

Week-6 Activity

Anas

2025-06-21

Activity 1 :- Change sample size and observe changes in shape

```
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)

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

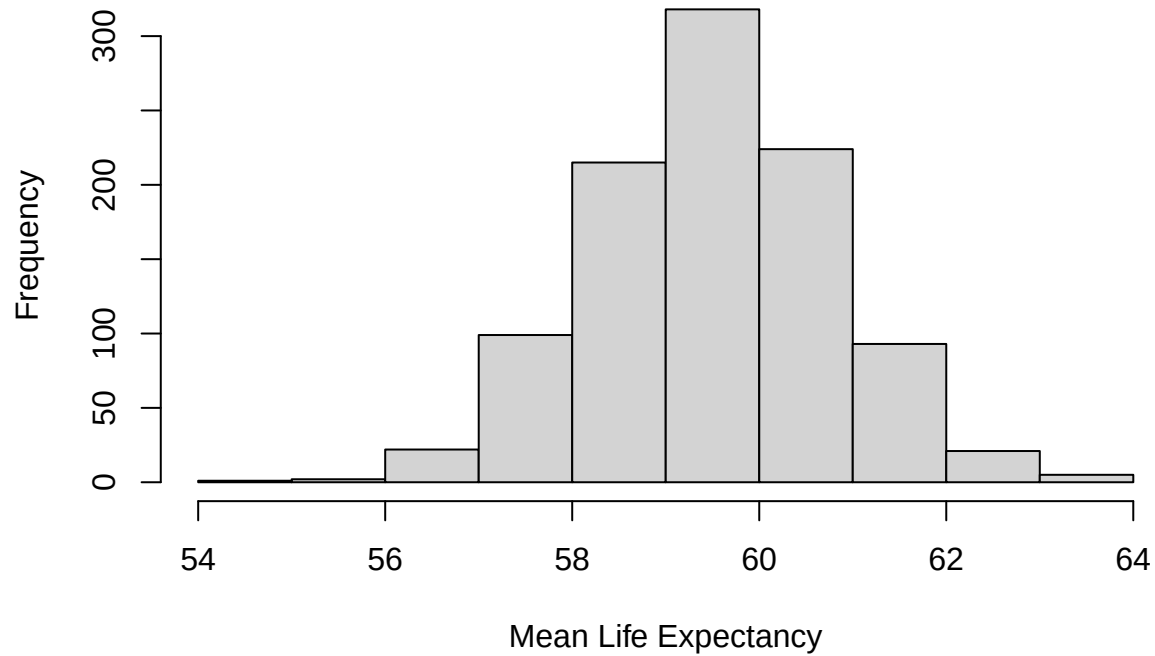
# Load the dataset
data("gapminder")

# Extract life expectancy data
life_data <- gapminder$lifeExp

sample_means <- replicate(1000, mean(sample(life_data, size = 100, replace = TRUE)))

# Visualise the sampling distribution
hist(sample_means,
      main = "Sampling Distribution of Sample Mean",
      xlab = "Mean Life Expectancy")
```

Sampling Distribution of Sample Mean

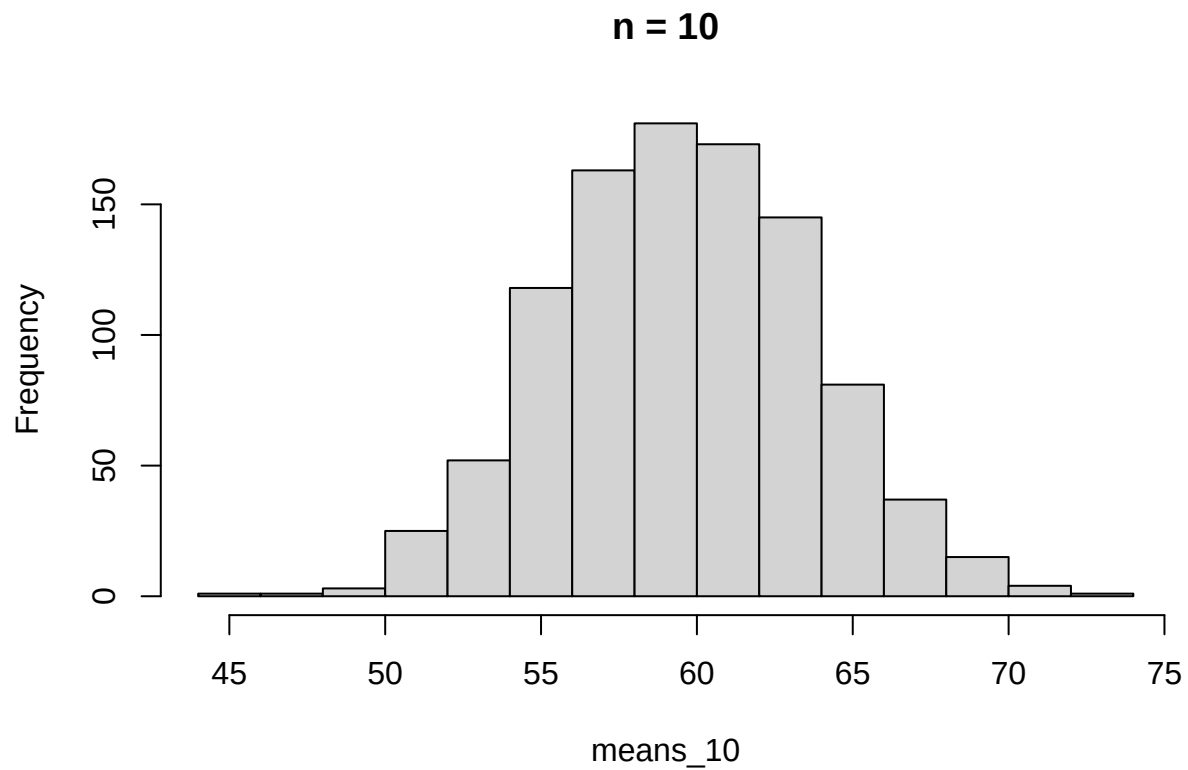


##

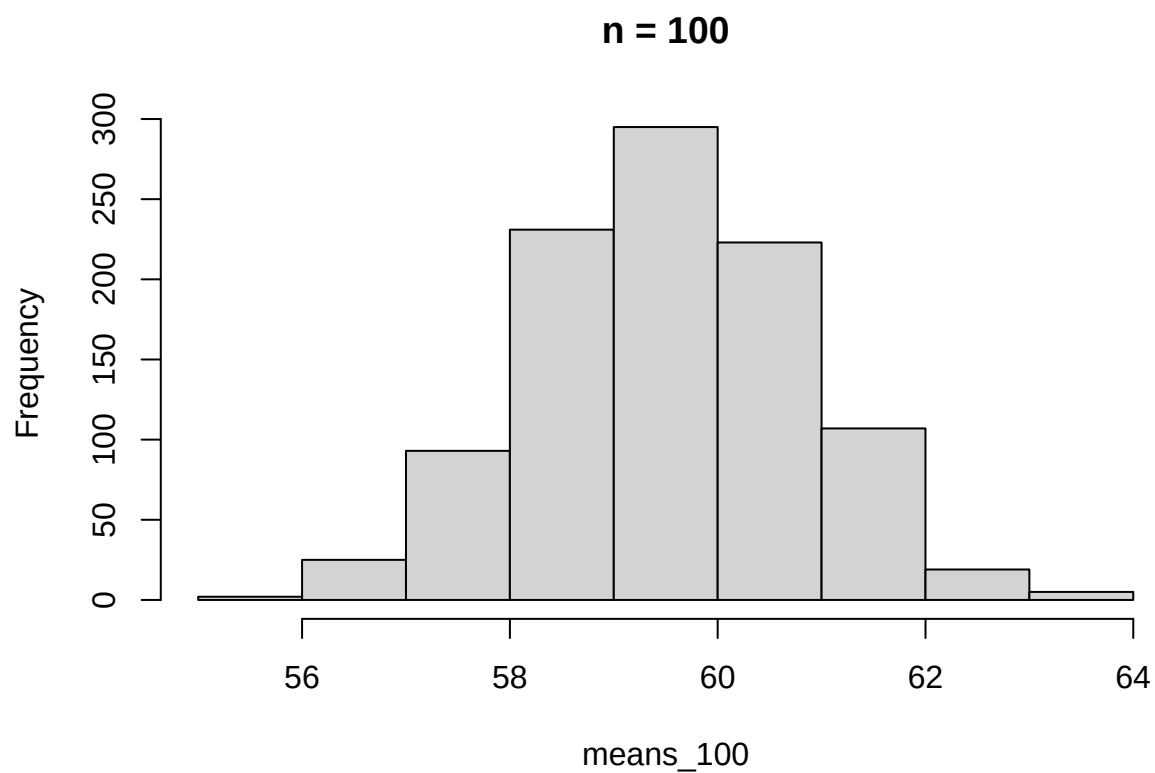
Effect of Sample Size

```
# Try with n = 10 and n = 100
means_10 <- replicate(1000, mean(sample(life_data, 10, replace = TRUE)))
means_100 <- replicate(1000, mean(sample(life_data, 100, replace = TRUE)))

hist(means_10, main="n = 10")
```



```
hist(means_100, main="n = 100")
```



Activ-

ity 2 :- Sampling Distribution Exploration

```
# Function to simulate sampling means
simulate_means <- function(n, reps = 1000) {
```

```

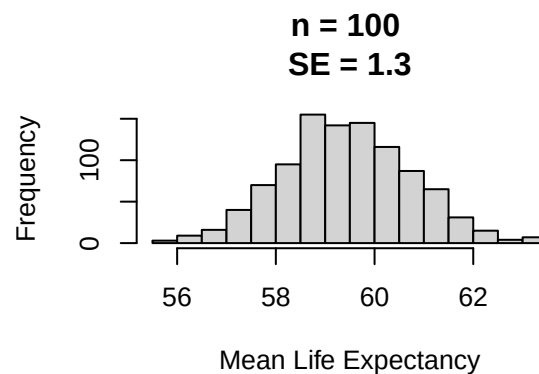
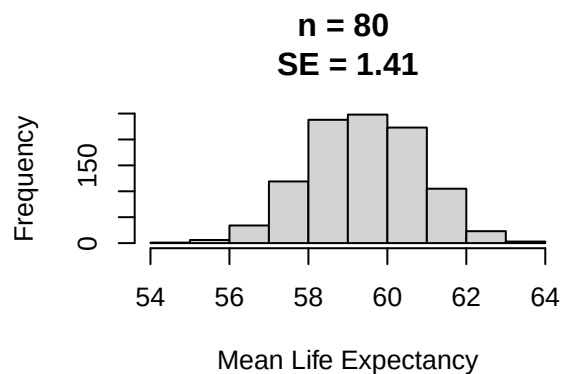
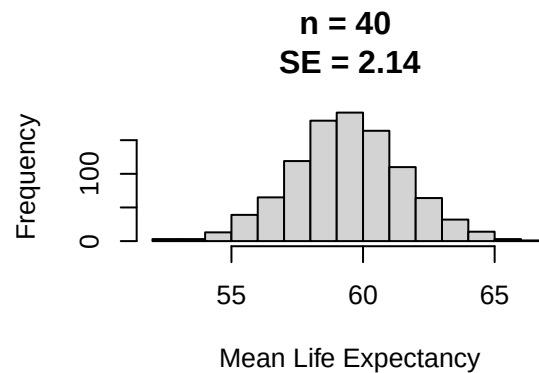
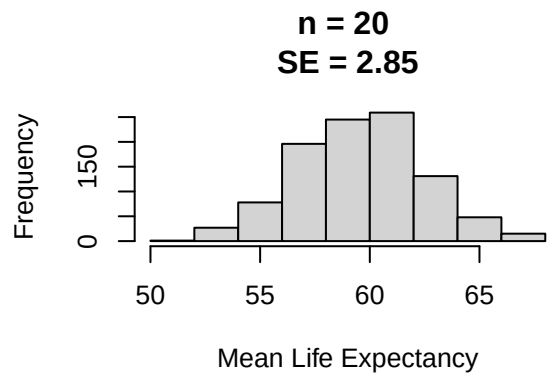
replicate(reps, mean(sample(life_data, size = n, replace = TRUE)))
}

# Define sample sizes
sample_sizes <- c(20, 40, 80, 100)

# Set plot layout for histograms
par(mfrow = c(2, 2))

# Loop over each sample size
for (n in sample_sizes) {
  sample_dist <- simulate_means(n)
  se <- sd(sample_dist) # Standard Error = SD of sample means
  hist(sample_dist,
        main = paste("n =", n, "\nSE =", round(se, 2)),
        xlab = "Mean Life Expectancy")
}

```



```

par(mfrow = c(1, 1)) # Reset plot layout

```

Exploring the Data

```

install.packages("ggplot2")

```

```

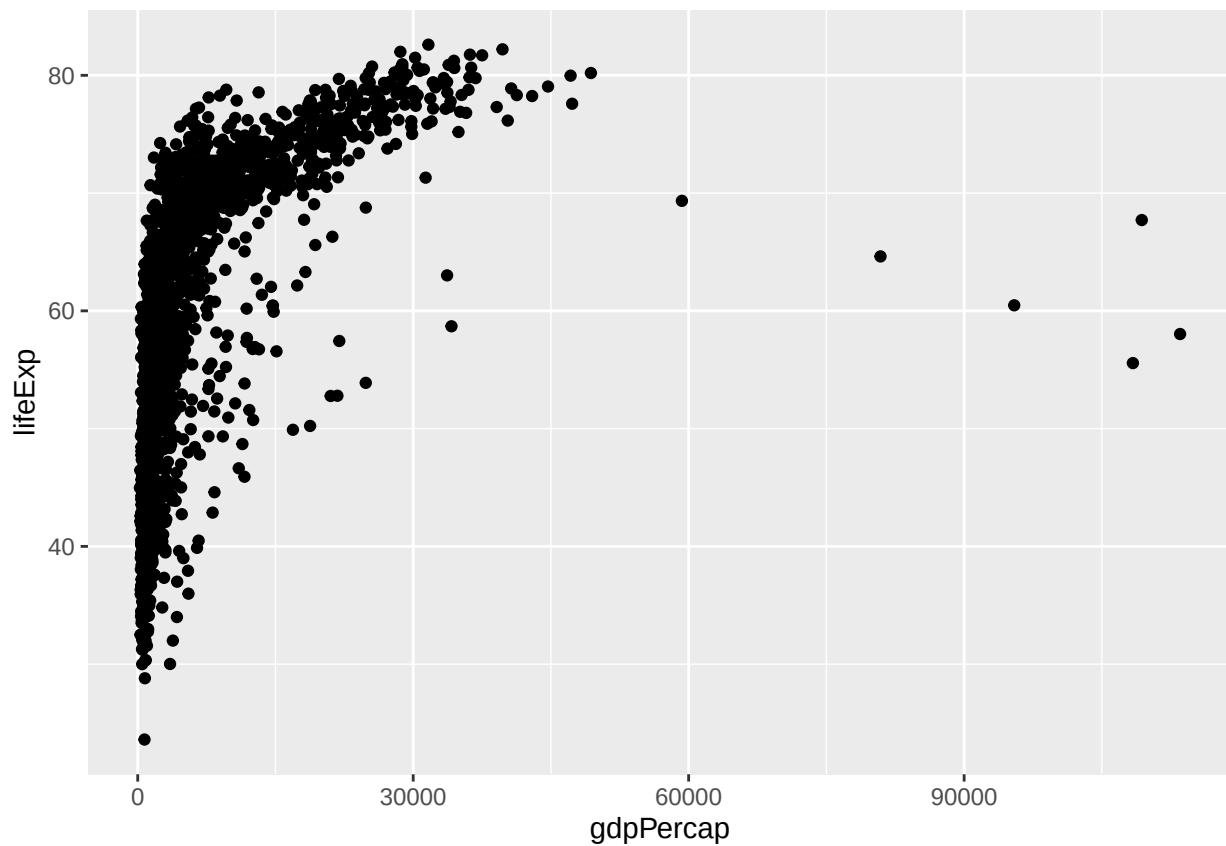
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)

```

```
library(ggplot2)
summary(gapminder)
```

```
##      country      continent      year      lifeExp
## Afghanistan: 12 Africa :624 Min. :1952 Min. :23.60
## Albania : 12 Americas:300 1st Qu.:1966 1st Qu.:48.20
## Algeria : 12 Asia :396 Median :1980 Median :60.71
## Angola : 12 Europe :360 Mean :1980 Mean :59.47
## Argentina : 12 Oceania : 24 3rd Qu.:1993 3rd Qu.:70.85
## Australia : 12 Max. :2007 Max. :82.60
## (Other) :1632
##      pop      gdpPercap
## Min. :6.001e+04 Min. : 241.2
## 1st Qu.:2.794e+06 1st Qu.: 1202.1
## Median :7.024e+06 Median : 3531.8
## Mean :2.960e+07 Mean : 7215.3
## 3rd Qu.:1.959e+07 3rd Qu.: 9325.5
## Max. :1.319e+09 Max. :113523.1
##
```

```
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp)) + geom_point()
```



Correlation Analysis - Pearson

Objective: Measure relationships between numeric variables.

```
cor(gapminder$lifeExp, gapminder$gdpPerCap, method = "pearson")
```

```
## [1] 0.5837062
```

Correlation Analysis - Spearman

Objective: Measure relationships between numeric variables.

```
cor(gapminder$lifeExp, gapminder$gdpPerCap, method = "spearman")
```

```
## [1] 0.8264712
```

```
# Visualise
```

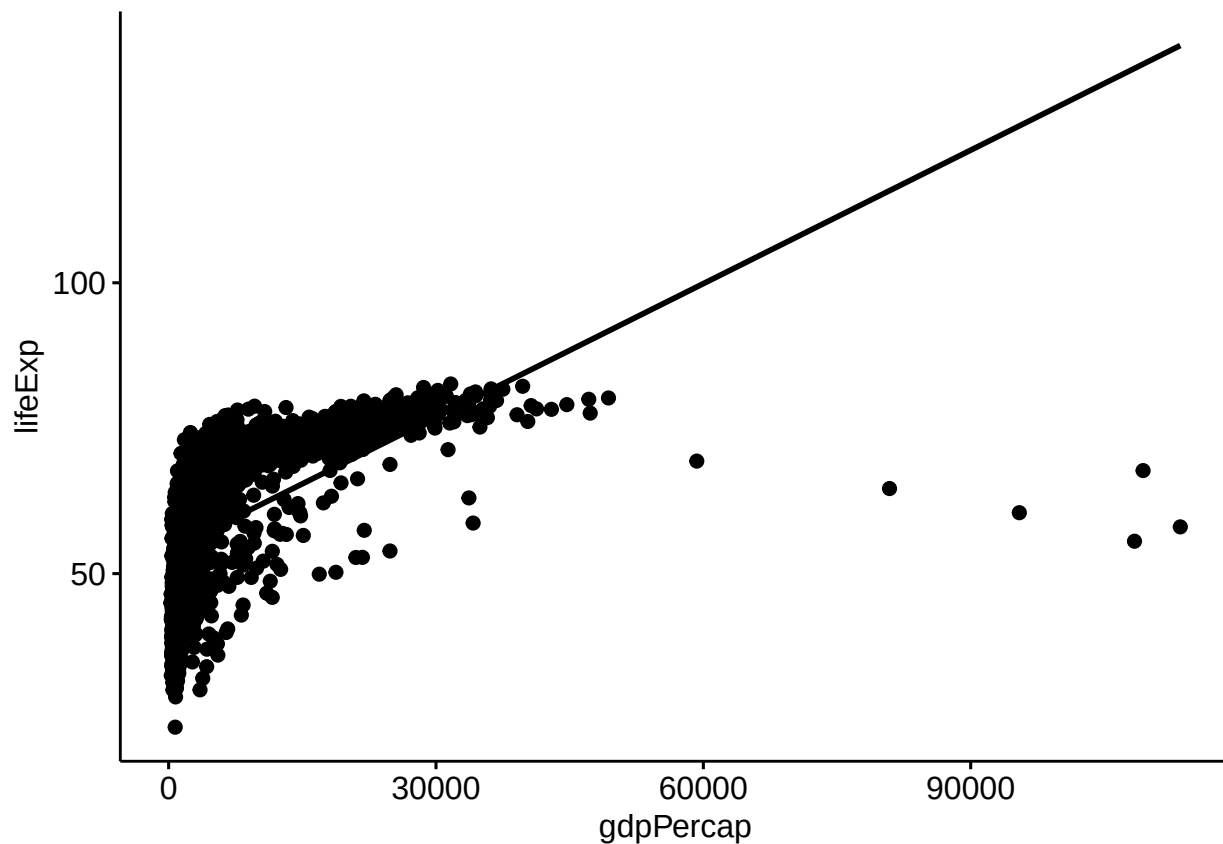
```
install.packages("ggpubr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
```

```
## (as 'lib' is unspecified)
```

```
library(ggpubr)
```

```
ggscatter(gapminder, x = "gdpPerCap", y = "lifeExp", add = "reg.line")
```



Activity 3: Compare Pearson vs. Spearman for lifeExp vs. pop. Which is more appropriate? Why?

Pearson correlation

```
pearson_corr <- cor(gapminder$lifeExp, gapminder$gdpPercap, method = "pearson")  
pearson_corr
```

```
## [1] 0.5837062
```

Spearman correlation

```
spearman_corr <- cor(gapminder$lifeExp, gapminder$gdpPercap, method = "spearman")  
spearman_corr
```

```
## [1] 0.8264712
```

Because of this :- If the relationship between life expectancy and population is non-linear or has extreme population outliers, Spearman is more appropriate. If the relationship is roughly linear without strong outliers, Pearson is suitable.