

ECON 2250: Statistics for Economists

Jim Crozier

Fall 2022

Mid term practice 1

Suppose X is a column in a panda's dataframe and x_1, x_2, \dots, x_n are the rows of X where n is the number of observations. For reference, the python code for the math we are about to type is.

```
import pandas as pd
import numpy as np
df = pd.DataFrame({'X': [5, 7, 10]})
n_obs = len(df)
mean_x = df["X"].mean()
np.sqrt(sum((df["X"] - mean_x)**2)/(n_obs-1))
```

0.1 summary stats

- what is the sum of the list above (5,7,10)?
- how many observations in the list?
- what is the mean of the list?
- demean each row (subtract off the mean), show your work to one decimal point.
- square the demeaned values, show your work.
- get the sum of squared values, show your work.
- divide this by n (number of observations).
- just put a square root over it (no need to calculate the sqrt), what is this value called and why does it matter?

0.2 jargony jargon

Define the following words:

- scalar
- vector

- dataframe
- column
- row
- continuous variable, give example
- discrete variable, give example
- dependent variable, give example
- what is pandas, how do you load it into python
- what the sum of a column (vector)
- mean
- quantile
- median
- summation operator
- marginal probability
- bayes rule
- $P(A \cup B)$
- $P(A \cap B)$
- $P(A, B)$
- $P(A|B)$
- independent events, what does it imply wrt $P(A|B)$
- mutually exclusive event
- what is a type error?

0.3 Mean to Standard Deviation

Use the summation operator (\sum) to express the sum of x from 1 to n . I just want the mathematical expression here.

What does the scalar from above represent?

Use the summation operator (\sum) to express the average of X . I just want the mathematical expression here.

What does the scalar from above represent?

Refer to the above scalar as $mean_x$ and write an equation for subtracting off $mean_x$ from element of X . Refer to the elements as x_i , and I only want the equation.

What does the list (vector) from above represent? How many rows will it have?

Write an equation for squaring the above equation.

What does the list (vector) from above represent? How many rows will it have?

Write an equation for summing the squared elements above and dividing by n . This should look kind of similar to taking the average above.

Now take the square of the above equation. What do we call this observation, and what does it represent.

0.4 Quantiles : the sorting hat

Describe how you would find the min and max, and what they represent.

Describe how you find the median? What the core idea behind median and quantiles? What quantile does the median represent?

What is the difference in finding the median of an even or odd number of observations.

Why do we care about the difference between the mean or median or other quantiles?

0.5 Probability

Explain the rule for finding $P(A \cup B)$. Should this be larger than $P(A)$? Could it be the same size as $P(A)$, if so how?

Explain the term "mutually exclusive" in your own words with regards to $P(A \cup B)$.

What is the $P(A \cup B)$ of two mutually exclusive events with $P(A) = 0.1$ and $P(B) = 0.2$? Draw a venn diagram to show this.

Referring to the above question, now what if I tell you that the $P(A|B)$ is 0.05, now what the $P(A \cup B)$? Draw a venn diagram to show this. Draw a venn diagram to show this.

Explain the rule for finding $P(A \cap B)$. Should this be larger than $P(A)$? Could it be the same size as $P(A)$, if so how?

What is the $P(A \cap B)$ of two independent events with $P(A) = 0.1$ and $P(B) = 0.2$? Draw a venn diagram to show this.

Ok, now for some bayes. You might find these familiar. 75% of the children in Exton school have a dog, and 30% have a cat. Given that 60% of those that have cat also have a dog, what percent of those that have a dog also have a cat?

In a factory, machine X produces 60% of the daily output and machine Y produces 40% of the daily output.

2% of machine X's output is defective, and 1.5% of machine Y's output is defective.

One day, an item was inspected at random and found to be defective. What is the probability that it was produced by machine X?

0.6 Correlation

What is the upper and lower bounds of correlation?

If we take the row wise deviations of x and y , explain in words how the algorithm enforces these bounds?

Explain in words what a correlation of 0.8 means?

Explain in words what a correlation of -0.8 means?

Explain in words what a correlation of -0.08 means?

Explain in words what a correlation of 0.08 means?

We often say that "correlation is not causation", briefly explain what that means (briefly = a few sentences)?

Use the data below the following questions.

y	x
5	3
3	8
7	1
1	4

What is the mean of y and x?

Create a grid below and demean x and y, and then square those columns.

Sum the squares of those two columns (demeaned x squared, and demeaned y squared), and show the square root (no need to calculate) of each multiplied by each other.

In your grid create new column and calculate $(x_i - \bar{x})(y_i - \bar{y})$ where \bar{x} is the mean and \bar{y} is the mean of y. Note that this will be n long.

Now sum up the numerator.

If the you were able to calculate the numerator and denominator and the correlation came out to -0.61394061, what would be the interpretation?

Draw an y axis and an x axis and plot the points. Does the correlation make sense.