

Week-8 Activity

Anas

2025-07-01

Activity – Check Assumptions

```
install.packages("gapminder")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)

install.packages("dplyr")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)

install.packages("ggplot2")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)

library(gapminder)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(ggplot2)

# Subset data for 2002
data_2002 <- gapminder %>% filter(year == 2002, continent %in% c("Europe", "Asia"))

# Check sample size
table(data_2002$continent)

##
##   Africa Americas      Asia  Europe Oceania
##        0         0       33       30        0

# Check normality (Shapiro-Wilk) - sample size is large, so approximation OK
by(data_2002$lifeExp, data_2002$continent, shapiro.test)

## data_2002$continent: Africa
## NULL
## -----
## data_2002$continent: Americas
## NULL
## -----
## data_2002$continent: Asia
##
## Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.93423, p-value = 0.04625
##
```

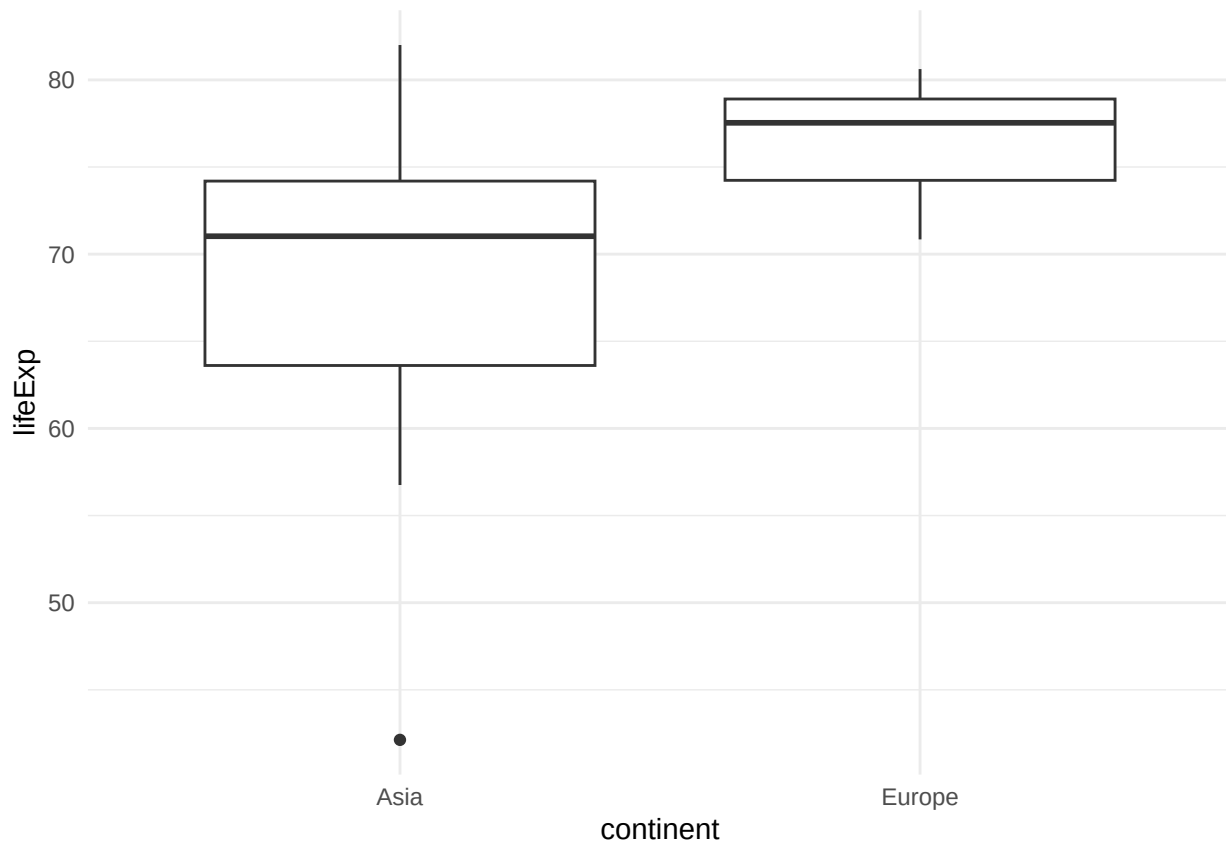
```
## -----
## data_2002$continent: Europe
##
## Shapiro-Wilk normality test
##
## data:  dd[, ]
## W = 0.92943, p-value = 0.04743
##
## -----
## data_2002$continent: Oceania
## NULL

# Check homogeneity of variances
var.test(lifeExp ~ continent, data = data_2002)

##
## F test to compare two variances
##
## data:  lifeExp by continent
## F = 8.2132, num df = 32, denom df = 29, p-value = 1.479e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  3.955443 16.834032
## sample estimates:
## ratio of variances
##      8.213228
```

Are the groups different? Proceed with the test.

```
# Boxplot for visual inspection
ggplot(data_2002, aes(x = continent, y = lifeExp)) +
  geom_boxplot() +
  theme_minimal()
```



```
# Run t-test
t.test(lifeExp ~ continent, data = data_2002, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data: lifeExp by continent
## t = -4.631, df = 61, p-value = 1.956e-05
## alternative hypothesis: true difference in means between group Asia and group Europe is not equal to 0
## 95 percent confidence interval:
## -10.690779 -4.242663
## sample estimates:
## mean in group Asia mean in group Europe
## 69.23388 76.70060
```

Activity – Match the Test

```
# Independent t-test
# lifeExp between two continents
subset_data <- gapminder %>% filter (continent %in% c("Europe", "Asia"))
t.test(lifeExp ~ continent, data = subset_data)
```

```
##
## Welch Two Sample t-test
##
## data: lifeExp by continent
## t = -17.899, df = 565.16, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means between group Asia and group Europe is not equal to
## 95 percent confidence interval:
## -13.13792 -10.53965
## sample estimates:
## mean in group Asia mean in group Europe
## 60.06490 71.90369
```

```
# ANOVA
```

```
anova_model <- aov (lifeExp ~ continent, data = gapminder)
summary(anova_model)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## continent    4 139343    34836   408.7 <2e-16 ***
## Residuals 1699 144805      85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Kruskal-Wallis
```

```
kruskal.test(lifeExp ~ continent, data = gapminder)
```

```
##
## Kruskal-Wallis rank sum test
##
## data: lifeExp by continent
## Kruskal-Wallis chi-squared = 843.38, df = 4, p-value < 2.2e-16
```