

Inferential Statistics Part 1

Setup: Load packages and view dataset

R Code:

```
# Install if needed:
# install.packages(c("gapminder", "dplyr", "ggplot2", "ggpubr"))

library(gapminder)
library(dplyr)
library(ggplot2)

head(gapminder)
summary(gapminder)

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

Output:

Output: head() prints first rows of the dataset; summary() prints descriptive summaries; the ggplot code produces a scatterplot.

Sampling Distribution Simulation (CLT)

R Code:

```
library(gapminder)
library(dplyr)

# Extract life expectancy data
life_data <- gapminder$lifeExp

set.seed(123) # For reproducibility
# Generate 1000 sample means (n = 30)
sample_means <- replicate(1000, mean(sample(life_data, size = 30, replace =
TRUE)))

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

Output:

Output: histogram of 1000 sample means (n=30).

Effect of Sample Size (n=10 vs n=100)

R Code:

```
set.seed(123)

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")
```

Output:

Output: two histograms. The n=100 sampling distribution appears more normal and has smaller spread (standard error).

Activity: Sampling Distribution Exploration (function + multiple histograms)

R Code:

```
simulate_means <- function(n, reps = 1000) {
  replicate(reps, mean(sample(life_data, size = n, replace = TRUE)))
}

sample_sizes <- c(10, 30, 50, 100)

par(mfrow = c(2,2))
for (n in sample_sizes) {
  hist(simulate_means(n), main = paste("n =", n))
}
```

Output:

Output: 2x2 panel of histograms for n = 10, 30, 50, 100.

Correlation Analysis (Pearson + plot)

R Code:

```
library(ggpubr)

cor(gapminder$lifeExp, gapminder$gdpPercap, method = "pearson")

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

Output:

Output: Pearson correlation coefficient (a single numeric value) and a scatterplot with regression line.

Correlation Analysis (Spearman + activity for pop)

R Code:

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

# Activity: compare Pearson vs Spearman for lifeExp vs pop
cor(gapminder$lifeExp, gapminder$pop, method = "pearson")
cor(gapminder$lifeExp, gapminder$pop, method = "spearman")
```

Output:

Output: Spearman correlation coefficient for lifeExp vs gdpPercap, and Pearson/Spearman coefficients for lifeExp vs pop.

Two-sample t-test (Europe vs Africa) – Welch t-test

R Code:

```
europe <- gapminder %>% filter(continent == "Europe") %>% pull(lifeExp)
africa <- gapminder %>% filter(continent == "Africa") %>% pull(lifeExp)

t.test(europe, africa)
```

Output:

Output (as shown in slides):

- t = 49.551
- df = 981.2
- p-value < 2.2e-16
- 95% CI for mean difference (Europe – Africa): 22.13 to 23.95
- mean(Europe) $\approx$ 71.9, mean(Africa) $\approx$ 48.9

Conclusion: Reject H0; Europe's mean life expectancy is significantly higher.

Simple Linear Regression (lifeExp ~ gdpPercap)

R Code:

```
model <- lm(lifeExp ~ gdpPercap, data = gapminder)
summary(model)
```

Output:

Output (key coefficients as shown in slides):

- (Intercept) $\approx$ 53.96

- gdpPercap coefficient ≈ 0.0007649

Interpretation: +\$1000 GDP per capita $\rightarrow$ lifeExp increases by ~ 0.7649 years (0.0007649×1000).

Multiple Regression (lifeExp ~ gdpPercap + year + continent)

R Code:

```
model_multi <- lm(lifeExp ~ gdpPercap + year + continent, data = gapminder)
summary(model_multi)
```

Output:

Output: regression summary including coefficients for GDP per capita, year, and continent (categorical).