Introduction to Hypothesis Testing, Exam 2025

Duration: 2 hours, no document allowed. Special attention will be given to clarity of writing.

Exercise 1: Testing a Preference for Renewable or Non-Renewable Energy Sources

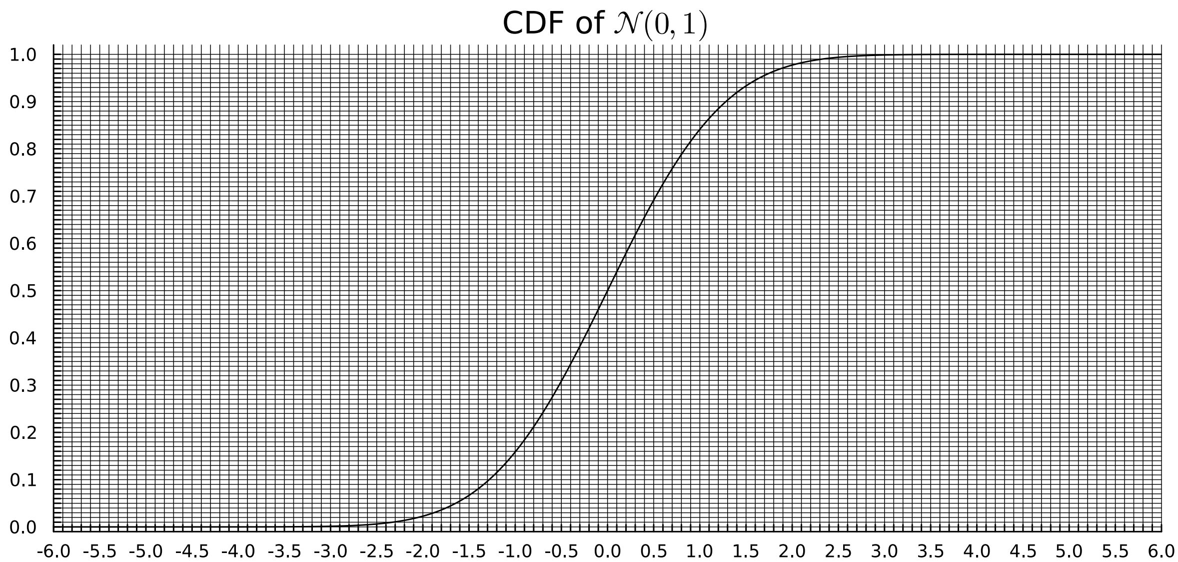
We aim to determine whether citizens in a region have **any** preference for **renewable energy sources** (e.g., solar, wind) or **non-renewable energy sources** (e.g., coal, natural gas). We assume that, **a priori**, there is no preference on average. We survey n individuals, and let X be the number of respondents who prefer renewable energy.

Questions:

1. Formalize the hypothesis testing problem, and define H_0 and H_1. Indicate whether this test is one-tailed or two-tailed.
2. We survey $n = 100$ individuals, and $X = 58$ prefer renewable energy sources. Write the p_{value} in function of F, the cdf of $\text{Bin}(100, 0.5)$ (binomial distribution with parameter $p = 0.5$).
3. Write a line of code that would compute the exact p-value in **Julia**, **Python**, or **R**.
4. Give an approximation of the p-value using a Gaussian approximation and the graph of the cdf of $\mathcal{N}(0, 1)$ given bellow. What do you conclude?
5. Redefine the alternative hypothesis H_1 and compute an approximated p-value if we aim to determine whether citizens have a preference for:

   a. renewable energy sources.
   b. Citizens prefer non-renewable energy sources.

   What do you conclude for these two other problems?



Exercise 2: Environmental Monitoring of River Pollution

An environmental agency is monitoring the pollution levels of a river to determine whether a nearby factory is causing an **increase** in harmful chemical concentration. The target concentration for a specific chemical is 15 ppm (parts per million), which is considered safe for aquatic life. For a sample of $n = 20$ water samples taken downstream from the factory, the empirical mean concentration is $\bar{X}_n = 16.3\,\text{ppm}$, and the empirical variance is $S_n^2 = 2.4\,\text{ppm}^2$. **A priori**, the river is assumed to meet the safe pollution threshold of 15 ppm.

We aim to test at the level of significance $\alpha = 0.05$ whether the chemical concentration downstream exceeds the safe threshold, indicating pollution from the factory.

Questions:

1. Using a Gaussian assumption, formalize the hypothesis testing problem, and define H_0 and H_1. Is this a one-tailed or two-tailed testing problem? Precise what the unknown parameters are.
2. Define the test statistic. What is its distribution under H_0?
3. Determine the rejection region. You can use a Gaussian approximation and the cdf of Exercise 1.
4. Write a line of code do compute the exact rejection threshold.
5. Does the river exhibit an increased chemical concentration that could indicate pollution from the factory?

Exercise 3: Bird Migration Habitat Distribution Analysis

A wildlife researcher is studying the behavior of a certain species of birds that migrate to a nature reserve. The researcher has a hypothesis about how the birds distribute themselves across different types of habitats in the reserve. The expected distribution, based on historical data, is as follows:

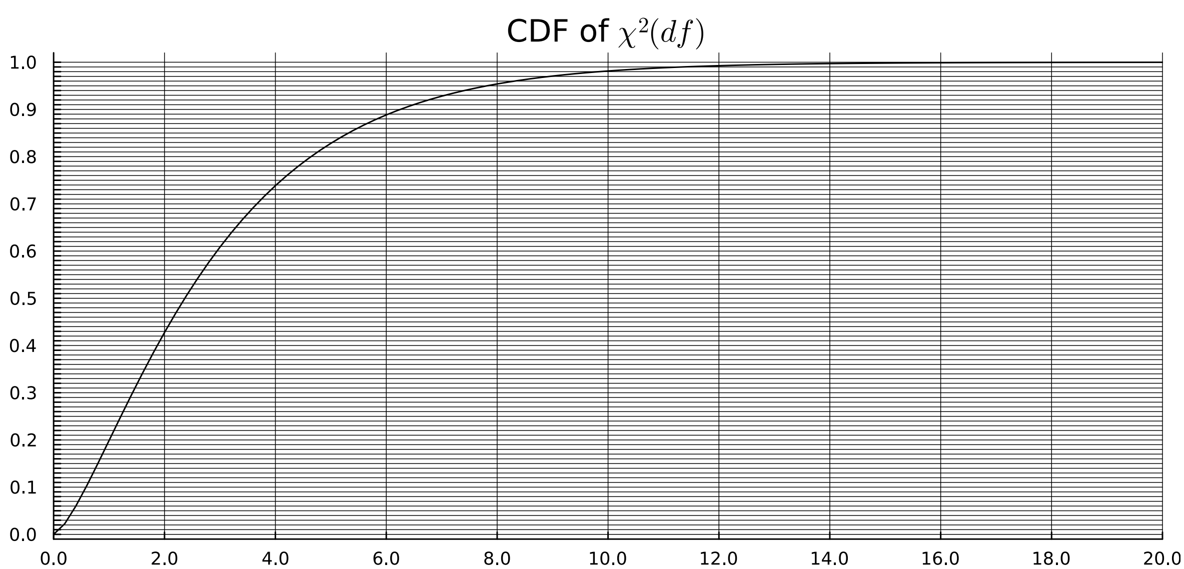
- **Grassland**: 40%
- **Wetlands**: 30%
- **Forests**: 20%
- **Rocky Areas**: 10%

To test this hypothesis, the researcher surveys 200 birds and records their habitat preferences. The observed counts are as follows:

Habitat	Grassland	Wetlands	Forests	Rocky Areas
Observed	90	60	30	20

Questions:

1. Formalize the hypothesis testing problem, and define H_0 and H_1.
2. Compute the expected counts.
3. Compute the chi-square statistic.
4. Determine the degree of freedom df of the chi-square statistic, and read the p-value on the following graph of the cdf.
5. What do you conclude?



Exercise 4

Employee Productivity Across Departments

A company wants to evaluate whether a new management style has had a consistent effect on employee productivity across five departments. Each department has adopted a specific variation of the management style for three months, and the company has recorded the average number of tasks completed per employee during that period.

Data:

Department	1	2	3	4	5
Number of employees	12	10	8	9	11
Average tasks completed	72.4	68.9	75.6	74.3	69.7
Variance of tasks	8.5	9.2	10.1	7.8	9.6

The company seeks to understand whether productivity levels vary significantly across departments, indicating that the management styles might have different impacts.

Let $d = 5$ be the number of departments and $N_{\text{tot}} = 50$ the total number of employees. For any department j, we denote N_k the number of employees in department k, and P_{ik} the number of tasks completed by employee i in department k. We assume that the P_{ik}'s are independent and normally distributed with mean μ_k and variance σ^2.

We write

$$\begin{aligned} \overline{P}_k &= \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} P_{ik} \\ \overline{P} &= \tfrac{1}{N_{\text{tot}}}\textstyle\sum_{k=1}^{d} N_k \overline{P}_k \end{aligned} \qquad \begin{aligned} V_k &= \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k}(P_{ik} - \overline{P}_k)^2 \\ V_W &= \tfrac{1}{N_{\text{tot}}}\textstyle\sum_{k=1}^{d} N_k V_k \\ V_B &= \tfrac{1}{N_{\text{tot}}}\textstyle\sum_{k=1}^{d} N_k(\overline{P}_k - \overline{P})^2 \\ V_T &= \tfrac{1}{N_{\text{tot}}}\textstyle\sum_{k=1}^{d}\sum_{i=1}^{N_k}(P_{ik} - \overline{P})^2 \end{aligned}$$

Questions

1. Define the hypotheses of the problem to test whether the management styles had a uniform impact on productivity.
2. Give a brief interpretation of each one of the quantities $\overline{P}_k$, $\overline{P}$, V_k, V_W, V_B, V_T.
3. Prove the analysis of variance formula $V_T = V_W + V_B$
4. Calculate $\overline{P}$, V_W, V_B, and V_T.
5. Express the ANOVA test statistic in terms of V_W and V_B.

6. What are the distributions of $N_k V_k$ and of $N_{\text{tot}} V_W$ under H_0? Do they change under H_1?
7. Recall the definition of ANOVA test statistic, and perform the ANOVA test at significance level $\alpha = 0.05$. We give the 0.05 and 0.95-quantiles of $\mathcal{D}$ which are approximately 0.18 and 2.58.
   Conclude whether productivity differs significantly between departments.

Course Questions

1. Recall the definition of a test statistic ψ and a test (or decision rule) T.
2. What are the two types of errors that we can commit?
3. For a given test statistic ψ, recall the definition of the p-value in the context of a **two-sided** test.
4. State the Neyman-Pearson theorem.

French Version

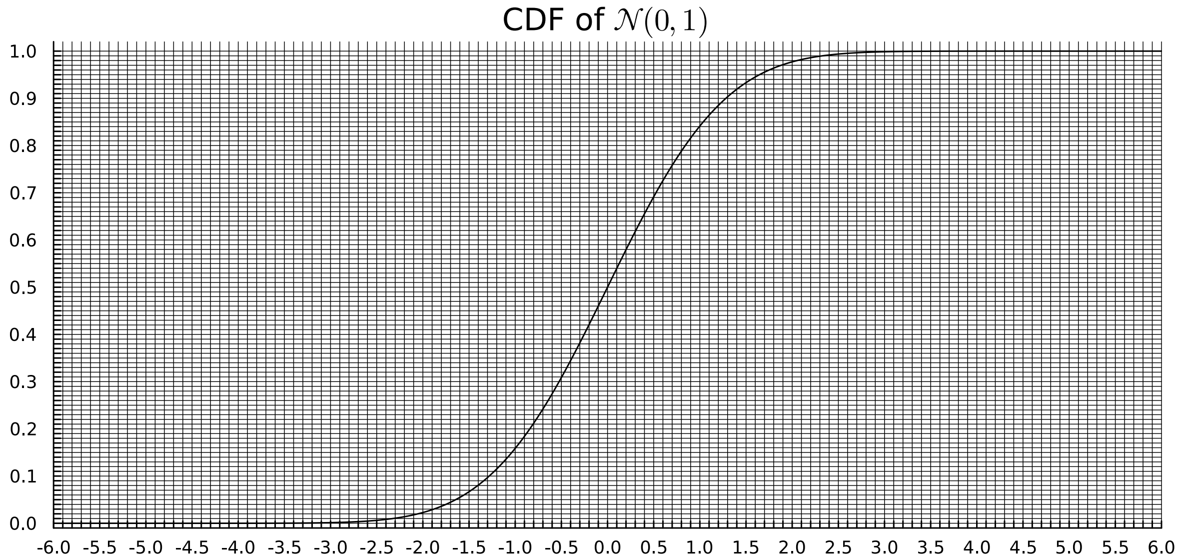
Exercice 1 : Test d'une préférence pour les sources d'énergie renouvelables ou non renouvelables

Nous cherchons à déterminer si les citoyens d'une région ont une quelconque préférence pour les sources d'énergie renouvelables (par exemple, solaire, éolienne) ou les sources d'énergie non renouvelables (par exemple, charbon, gaz naturel). Nous supposons que, a priori, il n'y a aucune préférence en moyenne. Nous interrogeons n individus, et nous notons X le nombre de répondants qui préfèrent les énergies renouvelables.

Questions :

1. Formalisez le problème de test d'hypothèse, et définissez H_0 et H_1. Indiquez si ce test est unilatéral ou bilatéral.
2. Nous interrogeons $n = 100$ individus, et $X = 58$ préfèrent les sources d'énergie renouvelables. Écrivez la p-valeur en fonction de F, la fonction de répartition d'une distribution Binomiale $\mathrm{Bin}(100, 0.5)$.
3. Écrivez une ligne de code qui calculerait la p-valeur exacte en Julia, Python, ou R.
4. Donnez une approximation de la p-valeur en utilisant une approximation gaussienne et le graphique de la fonction de répartition de $\mathcal{N}(0, 1)$ fourni dans le sujet. Quelle est votre conclusion ?
5. Redéfinissez l'hypothèse alternative H_1 et calculez une p-valeur approximative si nous cherchons à déterminer si les citoyens ont une préférence pour :

a. les sources d'énergie renouvelables.
b. les sources d'énergie non renouvelables. Quelles sont vos conclusions pour ces deux autres problèmes ?



Exercice 2 : Surveillance environnementale de la pollution fluviale

Une agence environnementale surveille les niveaux de pollution d'une rivière pour déterminer si une usine à proximité provoque une **augmentation** de la concentration de produits chimiques nocifs. La concentration cible pour un produit chimique spécifique est de 15 ppm (parties par million), ce qui est considéré comme sûr pour la vie aquatique. Pour un échantillon de $n = 20$ prélèvements d'eau effectués en aval de l'usine, la concentration moyenne empirique est $\bar{X}_n = 16,3\,\text{ppm}$, et la variance empirique est $S_n^2 = 2,4\,\text{ppm}^2$.

A priori, on suppose que la rivière respecte le seuil de pollution sans danger de 15 ppm.

Nous visons à tester, avec un niveau de signification $\alpha = 0,05$, si la concentration chimique en aval dépasse le seuil de sécurité, indiquant une pollution provenant de l'usine.

Questions :

1. En utilisant une hypothèse Gaussienne, formalisez le problème de test d'hypothèse et définissez H_0 et H_1. S'agit-il d'un test unilatéral ou bilatéral ?
2. Définissez la statistique de test. Quelle est sa distribution sous H_0 ?
3. Déterminez la zone de rejet. Vous pouvez utiliser une approximation Gaussienne et le graphe de l'exercice 1.
4. Ecrivez une ligne de code permettant de calculer le seuil de rejet exact.

5. La rivière présente-t-elle une concentration chimique accrue qui pourrait indiquer une pollution provenant de l'usine ?

Exercice 3 : Analyse de la distribution des habitats lors de la migration des oiseaux

Un chercheur en faune sauvage étudie le comportement d'une certaine espèce d'oiseaux qui migrent vers une réserve naturelle. Le chercheur a une hypothèse sur la façon dont les oiseaux se répartissent entre différents types d'habitats dans la réserve. La distribution attendue, basée sur des données historiques, est la suivante :

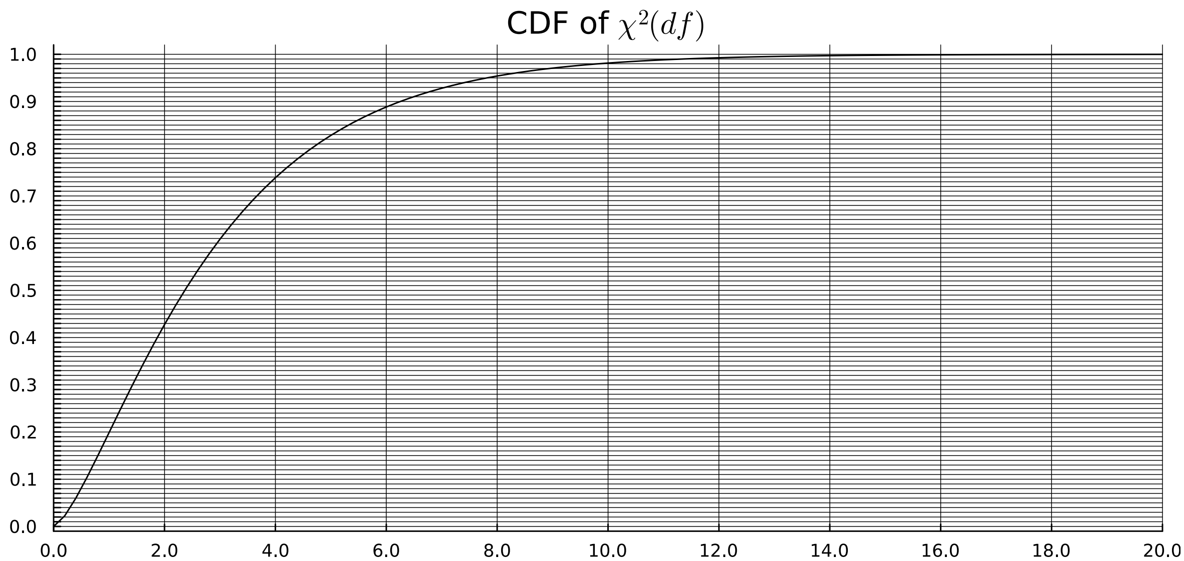
- **Prairie** : 40%
- **Zones humides** : 30%
- **Forêts** : 20%
- **Zones rocheuses** : 10%

Pour tester cette hypothèse, le chercheur observe 200 oiseaux et enregistre leurs préférences d'habitat. Les comptages observés sont les suivants :

Habitat	Prairie	Zones humides	Forêts	Zones rocheuses
Observé	90	60	30	20

Questions :

1. Formalisez le problème de test d'hypothèse et définissez H_0 et H_1.
2. Calculez les effectifs attendus.
3. Calculez la statistique du chi-deux.
4. Déterminez le degré de liberté df de la statistique du chi-deux, et lisez la p-value sur le graphique suivant de la fonction de répartition.
5. Quelle est votre conclusion ?



Exercice 4

Productivité des employés entre départements

Une entreprise souhaite évaluer si un nouveau style de management a eu un effet uniforme sur la productivité des employés dans cinq départements. Chaque département a adopté une variation spécifique du style de management pendant trois mois, et l'entreprise a enregistré le nombre moyen de tâches accomplies par employé durant cette période.

Données :

Département	1	2	3	4	5
Nombre d'employés	12	10	8	9	11
Moyenne des tâches accomplies	72,4	68,9	75,6	74,3	69,7
Variance des tâches	8,5	9,2	10,1	7,8	9,6

L'entreprise cherche à comprendre si les niveaux de productivité varient significativement entre les départements, indiquant que les styles de management pourraient avoir des impacts différents.

Soit $d = 5$ le nombre de départements et $N_{\text{tot}} = 50$ le nombre total d'employés. Pour tout département j, nous notons N_k le nombre d'employés dans le département k, et P_{ik} le nombre de tâches accomplies par l'employé i dans le département k. Nous supposons que les P_{ik} sont indépendants et suivent une distribution normale de moyenne μ_k et de variance σ^2.

Nous écrivons

$$\begin{aligned}
&& V_k &= \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k}(P_{ik}-\overline{P}_k)^2\\
\overline{P}_k &= \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} P_{ik} & V_W &= \tfrac{1}{N_{\text{tot}}}\textstyle\sum_{k=1}^{d} N_k V_k\\
\overline{P} &= \tfrac{1}{N_{\text{tot}}}\textstyle\sum_{k=1}^{d} N_k \overline{P}_k & V_B &= \tfrac{1}{N_{\text{tot}}}\textstyle\sum_{k=1}^{d} N_k(\overline{P}_k-\overline{P})^2\\
&& V_T &= \tfrac{1}{N_{\text{tot}}}\textstyle\sum_{k=1}^{d}\sum_{i=1}^{N_k}(P_{ik}-\overline{P})^2
\end{aligned}$$

Questions

1. Définissez les hypothèses du problème pour tester si les styles de management ont eu un impact uniforme sur la productivité.
2. Donnez une brève interprétation de chacune des quantités $\overline{P}_k$, $\overline{P}$, V_k, V_W, V_B, V_T.
3. Démontrez la formule d'analyse de la variance : $V_T = V_W + V_B$
4. Calculez $\overline{P}$, V_W, V_B, et V_T.
5. Exprimez la statistique de test ANOVA en termes de V_W et V_B.
6. Quelles sont les distributions de $N_k V_k$ et de $N_{\text{tot}} V_W$ sous H_0 ? Changent-elles sous H_1 ?
7. Rappelez la définition de la statistique de test ANOVA, donnez sa distribution $\mathcal{D}$ sous H_0 et effectuez le test ANOVA au niveau $\alpha = 0,05$. On donne les quantiles 0.05 et 0.95 de $\mathcal{D}$: 0.18 and 2.58. Concluez si la productivité diffère significativement entre les départements.

Questions de cours

1. Rappelez la définition d'une statistique de test ψ et d'un test (ou règle de décision) T.
2. Quels sont les deux types d'erreur que nous pouvons commettre ?
3. Pour une statistique de test ψ donnée et un problème de test bilatéral, rappelez la définition de la p-valeur.
4. Énoncez le théorème de Neyman-Pearson.