**HUYNH DO**

**Module 4C, Part 1/2**

1. **Specify which variable is the dependent and independent variable**

Age and total of weeks worked last year qualify for the Pearson test since they both are continuous variables and a potential linear association between the 2 subjects.

With the opening explanation, we can construct the following 2 variables:

* **Independent Variable (IV)** - **AGE**: Age (Predictor) doesn't depend on other variables in this investigation. It can influence other factors but is not influenced by them in the context of this analysis.
* **Dependent Variable (DV)** - **WEEKSWRK**: Age could be a factor that influences the Outcome of total weeks worked last year of a person.

1. **Using the variables that you’ve chosen, state the null and alternative hypotheses.**

With the 2 declared above variables, we can construct these 2 Hypotheses:

1. **Null Hypothesis (H₀):**

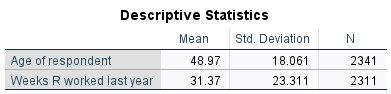
* There is **no** linear relationship between the number of weeks worked last year and the age of the respondent. In other words, the correlation coefficient (ρ) is equal 0
* **Mathematical Representation**: H0​:ρ = 0

1. **Alternative Hypothesis (H₁):**

* There is a linear relationship between the number of weeks worked last year and the age of the respondent.
* **Mathematical Representation**: H1​:ρ <> 0

1. **Perform the Pearson correlation and interpret the results.**

* **Descriptive Statistic**

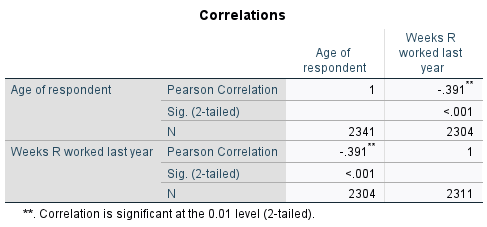
****

1. **Age of Respondent:**

* Mean Age: The average of the participants is 48.97 years.
* Standard Deviation: The number 18.061 indicates a wide range of ages among the respondents.
* Sample Size (**N**): A total of 2,341 random people were invited to participate in the survey.

1. **Weeks Worked Last Year**

* Mean Weeks Worked Last Year: On average, employees put 31.37 weeks worked in the last year.
* Standard Deviation: The number 23.311 suggests a noticeable variability in the number of weeks worked.
* Sample Size (**N**): A total of 2,311 employees answered how many weeks worked last year.
* **Correlation Analysis**

****

* 1. **Pearson Correlation Coefficient (r):**
* The correlation between Age of Respondent and Weeks Worked Last Year is **negative** 0.391 **(-0.391**).
* The magnitude of 0.391 (absolute value) suggests a moderate correlation strength.
* The **negative sign** indicates that the relationship is **inverse**, and it is translated that the number of weeks worked tends to **decrease** when the employee gets older (increased age).
  1. **Statistical Significance**
* The correlation is marked with **"−.391**" and "\*\*", indicating that it is significant at the 0.01 level (2-tailed).
* The extreme p-value < .001 probability means that this correlation occurred by chance is less than 0.1%. This high level of significance strongly proves that the observed correlation is not resulted by random variation in the sample data.
  1. **Sample Size (N):**
* The number of calculated paired observations is 2,311 pairs.
  1. **Correlation Analysis Findings:**
* There is a moderate significant negative correlation between ages and weeks worked, indicating that as people get older, they tend to work fewer weeks in entire year.
* The data shows that many young adults and middle-aged employees working (from part-time to full-time), while senior individuals are more likely to have reduced or no work for entire year possibly due to retirement, health status, occupation type, or even personal choice.

1. **Make sure to assess the test assumption**

The following assumptions must be considered for the Pearson test:

1. **Linearity**:

* Assumption:

There should be a linear relationship between the variables.

* Assessment:

Although the Pearson correlation measures may show whether there is a linear association and its strength, a scatterplot should be used to further visualize the relationship.

* **Evaluation**:

The moderate correlation (-0.391) suggests a linear trend, but scatterplot observation and analysis are recommended for a more precise evaluation.

1. **Independence**:

* Assumption:

The observations should be independent of each other.

* Assessment:

If the data is collected from a random sample and each participant represents an independent case, this assumption is met.

* Evaluation:

Each employee is independent from each other when reporting age and number of work weeks, this assumption is met.

1. **No Outliers**:

* Assumption:

The data should not contain from extreme outliers.

* Assessment:

Outliers can be observed and indentified using scatterplots, if found, the interpretation **should be carefully noted later on**, however, in some cases, we can apply the robustness of the statistical method.

* Evaluation:

Checking for outliers in both “Age” and “Weeks Worked Last Year” will help to confirm.

1. **Consistency**:

* Assumption:

The variability of the dependent variable (Weeks Worked Last Year) should be consistent across all levels of the independent variable (Age).

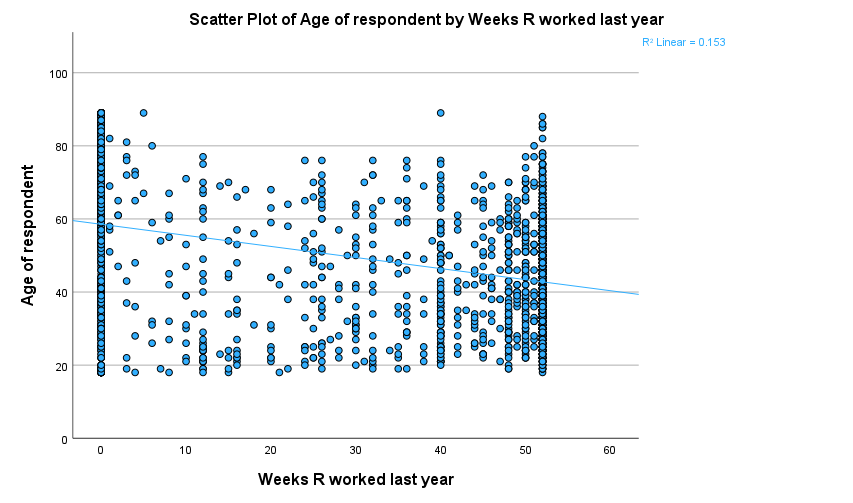
* Assessment:

This can be evaluated using a scatterplot to check if they are evenly spread across the range of the independent variable.

* Evaluation:

From observation, if wider spread for older or younger respondents is observed (Which against the correlation analysis findings), this assumption might be violated, affecting the reliability of the correlation.

1. **Using a scatterplot, visually display the results of the Pearson correlation and explain your findings.**

****

1. **Observations:**
2. **Clusters in the Data:**

**1st Impression:**

Visually, there are 2 significant dense clusters at both the 0 weeks and 52 weeks worked marks of the graph which suggest that a considerable number of individuals, across the spectrum of age, are either reported either not working at all or the entire year (likely full-time workers)

**2nd Impression:**

The third noticeable concentrated stack of plots, but less tall than the 2 edges in terms of age and discontinued at the age of 75 (y-axis), lies at the 40 (x-axis) weeks worked mark. This shows that 40 weeks worked probably is a favorable work length in a year for the young and middle-age including some seniors.

1. **Patterns in the Data:**
   * For younger and middle ages (18-40), there is a spread across weeks worked spectrum, from a few weeks to full year. This could reflect a mix of full-time workers, part-time workers, and those still in education. This reflex a mix of full-time employees and par-time students.
   * From the range of 40-60 years of age, in this section, the density of the plots is somewhat the same with the lower section but the concentration rises toward the upper range of weeks work, indicating that most are likely in stable employment.
   * For the older people (60+), the data shows a higher frequencies of fewer weeks worked likely due to retirement or semi-retirement (which is typical for women when reaching this age), or health-related issues.
2. **The Regression Line:**

The downward slope of the regression line also confirms this theory of **negative or disproportion** relationship between the numbers of weeks worked and age.

1. **The Outlier Explanations – Special Notes:**

* Zero Weeks Work Outliers:

As described in the 1st impression, a significant number of young adults to seniors (most left edge) didn’t work at all for the entire year may **not be** necessarily are statistical outliers; these groups, instead, simply are just exceptional subgroups of students, retirees, or maybe unemployed individuals.

* Full-Year Work Outliers

On the right most edge of the graph, a large number of plots rise on the 52-week mark, indicating these are full-time workers. Again, this is still perfectly normal data, due to possible financial obligations such as: Car payments, mortgage, or support.

1. **Findings:**

The scatterplot shows a negative relationship between age and weeks worked, with older individuals tending to work fewer weeks. Visually, there are 2 exceptional clusters at 0 weeks and 52 weeks but these clusters are simply showing a different pattern of workforce. The statistical data and the regression line together not only to confirm the alternative hypothesis (**H1**) but also, specifically indicate that age is **inversely** related to the number of weeks worked last year.