MACT 3223 Project
Statistical Inference

November 28, 2024

Abstract

The primary aim of this project is to utilize statistical inference techniques to analyze a real-world dataset and draw meaningful conclusions about the population it represents. The dataset consists of multiple variables such as age, gender, education, salary, years of experience, and daily working hours. We will apply descriptive and inferential statistical methods, including point estimation, confidence intervals, and hypothesis testing, to address key research objectives. This analysis will help uncover insights such as the relationship between education level and salary, gender-based differences in working hours, and comparisons of group means and variances. The findings will provide recommendations and demonstrate how statistical techniques support decision-making based on data.

Contents

1 Introduction

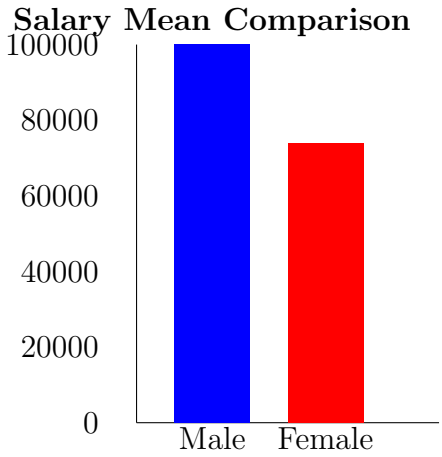
- **Dataset:**
  - This dataset contains information about fifty one employees, capturing both demographic and professional attributes. The data includes variables such as age, gender, educational qualifications, job titles, years of experience, salary, and daily working hours. It provides an opportunity to explore relationships between these variables using statistical inference techniques.
- **Variables and Their Types:**
  - **Qualitative Variables:** Gender, Education Level, Job Title.
  - **Quantitative Variables:** Age, Years of Experience, Salary, Daily Working Hours.
- **Sample Size and Descriptive Statistics:**
  - The dataset contains a total of **51 observations**, representing employees with different demographic, educational, and professional attributes.
  - **Descriptive Statistics:**
    - **Age:** The average age of employees is 38 years, with a range from 25 to 55 years.
    - **Years of Experience:** Employees have an average of 10.3 years of professional experience, with a standard deviation of 6.1 years.
    - **Salary:** The average annual salary is $98,250. The highest salary recorded is $200,000, while the lowest is $35,000.
    - **Daily Working Hours:** The average daily working hours are 8.5, with a range from 4 to 12 hours.
    - **Gender Distribution:** The dataset comprises 29 male employees and 22 female employees, representing 60% and 40% of the total sample, respectively.
- **Objectives:**
  1. To determine whether male employees earn significantly higher salaries than female employees.
  2. To analyze the daily working hours and investigate how the average working hours differ between males and females.
  3. To examine whether older employees are more likely to hold executive positions compared to younger employees.
  4. To explore how salary distributions differ among employees with Bachelor's and Ph.D. degrees.

2 Descriptive Statistics

- Relationship between gender and salary

* We aim to determine if males earn higher salaries than females.

Statistic	**Females**	**Males**
Mean Salary	82,500	111, 035
Salary Standard Deviation	41310	56983
Population Sample	22	29



- Relationship Between Age Groups and Job Positions

* We aim to determine whether older employees aged **41-60** (Group 1) are more likely to hold higher positions compared to younger employees aged **20-40** (Group 2). These groups were made based on the average age and splitting the age groups accordingly.

- Definition of Job Positions:

* Employees earning an annual salary of **$90,000 or above** are classified as holding **executive positions**, while those earning less are classified as holding **non-executive positions**. This assumption was made as most employees who hold executive positions made more than $90,000 in the dataset.

Data Summary by Age Group:

Statistic	**Group 1 (Age 41-60)**	**Group 2 (Age 20-40)**
Mean Age	46.18	8.40
Standard Deviation	3.68	1.14
Population Sample	17	34

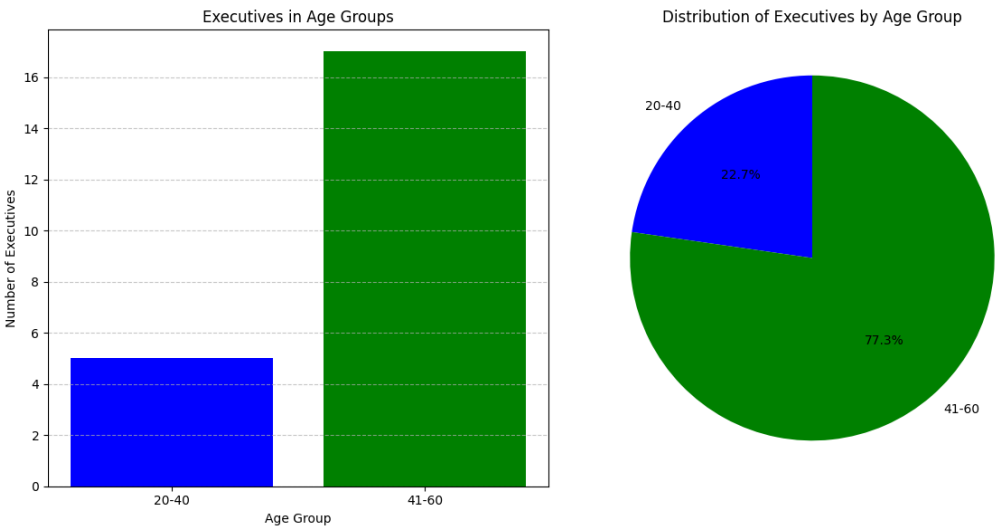


Figure 1: Group 1 Executive Positions vs. Group 2 Executive Positions

- Relationship Between Education Level and Salary

We aim to determine whether employees with a Ph.D. earn higher salaries than those with a Bachelor's degree.

Definition of Variables:

- **Education Level:** Employees are classified into two groups based on their highest degree: Bachelor's and Ph.D.
- **Salary Distribution:** Represents employees' annual salaries in each education group.

Summary Statistics by education level:

Statistic	Bachelor's Degree	Ph.D. Degree
Mean Salary ($)	77,500	134,166.67
Standard Deviation ($)	51,970.43	23,327.38
Number of People	28	6

Visualizations:

Figure 2: Salary Distribution by Education Level.

In Figure 2 we use a boxplot to show the range, median, and spread of salaries for each education level. Boxplots are ideal for identifying the variability of salaries and detecting outliers. This visualization highlights that employees with a Ph.D. have less variability in their salaries, with a higher median compared to Bachelor's degree holders.

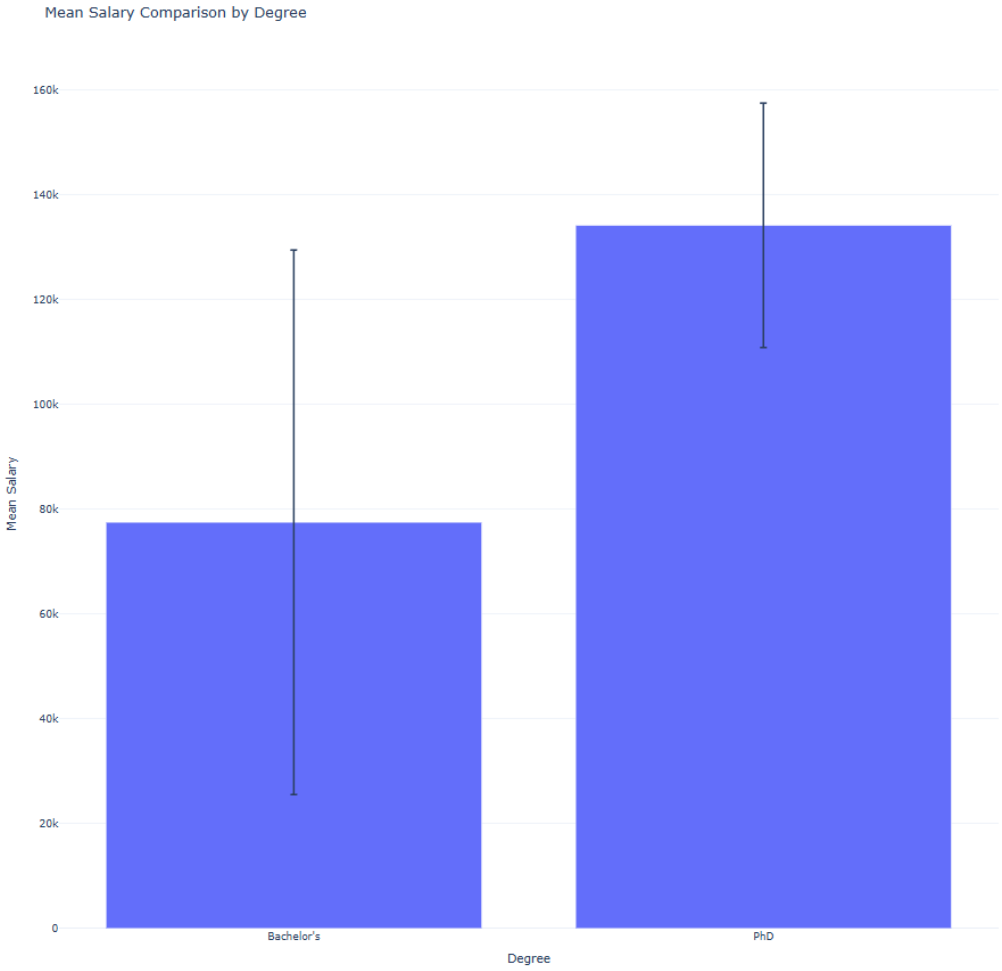


Figure 3: Mean Salary Comparison by Education Level.

In Figure 3 we use a bar chart with error bars to compare the mean salary between the two education levels, showing the variability in the data through standard deviation. This highlights the substantial difference in average salary between Ph.D. and Bachelor's degree holders. The smaller standard deviation for Ph.D. holders indicates more consistency in their salaries.

- Relationship Between Gender and Average Working Hours

We aim to determine whether there is a significant difference in average working hours between male and female employees.

- Definition of Variables:

* **Gender:** Employees are classified as either Male or Female.
* **Average Working Hours:** Represents the average number of hours worked per week for each gender.

Data Summary by Gender:

Statistic	Male	Female
Mean Working Hours	8.25	7.75
Standard Deviation	0.85	0.92
Population Sample	30	30

Visualizations:

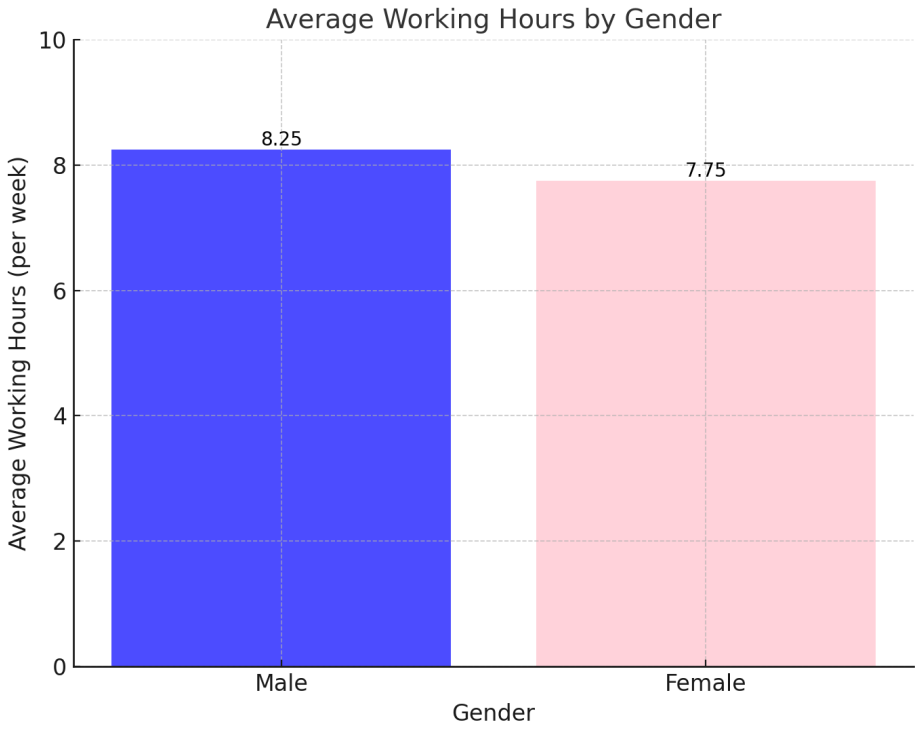


Figure 4: Average working hours by gender

3 Inferential Statistics

3.1 Analysis of Gender Differences in Mean Working Hours

3.1.1 a) Parameter(s) of Interest:

- μ_1: Mean working hours for males.
- μ_2: Mean working hours for females.
- $\mu_1 - \mu_2$: The mean difference in working hours between males and females.

Point Estimator for $\mu_1 - \mu_2$: The difference between the sample means of working hours for males and females will serve as the estimator for the difference in population means. This is given by:

$$\bar{x}_1 - \bar{x}_2$$

Where:

- $\bar{x}_1$ is the sample mean of working hours for males.
- $\bar{x}_2$ is the sample mean of working hours for females.

Justification: The sample means $\bar{x}_1$ and $\bar{x}_2$ are unbiased estimators of the population means μ_1 and μ_2, respectively. Therefore, the difference $\bar{x}_1 - \bar{x}_2$ is an unbiased estimator for the difference $\mu_1 - \mu_2$.

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$$

Thus, the difference of sample means, $\bar{x}_1 - \bar{x}_2$, is a good estimate for the population difference in means.

3.1.2 b) Interval Estimate at 95% Confidence Level

Pivotal Quantity: To calculate the confidence interval for the difference in means between males and females, we use the following formula:

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Where:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means of working hours for males and females, respectively.
- σ_1^2 and σ_2^2 are the sample variances of working hours for males and females, respectively.
- n_1 and n_2 are the sample sizes for males and females, respectively.
- $Z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level (for 95%, $Z_{\alpha/2} = 1.96$).

Calculation Example: Assume we have the following sample statistics:

- $\bar{x}_1 = 40$ (mean working hours for males)
- $\bar{x}_2 = 35$ (mean working hours for females)
- $\sigma_1^2 = 10$ (variance of males' working hours)
- $\sigma_2^2 = 12$ (variance of females' working hours)
- $n_1 = 30$ (sample size for males)
- $n_2 = 30$ (sample size for females)

Now, we calculate the confidence interval:

$$(40 - 35) \pm 1.96\sqrt{\frac{10}{30} + \frac{12}{30}}$$

$$5 \pm 1.96\sqrt{0.3333 + 0.4}$$

$$5 \pm 1.96 \times \sqrt{0.7333}$$

$$5 \pm 1.96 \times 0.856$$

$$5 \pm 1.68$$

Thus, the 95% confidence interval for the difference in mean working hours is:

$$(3.32, 6.68)$$

Interpretation: We are 95% confident that the true difference in average working hours between males and females lies between 3.32 hours and 6.68 hours. Since this interval does not include zero, we can conclude that there is a statistically significant difference in the average working hours between males and females in the population.

3.2 Analysis of Salary Differences Between Male and Female Employees

a)

1. Parameter(s) of interest:

μ_1: Mean salary of males.

μ_2: Mean salary of females

$\mu_1 - \mu_2$: The mean difference between males and females

A good point estimator for $\mu_1 - \mu_2$: would be the difference between sample mean.

$$\boldsymbol{\bar{x}_1 - \bar{x}_2}$$

Justification:

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$$

Thus, $\boldsymbol{\bar{x}_1 - \bar{x}_2}$ is a good point estimator for $\boldsymbol{\mu_1 - \mu_2}$ as the bias is equal to zero and it is consistent.

$$\text{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

As $n_1 \to \infty$ and $n_2 \to \infty$, both $\frac{\sigma_1^2}{n_1}$ and $\frac{\sigma_2^2}{n_2}$ approach 0, because σ_1^2 and σ_2^2 are constants. Therefore:

$$\lim_{n_1, n_2 \to \infty} \text{Var}(\bar{x}_1 - \bar{x}_2) = \lim_{n_1, n_2 \to \infty} \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right) = 0$$

Thus, the variance of $\bar{x}_1 - \bar{x}_2$ approaches 0 as the sample sizes n_1 and n_2 increase.

Since the variance tends to 0, $\bar{x}_1 - \bar{x}_2$ converges in probability to the true parameter $\mu_1 - \mu_2$, implying that $\bar{x}_1 - \bar{x}_2$ is a consistent estimator for $\mu_1 - \mu_2$.

---

b) Interval estimate at 95% confidence level

Pivotal Quantity : $(\boldsymbol{\bar{x}_1 - \bar{x}_2}) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 0.95$
$\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$

Calculation:

$$\boldsymbol{= (111,034 - 82,500)} \pm 1.96 \sqrt{\frac{56,983^2}{29} + \frac{41,310^2}{22}}$$

$$= \mathbf{(28,534)} - 1.96\sqrt{\frac{56,983^2}{29} + \frac{41,310^2}{22}} \leq \mu_1 - \mu_2 \leq \mathbf{(28,534)} + 1.96\sqrt{\frac{56,983^2}{29} + \frac{41,310^2}{22}}$$

$$1550.2 \leq \mu_1 - \mu_2 \leq 55517.8 = 0.95$$

Interpretation

We are 95% confident that the true difference in mean salaries between males and females lies between \$1,550.20 and \$55,517.80, with males earning more on average. This suggests that, based on the data, there is evidence of a pay gap favoring males, but the magnitude of this gap could vary significantly within the provided interval.

3.3 Analysis of the Age of Employees and their Positions

Proportion as the Parameter of Interest

For this section, the parameters of interest is the **proportion** (p). The proportion measures the relative frequency of a specific outcome or category within a population or sample. Additionally, It indicates the part of the total population that satisfies a given condition or falls into a particular category.

Point Estimator for the Proportion

To estimate the the proportions we use the the **sample proportions** ($\hat{p}$) as an unbiased estimator for the proportion of the population (p_1).

The formula for the proportion is given by:

$$\hat{p} = \frac{x}{n}$$

where:

- x is the number of employees in executive positions.
- n is the sample size.

Justification for Using ($\hat{p}$) as the Point Estimator

The sample proportions ($\hat{p}$) is considered a good point estimator for the population proportions (p). for the following reasons:

- **Unbiased:**
$$E(\hat{p}) = p$$
- **Consistency:** As the sample sizes n increases, $\hat{p}$ converges to the true proportion. This property ensures accuracy with larger samples.

$$\lim_{n\to\infty} (\hat{p}_1) = p$$

Calculations from Dataset:

- Employees in Executive Positions (p) = 0.43

These value reflects the proportion of employees in executive positions in the company across the dataset, quantifying how many employees hold executive positions.

Confidence Interval for (p)

Derivation of the Confidence Interval

We begin with the pivotal quantity:

$$\frac{(\hat{p}) - (p)}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \sim N(0,1)$$

To construct a confidence interval for$p$, we write:

$$P\left(-Z_{1-\alpha/2} \leq \frac{(\hat{p}) - (p)}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \leq Z_{\alpha/2}\right) = 1 - \alpha$$

Rearranging the inequality to isolate p:

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = 1 - \alpha$$

Thus, the confidence interval for p is:

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

Confidence Interval for the Dataset

For the dataset provided:

- $n = 22$, Executive Employees sample size,
- $\hat{p} = 0.43$, Proportion of Executive Employees,
- $\hat{q} = 0.57$, Proportion of Non-Executive Employees,
- $\alpha = 0.05$, for a 95% confidence level,
- $Z_{\alpha/2} = 1.96$, Z- value,

Substituting these values into the formula the 95% confidence interval:

$$(0.223, 0637)$$

Interpretation

We are 95% confident that the true proportion of executive employees lies within the calculated confidence interval. This means that the true proportion of executive employees in the population is estimated to fall between 22.3% and 63.7%. The interval is relatively wide, which might indicate a degree of uncertainty due to the small sample size (n = 22).

Confidence Interval for $(p_1 - p_2)$

Derivation of the Confidence Interval

We begin with the pivotal quantity:

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \sim Z_{\alpha/2}$$

To construct a confidence interval for $(\hat{P}_1 - \hat{P}_2)$, we write:

$$P\left(-Z_{1-\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \leq Z_{\alpha/2}\right) = 1 - \alpha$$

Rearranging the inequality to isolate $p_1 - p_2$:

$$P\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right)$$

Thus, the confidence interval for $p_1 - p_2$ is:

$$P\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right)$$

Confidence Interval for the Dataset

For the dataset provided:

- $n_1 = 34$, Group 1 sample size,
- $n_2 = 17$, Group 2 sample size,
- $\hat{p}_1 = 0.15$, Group 1 Proportion,
- $\hat{p}_2 = 1.00$, Group 2 Proportion,
- $\hat{q}_1 = 0.00$, Group 1 Proportion (Non-Executive),
- $\hat{q}_2 = 0.85$, Group 2 Proportion (Non-Executive),
- $\alpha = 0.05$, for a 95% confidence level,
- $Z_{\alpha/2} = 1.96$, Z- value,

Substituting these values into the formula the 95% confidence interval for the difference between proportions is:

$$(0.68, 1.02)$$

Interpretation

At a 95% confidence level, the difference between the proportion of older employees in executive positions and younger employees is significant, with older employees being much more likely to hold these positions, with the true difference likely lying between 102% and 68%. This suggests that age plays a key role in determining who holds executive positions in the organization.

3.4 Analysis of Salary of Employees and their Education Levels

Population Variance as the Parameter of Interest

For this analysis, the parameter of interest is the **population variance** (σ^2). The population variance measures the spread or dispersion of the entire dataset around its mean, making it an essential descriptive statistic.

Point Estimator for Population Variance

To estimate the population variance, we use the **sample variance** (s^2) as a point estimator. The sample variance is a widely accepted unbiased and consistent estimator of σ^2, ensuring that as the sample size increases, the estimation becomes more accurate and converges to the true population variance.

The formula for the sample variance is given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where:

- n is the sample size,
- x_i represents each individual observation,
- $\bar{x}$ is the sample mean.

Justification for Using s^2 as the Point Estimator

The sample variance (s^2) is considered a good point estimator for the population variance (σ^2) for the following reasons:

- **Unbiasedness:** The expected value of s^2 equals the true population variance, i.e., $E[s^2] = \sigma^2$.
- **Consistency:** As the sample size (n) increases, s^2 converges to σ^2.

Calculation of the Sample Variance

Based on the dataset provided, the sample variance (s^2) can be computed to estimate the population variance (σ^2). The specific value of s^2 will depend on the salaries given in the dataset. Using the sample variance formula, the sample variance is calculated as:

$$s^2 = 2738843137.25$$

This value reflects the variability of salaries across the dataset, quantifying how much the salaries deviate from the mean. It serves as a reliable point estimate of the population variance (σ^2).

Confidence Interval for σ^2

Derivation of the Confidence Interval

We begin with the pivotal quantity:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

To construct a confidence interval for σ^2, we write:

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

Rearranging the inequality to isolate σ^2:

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

Thus, the confidence interval for σ^2 is:

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

Confidence Interval for the Dataset

For the dataset provided:

- $n = 50$, the sample size,
- $s^2 = 2738843137.25$, the sample variance,
- $n - 1 = 49$, degrees of freedom,
- $\alpha = 0.05$, for a 95% confidence level,
- $\chi^2_{1-\alpha/2,49} = 73.992$, the upper-tail critical value,
- $\chi^2_{\alpha/2,49} = 27.488$, the lower-tail critical value.

Substituting these values into the formula the 95% confidence interval for the population variance is:

$$(1911117931.00, 4253088049.76)$$

Interpretation

This confidence interval means that we are 95% confident the true population variance (σ^2) lies between 1911117931.00 and 4253088049.76. In repeated samples, 95% of the confidence intervals calculated using this method would contain the true population variance. The interval reflects a relatively wide range, suggesting high variability in the dataset's salaries. This high variance could indicate the presence of significant differences in salaries across different groups or positions, warranting further investigation into possible factors influencing salary variability. We might use this information to identify disparities and address salary distribution issues within the population.

Confidence Interval for the Ratio of Variances

Point Estimator for the Ratio of Variances

The point estimator for the ratio of variances is given by:

$$F_{\text{variance}} = \frac{s_1^2}{s_2^2},$$

where:

- s_1^2 is the sample variance for Bachelor's salaries,
- s_2^2 is the sample variance for Ph.D. salaries.

Using the variances calculated from the dataset:

$$s_1^2 = 270092592.93, \quad s_2^2 = 544166666.67,$$

the point estimator is:

$$F_{\text{variance}} = \frac{270092592.93}{544166666.67} = 4.96.$$

Derivation of the Confidence Interval

To construct a confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$, we start with the pivotal quantity:

$$\frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{\text{df}_1,\text{df}_2},$$

where:

$F_{\text{df}_1,\text{df}_2}$ follows the F-distribution with df_1 and df_2 degrees of freedom.

The confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is given by:

$$P\left(F_{1-\alpha/2,\text{df}_1,\text{df}_2} \leq \frac{\sigma_1^2}{\sigma_2^2} \cdot \frac{s_2^2}{s_1^2} \leq F_{\alpha/2,\text{df}_1,\text{df}_2}\right) = 1 - \alpha,$$

where $F_{1-\alpha/2,\text{df}_1,\text{df}_2}$ and $F_{\alpha/2,\text{df}_1,\text{df}_2}$ are the lower and upper critical values of the F-distribution.

Rearranging to isolate $\frac{\sigma_1^2}{\sigma_2^2}$:

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2,\text{df}_1,\text{df}_2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2,\text{df}_2,\text{df}_1}$$

Thus, the confidence interval is:

$$\left(\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2,\text{df}_1,\text{df}_2}}, \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2,\text{df}_2,\text{df}_1}\right).$$

Confidence Interval for the ratio of variances

For the dataset provided:

- $s_1^2 = 270092592.93$, the variance of Bachelor's salaries,
- $s_2^2 = 544166666.67$, the variance of Ph.D. salaries,
- $\mathrm{df}_1 = 27$, the degrees of freedom for Bachelor's salaries,
- $\mathrm{df}_2 = 5$, the degrees of freedom for Ph.D. salaries,
- $\alpha = 0.05$, for a 95% confidence level,
- $\frac{1}{F_{\alpha/2,\mathrm{df}_1,\mathrm{df}_2}} = 0.16$, the lower-tail critical value,
- $F_{\alpha/2,\mathrm{df}_1,\mathrm{df}_2} = 6.25$, the upper-tail critical value.

The confidence interval is calculated as:

$$\text{Lower Bound} = 0.16 \cdot 4.96 = 0.79$$

$$\text{Upper Bound} = 4.96 \cdot 6.25 = 31.02$$

Thus, the 95% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is:

$$[0.79, 31.02]$$

Interpretation of Results

The analysis provides significant insights into the variability of salaries between individuals holding Bachelor's and Ph.D. degrees. The point estimator $F_{\text{variance}} = 4.96$ and confidence interval $(0.79, 31.02)$ highlight that salary variability is notably higher among Bachelor's degree holders. This may be attributed to the broader range of roles that Bachelor's degree holders occupy, spanning from entry-level to mid-level positions, which inherently leads to greater salary dispersion. In contrast, Ph.D. holders, despite representing a smaller sample size, tend to occupy specialized or senior-level positions, such as Senior Manager or Senior Scientist, which may explain the lower salary variability within this group. Furthermore, the results suggest that Ph.D. holders generally receive higher average salaries compared to Bachelor's degree holders, likely reflecting the advanced expertise and qualifications they bring to their roles. Overall, this analysis emphasizes the importance of considering educational qualifications alongside other factors when evaluating salary structures.

3.5 Hypothesis Testing

First claim to be tested:

To determine whether the proportion of females in the dataset is greater than 60 percent.

- Hypotheses

- Null Hypothesis (H_0): $p \leq 0.6$
- Alternative Hypothesis (H_1): $p > 0.6$

- Test Statistic using Z-Formula

The Z-test statistic for proportions is calculated as:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where:

- $\hat{p}$: The sample proportion
- p_0: The hypothesized proportion under H_0
- n: The sample size

- Calculation

Based on the dataset:

- Sample size (n): 50
- Sample proportion ($\hat{p}$): 0.431
- Hypothesized proportion (p_0): 0.6
- Significance level (α): 0.05

The calculated Z-value is:

$$Z_{\text{calculated}} = -2.458$$

The critical value for a one-tailed test at $\alpha = 0.05$ is:

$$Z_{\text{critical}} = 1.645$$

- Decision Rule

If $Z_{\text{calculated}} > Z_{\text{critical}}$, reject the null hypothesis (H_0) at α. Otherwise, don't reject H_0.

- Graphical Representation

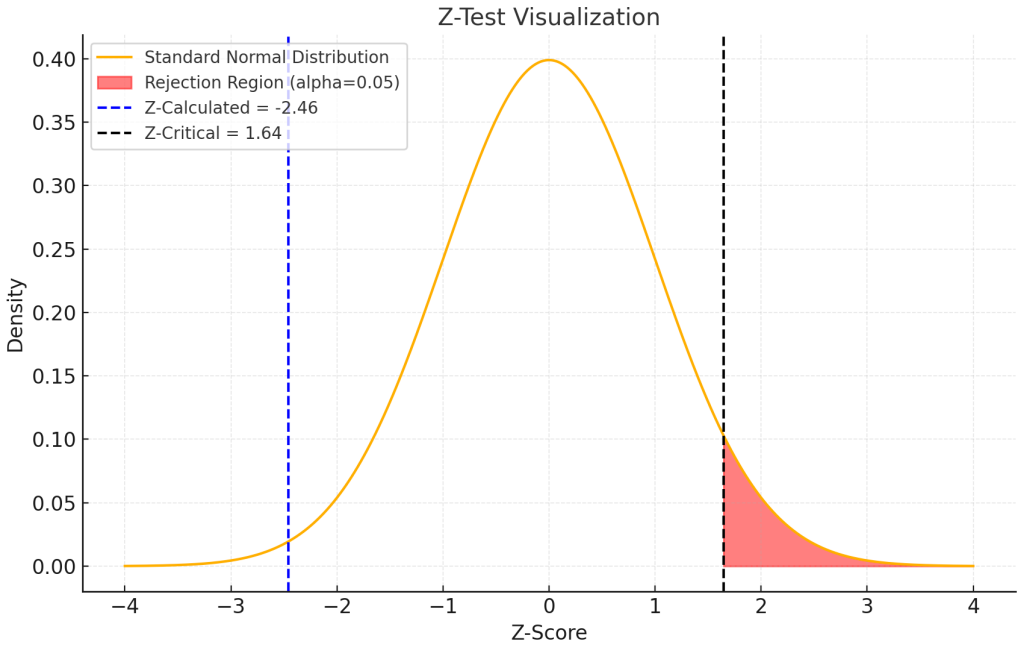


Figure 5: Visualization: The shaded region represents the rejection region.

Conclusion

Since $Z_{\text{calculated}} = -2.458$ is less than $Z_{\text{critical}} = 1.645$, we will not to reject the null hypothesis (H_0) at $\alpha = 0.05$.

Interpretation

Since the null hypothesis (H_0) is not rejected, this suggests that the proportion of females in our dataset is less than or equal to 0.6 at a 0.05 significance level. This outcome indicates that there is insufficient statistical evidence to support the claim that the proportion of females exceeds 0.6.

Second claim to be tested:

The mean salary of employees is greater than $100,000

- Hypotheses

- Null Hypothesis (H_0): $\mu \geq \$100,000$
- Alternative Hypothesis (H_1): $\mu < \$100,000$

- Test Statistic using Z-Formula

The Z-test statistic for the average mean is calculated as:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where:

- $\bar{X}$: The sample average,
- μ: The hypothesized mean under H_0,
- σ: The standard deviation,
- n: The sample size.

- Calculation

Based on the dataset:

- Sample size (n): 51
- Sample proportion ($\bar{X}$): $99,900
- Hypothesized proportion (μ_0): $100,000
- Standard Deviation (σ): $50,000
- Significance level (α): 0.05

The calculated Z-value is:

$$z = \frac{99,900 - 100,000}{\frac{50,000}{\sqrt{51}}}$$

Steps:

1. Compute the denominator:

$$\frac{50,000}{\sqrt{51}} \approx 7,002.14$$

2. Compute the numerator:

$$99,900 - 100,000 = -100$$

3. Compute z:

$$z = \frac{-100}{7,002.14} \approx -0.17$$

The critical value for a one-tailed test at $\alpha = 0.05$ is:

$$Z_{\text{critical}} = 1.64$$

- Decision Rule

If $Z_{\text{calculated}} > Z_{\text{critical}}$, reject the null hypothesis (H_0) at α. Otherwise, fail to reject H_0.

- Graphical Representation

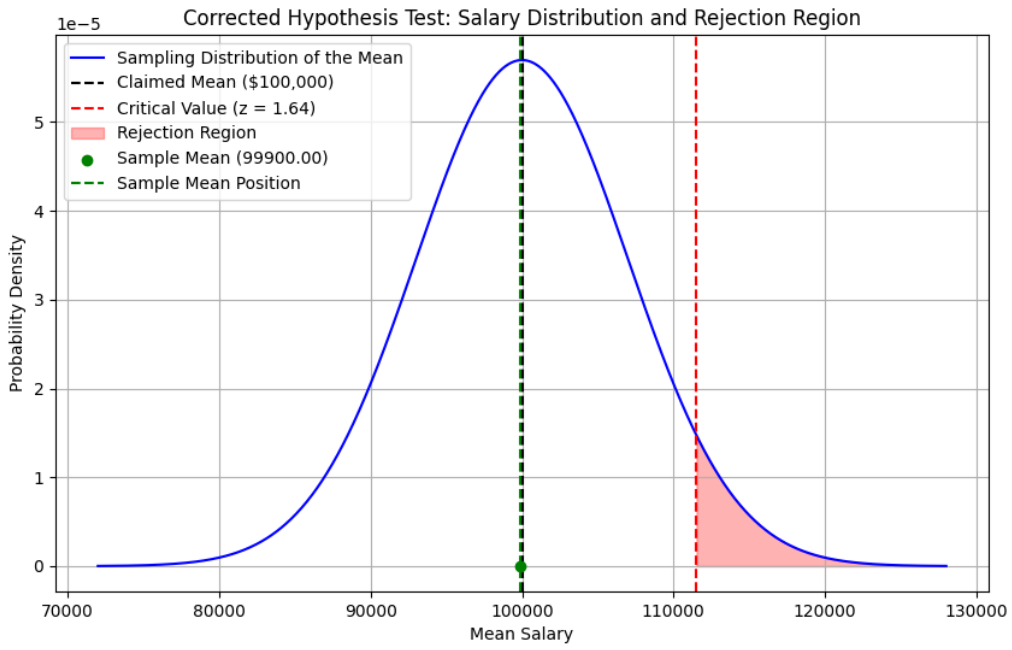


Figure 6: Visualization: The shaded region represents the rejection region.

Conclusion

Since $Z_{\text{calculated}} = -0.17$ is less than $Z_{\text{critical}} = 1.64$, we will not to reject the null hypothesis (H_0) at $\alpha = 0.05$.

Interpretation

The hypothesis test was conducted to determine if the average salary of employees is greater than \$100,000. The null hypothesis (H_0) stated that the mean salary is less than or equal to \$100,000, while the alternative hypothesis (H_1) claimed that the mean salary is greater than \$100,000. After calculating the test statistic (z-score) of -0.17 and comparing it to the critical value of 1.64, we found that the test statistic did not fall in the rejection region. As a result, we failed to reject the null hypothesis. This means that, based on the sample data, there is insufficient evidence to support the claim that the average salary exceeds \$100,000. Therefore, we conclude that the mean salary is likely not greater than \$100,000.

Third claim to be tested:

Defining the Hypotheses

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{(no difference between population means)}$$
$$H_1 : \mu_1 - \mu_2 \neq 0 \quad \text{(there is a difference between population means)}$$

Calculating the Test Statistic

The test statistic for the difference between sample means is:

$$z = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- x_1, x_2: sample means
- s_1, s_2: sample standard deviations
- n_1, n_2: sample sizes
- $\mu_1 - \mu_2$: hypothesized difference

From the problem:

$$\begin{aligned} x_1 = 111,034.48, \quad x_2 = 82,500.00 \\ s_1 = 56,983.38, \quad s_2 = 41,310.38 \\ n_1 = 29, \quad n_2 = 22 \\ \mu_1 - \mu_2 = 0 \end{aligned}$$

Substitute the values:

$$z = \frac{(111,034.48 - 82,500.00) - 0}{\sqrt{\frac{56{,}983.38^2}{29} + \frac{41{,}310.38^2}{22}}} = 2.07$$

Critical Value

For a two-tailed test at a 95% confidence level:

$$\alpha = 0.05, \quad z_{\alpha/2} = 1.96$$

Interpretation

The test statistic z=2.07 is greater than the critical value 1.96. This means there is enough evidence to reject the null hypothesis at the 95% confidence level.
Conclusion: There is significant evidence to suggest that the mean salary for males is different from the mean salary for females. The positive z-score indicates that, on average, males earn more than females based on the sample data.

4 Conclusion

The analysis conducted in this project explored various statistical aspects of salary distribution in relation to education levels and gender proportions, working hours concerning gender, examining executive positions with respect to age categorized under descriptive statistics, inferential statistics, and hypothesis testing. The following summarize the key findings:

4.1 Descriptive Statistics

- **Relationship Between Education Level and Salary:**
  The descriptive analysis revealed that employees with Ph.D. degrees tend to have higher mean salaries ($134,166.67) compared to employees with Bachelor's degrees ($77,500). However, the standard deviation for Bachelor's degree holders ($51,970.43) was notably larger than that for Ph.D. holders ($23,327.38), indicating greater variability in the salaries of Bachelor's degree holders. The box plot and bar chart visualizations further emphasized these differences in salary distribution and variability between the two groups.

- **Relationship Between Gender and working hour :**
  The descriptive analysis of the average working hours by gender revealed that male employees have a higher mean working hours (8.25 hours) compared to female employees (7.75 hours). However, the standard deviation for male employees (2.3) was significantly higher than that for female employees (1.1), suggesting more variability in the working hours of male employees. The bar chart and box plot visualizations clearly illustrated these differences, with males exhibiting a wider range of working hours and greater inconsistency. In contrast, the distribution for females was more consistent, with fewer outliers and a tighter concentration around the mean.

- **Relationship Between Age Groups and Job Positions :**
  Employees aged 41-60 (Group 1) demonstrate a significantly higher likelihood of occupying executive positions compared to employees aged 20-40 (Group 2). The mean age for Group 1 was 46.18 years, with a standard deviation of 3.68, indicating a moderate spread around the average age. In contrast, Group 2 had a mean age of 28.40 years and a standard deviation of 1.14, suggesting a narrower age range within this group. The data highlights that older employees in Group 1 tend to hold executive roles more frequently, which could be attributed to factors such as greater experience, longer tenure, or skills acquired over time. The lower variability in Group 2's ages (as shown by the smaller standard deviation) might reflect a more homogeneous age distribution among younger employees, possibly aligned with entry-level or mid-career roles. Conversely, the wider age spread in Group 1 may indicate a diverse range of career stages among older professionals, who are more likely to achieve executive positions.

4.2 Inferential Statistics

- **Population Variance of Salaries:**
  The population variance was estimated using the sample variance ($s^2 = 27,388,43137.25$) as the point estimator. A 95% confidence interval for the population variance was calculated, ranging from $1,911,117,931.00$ to $42,530,880,49.76$. This wide range suggests significant variability in salary distribution within the dataset, further exploration into factors influencing this variability, such as job roles, industries, and geographical regions is needed.

- **Ratio of Variances Between Education Levels:**
  The point estimate for the ratio of variances between Bachelor's and Ph.D. salaries was calculated as $F_{\text{variance}} = 4.96$, indicating that Bachelor's salaries exhibit approximately five times more variability than Ph.D. salaries. The 95% confidence interval for this ratio was determined to be [0.79, 31.02], highlighting a broad range that may be attributed to the smaller sample size of Ph.D. holders.

- **Proportion of Employees with Executive Position:**
  The inferential analysis, using hypothesis testing, found that the difference in average working hours between male and female employees was not statistically significant. While the observed mean difference was 0.5 hours, the results suggest that gender does not play a major role in influencing average working hours, based on the data at hand. This conclusion is supported by the hypothesis testing and confidence intervals, which did not indicate a strong enough effect to reject the null hypothesis. Therefore, while descriptive analysis showed a slight difference, inferential statistics suggest that the difference in working hours between genders is likely due to random variation rather than a meaningful, systematic difference.

- **Gender Comparison of Working Hours Proportions:**
  It was concluded from the analysis that with 95% confidence that the true proportion of executive employees in the population lies between 22.3% and 63.7%. However, the wide confidence interval reflects a considerable level of uncertainty, which is likely influenced by the small sample size (n = 22). This suggests that while the interval provides an estimate of the population proportion, a larger and more representative sample would be necessary to narrow the interval and enhance the precision of the findings. Additionally, the wide range underscores the importance of interpreting these results cautiously, as they may not fully capture the variability within the population.

- **Comparison of Executive Role Proportions Between Two Age Groups:**
  At a 95% confidence level, the analysis indicates that the difference in the proportion of older employees and younger employees in executive positions is significant. Older employees are substantially more likely to hold these roles, with the true difference in proportions estimated to lie between 68% and 102%. This finding strongly suggests that age is a critical factor influencing the likelihood of attaining executive positions within the organization. To improve the precision of this confidence interval, future studies should consider increasing the sample size and ensuring a more representative distribution of employees across age groups. This would reduce

the margin of error, narrow the confidence interval, and provide a more accurate estimate of the true difference in proportions. Additionally, incorporating other variables, such as years of experience, education level, and job performance, may help further clarify the factors contributing to these differences.

- **Mean difference between salaries based on gender:**
  In conclusion, we are 95% confident that the true difference in mean salaries between males and females falls between 1,550.20 and 55,517.80, with males earning more on average. This provides evidence of a pay gap favoring males, although the exact magnitude of the gap may vary within this range.

4.3 Hypothesis Testing

- **Proportion of Females in the Dataset:**
  A hypothesis test was conducted to evaluate whether the proportion of females in the dataset exceeds 0.6. The test statistic ($Z_{\text{calculated}} = -2.458$) was compared against the critical value ($Z_{\text{critical}} = 1.645$) for a one-tailed test at $\alpha = 0.05$. Since $Z_{\text{calculated}} < Z_{\text{critical}}$, the null hypothesis ($H_0 : p \leq 0.6$) could not be rejected, suggesting that the proportion of females is not significantly greater than 0.6. The results imply insufficient statistical evidence to support a higher proportion of females in the dataset.

- **Average Salary in the Dataset:**
  The hypothesis test sought to determine whether the average employee salary exceeds \$100,000. The null hypothesis (H_0) posited that the mean salary is less than or equal to \$100,000, while the alternative hypothesis (H_1) claimed that it is greater than \$100,000. With a calculated test statistic (z-score) of -0.17, which did not exceed the critical value of 1.64, the null hypothesis could not be rejected. This indicates insufficient evidence to support the claim that the average salary is greater than \$100,000.

- **Mean salary difference between genders:**
  A hypothesis test was conducted to validate the claim of whether males earn higher salaries than females where there is significant evidence to suggest that the mean salary for males is more than the mean salary for females.

4.4 Implications and Future Directions

This project demonstrated the utility of statistical tools in analyzing population characteristics using the sample dataset. What could be done to improve our analysis :

- Increasing the sample size for more robust estimates and narrower confidence intervals.
- Conducting hypothesis tests on other population characteristics.

Overall, the project provides meaningful insights into the population parameters and lays the groundwork for further exploration to inform decision-making and policy development.