HW3

Crystal Huang

2025-07-13

Make sure you put your name in the author and date above.

25 points total for this section.

```
#Please set your HSS where you have the NHANES.RData and this template saved, as your working directory
#If you continue to have trouble, you could just click on the "NHANES.RData" file name under "Files" at

load("NHANES.RData")
```

Study about vitamin D3, arthritis, sleep hours per wkday and thyroid problem among 30+ year olds"

First, set up your study dataset:

```
#[1 pt] Create a new dataset called NHANES_2 according to the flowchart in Canvas - 1 line:
NHANES_2 <- subset(NHANES,
                   age>=30 &
                     !is.na(vitamin_d3) &
                     !is.na(arthritis) &
                     !is.na(sleephrs_wkday) &
                     !is.na(thyroid_problem))
```

```
#[1 pt] Check the age range - first, show that the original NHANES includes ages 0-80; then, show that
range(NHANES$age)
```

```
## [1]  0 80
```

```
range(NHANES_2$age)
```

```
## [1] 30 80
```

```
#[1 pt] Check for missing data for vitamin_d3 - first, show the number of missing data for  in the orig
sum(is.na(NHANES$vitamin_d3))
```

```
## [1] 1391
```

```
sum(is.na(NHANES_2$vitamin_d3))
```

```
## [1] 0
```

```r
#[1 pt] Check for missing data for arthritis - first, show the number of missing data for  in the origi
sum(is.na(NHANES$arthritis))
```

```
## [1] 8222
```

```r
sum(is.na(NHANES_2$arthritis))
```

```
## [1] 0
```

```r
#[1 pt] Check for missing data for sleephrs_wkday - first, show the number of missing data for  in the
sum(is.na(NHANES$sleephrs_wkday))
```

```
## [1] 1839
```

```r
sum(is.na(NHANES_2$sleephrs_wkday))
```

```
## [1] 0
```

```r
#[1 pt] (v) Check for missing data for thyroid_problem - first, show the number of missing data for  in
sum(is.na(NHANES$thyroid_problem))
```

```
## [1] 2277
```

```r
sum(is.na(NHANES_2$thyroid_problem))
```

```
## [1] 0
```

Evaluate the impact of thyroid problem (2 categories) on vitamin D3 (continous variable) and sleep hours per work day (discrete variable)

```r
#[1 pt] Tabulate the number of people in the thyroid_problem yes vs. no groups (hint, don't forget to i
table(NHANES_2$thyroid_problem, exclude=F)
```

```
##
##   No  Yes
## 1247  349
```

```r
#[1 pt] [1 pt] Calculate Vitamin D3 mean by thyroid_problem yes vs. no groups - 1 line:
tapply(NHANES_2$vitamin_d3, NHANES_2$thyroid_problem, mean)
```

```
##       No      Yes
## 84.28292 93.86682
```

```r
#[1 pt] Calculate Vitamin D3 SD by thyroid_problem yes vs. no groups - 1 line:
tapply(NHANES_2$vitamin_d3, NHANES_2$thyroid_problem, sd)
```

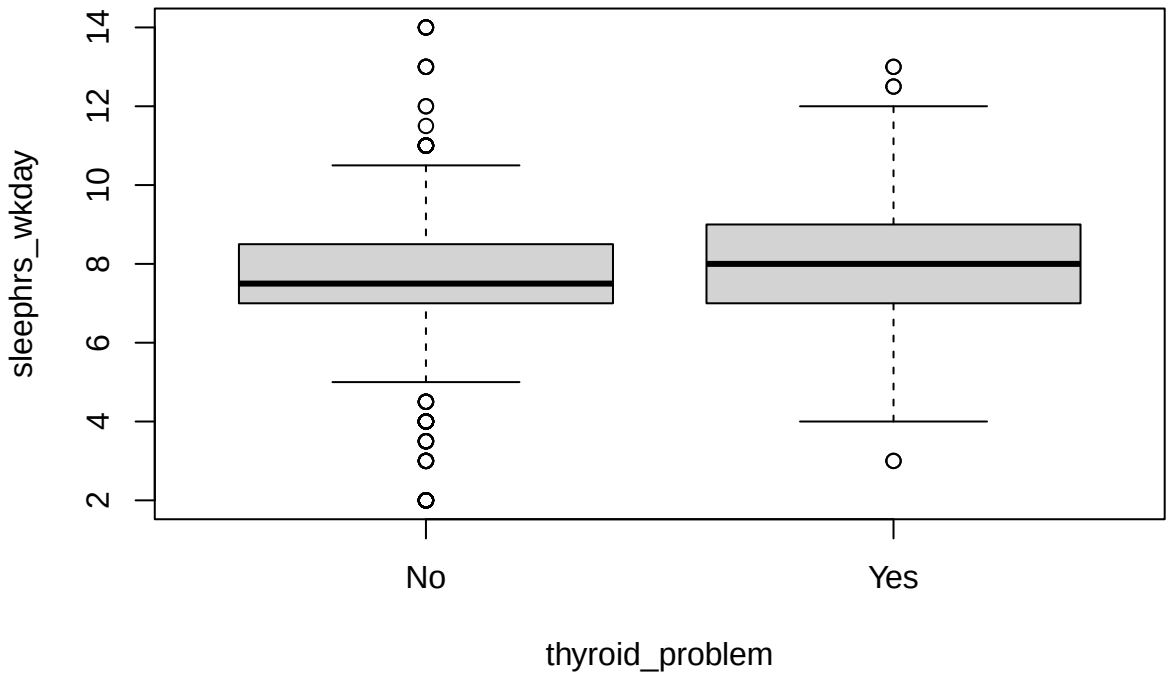
```
##      No     Yes
## 37.8223 42.3676
```

```r
#[1 pts] Show how you did your t-test for vitamin_d3 and thyroid_problem - 1 line:
t.test(vitamin_d3 ~ thyroid_problem, data=NHANES_2, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  vitamin_d3 by thyroid_problem
## t = -4.0726, df = 1594, p-value = 4.878e-05
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -14.199748  -4.968053
## sample estimates:
##  mean in group No mean in group Yes
##          84.28292          93.86682
```

This is how you could plot a box plot, if you are interested in learning more:

```r
boxplot(sleephrs_wkday ~ thyroid_problem, data=NHANES_2)
```




```r
#[1 pts] Compute the quantiles for sleephrs_wkday by thyroid_problem group - 1 line:
tapply(NHANES_2$sleephrs_wkday, NHANES_2$thyroid_problem, quantile)
```

```
## $No
##   0%  25%  50%  75% 100%
##  2.0  7.0  7.5  8.5 14.0
##
## $Yes
##   0%  25%  50%  75% 100%
##    3    7    8    9   13
```

```r
#[1 pts] Perform the Mann-Whitney U test to compare the median sleephrs_wkday of thyroid_problem Yes vs
wilcox.test(sleephrs_wkday ~ thyroid_problem, data = NHANES_2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sleephrs_wkday by thyroid_problem
## W = 192516, p-value = 0.0008948
## alternative hypothesis: true location shift is not equal to 0
```

##Assess the association of arthritis types (4 categories) with age (continuous variable) and sleep hours per workday (discrete variable)

```r
#Providing the table for arthritis types (bonus) -
table(NHANES_2$arthritis, exclude=F)
```

```
##
## Osteoarthritis or degenerative arthritis
##                                    1030
##                                   Other
##                                     191
##                     Psoriatic arthritis
##                                      48
##                    Rheumatoid arthritis
##                                     327
```

```r
#[1 pt] Provide the number of subjects in each arthritis group (hint: don't forget "exclude=F") - 1 lin
table(NHANES_2$arthritis, exclude = FALSE)
```

```
##
## Osteoarthritis or degenerative arthritis
##                                    1030
##                                   Other
##                                     191
##                     Psoriatic arthritis
##                                      48
##                    Rheumatoid arthritis
##                                     327
```

```r
#[1 pt] Provide the overall mean of age - 1 line:
mean(NHANES_2$age)
```

```
## [1] 61.92043
```

```r
#[1 pt] The overall SD of age - 1 line:
sd(NHANES_2$age)
```

```
## [1] 11.87777
```

```r
#[1 pt] The mean of age by arthritis - 1 line:
tapply(NHANES_2$age, NHANES_2$arthritis, mean)
```

```
## Osteoarthritis or degenerative arthritis
##                                 63.13204
##                                    Other
##                                 58.30366
##                      Psoriatic arthritis
##                                 58.25000
##                     Rheumatoid arthritis
##                                 60.75535
```

```r
#[1 pt] The SD of age by arthritis - 1 line
tapply(NHANES_2$age, NHANES_2$arthritis, sd)
```

```
## Osteoarthritis or degenerative arthritis
##                                11.032069
##                                    Other
##                                13.479885
##                      Psoriatic arthritis
##                                 9.557219
##                     Rheumatoid arthritis
##                                13.091283
```

```r
#[2pts] Conduct a one-way anova of age comparisons by arthritis - 2 lines
anova_model <- aov(age ~ arthritis, data = NHANES_2)
summary(anova_model)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## arthritis      3   5101  1700.3   12.31 5.83e-08 ***
## Residuals   1592 219924   138.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#[1pts] Posthoc analysis for the one-way anova test you conducted above - 1 line
TukeyHSD(anova_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = age ~ arthritis, data = NHANES_2)
##
## $arthritis
##                                                                    diff
## Other-Osteoarthritis or degenerative arthritis               -4.82837391
## Psoriatic arthritis-Osteoarthritis or degenerative arthritis -4.88203883
## Rheumatoid arthritis-Osteoarthritis or degenerative arthritis -2.37668715
## Psoriatic arthritis-Other                                    -0.05366492
## Rheumatoid arthritis-Other                                    2.45168676
## Rheumatoid arthritis-Psoriatic arthritis                      2.50535168
##                                                                     lwr
```

```
## Other-Osteoarthritis or degenerative arthritis                -7.2096810
## Psoriatic arthritis-Osteoarthritis or degenerative arthritis  -9.3454101
## Rheumatoid arthritis-Osteoarthritis or degenerative arthritis -4.2953130
## Psoriatic arthritis-Other                                     -4.9340547
## Rheumatoid arthritis-Other                                    -0.3010658
## Rheumatoid arthritis-Psoriatic arthritis                      -2.1667684
##                                                                      upr
## Other-Osteoarthritis or degenerative arthritis                -2.4470669
## Psoriatic arthritis-Osteoarthritis or degenerative arthritis  -0.4186676
## Rheumatoid arthritis-Osteoarthritis or degenerative arthritis -0.4580613
## Psoriatic arthritis-Other                                      4.8267249
## Rheumatoid arthritis-Other                                     5.2044394
## Rheumatoid arthritis-Psoriatic arthritis                       7.1774718
##                                                                    p adj
## Other-Osteoarthritis or degenerative arthritis                0.0000012
## Psoriatic arthritis-Osteoarthritis or degenerative arthritis  0.0255733
## Rheumatoid arthritis-Osteoarthritis or degenerative arthritis 0.0080157
## Psoriatic arthritis-Other                                     0.9999919
## Rheumatoid arthritis-Other                                    0.1006402
## Rheumatoid arthritis-Psoriatic arthritis                      0.5126962
```

```
#[1 pt] The overall quantiles of sleephrs_wkday - 1 line:
quantile(NHANES_2$sleephrs_wkday)
```

```
##   0%  25%  50%  75% 100%
##  2.0  7.0  7.5  8.5 14.0
```

```
#[1 pt] The quantiles of sleephrs_wkday by arthritis - 1 line:
tapply(NHANES_2$sleephrs_wkday, NHANES_2$arthritis, quantile)
```

```
## $`Osteoarthritis or degenerative arthritis`
##   0%  25%  50%  75% 100%
##  2.0  7.0  8.0  8.5 13.0
##
## $Other
##   0%  25%  50%  75% 100%
##  2.0  6.5  7.5  8.5 14.0
##
## $`Psoriatic arthritis`
##   0%  25%  50%  75% 100%
##  4.0  6.0  7.5  8.0 14.0
##
## $`Rheumatoid arthritis`
##   0%  25%  50%  75% 100%
##  3.0  7.0  7.5  8.5 13.0
```

```
#[1pts] Conduct a Kruskal Wallis test to compare the median sleephrs_wkday across arthritis categories
kruskal.test(sleephrs_wkday ~ arthritis, data = NHANES_2)
```

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  sleephrs_wkday by arthritis
## Kruskal-Wallis chi-squared = 12.803, df = 3, p-value = 0.005082
```

```
#[1pts] Posthoc analysis for the Kruskal-Wallis test you conducted above - 1 line
pairwise.wilcox.test(NHANES_2$sleephrs_wkday, NHANES_2$arthritis, p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  NHANES_2$sleephrs_wkday and NHANES_2$arthritis
##
##                      Osteoarthritis or degenerative arthritis Other
## Other                0.023                                    -
## Psoriatic arthritis  0.077                                    0.551
## Rheumatoid arthritis 0.221                                    0.155
##                      Psoriatic arthritis
## Other                -
## Psoriatic arthritis  -
## Rheumatoid arthritis 0.155
##
## P value adjustment method: BH
```

```
#Create a new categorical variable, called arthritis_cat - where you re-categorise the arthritis variab

NHANES_2$arthritis_cat <- ifelse(NHANES_2$arthritis=="Osteoarthritis or degenerative arthritis", "osteo

NHANES_2$arthritis_cat <- factor(NHANES_2$arthritis_cat)

data.class(NHANES_2$arthritis_cat)
```

```
## [1] "factor"
```

```
table(NHANES_2$arthritis_cat, exclude=F)
```

```
##
##           osteo other_psoriatic      rheumatoid
##            1030             239             327
```

```
#Let's create a table called chi, where you store a 2-way table of thyroid_problem with the new arthrit

chi <- table(NHANES_2$thyroid_problem, NHANES_2$arthritis_cat, exclude=F)
chi
```

```
##
##       osteo other_psoriatic rheumatoid
##   No    774             193        280
##   Yes   256              46         47
```

```
#[1pts] Please use the table "chi" created above to perform a chi-sq. test - 1 line
chisq.test(chi)
```

```
##
##  Pearson's Chi-squared test
##
## data:  chi
## X-squared = 17.089, df = 2, p-value = 0.0001946
```