ANOVA

Sophie Marion de Proce

2023-01-17

Introduction

In this self-guided lab, you will be running an ANOVA. We will be using a significance level of 5% throughout.

We'll be using functions from the tidyverse collection of packages, as well as from the emmeans package, and we will use the NHANES dataset again, so let's load them in.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.3      v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(emmeans)
library(NHANES)
```

The messages above are telling us which functions from the base R are overwritten by functions from the Tidyverse, and that some packages were set up using a previous version of R.

ANOVA exercise

We will explore whether people who watch more TV also tend to have a higher body mass index (BMI), indicating that they are more likely to be overweight or obese. We will use a one-way ANOVA to answer this question.

In order to meet the assumptions for ANOVA, we will restrict the NHANES dataset a little bit. This is just for demonstration purposes, if you were given a real dataset, you would need to analyse the whole dataset, rather than restricting it.

Let's create a new dataset, called NHANES_ANOVA, where we will create a new version of the variable about watching TV - my TV_categorical variable will only have 3 levels, and it will indicate whether someone watches TV less than the median (i.e.less than 2 hours a day), at the median (2 hours a day), or more than the median (3 or more hours). We will also filter out people with BMI over 40 (again, this is for demonstration purposes only).

```
NHANES_ANOVA <- NHANES %>%
  # creating a TV_categorical variable with 3 levels
  mutate(TV_categorical = case_when(TVHrsDay == "0_hrs"|TVHrsDay == "0_to_1_hr"|
                                      TVHrsDay == "1_hr" ~ "less_than_median",
                                    TVHrsDay == "2_hr" ~ "median",
                                    TVHrsDay == "3_hr"|TVHrsDay == "4_hr"|
                                      TVHrsDay == "More_4_hr"
                                    ~ "more_than_median")) %>%
  # removing people with a missing value for TV_categorical
  drop_na(TV_categorical) %>%
  # filtering out people with BMI of 40 or more
  filter(BMI < 40)
```

Now let's have a quick look at our NHANES_ANOVA subset to see how many people there are in each category.

```
table(NHANES_ANOVA$TV_categorical)
```

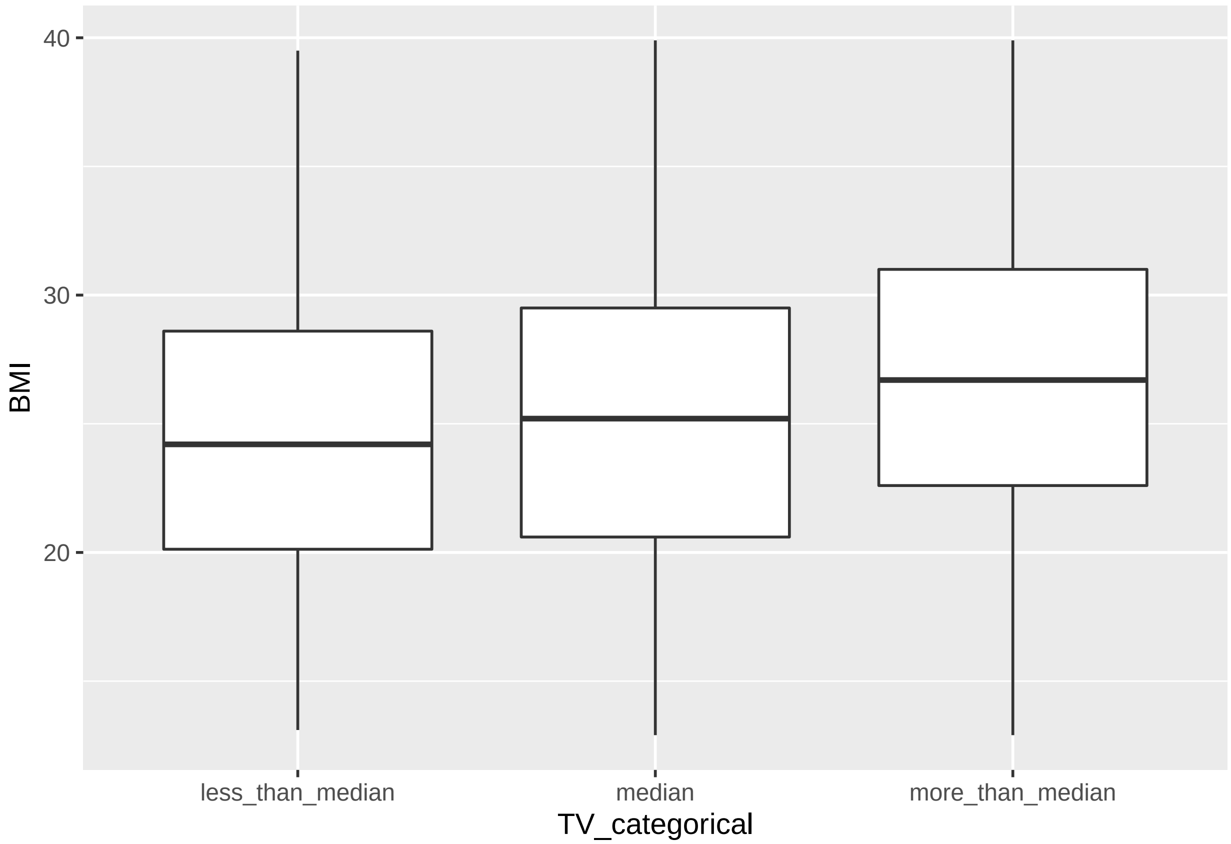
```
##
## less_than_median           median more_than_median
##             1590             1211             1792
```

Task 1: Dataset exploration.

Let's create a box plot and a table of summary statistics including mean and standard deviation to see what the pattern is for BMI across TV_categorical groups. Based on the descriptive statistics and the boxplot, what pattern do you see in the data?

Solution 1

```
NHANES_ANOVA %>%
  ggplot(aes(x = TV_categorical, y = BMI)) +
  geom_boxplot()
```



```
NHANES_ANOVA %>%
  group_by(TV_categorical) %>%
  summarise(mean_BMI = mean(BMI, na.rm = TRUE),
            sd_BMI = sd(BMI, na.rm = TRUE))
```

```
## # A tibble: 3 x 3
##   TV_categorical   mean_BMI sd_BMI
##   <chr>               <dbl>  <dbl>
## 1 less_than_median     24.4   5.97
## 2 median               25.2   5.99
## 3 more_than_median     26.8   5.86
```

The BMI variable has a higher median in the group with the individuals who watch TV for more hours than the median than in both other groups, with a larger difference between the group with the least hours of TV watched and the group with the most hours of TV watched. The variance seems similar between the groups.

Task 2: ANOVA

Let's run an ANOVA to test if what we're seeing is statistically significant. How would you interpret the results of the ANOVA?

Solution 2

```
ANOVA1 <- aov(BMI~TV_categorical, data = NHANES_ANOVA)
summary(ANOVA1)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)    
## TV_categorical    2   4838  2418.9   68.76 <2e-16 ***
## Residuals      4590 161467    35.2                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the F-statistic of the ANOVA is highly significant (p<2.2x10-16), suggesting that there is an overall effect of the number of hours of TV watched on the BMI of individuals.

Task 3: Post-hoc tests

As you know, ANOVA is an omnibus test, so it doesn't tell us where the significant differences lie. We need to look at post-hoc comparisons to tell us which groups are different. Let's do it using Tukey's HSD test for each pairwise comparison. We would like to get the differences in means, their test statistic and associated p-value, standard error and confidence interval. How would you interpret the results of this test?

Solution 3

```
anova1_emmeans <- emmeans(ANOVA1,"TV_categorical")
anova1_emmeans
```

```
##  TV_categorical   emmean    SE   df lower.CL upper.CL
##  less_than_median   24.4 0.149 4590     24.1     24.7
##  median             25.2 0.170 4590     24.9     25.5
##  more_than_median   26.8 0.140 4590     26.5     27.0
## 
## Confidence level used: 0.95
```

```
anova1_pairs <- pairs(anova1_emmeans)
anova1_pairs
```

```
##  contrast                            estimate    SE   df t.ratio p.value
##  less_than_median - median             -0.781 0.226 4590  -3.454  0.0016
##  less_than_median - more_than_median   -2.348 0.204 4590 -11.493  <.0001
##  median - more_than_median             -1.567 0.221 4590  -7.103  <.0001
## 
## P value adjustment: tukey method for comparing a family of 3 estimates
```

```
confint(anova1_pairs)
```

```
##  contrast                            estimate    SE   df lower.CL upper.CL
##  less_than_median - median             -0.781 0.226 4590    -1.31   -0.251
##  less_than_median - more_than_median   -2.348 0.204 4590    -2.83   -1.869
##  median - more_than_median             -1.567 0.221 4590    -2.08   -1.050
##
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
```

All three pairwise comparisons are significant at the 5% significance level, with all p-values smaller than 0.002 and 95% confidence intervals not overlapping 0. The largest difference in means is -2.348 for the comparison between the more_than_median and the less_than_median groups.

Task 4: Checking the assumptions.

Let's check the assumptions.

- 4.1. Independence. Would you say that the data meet the independence assumption? Are the groupings independent of one another? Is each observation independent of the others?
- 4.2. Homogeneity of variance. Does the BMI variable have a similar variance among the three groups? Run a Bartlett's test to check. How would you interpret the results of this test?
- 4.3. Normality. Does the BMI variable come from a normal distribution? Use both visual inspection of the distribution and hypothesis test. How would you interpret the plot and the test results?

Solution 4.1

The NHANES dataset comes from surveys of independent individuals, therefore observations should be independent from each other. Similarly, the groups should be independent of each other, assuming that only one person per household was surveyed. This should be thoroughly checked when analysing a real dataset.

Solution 4.2

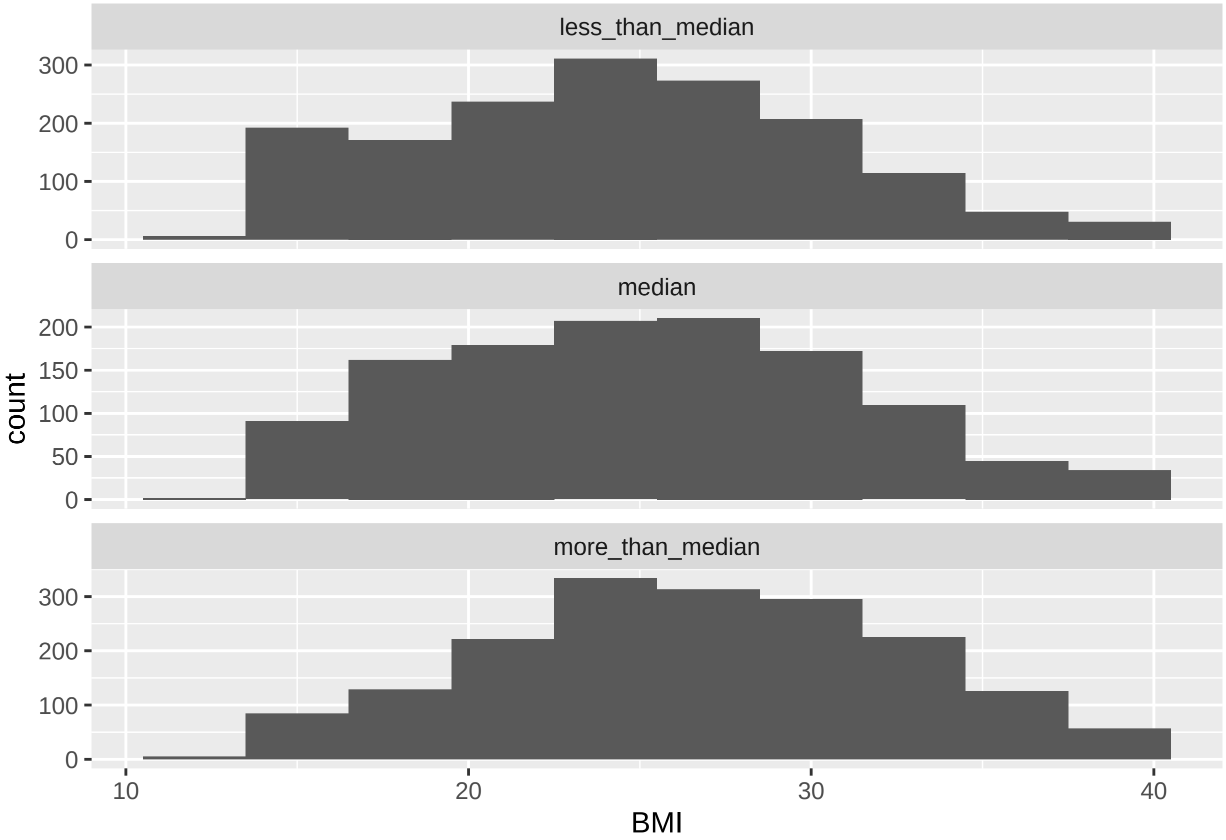
```r
bartlett.test(BMI~TV_categorical, data = NHANES_ANOVA)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  BMI by TV_categorical
## Bartlett's K-squared = 0.93301, df = 2, p-value = 0.6272
```

The p-value is very high, much higher than the significance level 0.05, so we conclude that the variances are equal across the groups.

Solution 4.3

```r
# Create a histogram of BMI at each level of TV_categorical.
NHANES_ANOVA %>%
  ggplot(aes(x = BMI)) + geom_histogram(bins = 10) +
  facet_wrap(facets = ~TV_categorical, ncol = 1, scales = "free_y")
```



The histograms suggest that BMI is normally distributed in each category of the TV_categorical variable.

```r
# Let's run the Shapiro-Wilk normality test.
shapiro.test(NHANES_ANOVA$BMI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  NHANES_ANOVA$BMI
## W = 0.98757, p-value < 2.2e-16
```

The Shapiro-Wilk test has a very low p-value (p<2.2e-16), suggesting that the BMI variable doesn't come from a normal distribution. The large sample size may be picking a small deviation from the normal distribution, so this result alone doesn't mean that the assumption of normality is not met.