CS 5A

Nov. 14th

Correlation

Correlation measures the strength and direction of relationship between two variables
Ranges from -1 to +1

- +1: Perfect positive correlation
- -1: Perfect negative correlation
- 0: No correlation

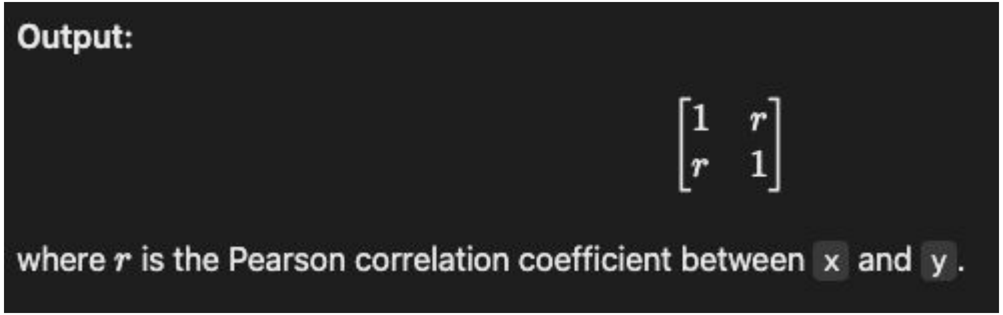
Correlation

Pearson correlation formula:

$r = \Sigma((x - \mu_x)(y - \mu_y)) / (\sigma_x \sigma_y)$

Code:

```
np.corrcoef(column_1, column_2)[0][1] -> r
```

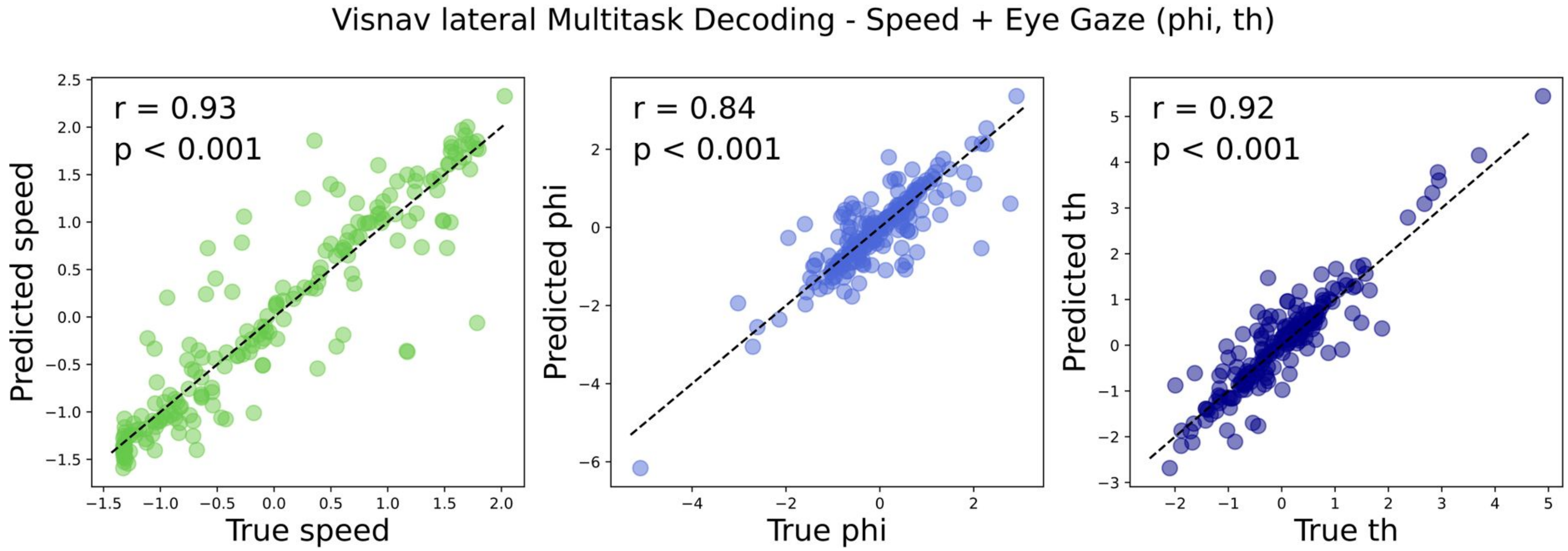


Output:

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

where r is the Pearson correlation coefficient between `x` and `y`.

Correlation



Sampling Bias

1. Convenience Sampling Bias
    - Selectively choosing data points can artificially increase correlation
    - Can lead to misleading conclusions
    - Common issue in research when data collection isn't random

```
np.corrcoef(data.column('quality'), data.column('alcohol'))
```

```
Correlation: 0.4
```

```
# Convenience sampling (selecting only high quality wines) biased_sample =
data.where('quality',are.above(6)).where('alcohol', are.above(10)).take(np.arange(100))
```

```
New correlation after convenience sampling
```

```
biased_corr = np.corrcoef(biased_sample.column('quality'), biased_sample.column('alcohol'))
```

```
Correlation: 0.7
```

Random Sampling creates a more representative correlation coefficient

Probabilities

P(Event) = **Number of Favorable Outcomes / Total Number of Possible** Outcomes

- Or more formally: **P(A) = n(A) / n(S)** where **n(A) is count of event A**, and **n(S) is total sample size**

```
matching_students = students.where('height', are.above(175)) \
.where('weight', are.between(60, 70)) \ .where('age',are.below(22))

probability = matching_students.num_rows / students.num_rows
```

Can sampling bias affect probabilities?

Law of Averages

Law of Averages:

- Empirical probability approaches theoretical probability as trials increase

Example - coin flipping

- theoretical probability is 50% heads, 50% tails
- if you flip 5 times, you might get 4 heads and 1 tail - 80%/20%!
- does this scale to 5000 coin flips?

Lab05 Examples

Simulating an event

```
def simulate_probability(n_samples):
      return (students.sample(n_samples).where('height', are.above(175)) \
      .where('weight', are.above(70)).num_rows / n_samples)


# Run multiple simulations
n_simulations = 10
results = [] for _ in range(n_simulations):results.append(simulate_probability(100))


true_prob = (students.where('height', are.above(175)) .where('weight', are.above(70)) .num_rows /
students.num_rows)
```

Simulation

https://www.youtube.com/watch?v=SCNr_Lom5z8

The distribution will always tend towards a bell curve with more samples like this!