6: Part 3 - Generalized Linear Models

Environmental Data Analytics | John Fay and Luana Lima | Developed by Kateri Salk

Spring 2023

Objectives

1. Describe t-test under GLM framework

Set up your session

```
#install.packages('formatR')
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)

library(tidyverse)
library(lubridate)
library(here)
here()
```

```
## [1] "/Users/sokna/Documents/EDA-Spring2023"
```

T-Test

Continuous response, one categorical explanatory variable with two categories (or comparison to a single value if a one-sample test).

Formulating Hypothesis for μ

Two hypotheses are formed – the null hypothesis and the alternative hypothesis. The null hypothesis and the alternative hypothesis combine to cover all possible values for the population mean. The null hypothesis must have the equality. The null and alternative hypotheses are always stated in terms of the population mean (mu).

One-sample t-test

The object of a one sample test is to test the null hypothesis that the mean of the group is equal to a specific value.

Function t.test() **x** a (non-empty) numeric vector of data values. **alternative** a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter. **mu** a number indicating the true value of the mean (or difference in means if you are performing a two sample test). **formula** a formula of the form lhs ~ rhs where lhs is a numeric variable

giving the data values and rhs either 1 for a one-sample or paired test or a factor with two levels giving the corresponding groups. If lhs is of class "Pair" and rhs is 1, a paired test is done.

The one-sample t-test relies on the assumption that the variable is normally distributed in the population. However, the t-test is robust to mild departures from normality when the sample size is small, and when the sample size is large the normality assumption becomes less important.

For example, we might ask ourselves (from the EPA air quality processed dataset): Are Ozone levels below the threshold for "good" AQI index (0-50)?

Import data set

```
EPAair <- read.csv(here("Data/Processed_KEY/EPAair_O3_PM25_NC1819_Processed.csv"),
    stringsAsFactors = TRUE)
# Set date to date format
EPAair$Date <- as.Date(EPAair$Date, format = "%Y-%m-%d")
```

Exercise 1: State the hypotheses for testing mean of AQI index.

Answer: H0: mu >= 50 Ha: mu < 50

```
summary(EPAair$Ozone)  #mean = 40.88 and median = 40
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.00   32.00   40.00   40.88   46.00  129.00    2146
```

```
length(EPAair$Ozone)  #8976 observations
```

```
## [1] 8976
```

```
O3.onesample <- t.test(EPAair$Ozone, mu = 50, alternative = "less")
O3.onesample
```

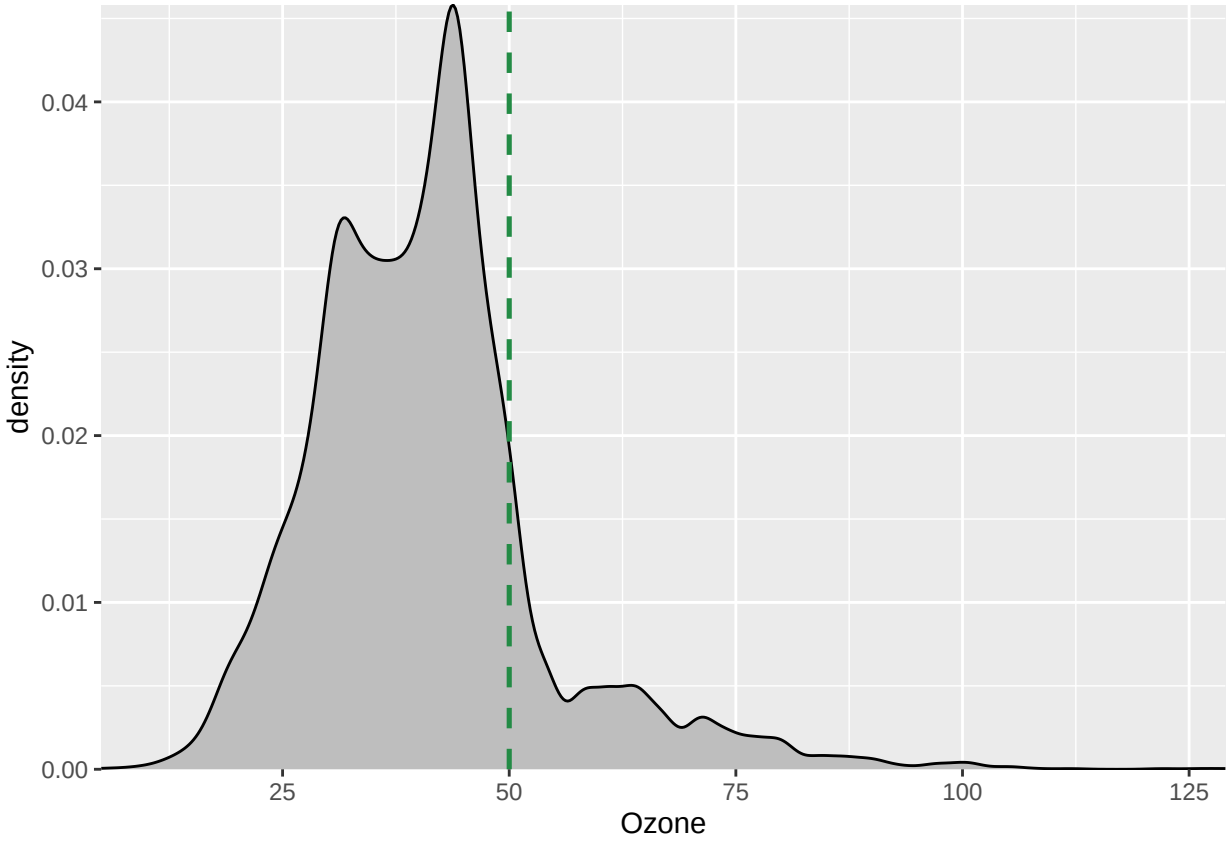
```
##
##  One Sample t-test
##
## data:  EPAair$Ozone
## t = -57.98, df = 6829, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 50
## 95 percent confidence interval:
##      -Inf 41.13416
## sample estimates:
## mean of x
##  40.87526
```

```
Ozone.plot <- ggplot(EPAair, aes(x = Ozone)) + geom_density(fill = "gray") + geom_vline(xintercept = 50
    color = "#238b45", lty = 2, size = 0.9) + scale_x_continuous(expand = c(0, 0)) +
    scale_y_continuous(expand = c(0, 0))
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```

```
print(Ozone.plot)
```

```
## Warning: Removed 2146 rows containing non-finite values (‘stat_density()‘).
```




```
# Format as a GLM
O3.onesample2 <- lm(Ozone ~ 1, EPAair)
O3.onesample2
```

```
##
## Call:
## lm(formula = Ozone ~ 1, data = EPAair)
##
## Coefficients:
## (Intercept)
##       40.88
```

Write a sentence or two about the results of this test. Include both the results of the test and an interpretation that puts the findings in context of the research question.

Two-sample t-test

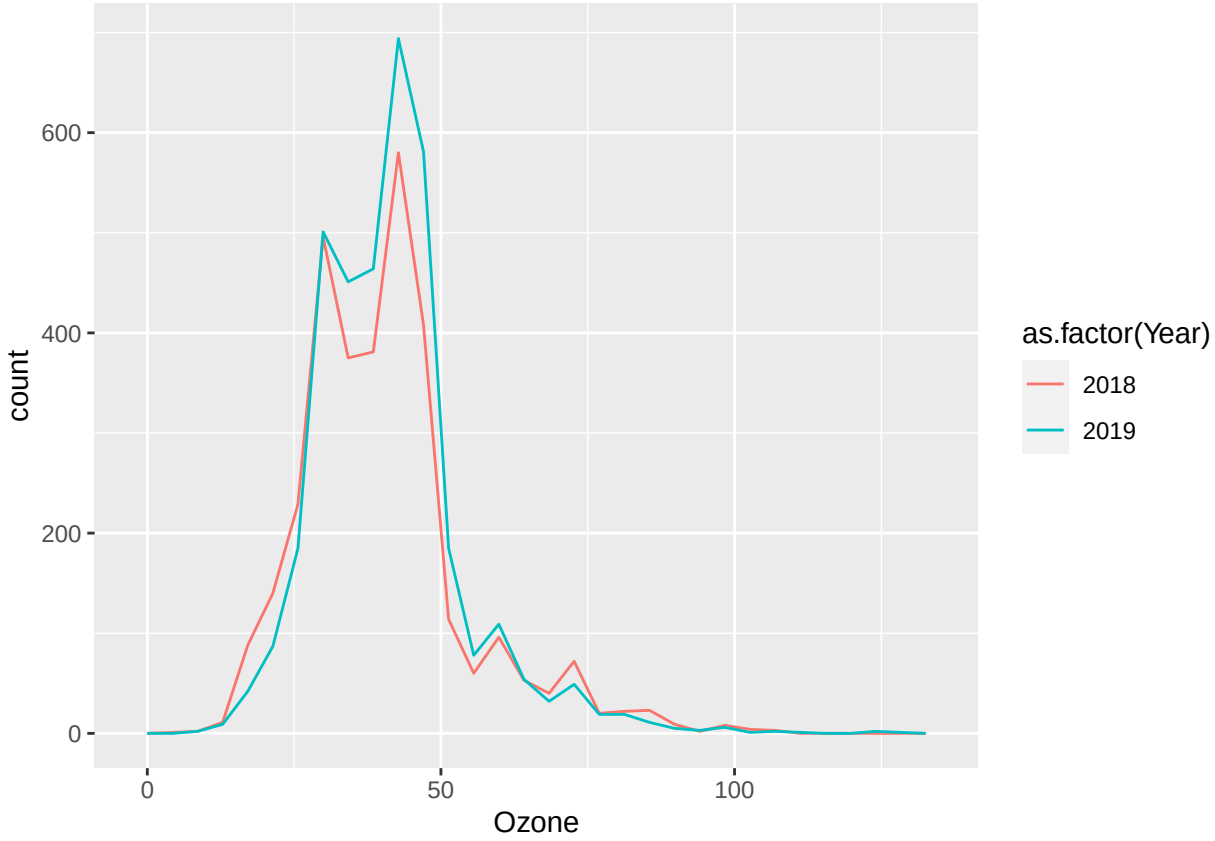
The two-sample t test is used to test the hypothesis that the mean of two samples is equivalent. Unlike the one-sample tests, a two-sample test requires a second assumption that the variance of the two groups is equivalent.

For example, we might ask ourselves (from the EPA air quality processed dataset): Are Ozone levels different between 2018 and 2019?

```
# First let's look at the data
ggplot(EPAair, aes(x = Ozone, color = as.factor(Year))) + geom_freqpoly()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2146 rows containing non-finite values (`stat_bin()`).
```




```
# Format as a t-test EPAair$Ozone will be our continuous dependent variable
# EPAair$Year will be our categorical variable with two levels (2018 and 2019)
O3.twosample <- t.test(EPAair$Ozone ~ EPAair$Year)
O3.twosample
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  EPAair$Ozone by EPAair$Year
## t = -2.6642, df = 6467.7, p-value = 0.007736
## alternative hypothesis: true difference in means between group 2018 and group 2019 is not equal to 0
## 95 percent confidence interval:
##  -1.4670426 -0.2232942
## sample estimates:
## mean in group 2018 mean in group 2019
##           40.43065           41.27581
```

```
# Format as a GLM
O3.twosample2 <- lm(EPAair$Ozone ~ EPAair$Year)
summary(O3.twosample2)
```

```
##
## Call:
## lm(formula = EPAair$Ozone ~ EPAair$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.431  -8.431  -0.431   5.569  87.724
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1665.1192   635.9203  -2.618  0.00885 **
## EPAair$Year     0.8452     0.3150   2.683  0.00732 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13 on 6828 degrees of freedom
##   (2146 observations deleted due to missingness)
## Multiple R-squared:  0.001053,   Adjusted R-squared:  0.0009066
## F-statistic: 7.197 on 1 and 6828 DF,  p-value: 0.00732
```