EBU5601
Data Design

Correlation

Dr Chao Shu
School of Electronic Engineering and Computer Science
Queen Mary University of London
Dec. 2024

Learning Outcomes

- The main outcomes are:
  - [**LO8.1**] Correctly interpret correlation coefficients
  - [**LO8.2**] Analyse the relationship between two variables using the appropriate correlation coefficient and Python codes
  - [**LO8.3**] Understand the difference between Pearson's correlation coefficient and Spearman's correlation coefficient
  - [**LO8.4**] Apply t-tests to test for the significance of correlation coefficients
  - [**LO8.5**] Apply permutation tests to test for the significance of correlation coefficients
  - [**LO8.6**] Understand correlation does not imply causation

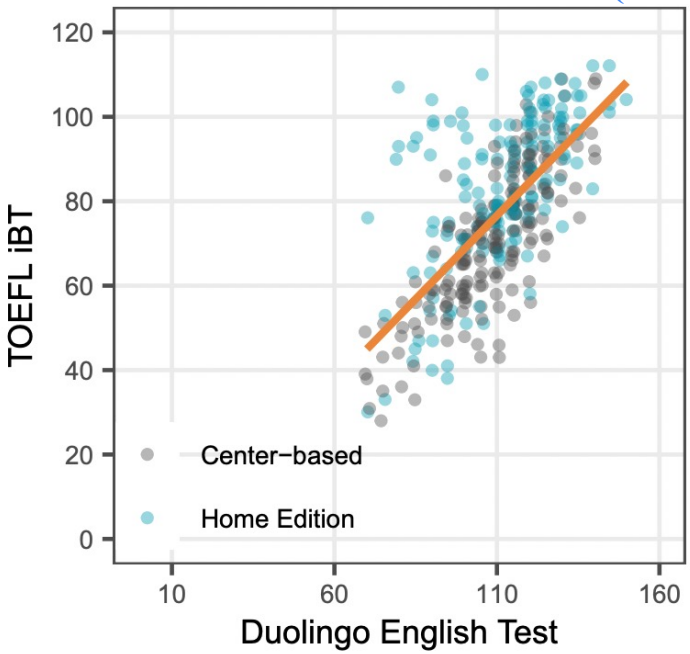
INTRODUCTION

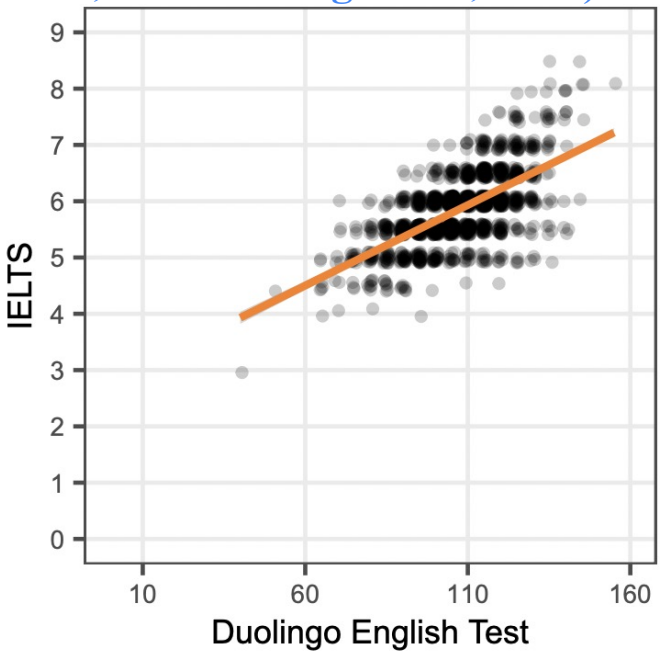
Introduction

- *Pearson's correlation coefficients* were estimated from official score data to evaluate the relationship between the DET and the TOEFL iBT and IELTS Academic.
- The correlation coefficients show *strong, positive relationships* of DET scores with TOEFL iBT scores and with IELTS scores

Correlations Between DET Scores and TOEFL / IELTS Scores (March 29, 2022 — August 05, 2022)

	TOEFL	IELTS
All candidates	.71 (328)	.65 (1,643)
Center-based	.82 (183)	—
Home Edition	.61 (145)	—



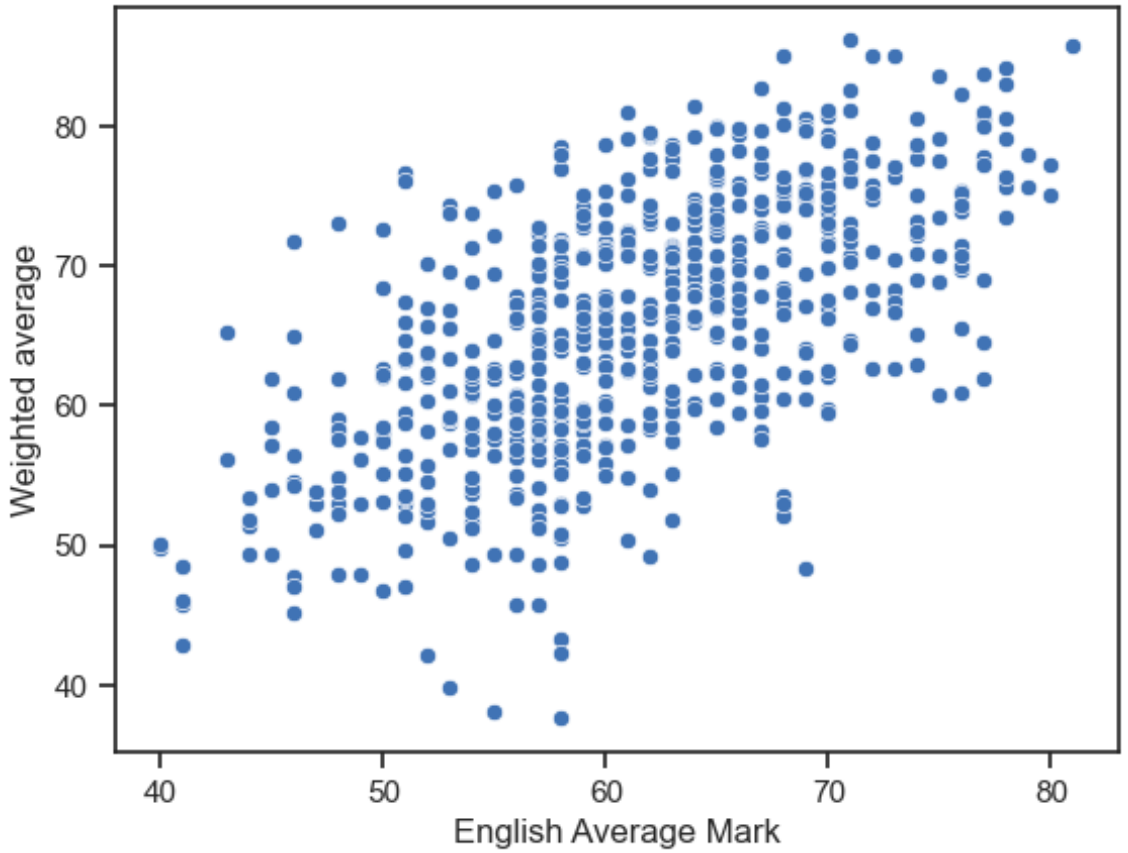


Source: https://duolingo-testcenter.s3.amazonaws.com/media/resources/technical_manual.pdf

Introduction

- Relationships between two variables.
  - x = explanatory/independent variable
  - y = response/dependent variable
  - Scatter Plot

- Student **English Marks** vs. **Weighted Average**
  - data from 2015 – 2018 Cohorts of one engineering programme in JP
  - x = English average marks
  - y = Overall weighted Average (determine the degree classification)

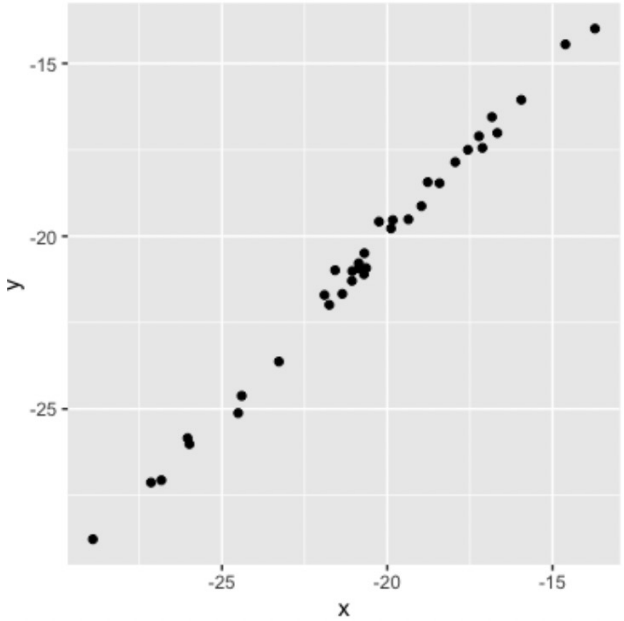


COMMON CORRELATION COEFFICIENTS

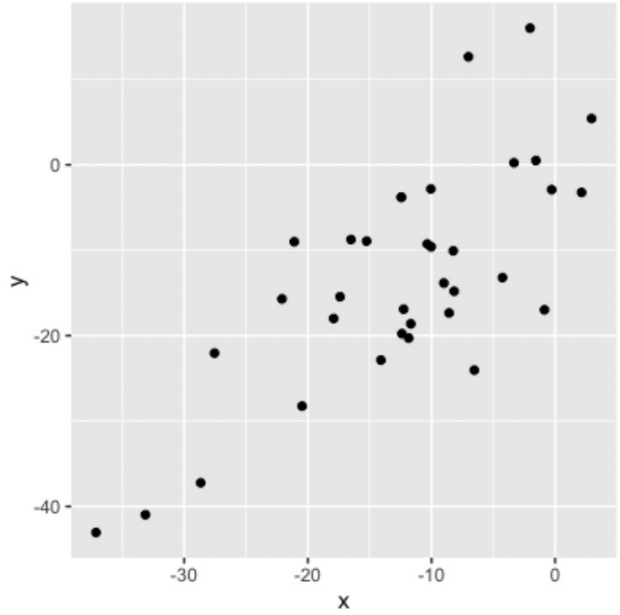
Pearson Correlation Coefficient

- the **Pearson correlation coefficient** is a correlation coefficient that measures linear correlation between two sets of data
  - Number between -1 and 1
  - Magnitude corresponds to strength of relationship
  - Sign (+ or -) corresponds to direction of relationship

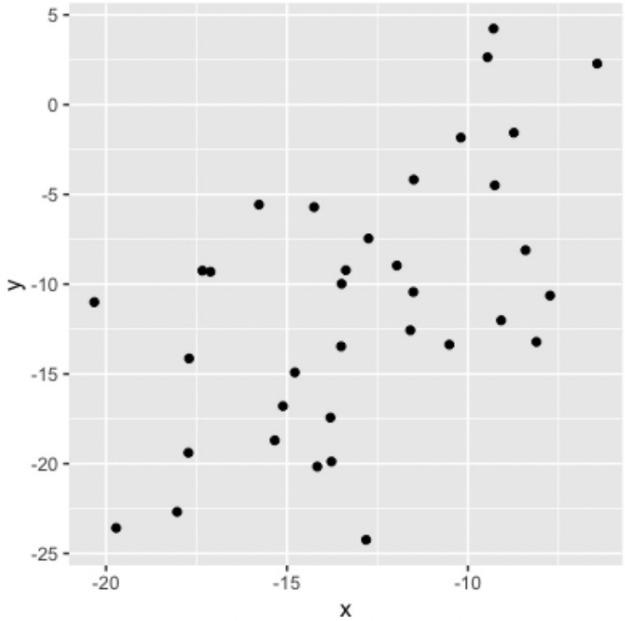
$r = 0.99$ (very strong)



$r = 0.75$ (strong)



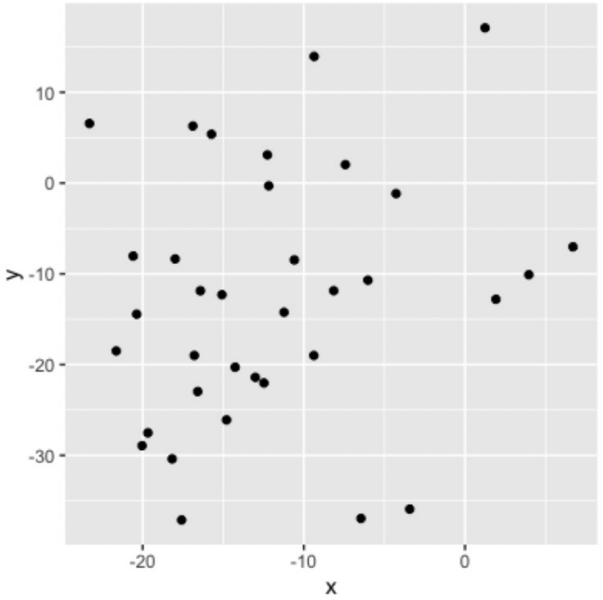
$r = 0.56$ (moderate)



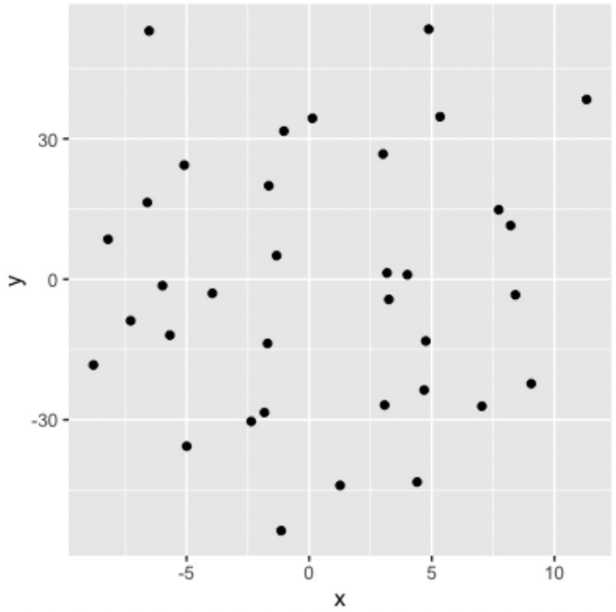
Pearson Correlation Coefficient

- the **Pearson correlation coefficient** is a correlation coefficient that measures linear correlation between two sets of data
  - Number between -1 and 1
  - Magnitude corresponds to strength of relationship
  - Sign (+ or -) corresponds to direction of relationship

$r = 0.21$ (weak)



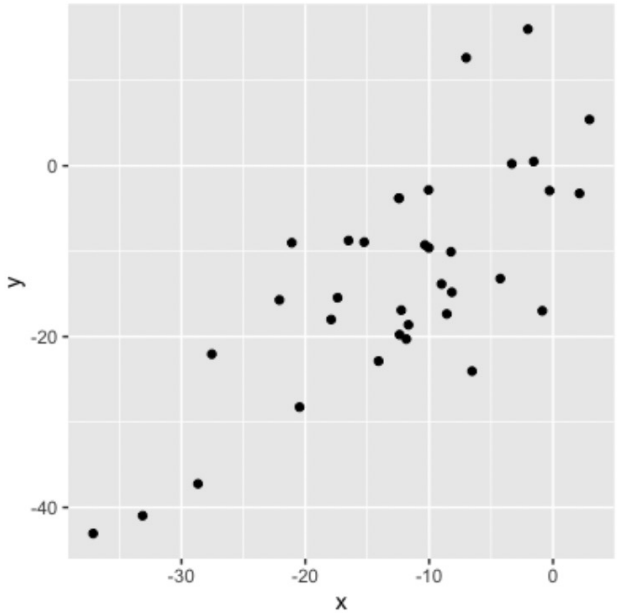
$r = 0.04$ (no relation)



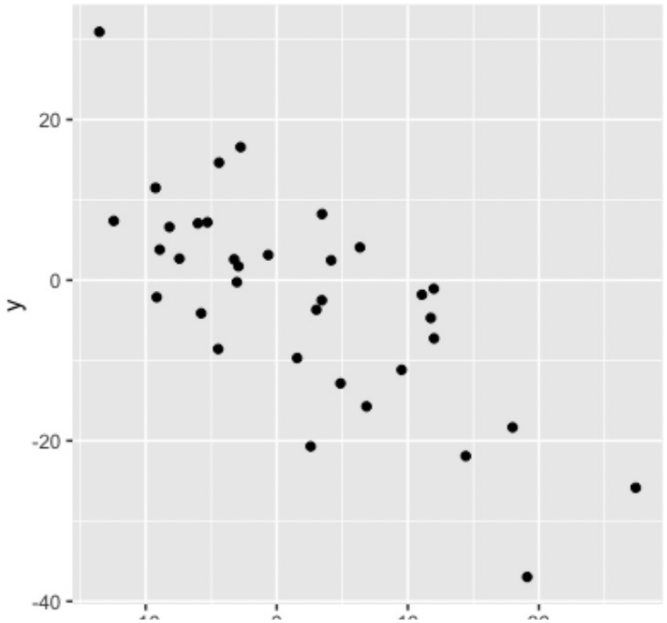
Pearson Correlation Coefficient

- the **Pearson correlation coefficient** is a correlation coefficient that measures linear correlation between two sets of data
  - Number between -1 and 1
  - Magnitude corresponds to strength of relationship
  - Sign (+ or -) corresponds to direction of relationship

$r = 0.75$ (positive strong)



$r = -0.75$ (negative strong)



Pearson Correlation Coefficient

- the **Pearson correlation coefficient** is a correlation coefficient that measures linear correlation between two sets of data

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Using pandas

```
1  stu_english_final_df['English Average Mark'].corr(stu_english_final_df['Weighted average'], method='pearson')
✓ 0.0s
0.630208873533884
```

- Using scipy.stats

```
1  stats.pearsonr(stu_english_final_df['English Average Mark'], stu_english_final_df['Weighted average'], alternative='greater')
✓ 0.0s
PearsonRResult(statistic=0.6302088735338832, pvalue=3.3553095102857985e-75)
```

Spearman's Rank Correlation Coefficient

- Spearman's rank correlation coefficient is a nonparametric measure of rank correlation, which is used to assess how well the relationship between two variables can be described using a monotonic function.
  - The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables;
  - Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships.

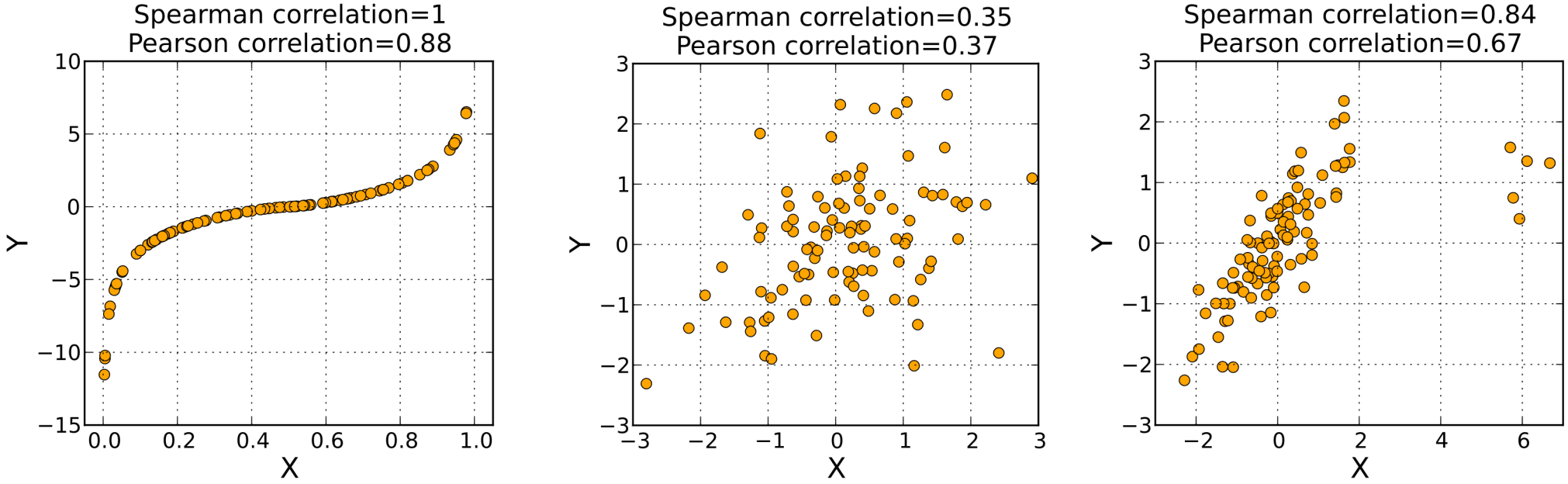


Image Source: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

TEST FOR SIGNIFICANCE

t-Test for Correlation Coefficient

- t-Test can be used to test whether an observed value of ρ (population correlation coefficient) is significantly different from zero (the observed value r will always between $-1 \leq r \leq 1$).
- The t-value under the null hypothesis of statistical independence ($\rho = 0$) can be calculated as:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where n is the sample size and $n - 2$ is the degree of freedom.

t-Test for Correlation Coefficient

Case Study

- Consider the following data from [1], which studied the relationship between free proline (an amino acid) and total collagen (a protein often found in connective tissue) in unhealthy human livers.

```
# total collagen (mg/g dry weight of liver)
x = np.array([7.1, 7.1, 7.2, 8.3, 9.4, 10.5, 11.4])
# free proline (μ mole/g dry weight of liver)
y = np.array([2.8, 2.9, 2.8, 2.6, 3.5, 4.6, 5.0])
```

[1] Kershenobich, D., Fierro, F. J., & Rojkind, M. (1970). The relationship between the free pool of proline and collagen content in human liver cirrhosis. The Journal of Clinical Investigation, 49(12), 2246-2249.

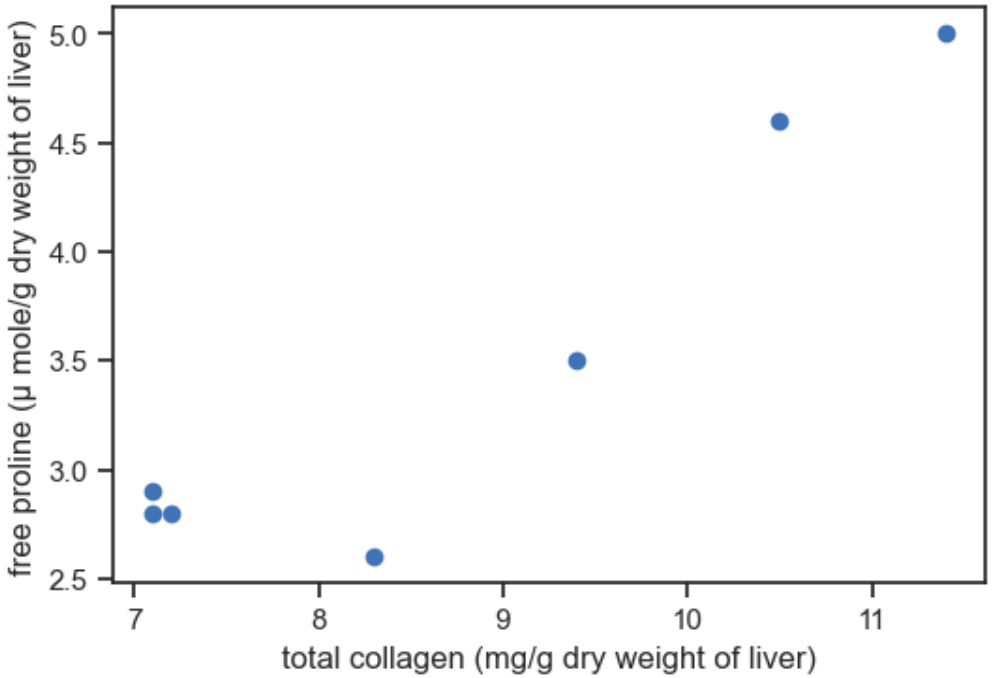
t-Test for Correlation Coefficient

Case Study

- The scatter plot shows that there is a great chance that the two variables are correlated.
- Let's use Scipy to calculate the Spearman's correlation coefficient. Since we expect a positive correlation between x (total collagen) and y (free proline), we can set up a one-sample one-tailed t-test.

$$H_0: \rho \leq 0$$

$$H_1: \rho > 0$$



t-Test for Correlation Coefficient

Case Study

- According to the result, the Spearman's correlation coefficient (statistic is 0.7, which suggests a strong positive correlation. The p-value (pvalue) is about 4%, which is also below the 5% alpha level.
- So, we can reject the null in favour of the alternative, i.e., the positive correlation between total collagen and free proline is statistically significant.

```
# Calculate Spearman's correlation coefficient between x and y,
# as well as the p-value of the correlation coefficient under a one-sample one-tailed t-test
res = stats.spearmanr(x, y, alternative='greater')
res
```

```
SignificanceResult(statistic=0.7000000000000001, pvalue=0.03995834515444955)
```

t-Test for Correlation Coefficient

Case Study

- We can perform the t-test by ourselves to verify the p-value result.
- The t-value under the null hypothesis of statistical independence ($\rho = 0$) can be calculated as:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.7\sqrt{\frac{7-2}{1-0.7^2}} = 2.192$$

- By using scipy.stats.t.cdf(), we can find p-value = 0.03995
- Or, we can get the critical value for 5% significance level is 2.015.
- Because p-value = 0.03995 < 0.5 or t=2.192 > 2.015, we reject the null in favour of the alternative, i.e., there is a positive correlation between free proline (an amino acid) and total collagen (a protein often found in connective tissue) in unhealthy human livers.

PERMUTATION TESTS

Introduction

- **z/t-Tests**
  - **parametric tests** where null distributions are obtained from theoretical probability distributions.
  - rely on the assumption that the populations should follow the normal distribution
- **Permutation Tests**
  - also known as a randomization test or re-randomization test, is a **non-parametric** statistical test used to assess whether the observed differences between groups or the observed relationship between variables in a dataset are statistically significant.
  - does not rely on specific assumptions about the underlying population distribution. It is particularly useful when the assumptions of traditional parametric tests, such as the z/t-test, are not met, or when you want to perform hypothesis testing without making distributional assumptions.
  - involve two or more samples

Introduction

- **The Null Distribution in Permutation Tests**
  - The null hypothesis is that all samples come from the same distribution.
  - Under the null hypothesis, the distribution of the test statistic is obtained by calculating all possible values of the test statistic under possible rearrangements of the observed data.
  - For example, when we want to test the mean difference between two populations, the permutation test simply generates the distribution of mean differences under the assumption that the two groups are not distinct in terms of the measured variable.
  - From this, one then uses the observed statistic to see to what extent this statistic is special, i.e., the likelihood of observing such a value (or larger) if the treatment labels had simply been randomized after treatment
- **Permutation Test Procedure**
  - The overall steps of a permutation test are the same as a z/t-test (**step 1** to **step 6**). The differences are in **step 4** (Generate the null distribution) and **step 5** (Calculate p-value).

Permutation Procedure

The permutation procedure in **step 4** (Generate the null distribution) and **step 5** (Calculate p-value) is listed as follows:

- **Step 4d**: Generate the null distribution
  - 1. Combine the results from the different groups/samples into a single dataset.
  - 2. Shuffle the combined data and then randomly draw (without replacement) a resample of the same size as group A.
  - 3. From the remaining data, randomly draw (without replacement) a resample of the same size as group B.
  - 4. Do the same for groups C, D, and so on. You have now collected one set of resamples that mirror the sizes of the original samples.
  - 5. Calculate the statistic of interest for the permuted samples; this constitutes one permutation iteration.
  - 6. Repeat the previous steps R times to yield a permutation distribution of the test statistic (null distribution).

Permutation Procedure

- **Step 5d**: Calculate p-value
  - Compare the observed test statistic to the distribution of test statistics obtained from the permutations (step 4).
  - Calculate the proportion of permuted test statistics that are as extreme as or more extreme than the observed test statistic. This proportion is the p-value.
- **Step 6**: Draw conclusions
  - Same as other tests, If the p-value is smaller than a predetermined significance level (α), you reject the null hypothesis infavourr of the alternative hypothesis.
  - Otherwise, you fail to reject the null hypothesis.

Permutation Tests

Example:

- Suppose we have two random samples drawn from two populations. The 4 values in red are drawn from one distribution, and the 5 values in blue from another. We'd like to test whether the means of the two populations are different.
  - The null hypothesis is that both groups of samples are drawn from the same distribution. $H_0: \mu_1 = \mu_2$ or $H_0: \mu_1 - \mu_2 = 0$
  - The alternative hypothesis is that the mean of the first distribution is higher than the mean of the second. $H_1: \mu_1 \neq \mu_2$ or $H_1: \mu_1 - \mu_2 \neq 0$

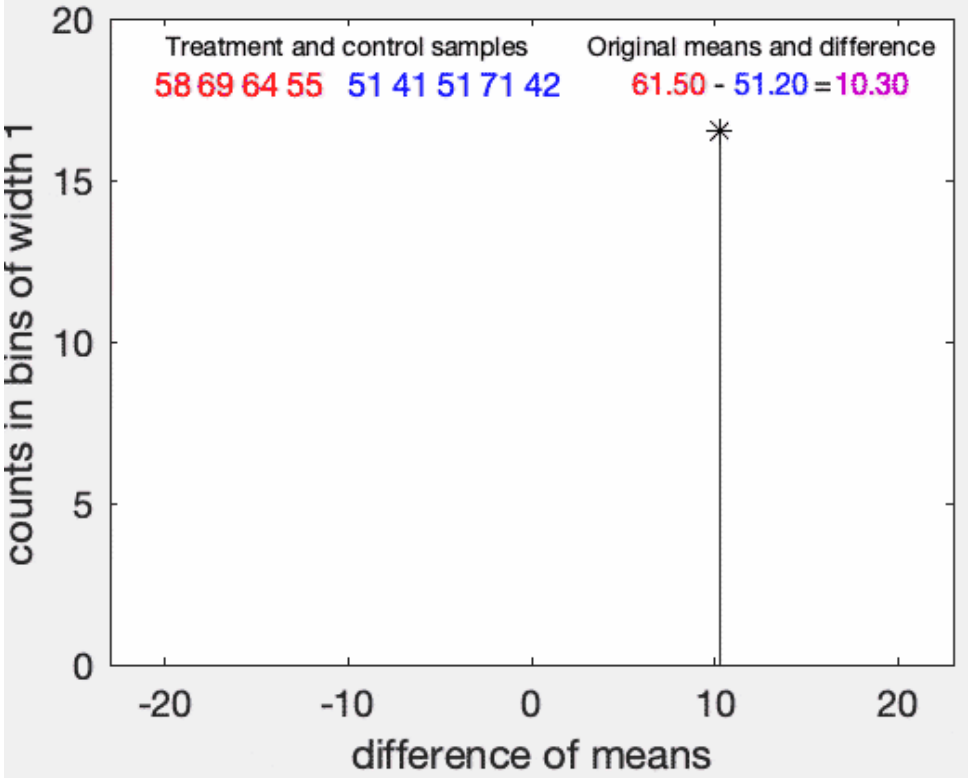
{58, 59, 64, 55}: $\bar{x}_1 = 61.5$

{51, 41, 51, 71, 42}: $\bar{x}_2 = 51.2$

Permutation Tests

Example:

- A permutation test is computed on sets of 4 and 5 random values.
  - There are 126 distinct ways to put 4 values into one group and 5 into another (9-choose-4 or 9-choose-5).
  - The p-value of the hypothesis is estimated as the proportion of permutations that give a difference as large or larger than the difference of means of the original samples.
  - In this example, we failed to reject the null at the p = 5% level.



Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 1**: Set up a hypothesis.

- Suppose we are interested to know whether male university students' average sleep hours per day is significantly different from the average sleep hours per day of the female students or not. We can set up the hypothesis test as follows.

$$H_0: \mu_f - \mu_m = 0$$
$$H_1: \mu_f - \mu_m \neq 0$$

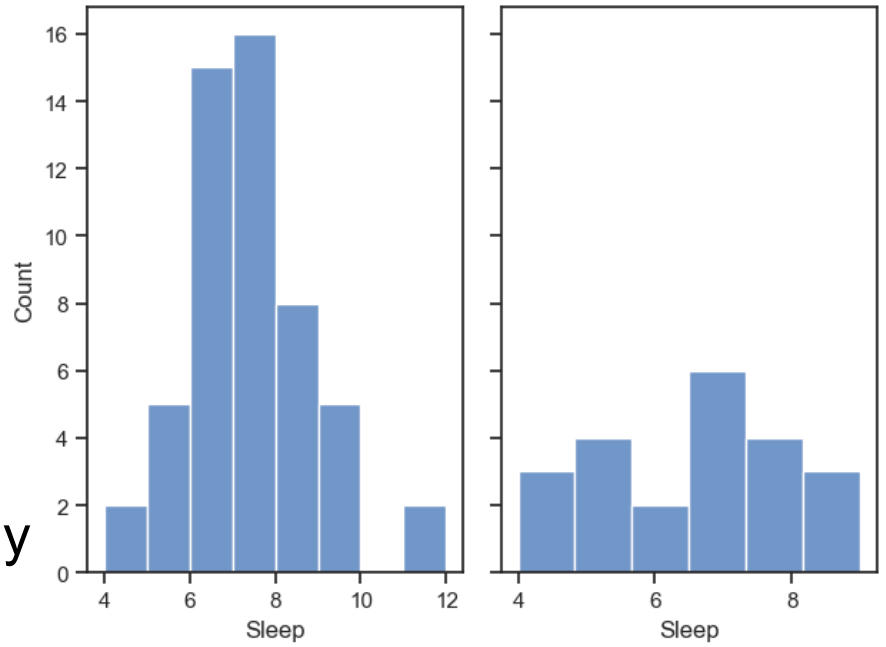
Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 2**: Collect a sample and find out the summary statistics of the sample
- We already have the data and the summary statistics.
  - Sample sizes $n_f = 53, n_m = 22$
  - Sample mean $\bar{x}_f = 7.019, \bar{x}_m = 6.523$
  - Sample std $s_f = 1.535, s_m = 1.523$
  - Sample mean difference

    $\bar{x}_f - \bar{x}_m = 7.019 - 6.523 = 0.496$
  - the permutation test doesn't rely no any assumption of the distributions of two sample data



Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 3**: Define a significance level

- We keep the significance level the same as the last demonstration

$$\alpha = 0.05$$

Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 4**: Generate the null distribution

```python
# Define the method to calculate the test statistic of interest
def diff_statistic(x, y):
    x_mean = np.mean(x)
    y_mean = np.mean(y)
    mean_diff = x_mean - y_mean
    return mean_diff

# Obtain the two original samples
female_sleep = stu_survey_sleep_df[stu_survey_sleep_df['Gender'] == 'female']['Sleep']
male_sleep = stu_survey_sleep_df[stu_survey_sleep_df['Gender'] == 'male']['Sleep']

print(female_sleep.shape, male_sleep.shape)
```

✓ 0.0s Python

```
(53,) (22,)
```

statistic: the mean difference between two original samples, the null_distribution is the sampling distribution of the mean difference under the null hypothesis generated by the permutation method

```python
# Perform a permutation test for the two groups
pm_res = stats.permutation_test((female_sleep, male_sleep), diff_statistic, permutation_type='independent', alternative='two-sided')
pm_res
```

✓ 0.0s Python

```
PermutationTestResult(statistic=0.4961406518010296, pvalue=0.2188, null_distribution=array([-0.01843911,  0.1745283 , -0.17924528, ..., -0.661
       -0.01843911,  0.68910806]))
```

Permutation Tests

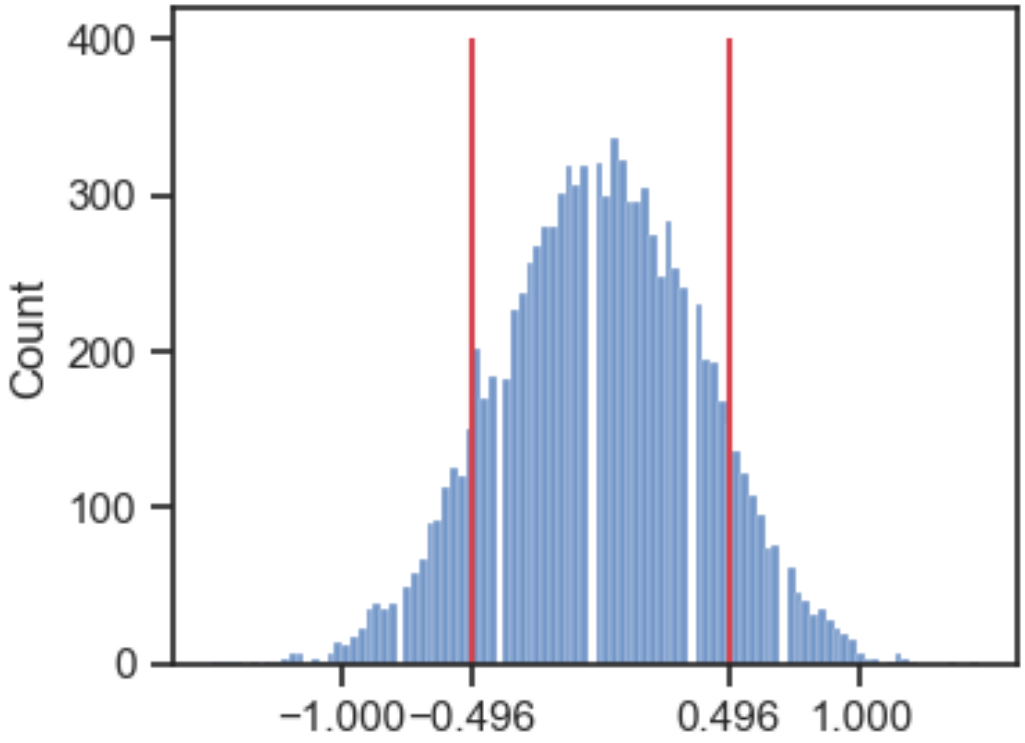
Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 4**: Generate the null distribution

```
PermutationTestResult(statistic=0.4961406518010296, pvalue=0.2188, null_distribution=array([-0.01843911,  0.1745283 , -0.17924528, ..., -0.661
       -0.01843911,  0.68910806]))
```

Visualise
null_distribution



Permutation Tests

Demo: Differences in Sleep Hours between Male and Female Students

Two-sample two-tailed test

- **Step 5**: Calculate the p-value (permutation test)
- Read the pvalue in the returned results

```
PermutationTestResult(statistic=0.4961406518010296, pvalue=0.2188, null_distribution=array([-0.01843911,  0.1745283 , -0.17924528, ..., -0.661
       -0.01843911,  0.68910806]))
```

- According to the definition.

```
  1  (pm_res.null_distribution >= pm_res.statistic).mean() + (pm_res.null_distribution <= -pm_res.statistic).mean()
✓ 0.0s
0.2208220822082208
```

- **Step 6**: Draw conclusions
- Because p-value > $\alpha = 0.05$, we fail to reject the null.

Permutation Tests for Correlation Coefficient

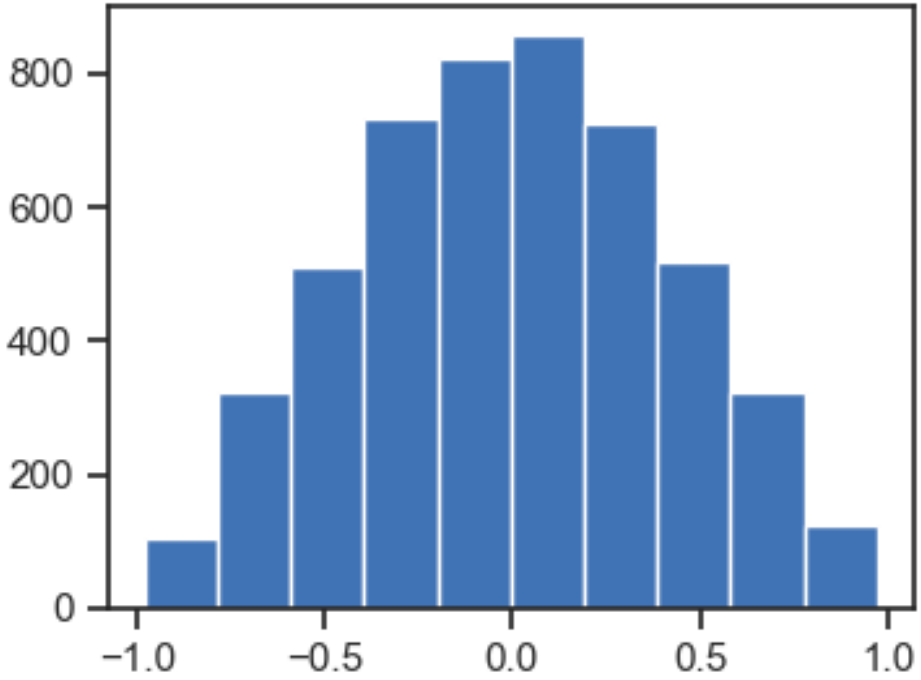
- As the sample size is small, we'll test for the significance by performing a permutation test based on the same data.
- We explore all possible pairings by only permuting one of the two variables.

```python
# Function called by permuation test to calculate the statistic (correlation coefficient)
def statistic(x):  # in this case we only explore all possible pairings by permuting `x`
    rs = stats.spearmanr(x, y).statistic  # ignore pvalue
    return rs

# Perform a permutation test
pm_res = stats.permutation_test((x,), statistic, alternative='greater', permutation_type='pairings')
r = pm_res.statistic
p_value_pm = pm_res.pvalue
null_distribution = pm_res.null_distribution
print("Permutation numbers: ", len(null_distribution))
print("statistic={}, p-value={}".format(r, p_value_pm))

# Plot the sampling distribution of the correlation coefficent under the null hypothesis
fig, ax = plt.subplots(figsize=(4, 3))
ax.hist(pm_res.null_distribution)
plt.show()
```

```
Permutation numbers:  5040
statistic=0.7000000000000001, p-value=0.04563492063492063
```

Permutation Tests for Correlation Coefficient

- The null distribution is generated by permuting x for 7!=5040 times and calculating the Spearman's correlation coefficient for each permutation of x.
- the p-value from the permutation test (4.56%) is higher than that from the t-test, but is still smaller than the 5% alpha level.
- We can still reject the null hypothesis based on the permutation test results.



- verify the p-value from the generated null distribution

```
1  (null_distribution >= r).mean()
```

```
0.04563492063492063
```

CORRELATION AND CAUSATION

Correlation and Causation

- Correlation does not imply causation
  - x is correlated with y **does not mean** x causes y

Example: Divorce rate in Maine vs. Per capita consumption of margarine

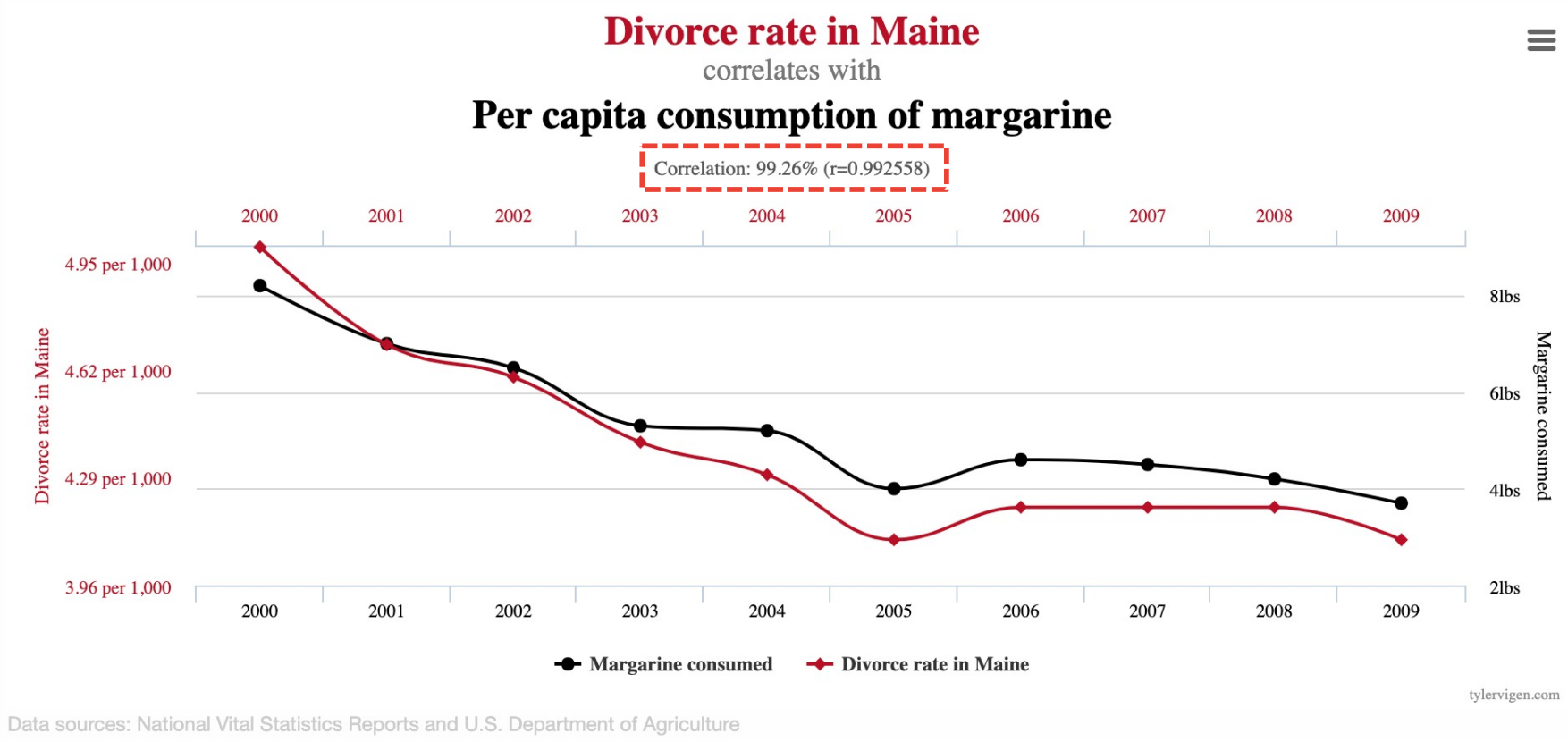


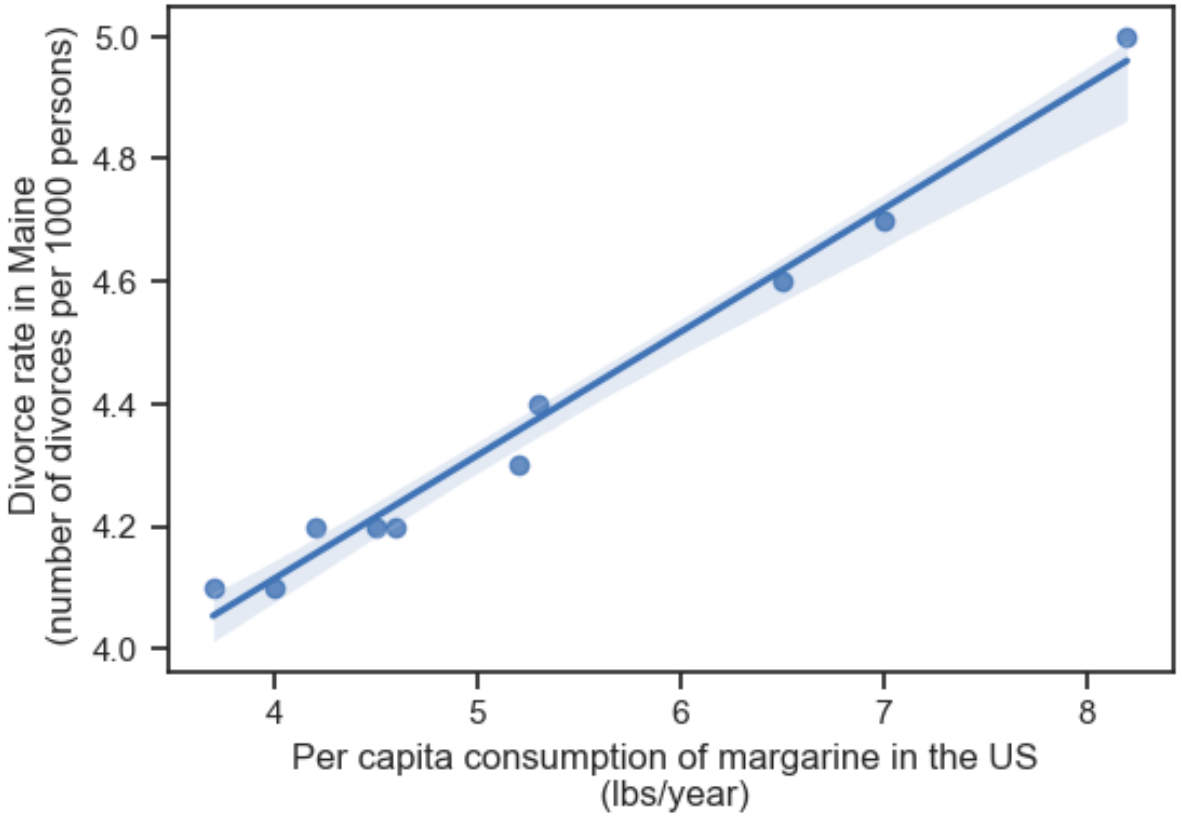
Image Source: https://www.tylervigen.com/spurious-correlations

Correlation and Causation

- Correlation does not imply causation
  - x is correlated with y **does not mean** x causes y

Example: Divorce rate in Maine vs. Per capita consumption of margarine

- Strong correlation
- No causation
- A spurious correlation



Correlation and Causation

- In causal inference, a **confounder** (also confounding variable, confounding factor, extraneous determinant or lurking variable) is a variable that influences both the dependent variable and independent variable, causing a spurious correlation.

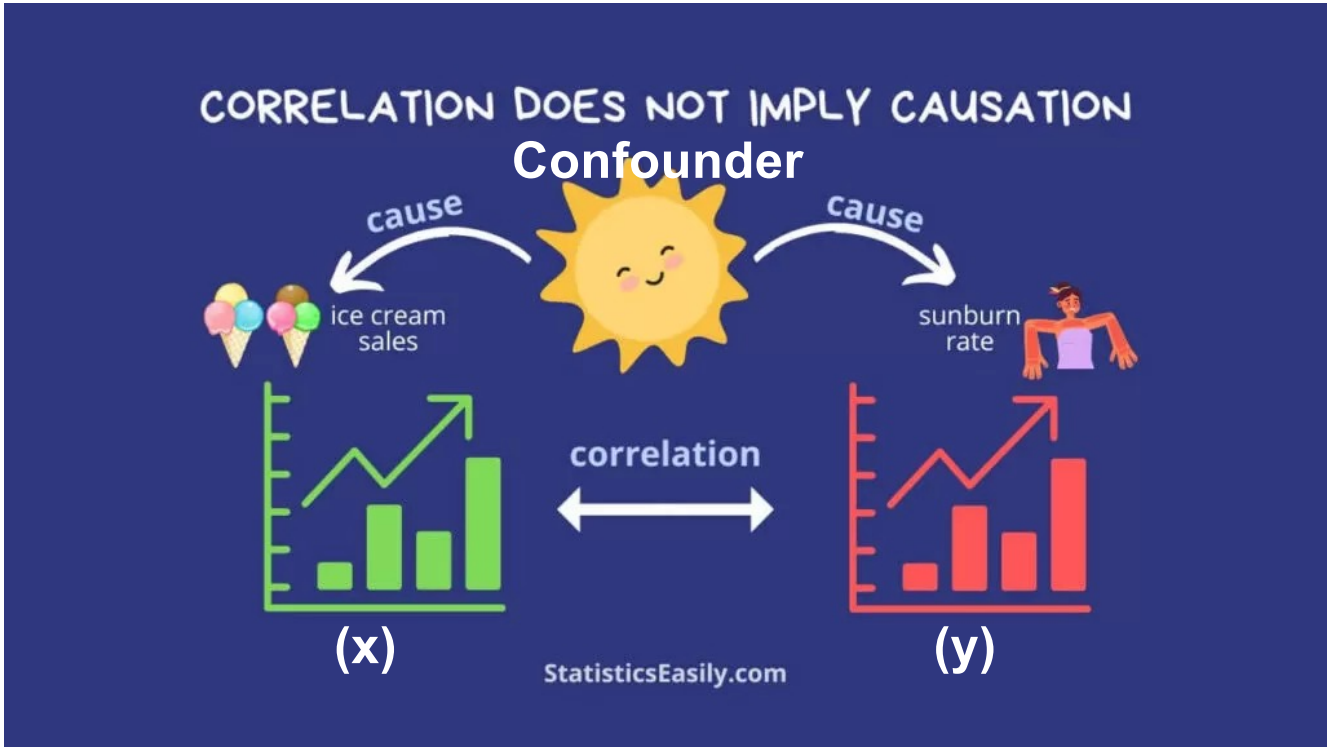


Image Source: https://statisticseasily.com/correlation-vs-causality/

RECAP

Correlation

Recap

- **Correlation and Causation:**
  - Number between -1 and 1
  - Relationship between two variables
    - Linear, Monotonic
    - Positive, Negative
    - Strong, moderate, weak, no relation
  - Pearson, Spearman
- **Test for Significance**
  - t-Test $$t = r\sqrt{\frac{n-2}{1-r^2}}$$
  - Permutation Test
    - explore all possible pairings by only permuting one of the two variables

Recap

- **Correlation Coefficients:**
  - Correlation does not imply causation
  - Spurious correlations
  - Confounder

Questions

Use student forum on QM+

chao.shu@qmul.ac.uk