalement, 72,2 % des attaques échouent, tandis que 27,8 % réussissent, reflétant une efficacité limitée mais non négligeable des attaques. L'email domine avec 59,2 % des attaques, dont 16,8 % des succès et 42,4 % des échecs, confirmant son rôle de canal principal. Les SMS représentent 19,4 % des attaques, avec 5,3 % des succès, tandis que les réseaux sociaux, bien que moins fréquents (11,2 %), enregistrent 3,5 % des succès, montrant une efficacité relative notable. Les attaques via sites web, représentant 10,2 % des attaques et seulement 2,2 % des succès, semblent les moins efficaces. Ainsi, comme dans le premier tableau, l'email reste le canal principal à surveiller en raison de sa fréquence élevée et de son nombre significatif de succès. Les réseaux sociaux et SMS, bien que moins fréquents, méritent également une attention particulière en raison de leur efficacité relative.
- **Distribution conditionnelle x|Y= xi :** Ce tableau présente la distribution des canaux de phishing (X) en fonction des résultats (Y), c'est-à-dire la proportion d'attaques réussies ou échouées par canal. Par exemple, lorsqu'une attaque échoue (failure), 58,73 % des attaques ont utilisé le canal email, et 60,43 % des attaques réussies (success) ont également utilisé ce canal. On observe que les emails dominent largement les deux résultats (succès et échec), en faisant le canal de phishing le plus fréquent et préoccupant. Les SMS représentent environ 19 % des échecs et des réussites, tandis que les réseaux sociaux, bien que moins utilisés, connaissent un taux de succès plus élevé (12,59 %) comparé à leur taux d'échec (10,66 %). Enfin, les attaques via les sites web montrent un taux d'échec de 11,08 % et un taux de succès de 7,91 %, suggérant que ce canal pourrait être mieux sécurisé. Ces observations indiquent qu'une attention particulière devrait être portée sur les canaux comme les emails et les réseaux sociaux, avec des stratégies de sécurité adaptées à chaque type de canal.
- **Distribution conditionnelle Y|X= xi** Ce tableau illustre la probabilité de réussite ou d'échec des attaques selon le canal utilisé. Les emails, proches des moyennes globales, affichent un taux de réussite de 28,38 % et un taux d'échec de 71,62 %. Les SMS montrent une efficacité légèrement inférieure (27,32 % de réussites) et un taux d'échec de 72,68 %. Les réseaux sociaux, bien que moins utilisés (11,2 % des attaques), sont les plus efficaces avec 31,25 % de réussites. En revanche, les sites web, avec seulement 21,57 % de réussites et 78,43 % d'échecs, sont les moins performants. Les réseaux sociaux méritent une attention particulière pour leur efficacité, tandis que la vigilance reste nécessaire pour tous les canaux.
- **Conclusion:** En conclusion, cette étude bivariée a identifié les canaux de phishing les plus utilisés et leurs niveaux d'efficacité. L'email, dominant avec 59,2 % des attaques, présente un taux de réussite de 28,38 % malgré un taux d'échec élevé (71,62 %). Les réseaux sociaux, moins fréquents (11,2 %), affichent un taux de réussite notable de 31,25 %, tandis que les SMS (19,4 % des attaques) ont un taux de réussite de 27,32 %. Les sites web, bien que rares (10,2 %), montrent la plus faible efficacité avec un taux de réussite de 21,57 %. Ces résultats soulignent l'importance d'adopter des stratégies de sécurité adaptées à chaque canal pour réduire leur vulnérabilité.

3.Estimation de la proportion d'attaques réussies :

Dans le cadre de notre étude sur la sécurité, nous avons estimé la proportion d'attaques réussies à partir d'un échantillon de 1000 attaques. Parmi ces 1000 attaques, 278 ont été considérées comme réussies. En utilisant les techniques statistiques suivantes, nous avons estimé la proportion d'attaques réussies et calculé l'intervalle de confiance correspondant.

Résultat : L'intervalle de confiance pour la proportion d'attaques réussies, au seuil de confiance de 95 %, est le suivant: [0.2502,0.3058]. Cela signifie que, avec 95 % de certitude, la proportion réelle d'attaques réussies dans la population étudiée se situe entre 25,02 % et 30,58 %. Cet intervalle suggère que la proportion d'attaques réussies est significativement inférieure à 50 %, ce qui pourrait indiquer que les mesures de sécurité en place ont un impact positif, mais qu'il reste encore des vulnérabilités à adresser."

4.Test des hypothèses :

Première hypothèse : plus de 50 % des attaques par email sont réussies

Ce test vise à évaluer si plus de 50 % des attaques par email aboutissent à un succès. Les hypothèses sont les suivantes : l'hypothèse nulle (Ho) stipule que la proportion de succès est inférieure ou égale à 50 % ($p \leq 0,5$), tandis que l'hypothèse alternative (H1) affirme qu'elle est supérieure à 50 % ($p > 0,5$). Un test unilatéral à droite est utilisé, car l'objectif est de vérifier si la proportion dépasse 50 %, en recherchant une différence dans une seule direction.

❖ **Interprétation des résultats:**

- **Proportion observée de succès :**La proportion de succès observée dans les attaques par email est de 28,38 %, ce qui est nettement inférieur à 50 %. Cela signifie que la proportion des attaques par email réussies est bien plus faible que ce que nous attendions si 50 % des attaques étaient réussies.
- **Statistique de test z :**La statistique de test z calculée est -10.52, ce qui est bien inférieur à la valeur critique de 1.645 pour un test unilatéral à droite (qui est utilisé pour vérifier si la proportion de succès est supérieure à 50 %). La statistique z est négative, ce qui indique que la proportion observée est bien plus faible que la proportion théorique de 50 %.
- **Valeur critique :**La valeur critique pour un test unilatéral à droite à un niveau de signification de alpha = 0,05 est 1.645. Si la statistique de test z était supérieure à cette valeur critique, nous rejetterions l'hypothèse nulle (Ho).

➢ **Conclusion du test** :Comme la statistique de test z (-10.52) est bien inférieure à la valeur critique de 1.645, nous ne rejetons pas l'hypothèse nulle (Ho). En d'autres termes, les données ne fournissent pas suffisamment de preuves pour soutenir l'hypothèse alternative (H1), qui suggère que plus de 50 % des attaques par email sont réussies.Les résultats du test indiquent que, à un niveau de signification de 5 %, la proportion des attaques par email réussies n'est pas supérieure à 50 %. En fait, la proportion observée est significativement inférieure à cette valeur, ce qui renforce l'idée que moins de 50 % des attaques par email sont réussies.

Deuxième hypothèse : plus que 40% des attaques sms sont réussies

Pour la même raison que l'hypothèse précédente, on procède par un test unilatéral. On a **Hypothèse nulle (Ho) :** La proportion des attaques par SMS réussies est égale ou supérieure à 40 % ($p \geq 0,40$) et **Hypothèse alternative (H1) :** La proportion des attaques par SMS réussies est inférieure à 40 % ($p < 0,40$).

➢ **Interprétation des résultats :** La proportion observée de succès dans les attaques par SMS est de 27,32 %, ce qui signifie qu'environ 27,32 % des attaques par SMS ont réussi. Les conditions de normalité sont remplies, avec un échantillon d'au moins 30 observations et un produit de la taille de l'échantillon, de la proportion observée et de la proportion d'échec supérieur à 5. La valeur critique pour un test unilatéral à gauche avec un niveau de signification de 5 % est -1,645. La statistique de test calculée est -3,61, ce qui est inférieur à la valeur critique, conduisant ainsi au rejet de l'hypothèse nulle. Cela fournit suffisamment de preuves pour conclure que la proportion de succès des attaques par SMS est significativement inférieure à 40 %.

Troisième hypothèse : la proportion de succès des attaques par sites web est inférieure ou égale à 35 %

On procède par un test unilatéral à gauche, Hypothèses :

❖ **Hypothèse nulle (Ho)** : La proportion de succès des attaques par sites web est inférieure ou égale à 35 %. Formellement, Ho: $p \leq 0.35$

❖ **Hypothèse alternative (H1) :** La proportion de succès des attaques par sites web est supérieure à 35 %. Formellement, H1: $p > 0.35$

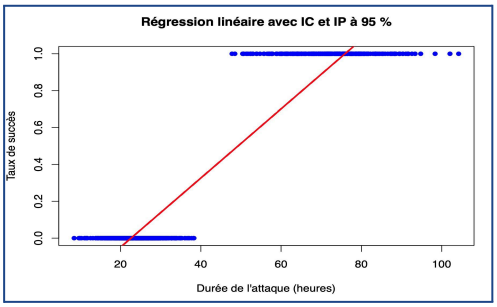
❖ **interprétation des résultats :** La proportion de succès observée des attaques par sites web (21,57 %) est nettement inférieur à la proportion théorique de 35 %. Les conditions de normalité nécessaires pour effectuer un test de proportion sont satisfaites, permettant ainsi l'analyse. Avec un niveau de signification de 5 %, la valeur critique pour rejeter l'hypothèse nulle (Ho) est $u_\alpha = -1,6449$. La statistique de test calculée ($u_{stat} = -2,844$) étant inférieure à cette valeur critique, l'hypothèse nulle est rejetée. Cela confirme, avec suffisamment de preuves statistiques, que la proportion de succès des attaques est significativement inférieure à 35 %.

5. Modélisation de la relation entre la durée de l'attaque et le taux de succès par régression linéaire

L'objectif de cette analyse est de modéliser la relation entre la **durée de l'attaque** et le **taux de succès** d'attaques de phishing dans une entreprise. En particulier, nous chercherons à comprendre comment la durée de l'attaque influence la probabilité de succès de l'attaque. Pour cela, nous utiliserons une régression linéaire, qui est un modèle statistique simple .

- Le taux de succès des attaques est: **0.278**
- la moyenne de la durée de l'attaque est: **37.43643**

❖ **Nuage de points** Le **nuage de points** a pour objectif principal de visualiser la **relation entre deux variables quantitatives**. Ce dernier a été créé pour visualiser la relation entre la **durée de l'attaque** et le **taux de succès**. Le graphe montre que, globalement, **plus la durée de l'attaque augmente, plus le taux de succès est élevé**. La pente positive de la ligne rouge traduit cette relation.Une tendance ou une relation linéaire semble exister entre les deux variables.



❖ **Covariance:** La covariance est: **COV xy = 9.745158.** Elle indique une **relation positive** entre le **taux de succès** et la **durée de l'attaque** *(detection_ time_hours)* . Cela signifie que, **lorsque le temps de détection augmente, la probabilité de succès a également tendance à augmenter.**

❖ **Coefficient de corrélation** Le coefficient de corrélation: **r = 0.95391**

➢ **Interprétation :** indique une **relation très forte et positive** entre le taux de succès et la durée de l'attaque **(detection_time_hours)** . Avec une valeur proche de 1, la corrélation montre que les deux variables sont presque parfaitement liées de manière linéaire.Cela signifie qu'**à mesure que le temps de détection augmente, la probabilité de succès augmente de manière cohérente.** La corrélation positive indique que **l'augmentation de la durée de l'attaque** (detection_time_hours) est associée à une **augmentation du taux de succès**.

❖ **Coefficient de pente (b1):** Avec une covariance de 9.745158 la valeur calculée pour b1 est : **b1=48.50342 .** Cela signifie qu'une augmentation de 1 unité dans la durée d'attaque entraîne une augmentation moyenne de **48.50342** dans la probabilité de succès.

❖ **Ordonnée à l'origine (bo) :**En utilisant les moyennes des deux variables, l'ordonnée à l'origine est calculée comme: **b0=23.95248.** Cela représente la valeur prédite pour le taux de succès lorsque la durée d'attaque est égale à 0.

❖ **Modèle de régression linéaire**

Équation estimée : **y=0.0188x−0.4243**

- **a=0.0188:** Chaque heure supplémentaire d'attaque augmente le taux de succès de **0,0188 (1,88 %)**.
- **b=−0.4243:** Lorsque la durée de l'attaque est nulle, le taux de succès est théoriquement négatif.

❖ **Qualité du modèle**

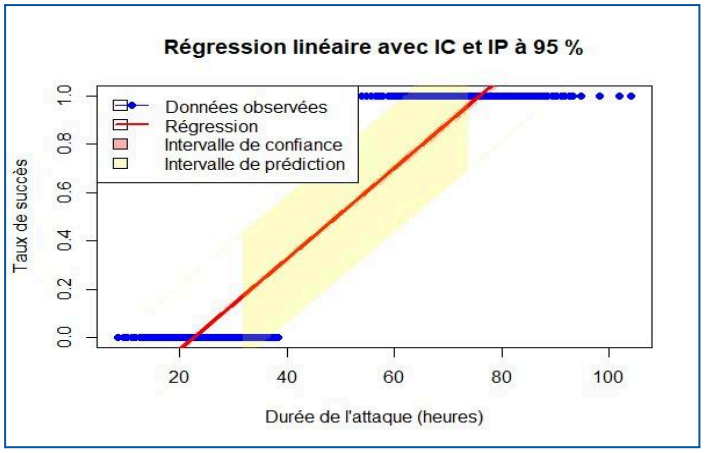
- **Coefficient de détermination (R^2)** Le **R^2=0.9099**, indiquant que **90,99** % de la variabilité du taux de succès est expliquée par la durée de l'attaque.
- **Somme des carrés des résidus (SCR)** La SCR est de **18.07564**, montrant que les écarts entre les valeurs observées et prédites restent faibles.
- **Estimation de la variance (σ^2):** La variance estimée des résidus **(σ^2=0.0181)** est également faible, renforçant la validité du modèle

❖ **Analyse des intervalles de confiance**

Intervalles pour les coefficients :

- **b0 :** [-0.4404, -0.4083]
- **b1 :** [0.0184, 0.0191]

➢ **Interprétation** : Les intervalles étroits montrent une estimation précise des coefficients. La pente (b1) étant strictement positive confirme une relation significative et positive entre la durée de l'attaque et le taux de succès.



Ce graphique montre que le **taux de succès** augmente avec la **durée de l'attaque (detection_time_hours)**. Une transition notable apparaît autour de **40 heures**, où le succès passe de faible (0) à élevé (1).