Data Report #1 - Age and Death Penality

Jaimie Chin

2023-05-05

```
# Load packages
library(tidyverse)
library(ggplot2)
library(DescTools)
```

1. Study Description

Age and Death Penalty

A survey of 400 voters in a congressional district in VA collected, among other things, the age of voters and their support for various policies. Is age related to voters supporting the death penalty more or less?

```
# Load the dataset from .csv
filepath = 'data/age and death penalty.csv'
df = read_csv(filepath, show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
# Let's take a look at what type of data we have
df
```

```
## # A tibble: 400 x 3
##     ...1   Age `Support for Death Penalty`
##    <dbl> <dbl>                       <dbl>
##  1     1    65                           3
##  2     2    53                           4
##  3     3    18                           2
##  4     4    47                           4
##  5     5    23                           5
##  6     6    63                           2
##  7     7    63                           5
##  8     8    83                           2
##  9     9    18                           3
## 10    10    18                           5
## # ... with 390 more rows
```

```r
# Let's rename our "Support for Death Penalty" for easier analysis later
colnames(df)[3] ="SupportforDeathPenalty"
df$SupportforDeathPenalty = as.factor(df$SupportforDeathPenalty)
```

```r
# Let's look at a summary of our data
summary(df)
```

```
##       ...1            Age         SupportforDeathPenalty
##  Min.   :  1.0   Min.   : 18.00   1: 31
##  1st Qu.:100.8   1st Qu.: 23.00   2: 55
##  Median :200.5   Median : 39.00   3: 90
##  Mean   :200.5   Mean   : 42.77   4:100
##  3rd Qu.:300.2   3rd Qu.: 57.00   5: 79
##  Max.   :400.0   Max.   :121.00   6: 36
##                                   7:  9
```

The Age and Death Penalty data describes a situation where a survey of 400 voters was conducted in a Virginian congressional district: specifically, the age of voters and their support for various policies.

It seems that the study was conducted to consider whether age is correlated with support for the death penalty. A precursive glance at the data reveals the age of the participants in the study (range: 18-121) and their support for the death penalty which was completed on a 7-point Likert scale (assuming 1 to be the least supportive and 7 being the most supportive).

A Pearson correlation analysis seems to be appropriate for this study, but first we must check if the assumptions for this analysis are met.

One of the assumptions of the Pearson correlation is that both variables should be continuous, which is the case here as age is a continuous variable and the support for the death penalty is measured on a 7-point Likert scale, which can be treated as a continuous variable.

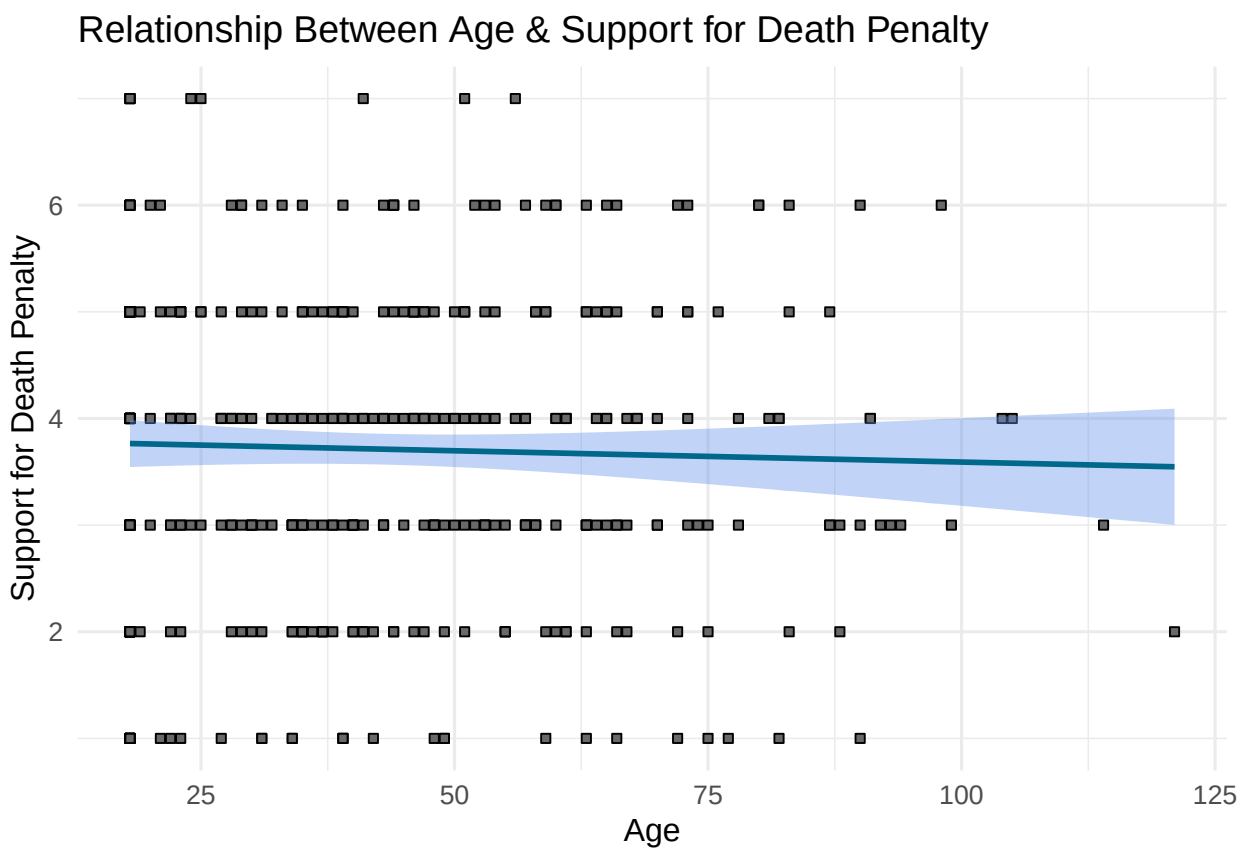
Another assumption is that the relationship between the two variables should be linear, and there should not be any outliers.

1.1 Scatterplot & Box Plot of the Data

Usually when data is appropriate to run a correlation, the corresponding plot is a scatterplot. However for this data, it's difficult to form conclusions about linearity and any patterns with a scatterplot:

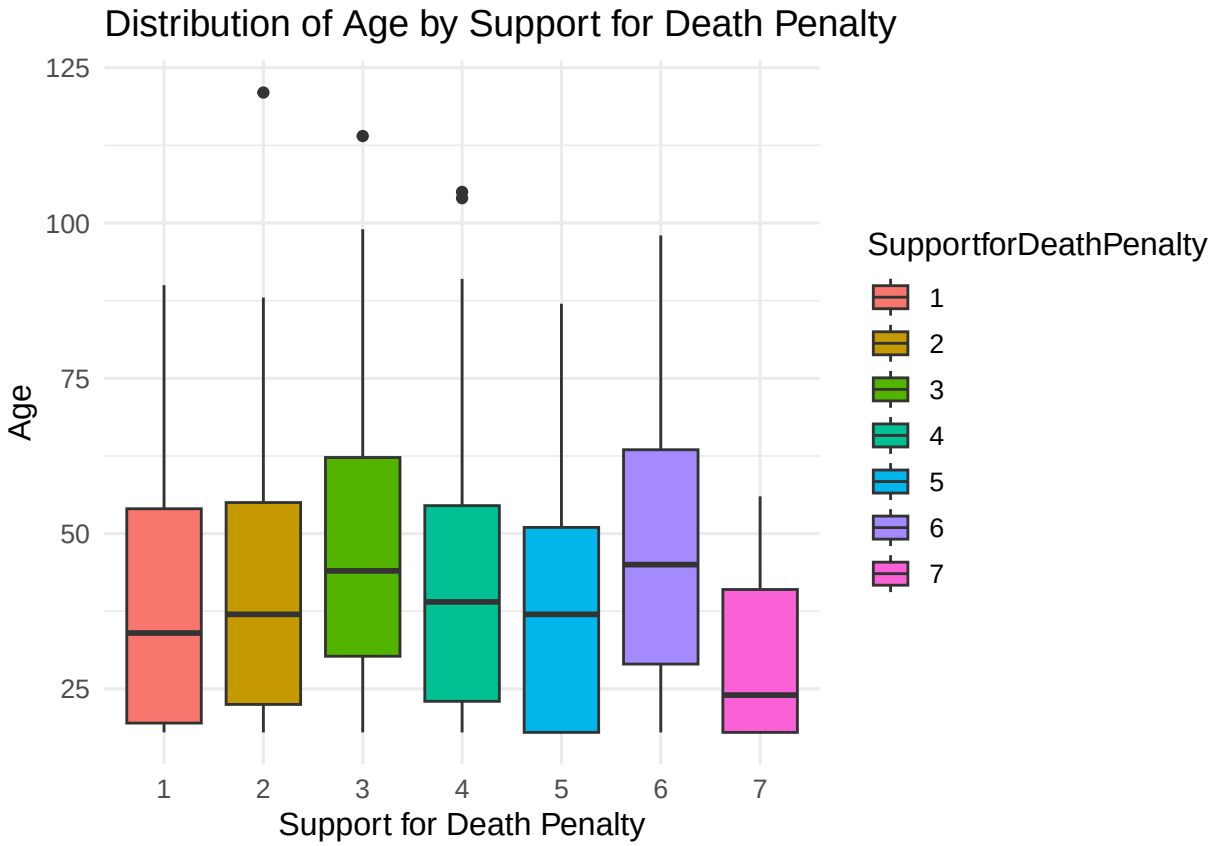
```r
# Create scatterplot to visualize relationship between age and support for the death penalty
ggplot(data = df) +
  geom_point(aes(x = Age, y = as.numeric(SupportforDeathPenalty)), shape = 22, fill = "dimgray") +
  geom_smooth(aes(x = Age, y = as.numeric(SupportforDeathPenalty)),
              color = "deepskyblue4",
              fill = "cornflowerblue",
              method = "lm") +
  theme_minimal(base_size = 12,
                base_line_size = 12/22,
                base_rect_size = 12/22) +
  labs(x = "Age",
       y = "Support for Death Penalty",
       title = "Relationship Between Age & Support for Death Penalty")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Again, its difficult to discern any patterns in our data from the scatterplot. Let's do a box plot instead to observe the distribution of ages per strength of support for the death penalty.

```
# Let's plot our features to see if they have a linear relationship
ggplot(df, aes(x = SupportforDeathPenalty, y = Age, fill = SupportforDeathPenalty)) +
  geom_boxplot() +
  theme_minimal(base_size = 12,
                base_line_size = 12/22,
                base_rect_size = 12/22) +
  xlab("Support for Death Penalty") +
  ylab("Age") +
  ggtitle("Distribution of Age by Support for Death Penalty")
```



A box and whisker plot of the data reveals the distribution of ages per strength of support for the death penalty. The plot reveals that there is unlikely a strong correlation or association between ages and support for the death penalty since the distribution of ages is relatively similar across all support strengths except for a very strong support which has a distribution among younger ages. It is also important to note the outliers present within the dataset (participants who are 100+ years of age) that may affect the results.

Since there does not seem to be a clear linear pattern between Age and Support for the Death Penalty, we should consider using a non-parametric test. The most common non-parametric test used to assess the relationship between two variables is the Spearman's rank correlation coefficient.

Spearman's rank correlation coefficient is a measure of the strength and direction of the monotonic relationship between two variables, which means that it measures the degree to which the two variables are related, regardless of whether the relationship is linear or not.

2. Spearman Rank Correlation

```
# Create new columns in the dataset that rank order existing columns
df$Age_rank = rank(df$Age)
df$Support_rank = rank(df$SupportforDeathPenalty)

# Conduct Spearman Rank Correlation
cor.test(df$Age_rank, df$Support_rank, method = "spearman", exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
```

```
##
## data:  df$Age_rank and df$Support_rank
## S = 11014744, p-value = 0.5151
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.03263866
```

For comparison, let's also do a Pearson Correlation to support our previous intuition that a Spearman Correlation would be more appropriate.

```
cor.test(df$Age, as.numeric(df$SupportforDeathPenalty))
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$Age and as.numeric(df$SupportforDeathPenalty)
## t = -0.61998, df = 398, p-value = 0.5356
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.12872171  0.06719433
## sample estimates:
##         cor
## -0.03106204
```

It looks like our intuition was reasonable as the Spearman's rank correlation seems to be more significant compared to the Pearson.

```
# Let's get our confidence intervals from our Spearman Correlation
SpearmanRho(df$Age_rank, df$Support_rank, conf.level = 0.95)
```

```
##         rho      lwr.ci      upr.ci
## -0.03263866 -0.13027347  0.06562307
```

3. Report of the Results

A Spearman's rank correlation was conducted to examine the relationship between age and support for the death penalty on a 7-point Likert scale. The sample consisted of 400 voters from a congressional district in Virginia. Age was measured as a continuous variable, while support for the death penalty was measured on a Likert scale. The correlation coefficient between age and support for the death penalty was found to be non-significant (r_s(398) = -0.0326, 95% CI [-0.130, 0.0656], p = 0.5151). These results suggest that there is insufficient evidence to suggest a significant correlation between age and support for the death penalty. The negative correlation coefficient indicates that as age increases, support for the death penalty tends to decrease, but this relationship is weak and not statistically significant. The resulting confidence interval is (-0.130, 0.0656). This indicates that we can be 95% confident that the true correlation coefficient for the population falls between -0.130 and 0.0656.

4. Conclusion

These results lead me to conclude that we fail to reject the null hypothesis of there being no correlation between age and support for the death penalty. Altogether, what this means for the original scenario is that

age is uncorrelated with support for the death penalty in a Virginian congressional district. However, the outliers present in the data may have affected the results. While an age of 100+ years is certainly possible, it is rather unlikely for four participants of 100+ years of age to be in the same Virginian congressional district.