PS: One Sample Z test for TAs

Pablo E. Gutiérrez-Fonseca

2023-08-09

R practice.

Install packages.

```
library(ggplot2)
```

Load the water pollution data into R.

```
##   State     Country Income_class Ethnicity Cancer_risk_mil
## 1    CA Los Angeles         High     White             290
## 2    MI     Clinton         High     White             240
## 3    MI     Clinton          Low     White             240
## 4    NH   Strafford          Low     White             235
## 5    FL  Palm Beach          Low     White             235
## 6    CA Los Angeles          Low     White             210
```

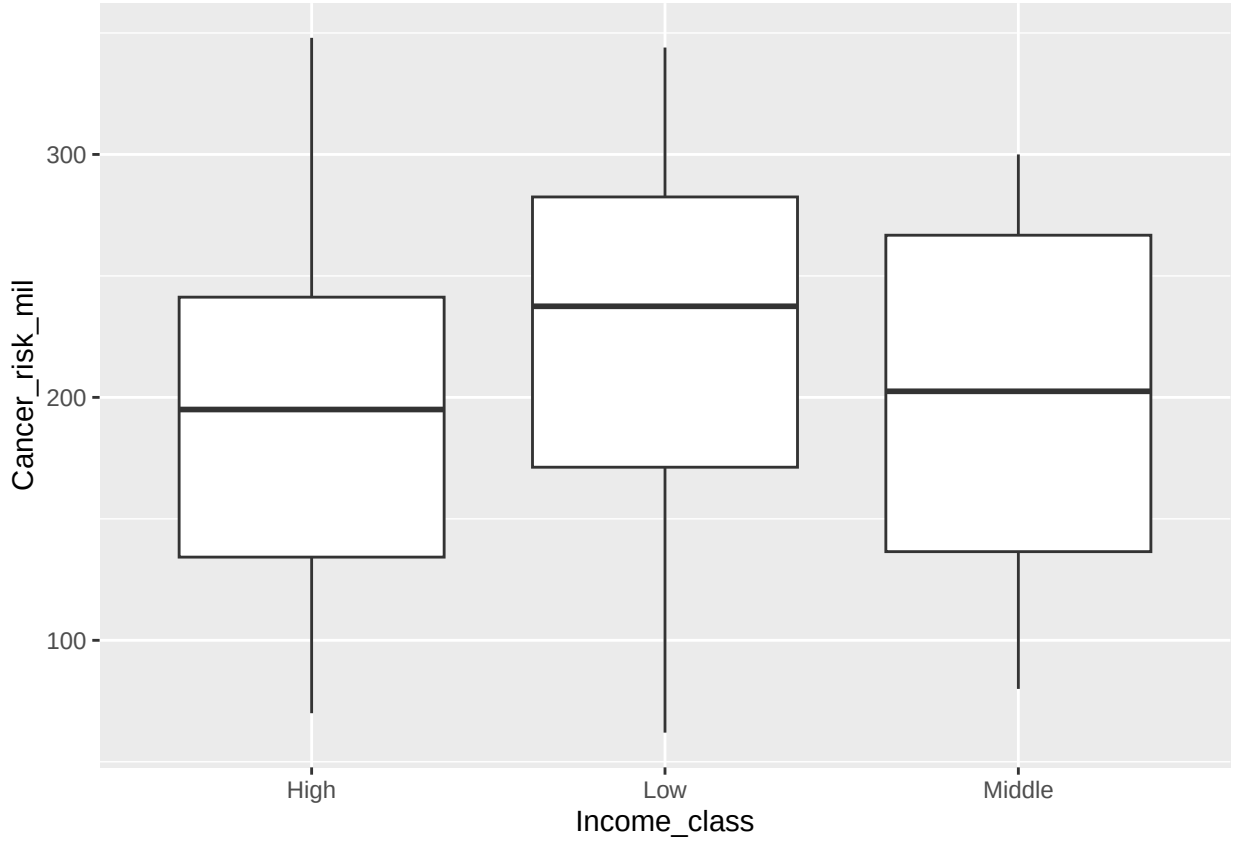
H 1 : Cancer risk differs by income level.

```
mod1 <- aov(Cancer_risk_mil~ Income_class, data=df_env_just)
summary(mod1)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Income_class  2  21165   10582   1.952  0.149
## Residuals    81 439074    5421
```

A one-way ANOVA to characterize the difference in CDC reported cancer rates across US counties indicated that there is no significant difference between income levels (F 2,81 = 1.95, p = 0.1486). The cancer rate for low, middle and high income groups differed only as one would expect due to random variation in the human population. This suggests that income level is not related to cancer rates, and rich are equally as susceptible as poor individuals.

```
ggplot(df_env_just, aes(x=Income_class, y=Cancer_risk_mil)) +
geom_boxplot()
```



H 1 : H 1 : Cancer risk differs by ethnic group.

```
mod2 <- aov(Cancer_risk_mil~ Ethnicity, data=df_env_just)
summary(mod2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Ethnicity    3  53425   17808   3.502 0.0192 *
## Residuals   80 406814    5085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A one-way ANOVA to characterize the difference in CDC reported cancer rates across 5 US counties indicated that there is a significant difference between ethnic groups (F 3,80 = 3.5020, p = 0.0192, r2 = 0.11). Post hoc Tukey's means comparison tests indicate that African Americans have significantly higher cancer risk than whites, but not significantly different from Latino and Native American. Latino and Native Americans were also similar to Whites. A very low R-square (0.11) indicates, that while significant, Ethnicity does not account for much of the overall variability in the model for cancer risk, and is likely not meaningful. Other factors likely contribute to cancer rates.

```
ggplot(df_env_just, aes(x=Ethnicity, y=Cancer_risk_mil)) +
geom_boxplot()
```