Experiment 5

Aim: To Explore the inferential statistics on the given dataset.

Theory:

What is Inferential Statistics?

Inferential statistics extends beyond the basic data summarization characteristic of descriptive statistics. By analyzing sample data, it equips us with the tools to make broader, population-level predictions, test hypotheses, estimate values, and calculate the margin of error or uncertainty inherent in those predictions.

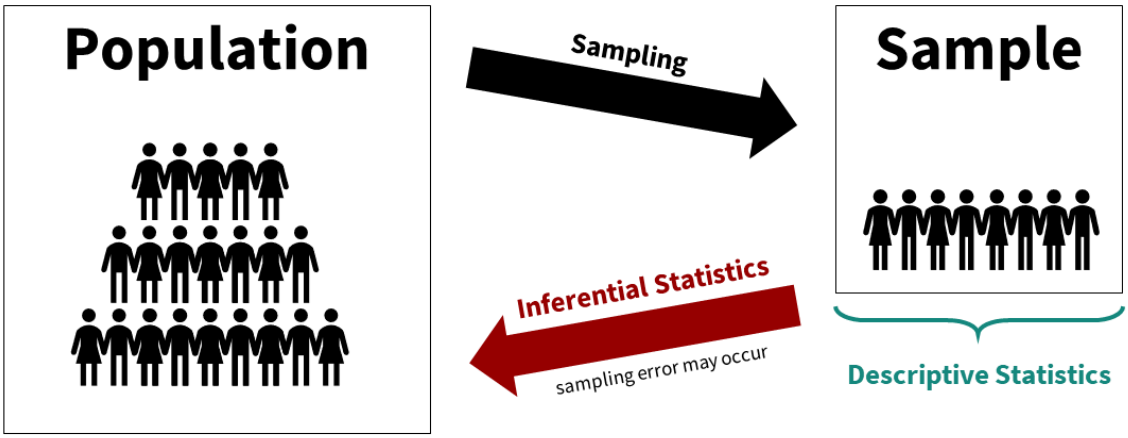

Fig 1: Overview of Statistics

1. **Techniques in Inferential Statistics**

Inferential statistics offers several key methods for testing hypotheses, estimating population parameters, and making predictions. Here are the major techniques:

Confidence Intervals:

It gives us a range of values that likely includes the true population parameter. It helps quantify the uncertainty of an estimate. The formula for calculating a confidence interval for the mean is:

$$CI = \bar{x} \pm Z_{(\alpha/2)} * (\sigma / \sqrt{n})$$

Equation 5.1

where:

- $\bar{x}$ *- is the sample mean*
- $Z_{(\alpha/2)}$ *- is the Z-value from the standard normal distribution (e.g., 1.96 for a 95% confidence interval)*

- *σ - is the population standard deviation*
- *n - is the sample size*

For example, if we measure the average height of 100 people, a 95% confidence interval gives us a range where the true population mean height is likely to fall. This helps gauge the precision of our estimate and compare models (like in A/B testing).

Hypothesis Testing:

Hypothesis testing is a formal procedure for testing claims or assumptions about data. It involves the following steps:

- **Null Hypothesis (H_0):** The default assumption, such as "there's no difference between two models."
- **Alternative Hypothesis (H_1):** The claim you aim to prove, such as "Model A performs better than Model B."

We collect data and compute a test statistic (such as Z for a Z-test or t for a T-test):

$$Z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$$

Equation 5.2

where:

$\bar{x}$ - is the sample mean

μ_0 - is the hypothesized population mean

σ - is the population standard deviation

n - is the sample size

After calculating the test statistic, we compare it with a critical value or use a p-value to decide whether to reject or accept the null hypothesis. If the p-value is smaller than the significance level α (usually 0.05), we reject the null hypothesis.

$$\text{p-value} = 2 * P(Z > |z_{obs}|)$$

Equation 5.3

where:

z_{obs} - is the observed test statistic.

A small p-value suggests strong evidence against the null hypothesis.

Central Limit Theorem:
It states that the distribution of the sample mean will approximate a normal distribution as the sample size increases, regardless of the original population distribution. This is important because many statistical methods assume that data is normally distributed. The CLT can be mathematically expressed as:

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$
Equation 5.4

where:
μ - is the population mean
σ - is the population standard deviation
n - is the sample size

This theorem allows us to apply normal distribution-based methods even when the original data is not normally distributed, such as in cases with skewed income or shopping behavior data.

2. **Errors in Inferential Statistics**

In hypothesis testing, Type I Error and Type II Error are key concepts:

- **Type I Error** occurs when we wrongly reject a true null hypothesis. The probability of making a Type I error is denoted by α (the significance level).
- **Type II Error** occurs when we fail to reject a false null hypothesis. The probability of making a Type II error is denoted by β and the power of the test is given by $1-\beta$.

Dataset - Employee Attribution Dataset

This dataset consists of 31 columns and 1,470 rows, providing a comprehensive snapshot of a corporate workforce.

- **Context:** The Employee dataset contains historical data regarding demographics, compensation, and professional engagement metrics for individual employees within an organization.
- **Objective:** The primary objective is to monitor and analyze workforce dynamics, salary structures, and the factors influencing employee performance and retention.
- **Identification Attributes:** Each entry is identified by an employee index, with categorical attributes such as **Department**, **JobRole**, and **EducationField** defining the organizational structure.
- **Core Target Variable:** The **MonthlyIncome** serves as a central metric, representing the primary financial compensation assigned to each employee.
- **Career Indicators:** Attributes such as **TotalWorkingYears**, **JobLevel**, and **YearsAtCompany** provide critical insights into the professional experience and tenure of the staff.
- **Engagement & Environment:** Metrics including **JobSatisfaction**, **WorkLifeBalance**, and **EnvironmentSatisfaction** help evaluate the qualitative aspects of the workplace and their potential impact on employee attrition.

Implementation:

Code:

```
import pandas as pd
import scipy.stats as stats

# Load the dataset
df = pd.read_csv("empattds.csv")

print("\n-----------------------------")
print("Part A: Pearson Correlation")
print("-----------------------------")
# Testing linear relationship between Total Working Years and Monthly Income
corr_coef, p_val_a = stats.pearsonr(df["TotalWorkingYears"],
df["MonthlyIncome"])
print("Variables: TotalWorkingYears vs MonthlyIncome")
print(f"Correlation Coefficient (r): {corr_coef:.4f}")
print(f"P-value: {p_val_a:.4f}")

print("\n-----------------------------")
print("Part B: One-Way ANOVA")
print("-----------------------------")
# Testing if Monthly Income differs significantly across different Departments
departments = df["Department"].unique()
groups = [df[df["Department"] == dept]["MonthlyIncome"] for dept in
departments]

f_stat, p_val_b = stats.f_oneway(*groups)
print("Variables: MonthlyIncome grouped by Department")
print(f"F-statistic: {f_stat:.4f}")
print(f"P-value: {p_val_b:.4f}")

print("\n-----------------------------")
print("Part C: Chi-Square Test")
print("-----------------------------")
# Testing association between Attrition and OverTime
contingency_table = pd.crosstab(df["Attrition"], df["OverTime"])
chi2_stat, p_val_c, dof, expected = stats.chi2_contingency(contingency_table)

print("Variables: Attrition vs OverTime")
print("Contingency Table:")
print(contingency_table)
print(f"\nChi-Square Statistic: {chi2_stat:.4f}")
print(f"P-value: {p_val_c:.4f}")
```

Output:

```
(.venv) [aditya@archlinux ss]$ python main.py

----------------------------
Part A: Pearson Correlation
----------------------------
Variables: TotalWorkingYears vs MonthlyIncome
Correlation Coefficient (r): 0.7729
P-value: 0.0000

----------------------------
Part B: One-Way ANOVA
----------------------------
Variables: MonthlyIncome grouped by Department
F-statistic: 3.2018
P-value: 0.0410

----------------------------
Part C: Chi-Square Test
----------------------------
Variables: Attrition vs OverTime
Contingency Table:
OverTime     No  Yes
Attrition
No          944  289
Yes         110  127

Chi-Square Statistic: 87.5643
P-value: 0.0000
(.venv) [aditya@archlinux ss]$
```

Fig 3: Testing using Python

Pearson Correlation	0.7728932463					
Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Sales	446	3103791	6959.172646	16473364.88		
Research and Development	961	6036284	6281.252862	23969201.2		
Human Resources	63	419234	6654.507937	33509428.83		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	141509923.1	2	70754961.53	3.201782929	0.04097409725	3.001858137
Within Groups	32418665115	1467	22098612.89			
Total	32560175038	1469				

COUNTA of Attri OverTime			
Attrition	No	Yes	Grand Total
	0		0
No	944	289	1233
Yes	110	127	237
Grand Total	**1054**	**416**	**1470**

COUNTA of Attri OverTime			
Attrition	No	Yes	Grand Total
	0		0
No	884.0693878	348.9306122	1233
Yes	169.9306122	67.06938776	237
Grand Total	0 1054	416	1470

P Value	0.0000

Fig 4: Testing using Excel

Insights:

1) Pearson Correlation (`TotalWorkingYears` vs. `MonthlyIncome`)

Why we used the Pearson Correlation

- **Variable 1:** `TotalWorkingYears` is a continuous quantitative variable representing the total duration an employee has been in the workforce.
- **Variable 2:** `MonthlyIncome` is also a continuous quantitative variable representing the employee's salary.
- When we want to mathematically measure the strength and direction of a linear relationship between exactly two continuous variables, the Pearson Correlation Coefficient is the standard mathematical tool. It tells us whether an increase in one variable is reliably associated with an increase (or decrease) in the other.

The Hypotheses (H0 and H1)

- **Null Hypothesis (H0):** There is no statistically significant linear relationship. The true correlation coefficient is zero. Any apparent relationship in our sample is purely due to random chance. (Mathematical representation: $r = 0$)
- **Alternative Hypothesis (H1 or Ha):** There is a statistically significant linear relationship. As total working years increase, monthly income genuinely changes in a predictable, linear way. (Mathematical representation: $r \neq 0$)

The Verdict Because our p-value (0.0000) is much smaller than our alpha level (0.05), we strongly **reject the Null Hypothesis (H0)**.

- **What this means:** We have overwhelmingly strong statistical evidence to say that total working years and monthly income are genuinely related. Furthermore, our correlation coefficient ($r = 0.7729$) indicates a **strong positive correlation**. As far as the math is concerned, as employees accumulate more working years, their monthly income reliably and significantly increases.

2) One-Way ANOVA (`MonthlyIncome` across `Department`)

Why we used the One-Way ANOVA

- **Independent Variable:** `Department` is a categorical variable with three or more independent groups (e.g., Sales, Research & Development, Human Resources).
- **Dependent Variable:** `MonthlyIncome` is a continuous quantitative variable.
- When we want to compare the mean (average) of a continuous variable across *three or more* distinct categorical groups, the One-Way ANOVA (Analysis of Variance) is the

correct mathematical tool. If we only had two departments, we would use a T-Test; because we have more than two, ANOVA prevents the mathematical errors that would occur if we ran multiple overlapping T-Tests.

The Hypotheses (H0 and H1)

- **Null Hypothesis (H0):** There is no statistically significant difference in salaries across departments. The true mean monthly income is equal across all departments. Any differences in the sample averages are just random chance. (Mathematical representation: Mean 1 = Mean 2 = Mean 3...)
- **Alternative Hypothesis (H1 or Ha):** There is a statistically significant difference. At least one department has a true mean monthly income that is genuinely different from the others.

The Verdict Because our p-value (0.0410) is smaller than our alpha level (0.05), we **reject the Null Hypothesis (H0)**.

- **What this means:** We have enough statistical evidence to confidently say that salaries are not uniform across the company. The Alternative Hypothesis (H1) is supported. Statistically speaking, the department an employee works in has a genuine, significant impact on their average monthly income.

3) Chi-Square Test of Independence (`Attrition` vs. `OverTime`)

Why we used the Chi-Square Test of Independence

- **Variable 1:** `Attrition` is a categorical variable with exactly two groups (Yes = Left the company, No = Stayed).
- **Variable 2:** `OverTime` is a categorical variable with exactly two groups (Yes = Works overtime, No = Does not work overtime).
- When we want to see if two categorical variables are related or dependent on one another—specifically, if belonging to a certain category in one variable changes your probability of belonging to a category in the other—the Chi-Square Test of Independence is the correct mathematical tool. It compares our *observed* counts against the *expected* counts if the two variables had absolutely nothing to do with each other.

The Hypotheses (H0 and H1)

- **Null Hypothesis (H0):** The two variables are completely independent. Working overtime has absolutely no effect on whether an employee leaves the company. The proportion of people quitting is mathematically identical regardless of their overtime status.

- **Alternative Hypothesis (H1):** The two variables are not independent. Working overtime is statistically associated with a change in the likelihood of an employee leaving the company.

The Verdict Because our p-value (0.0000) is massively smaller than 0.05, we decisively **reject the Null Hypothesis (H0)**.

- **What this means:** The differences in the contingency table (specifically, the high proportion of overtime workers who quit compared to non-overtime workers) are mathematically significant. Statistically speaking, `OverTime` and `Attrition` are highly dependent. Working overtime is genuinely associated with a significantly higher rate of employees leaving the company.

Conclusion

Inferential statistics elevates data analysis from mere summarization to predictive validation, allowing us to draw objective, population-level conclusions from sample data. By utilizing frameworks like hypothesis testing, confidence intervals, and the Central Limit Theorem, we can quantify uncertainty and rigorously distinguish genuine trends from random variance. This phase of analysis is crucial for validating assumptions, without which observed differences in a dataset would remain mathematically unsubstantiated and potentially misleading.

An inferential analysis was performed using Python to explore the underlying factors driving compensation and turnover within the Employee Attrition dataset. Through Python, the execution of core statistical tests—specifically the Pearson correlation, One-Way ANOVA, and Chi-Square test of independence—was automated to rigorously evaluate these organizational dynamics. The hypothesis testing revealed highly significant, actionable HR insights: the Pearson Correlation yielded a p-value of 0.0000, mathematically confirming a strong positive relationship between an employee's total working years and their monthly income. Furthermore, the One-Way ANOVA (p-value of 0.0410) demonstrated that this average monthly income differs significantly across various departments. Finally, the Chi-Square test (p-value of 0.0000) exposed a critical driver of turnover, proving that working overtime is definitively associated with higher rates of employee attrition. Collectively, these tools and tests provided mathematical proof of the key variables influencing the workforce, emphasizing the necessity of data-driven rigor over intuition in human resources and retention strategies.

Post Experiment Questions:

1. **A data science study is analysing customer churn using a sample of 500 customers. Explain how hypothesis testing can be applied to determine where the churn rates differ significantly between 2 regions. Discuss the assumptions, test selection and interpretation of results.**

- **Test Selection**

  Because churn is a binary outcome (Yes/No) and we are comparing two independent groups, the most appropriate statistical test is the Two-Sample Z-Test for Proportions. Alternatively, a Pearson's Chi-Square Test for Independence would yield the exact same mathematical conclusion in this 2x2 scenario, but the Z-test is more direct when we specifically want to look at the difference between two rates.

- **Setting Up the Hypotheses**
  a. **Null Hypothesis (H0):** There is no difference in churn rates between the two regions. Mathematically, $p1 - p2 = 0$.
  b. **Alternative Hypothesis (HA):** There is a significant difference in churn rates between the two regions. Mathematically, $p1 \neq p2$.
  Because we are checking if Region A could be higher or lower than Region B, this is a two-tailed test.

- **Setting the Assumptions**

  Before calculating anything, the data must meet specific criteria for the Z-test to be mathematically valid:
  a. **Independence:** The customers in Region A must be completely distinct from Region B. One customer's decision to churn cannot influence another's.
  b. **Random Sampling:** The 500 customers must be a random, representative sample of the broader population in those regions.
  c. **Sufficient Sample Size (Success/Failure Condition):** we must have at least 10 "successes" (churns) and 10 "failures" (retention) in both regions. For example, if Region A has 200 customers, but only 3 churned, the sample size is too small for a normal distribution approximation, and we would need to use a non-parametric alternative like Fisher's Exact Test.

- **Calculating the Test Statistic**

  Now we calculate the Z-score. First, we calculate the "pooled proportion" ($\hat{p}$) , which is the total number of churned customers across both regions divided by the total sample size (500). Then, we plug our region-specific proportions ($\hat{p}1$ and $\hat{p}2$) and sample sizes (n1 and n2) into the Z-test formula.

- Interpretation of Results
  Once we have the Z-score, we determine the p-value, which is the probability of observing a difference this large (or larger) if the null hypothesis were true. We compare this p-value to our chosen significance level (α), typically set at 0.05.
    a. **If p-value < 0.05:** we reject the null hypothesis. We can confidently conclude to stakeholders that the difference in churn rates between Region A and Region B is statistically significant, meaning it is highly unlikely to be the result of random sample noise.
    b. **If p-value ≧ 0.05:** we fail to reject the null hypothesis. The observed difference in the 500-customer sample is not large enough to rule out random chance. We cannot conclusively say one region churns more than the other.

2. **Critically analyze the issue of multiple hypothesis testing in inferential learning, explain how these techniques control and mitigate the risk of type 1 errors**
   In inferential learning, if we are evaluating 50 different state features to see which ones significantly impact the reward function of a reinforcement learning agent, testing each feature individually creates a massive statistical trap known as the Multiple Comparisons Problem.

   The Compounding Risk of Type I Errors
   A Type I Error (a false positive) occurs when you reject a true null hypothesis. If you set your significance level (α) to 0.05, you accept a 5% risk of this happening for a single test. However, probabilities compound. If you test 50 independent features to see if they predict an outcome, and none of them actually do (the null hypothesis is true for all 50), the math turns against you.

   The probability of making at least one Type I error across m tests is calculated as:

   P(at least one Type I error) = 1 - (1 - α)^m
   - Equation 5.5

   If α = 0.05 and m = 50: 1 - (0.95)^{50} = 0.923

   We now have a 92.3% chance of finding at least one "statistically significant" feature purely by random noise. If you feed those noisy, false-positive features into a predictive model, it will overfit to the training data and fail in production.

Mitigation Strategy 1: Controlling FWER

The traditional way to handle this is to control the Family-Wise Error Rate (FWER). This ensures that the probability of making even one false discovery across all your tests stays at your original α (e.g., 5%).

The Bonferroni Correction

The most common FWER technique is the Bonferroni correction. It simply divides your target α by the number of tests (m) to create a new, brutally strict threshold:

$$\alpha\{adjusted\} = \alpha/m$$

- Equation 5.6

If you run 50 tests with a desired FWER of 0.05, your new threshold for each individual test is 0.05 / 50 = 0.001. A p-value must be lower than 0.001 to be considered significant.

- **Strength:** It virtually guarantees you won't get false positives. It is highly effective in fields like clinical medicine, where a false positive could mean approving a deadly drug.
- **Limit:** It destroys statistical power. By making the threshold so strict, you drastically increase your Type II Errors (false negatives).

Mitigation Strategy 2: Controlling FDR

In modern data science and machine learning, FWER is often considered too draconian. Instead of trying to guarantee zero false positives, we control the False Discovery Rate (FDR). FDR accepts that false positives will happen, but it controls the proportion of false positives among the tests you actually declare significant.

The Benjamini-Hochberg (BH) Procedure

The BH procedure ranks your results and applies a sliding scale to the threshold.

1. *Sort all m of your p-values from smallest (most significant) to largest: $p1 \leqq p2 \leqq \ldots \leqq pm$.*
2. *Assign a rank i to each p-value (where i = 1 for the smallest, up to m).*
3. *Compare each p-value to a sliding threshold: (i/m)*Q, where Q is your acceptable false discovery rate (e.g., 0.05).*
4. *Find the largest p-value that is smaller than its sliding threshold. That value, and all p-values smaller than it, are deemed significant.*

- **Strength:** It preserves statistical power. If you are filtering features for a model, setting an FDR of 0.05 means you accept that 5% of the features you keep might be useless noise, but you successfully capture the vast majority of the truly useful features
- **Limit**: It is slightly riskier than Bonferroni, as you are actively permitting a known fraction of garbage data into your "significant" pool.