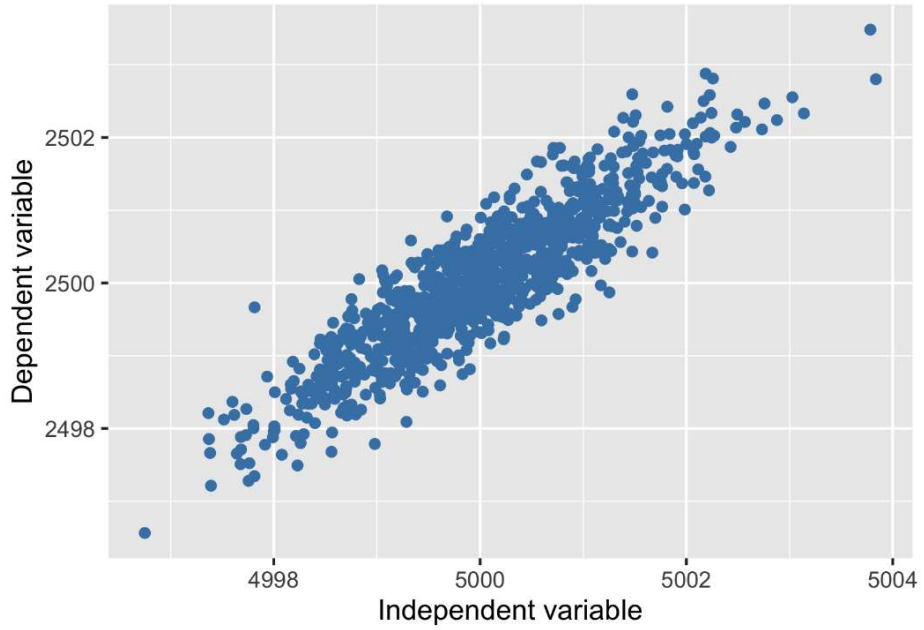
# Independent and dependent variables

- **Independent variable:**
  - Unaffected by other data
  - Vitamin C supplementation
- **Dependent variable:**
  - Affected by other data
  - Birth gender ratio

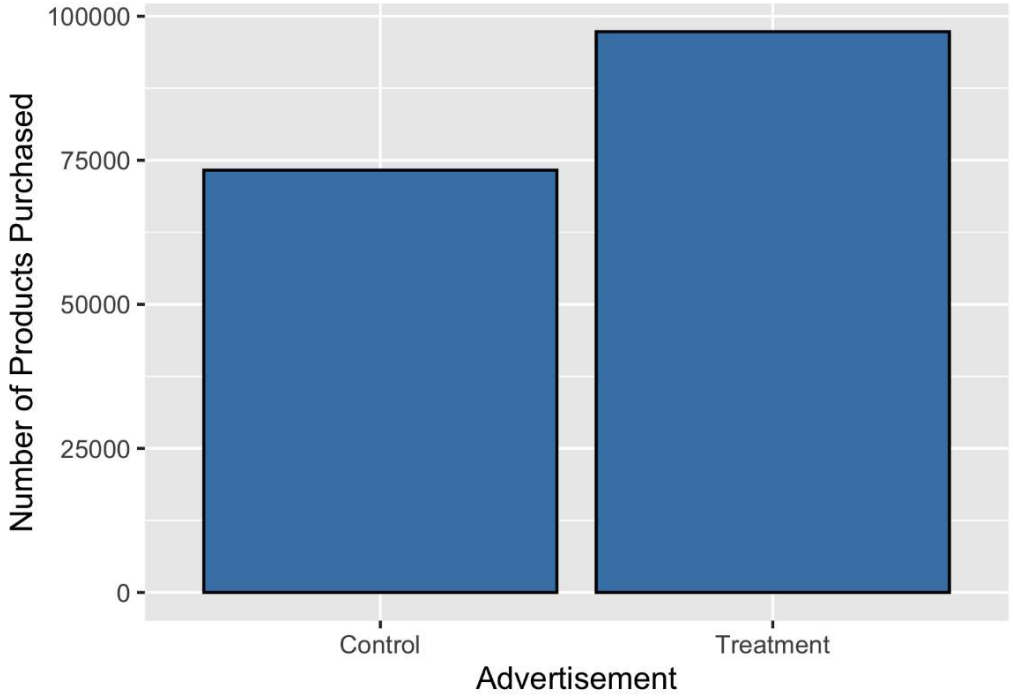
- Commonly used to describe hypothesis test results



# Advertising as a treatment

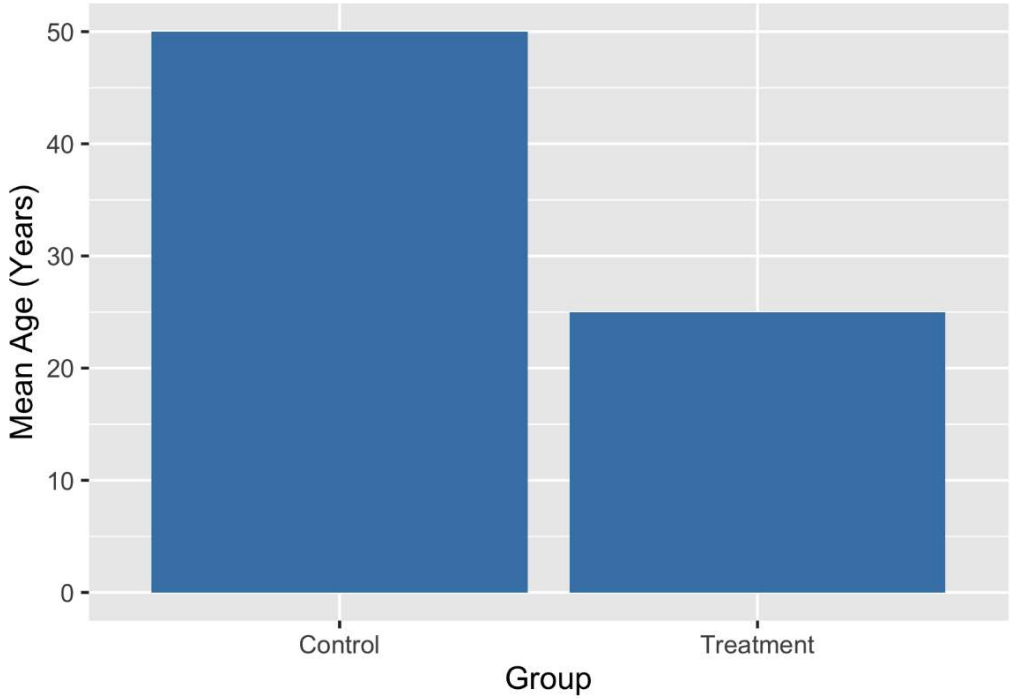
*What is the effect of an advertisement on the number of products purchased?*

- Treatment: advertisement
- Response: number of products purchased

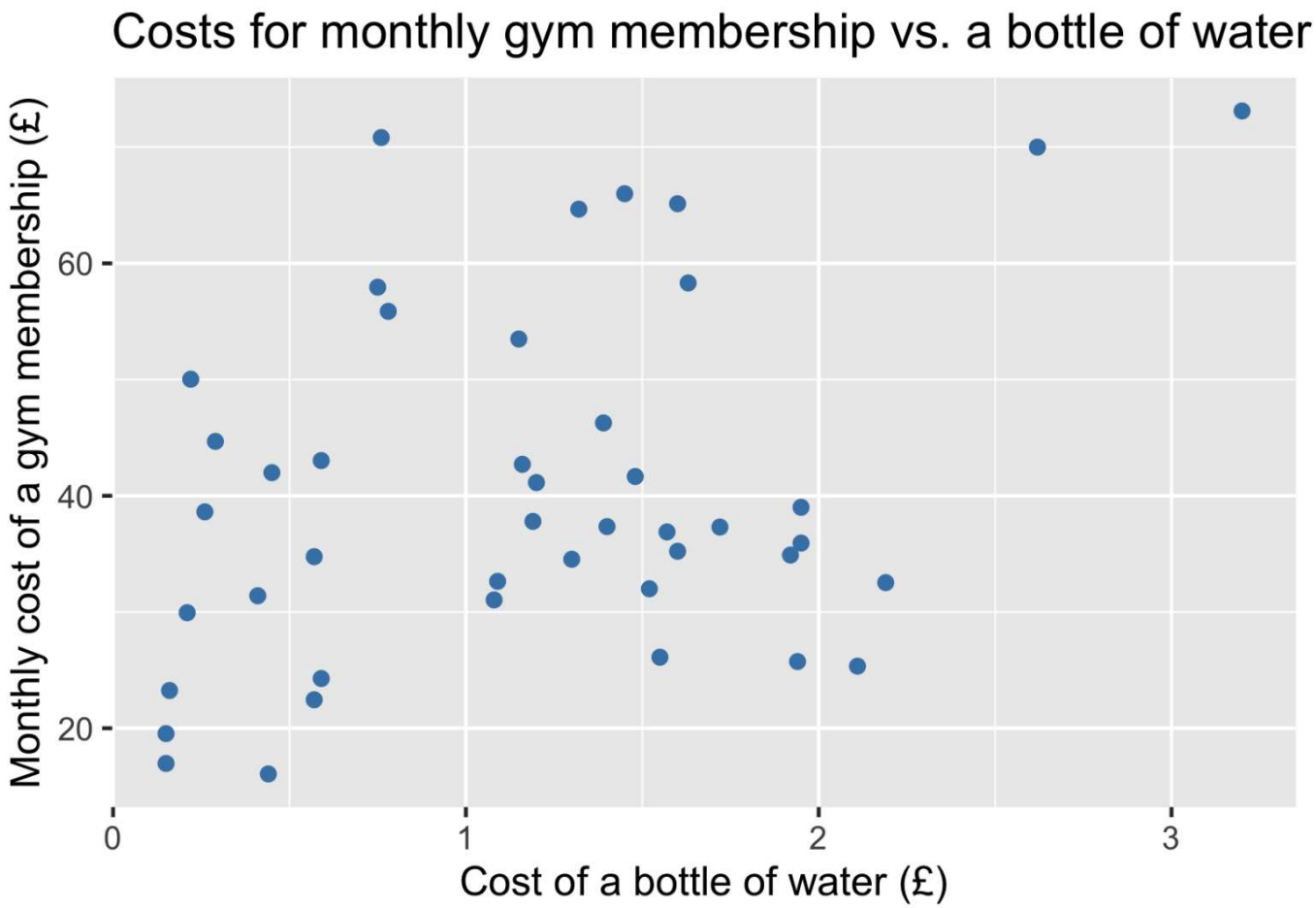


# Controlled experiments

- Participants are assigned to *either* the treatment group or the control group
  - ***Treatment group*** sees the advertisement
  - ***Control group*** does not see the advertisement
- Groups should be comparable to avoid introducing *bias*
- If groups are not comparable, this could lead to drawing incorrect conclusions

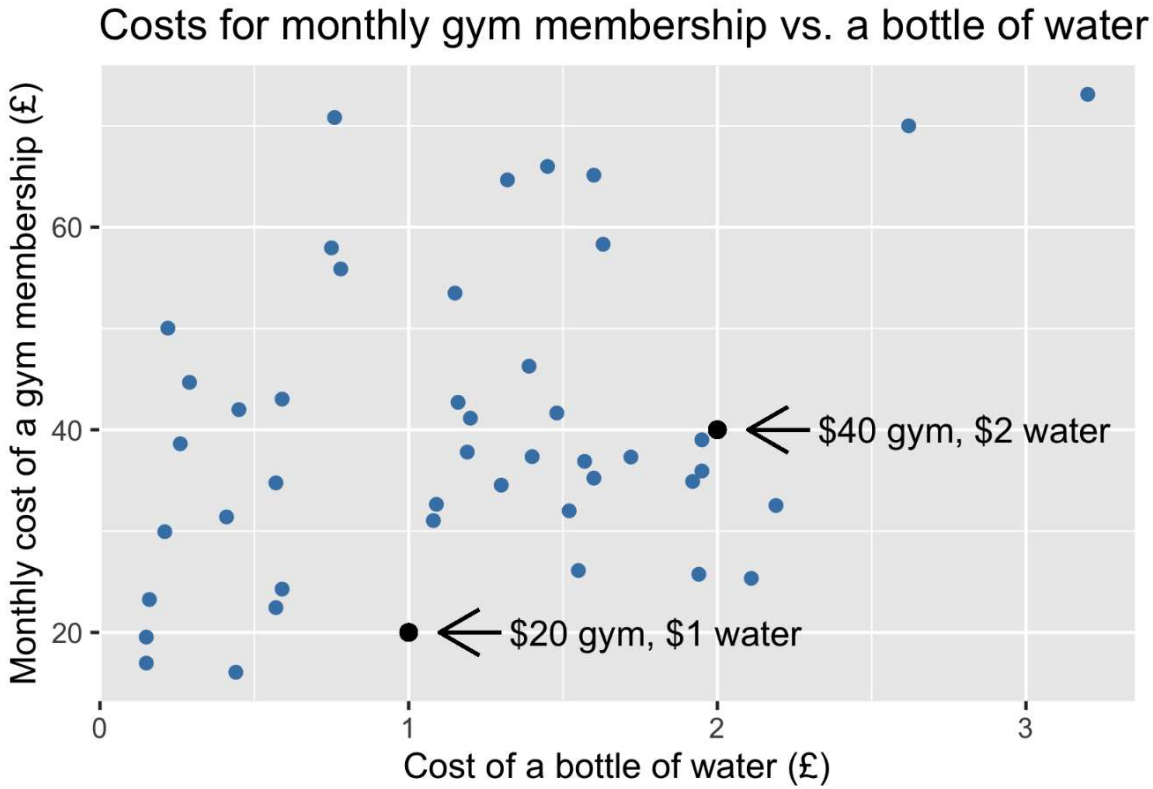


# Relationships between two variables



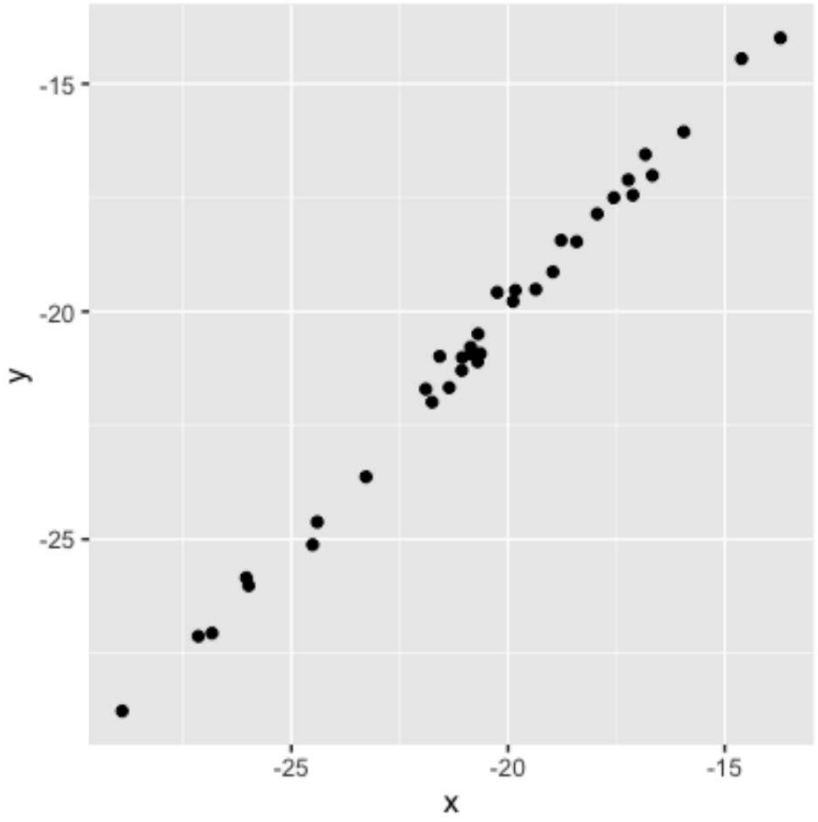
# Linear relationships

- Linear = proportionate changes between dependent and independent variables



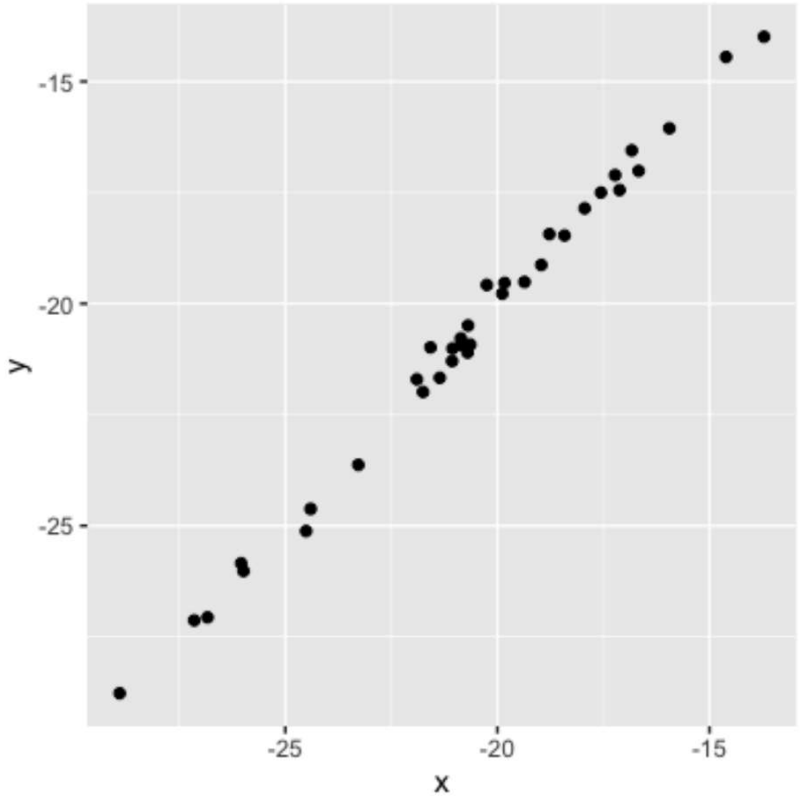
# Values = strength of the relationship

## 0.99 (very strong relationship)

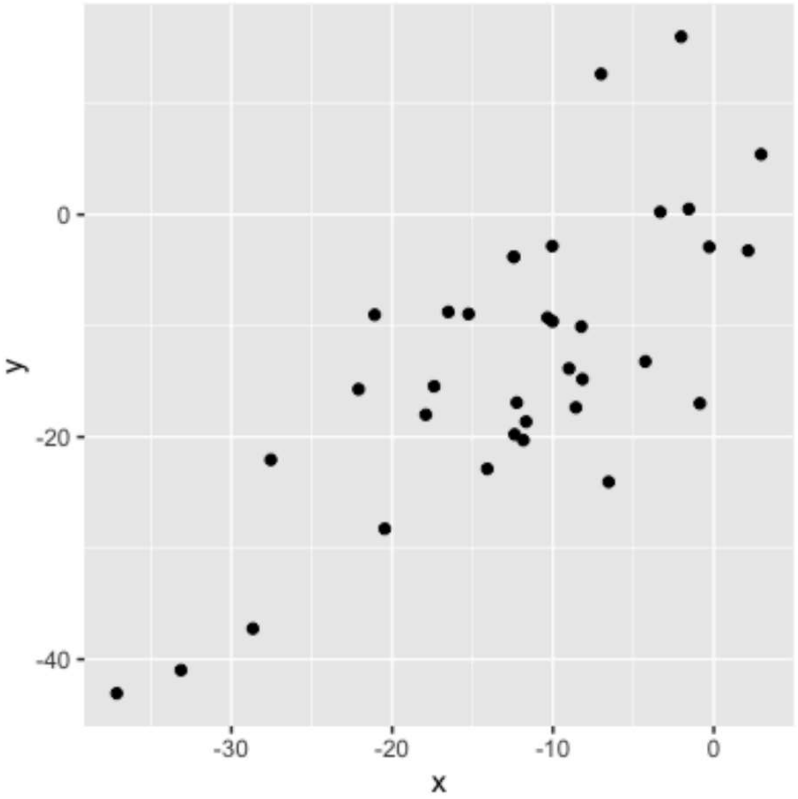


# Values = strength of the relationship

## 0.99 (very strong relationship)

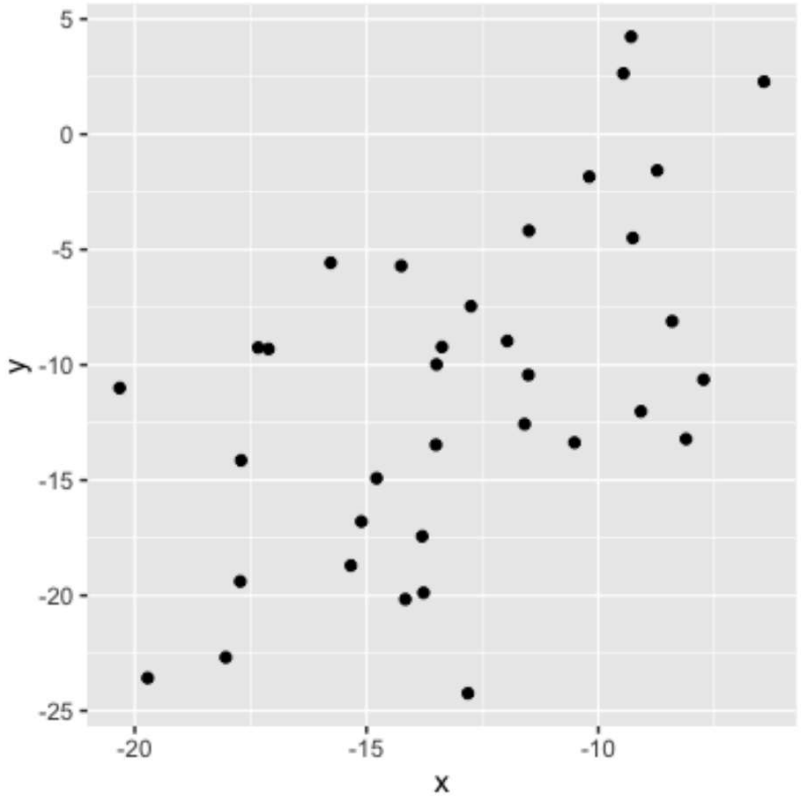



## 0.75 (strong relationship)



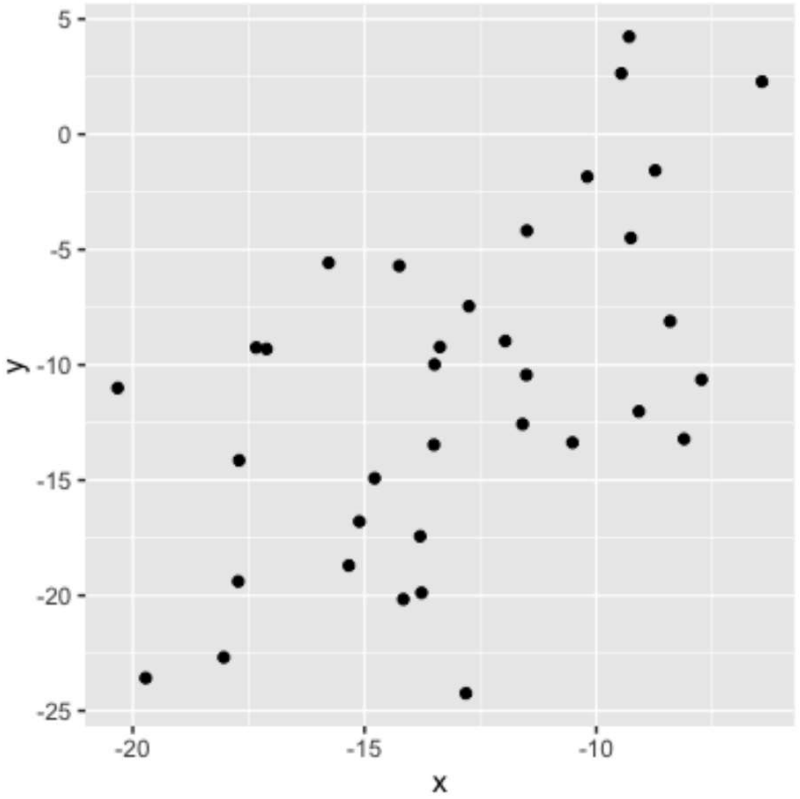
# Values = strength of the relationship

## 0.56 (moderate relationship)

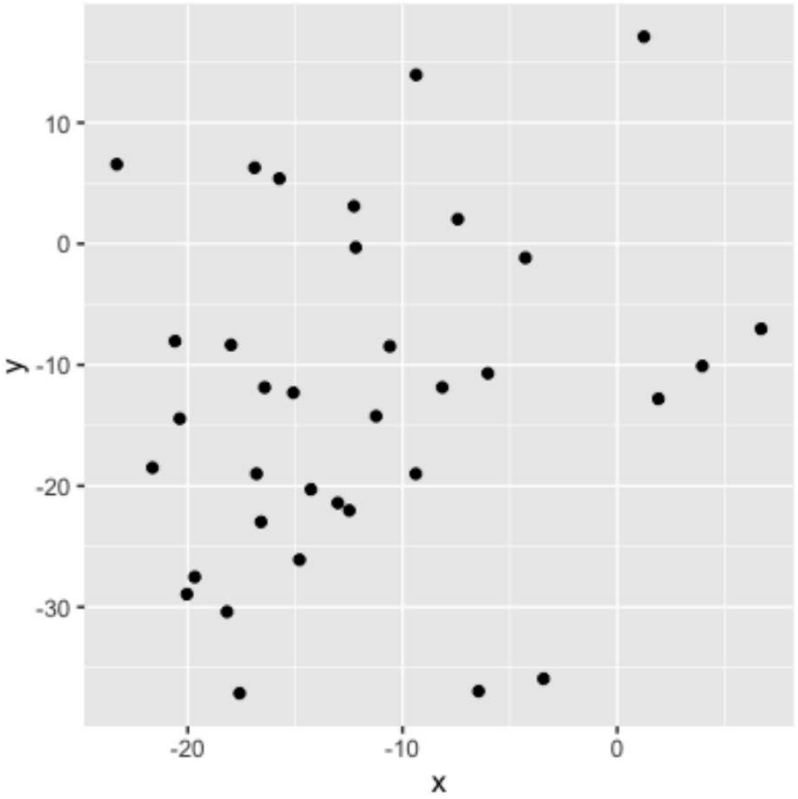


# Values = strength of the relationship

## 0.56 (moderate relationship)

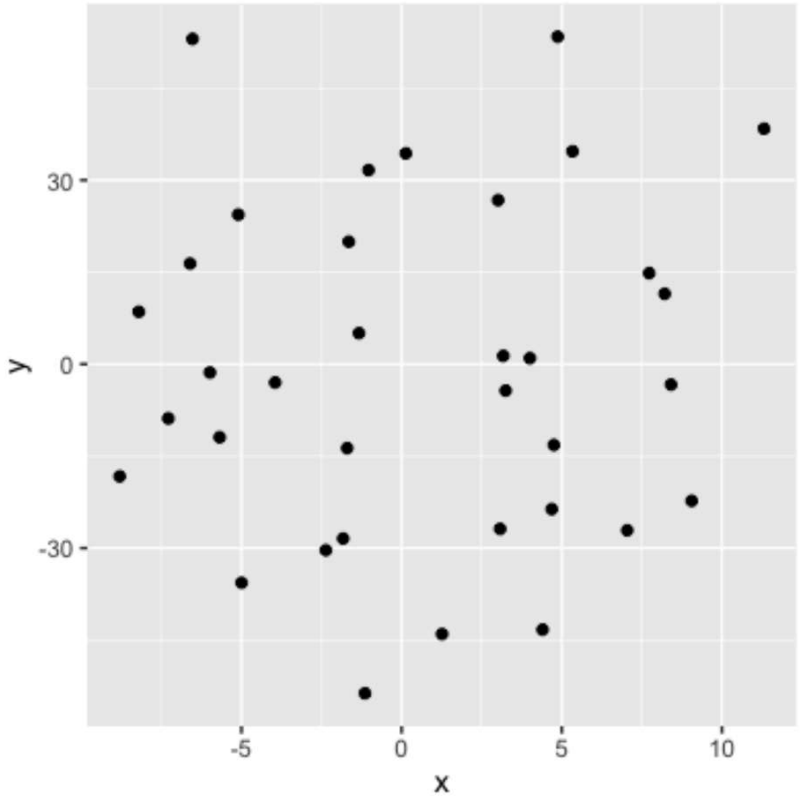



## 0.21 (weak relationship)



# Values = strength of the relationship

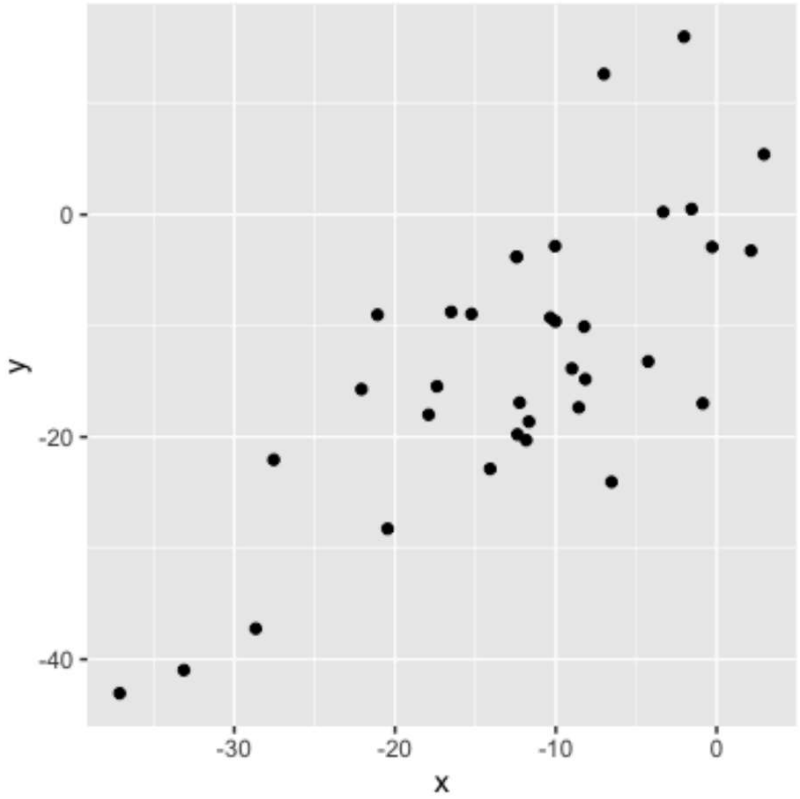
## 0.04 (no relationship)



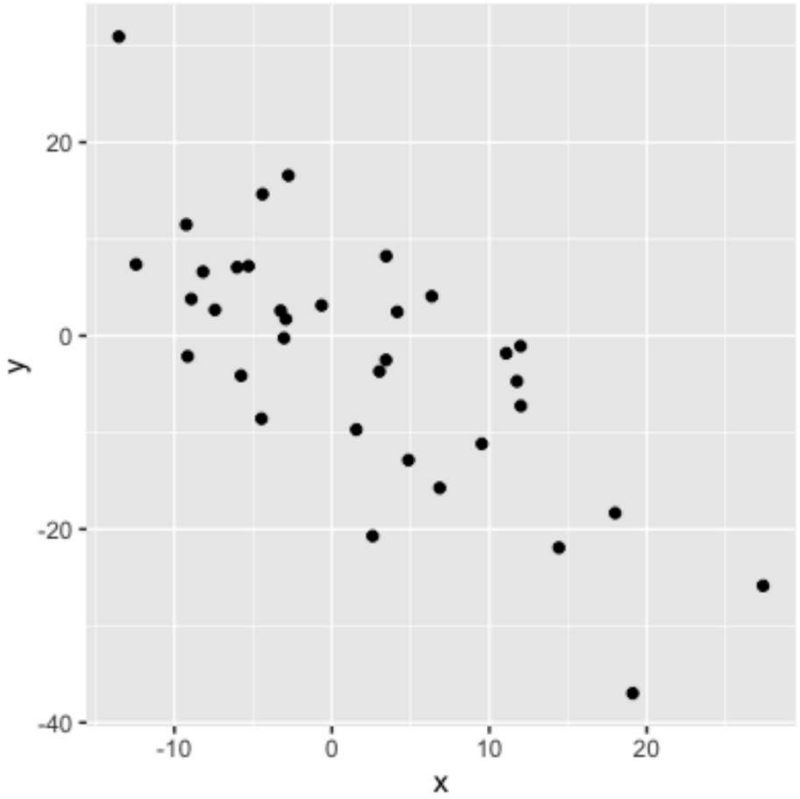

- Knowing the value of **x** doesn't tell us anything about **y**

# Sign = direction

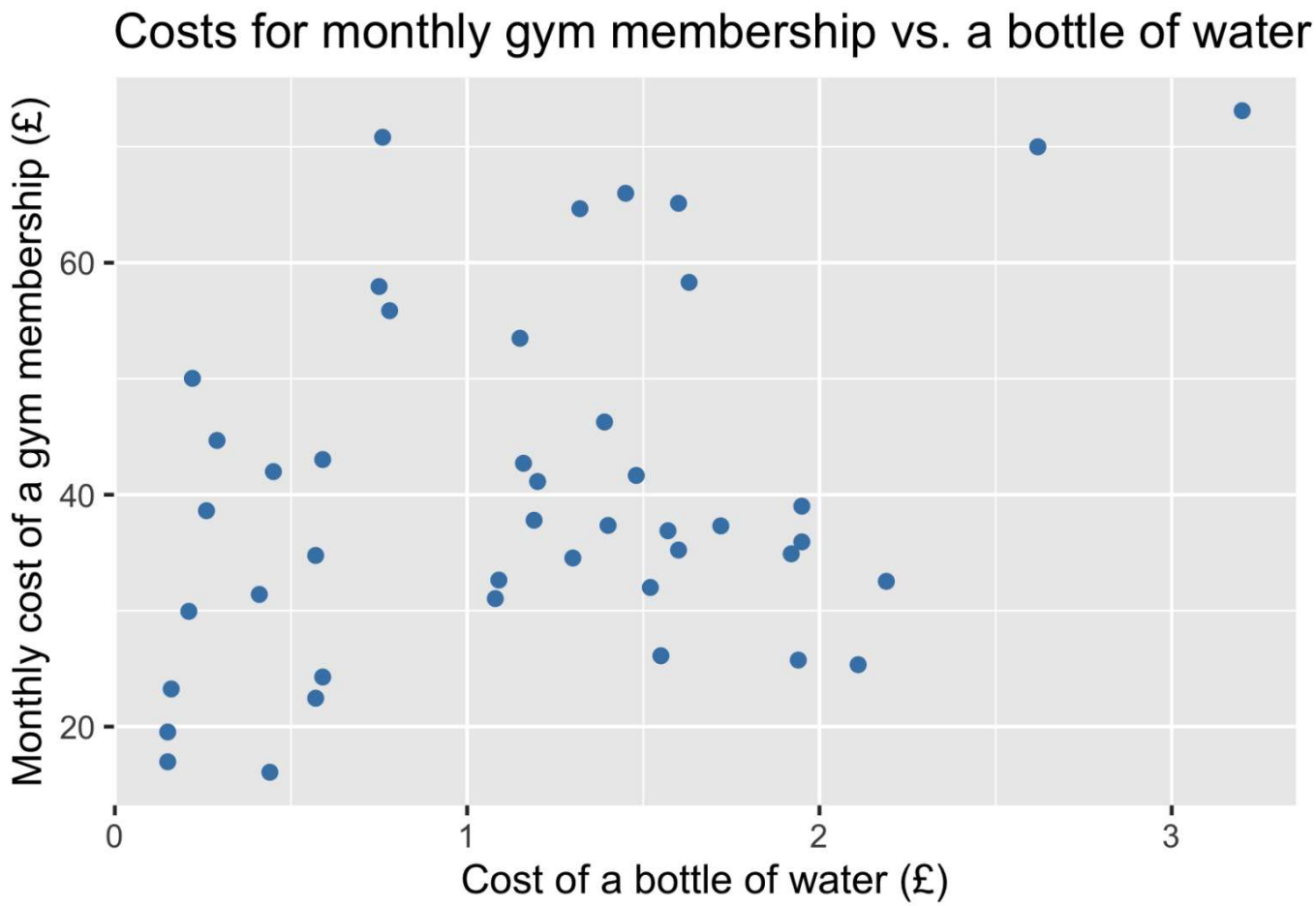
## 0.75: as x increases, y increases



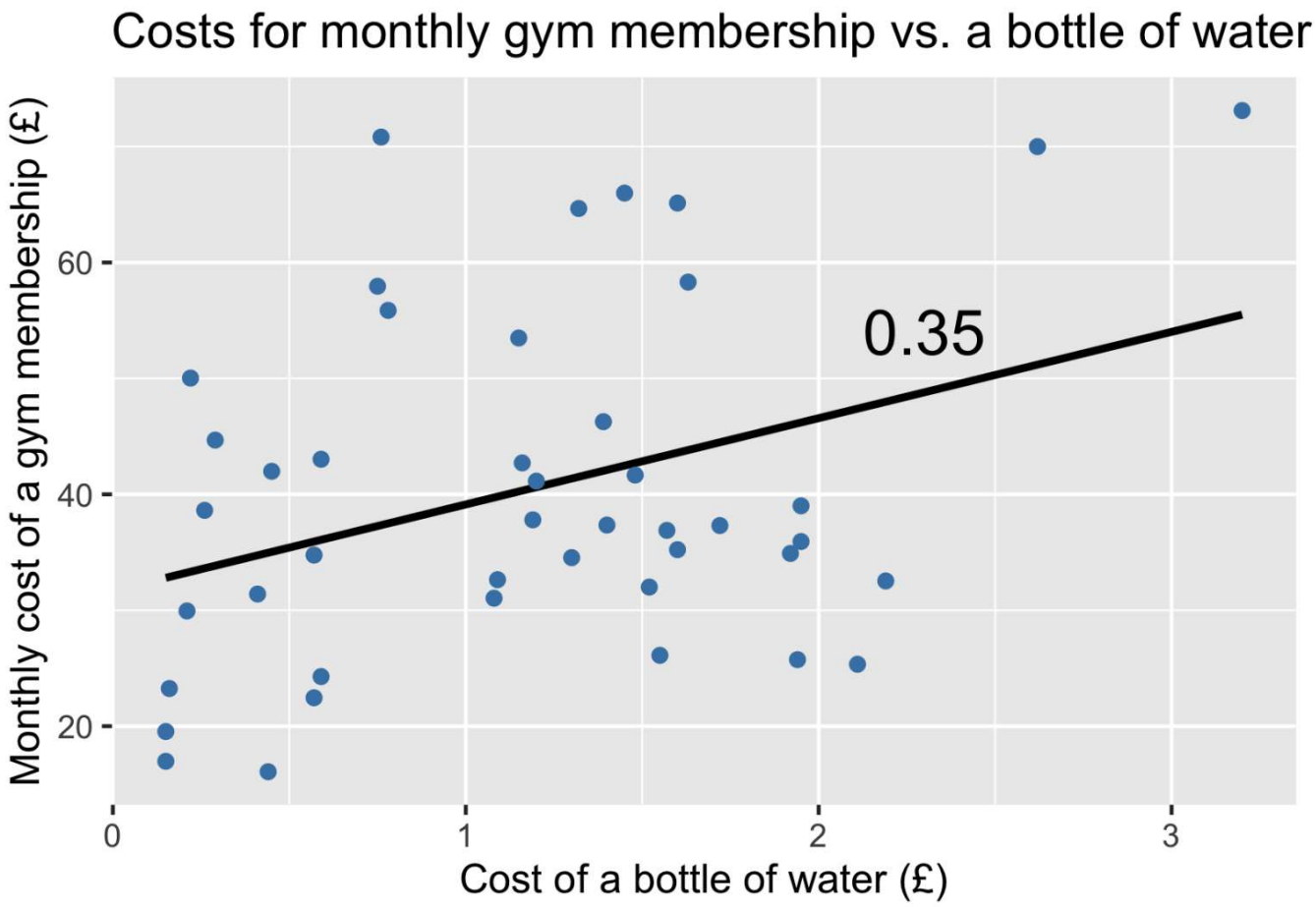

## -0.75: as x increases, y decreases



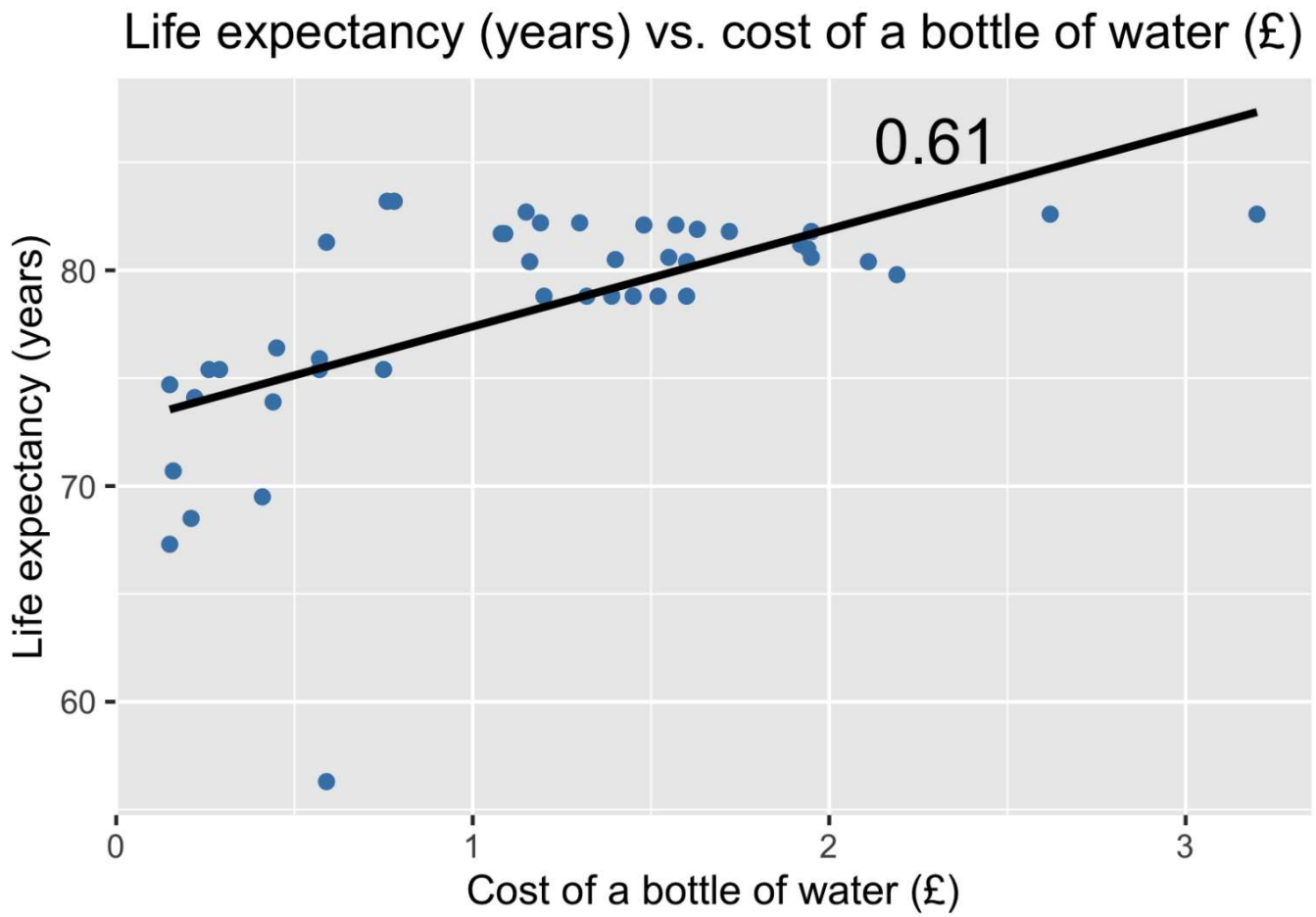
# Gym costs vs. water costs



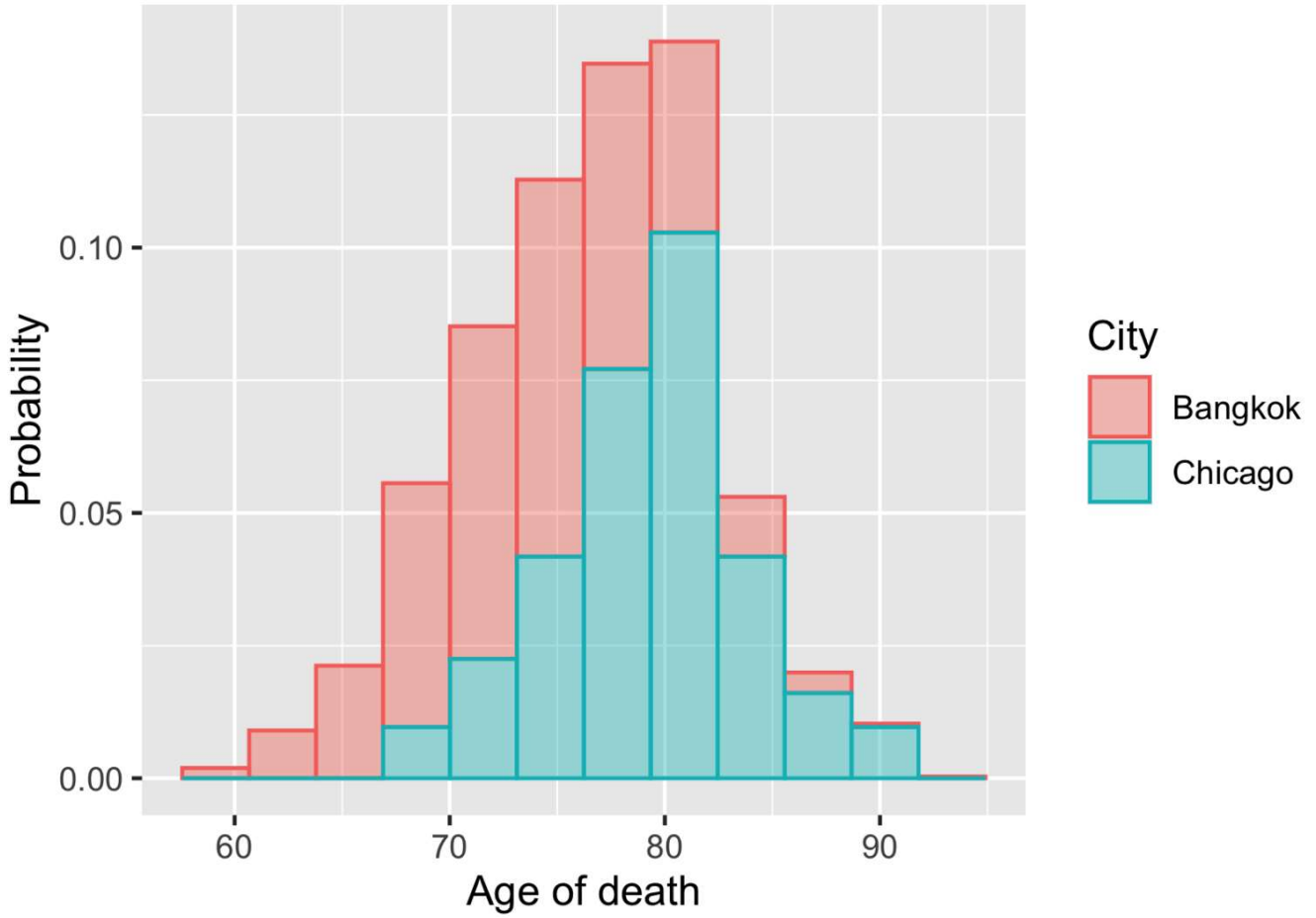
# Adding a trendline



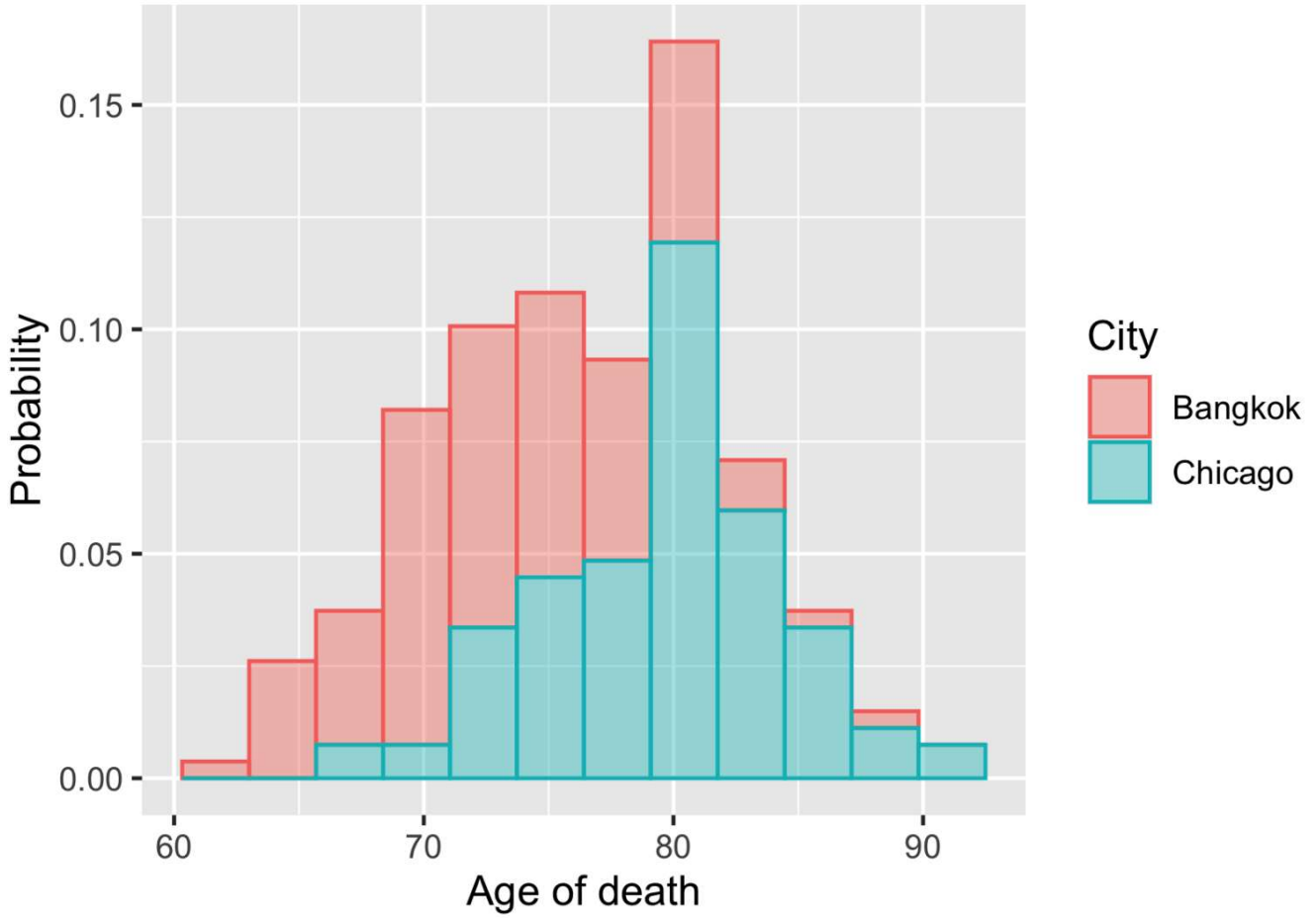
# Life expectancy vs. cost of a bottle of water



# Sampling distribution



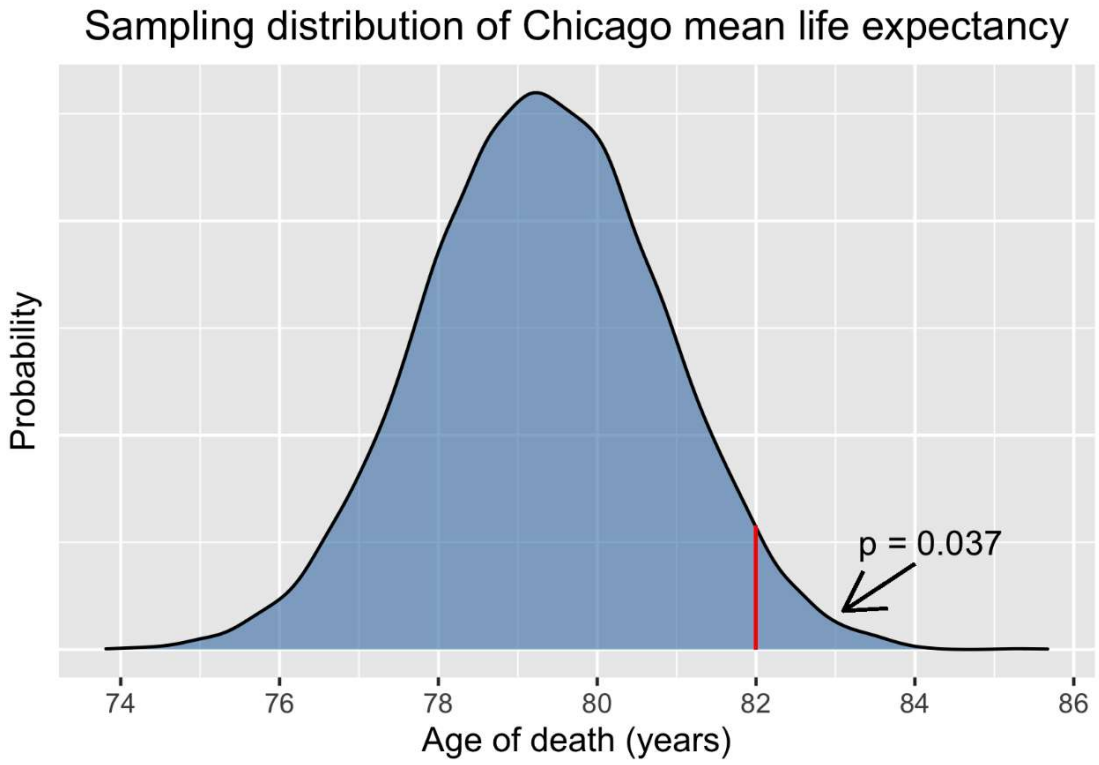
# Different samples



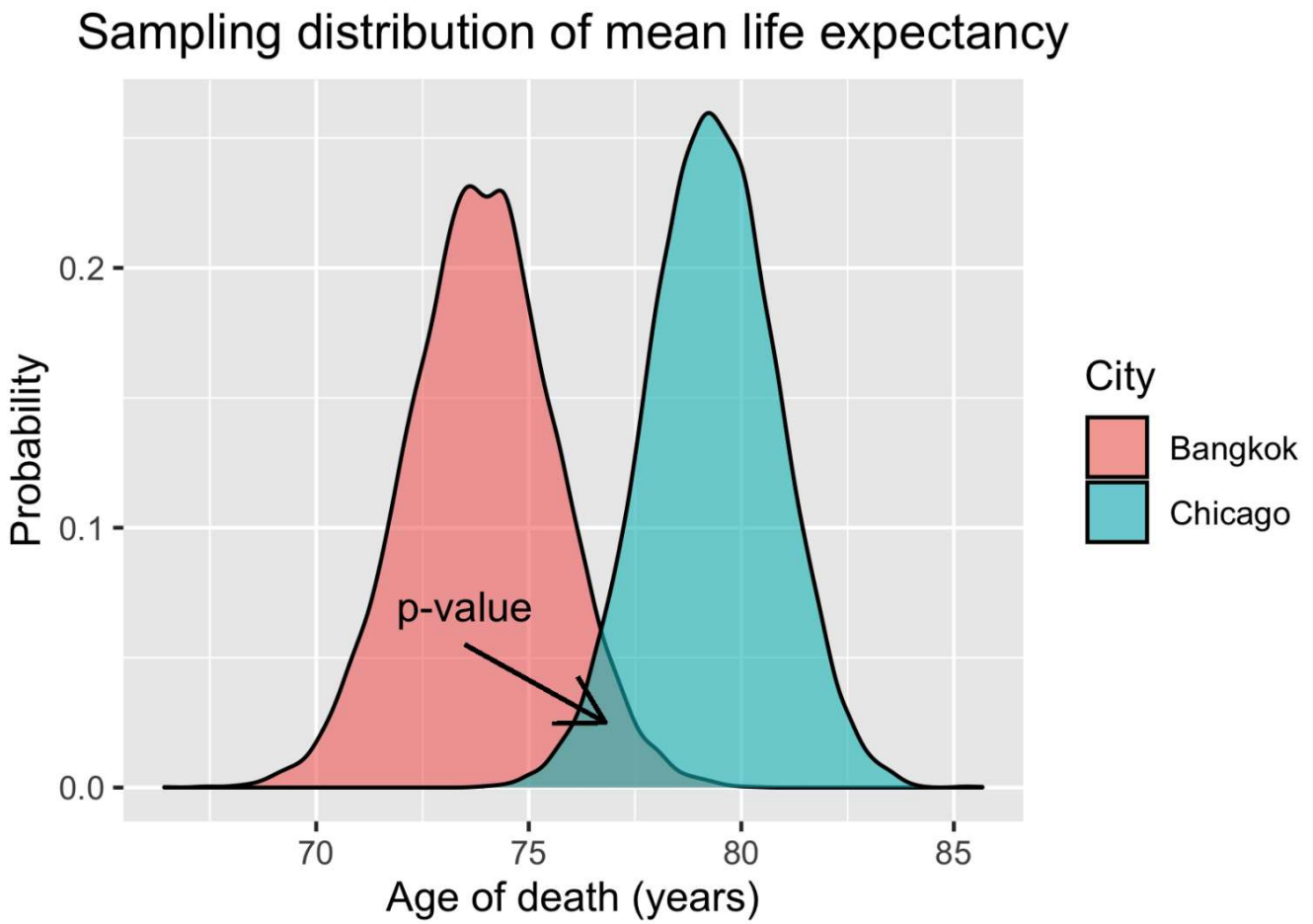
# Sampling distribution of mean life expectancy

# p-value

- $p$
  - Probability of achieving this result, assuming the null hypothesis is true



# p-value


Sampling distribution of mean life expectancy
Probability
0.2
0.1
0.0
p-value
70
75
80
85
Age of death (years)
City
Bangkok
Chicago

# Significance level ($\alpha$)

- To reduce the risk of drawing a false conclusion:
  - Set a probability threshold for rejecting the null hypothesis
- Known as $\alpha$ or *significance level*
- Decided before data collection to minimize bias:
  - Otherwise they could choose a different $\alpha$ to serve their interests
- A typical threshold is 0.05
  - 5% chance of wrongly concluding that Chicago residents live longer than Bangkok residents
- If $p \leq \alpha$, reject the null hypothesis
- These results are said to be *statistically significant*

# Type I/II error

| | Null hypothesis is TRUE | Null hypothesis is FALSE |
|---|---|---|
| Reject null hypothesis | Type I Error | |
| Accept null hypotheis | | |

# Type I/II error

| | Null hypothesis is TRUE | Null hypothesis is FALSE |
|---|---|---|
| Reject null hypothesis | Type I Error | |
| Accept null hypotheis | | Type II Error |

# Type I/II error

| | Null hypothesis is TRUE | Null hypothesis is FALSE |
|---|---|---|
| Reject null hypothesis | Type I Error | |
| Accept null hypotheis | Correct conclusion | Type II Error |

# Type I/II error

| | Null hypothesis is TRUE | Null hypothesis is FALSE |
|---|---|---|
| Reject null hypothesis | Type I Error | Correct conclusion |
| Accept null hypotheis | Correct conclusion | Type II Error |

# Drawing a conclusion