Quantitative Reasoning – LAB 413

Describing Data & Relationships

Week 2–3 Recap ▪ Lab 2

Instructor: **Subin Na**

September 10, 2025

Today's Plan

1. Week 2–3 Recap

- Measures of Center
- Spread of Data
- Visualizing Distributions
- Standardization (Z-scores)
- Correlation
- Linear Regression
- Least Squares Regression

2. RStudio Practice

- Data frames, plots, `ggplot2`
- Transformation, correlation, regression

01 Measures of Center

Key Concepts

- **Mean** = arithmetic average — *sensitive* to outliers.
- **Median** = midpoint — *resistant* to outliers.
- When distributions are symmetric $\Rightarrow$ mean $\approx$ median.

Incomes $= [40k, 45k, 50k, 55k, 1M]$

- Mean $\approx 238k \Rightarrow$ does not represent most people!

Test scores $= [65, 70, 75, 80, 85]$

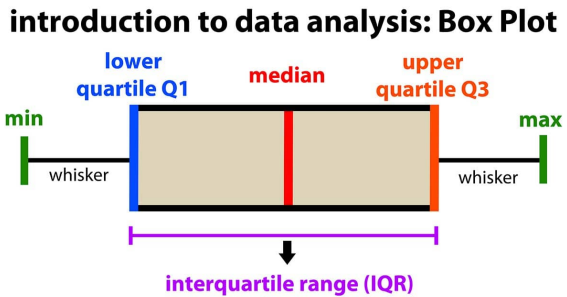
- Mean $= 75$, Median $= 75$

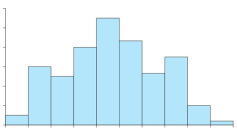
02 Spread of Data

- **Quartiles:** split the data into four equal parts.
- **IQR** ($Q_3 - Q_1$): the "middle spread" where most of the data sits.
- **Five-number summary:** min, Q_1, median, Q_3, max.
- **Standard deviation (SD):** on average, how far the values are from the mean.
- **Outliers:** a common rule is anything more than $1.5 \times$ IQR.

03 Visualizing Distributions (1)

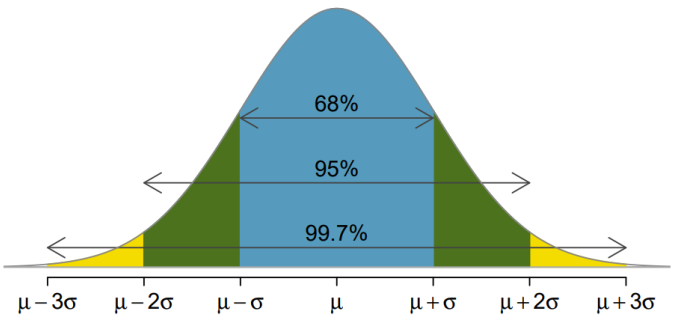
- **Boxplots** → visualize the five-number summary (min, Q_1, median, Q_3, max).
- **Histograms** → distribution shape (bins).





03 Visualizing Distributions (2)

- **Density curves / Normal distribution:** bell-shaped, symmetric, unimodal.
- **68–95–99.7:** rule for normal distributions.



04 Standardization (Z-scores)

Key Concepts

A **z-score** says how many standard deviations a value is from the mean, transforming different units into a common scale.

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x} = \frac{x - \mu}{\sigma}$$

- Interpret as how many SDs from the mean.
- Useful for **comparing across variables**.

05 Correlation (1)

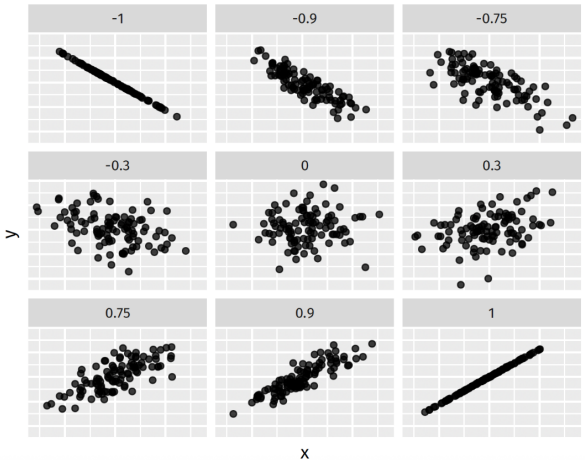
Key Concepts

A number that tells you how two variables move together — it measures **linear association** (r between -1 and 1).

- Positive $r \rightarrow$ variables move **together**.
- Negative $r \rightarrow$ variables move in **opposite** directions.
- $r = 0 \rightarrow$ no linear relationship.

05 Correlation (2)

The best way to see correlation is with a **scatterplot!**

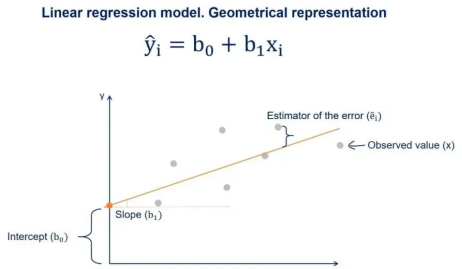


Key Concepts

Draw the **best-fitting line** through the data to predict Y from X.

$$Y_i = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} \cdot X_i + \underbrace{\epsilon_i}_{\text{error term}}$$

- **Intercept** (α): value of Y when $X = 0$.
- **Slope** (β): change in Y per one-unit change in X.
- **Residuals:** prediction errors.

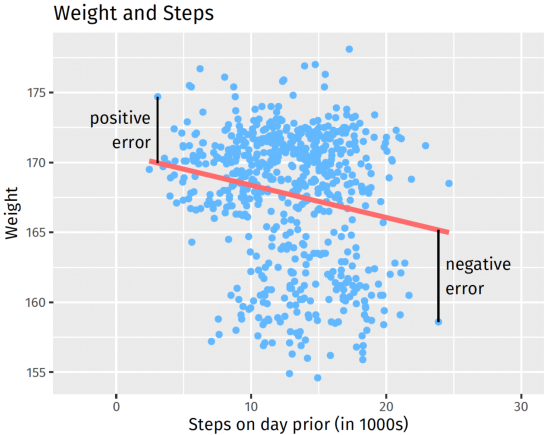


07 Least Squares Regression

- Find the line that minimizes the **sum of squared residuals (SSR)**.
- The best-fit line balances errors above and below the line ⇒ overall error is minimized!

R function:

```
lm(y ~ x, data = mydata)
```



RStudio Practice

Visualization, Transformation, Correlation & Regression

`summary()` for descriptive stats ▪ `ggplot2` for plots

`cor()` for correlation ▪ `lm()` for regression

See `Lab2.Rmd` / `Lab2_0910.R`