Bivariate Statistics Exploration

Jacob Bruner

September 26, 2022

1 Introduction

In 1973, British mathematician Francis Anscombe formulated four datasets, each with 11 points, in order to demonstrate shortcomings of common statistical measures.

In the real-world, statistics are often viewed as objective measures of datasets. Because of this, we often draw important, life-changing decisions on their basis alone. For instance, pharmacists employ statistics when researching new treatments for disease. Luckily for us, statistics often *are* helpful and insightful in disseminating trends and patterns in data. But, again, it is all too easy to forget their short-comings.

In this paper, I will explore the ways in which statistical measures can be misleading in interpreting the significance of, trends in, or general patterns within data.

2 The Statistical Mean

2.1 Computation

The mean is a commonly used measure of center for a given statistic. To calculate the mean, we can use the following formula: for a given indexed set of values, the mean is written:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Where N is the length of the data set. In other words, the mean is the sum of each value divided by the number of values for a given parameter.

Calculating these values for each dataset we obtain:

Data set	**I**		**II**		**III**		**IV**	
Parameter	x	y	x	y	x	y	x	y
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5

2.2 Interpretation

After computing the mean for each, it's clear that the means for each x, y of the respective datasets are equal. Without any other prior knowledge, one might infer that these datasets have a similar center, save variation or spread. This is because the mean doesn't provide information about the variation or spread of the data. Despite this, however, one might assume it to be a reasonable assumption that these datasets and parameters might be highly similar, especially because they match in two variables.

3 Exploring Other Statistical Measures

3.1 Computation

In addition to computing the mean, there are a number of other tools that provide insight into the properties of a dataset. For instance, the ***variance*** (σ^2) and ***standard deviation*** (σ) convey a measure

	$\bar{x}$	$\bar{y}$	σ_x^2	σ_y^2	σ_x	σ_y	r^2	$cov(x,y)$
set 1	9.0	7.5	11.0	4.13	3.32	2.03	0.67	5.5
set 2	9.0	7.5	11.0	4.13	3.32	2.03	0.67	5.5
set 3	9.0	7.5	11.0	4.12	3.32	2.03	0.67	5.5
set 4	9.0	7.5	11.0	4.12	3.32	2.03	0.67	5.5

of the 'spread' or variation in a dataset. Similarly, ***'Peterson's Correlation Coefficient'*** often denoted with its square, r^2, is measure of the *linear* coorelation of two parameters. (Unsquared) values of r range from -1 to 1, with numbers farther from zero denoting stronger coorelation. A closely related notion is the ***covarience***, which measures the linear coorelation, but with a magnitude not necessarily between $(-1, 1)$. These measures are related like so: $r = \frac{cov(x,y)}{\sigma_x \sigma_y}$ where σ denotes the standard deviation.

3.2 Plotting Data

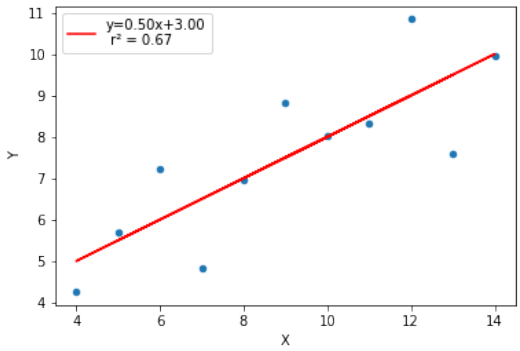
Using the python packages `matplotlib`, `numpy` and `pandas`, we can plot each dataset (as a dataframe contained in the array `dfs`). A linear regression is fit to the data using the `np.polyfit()` method with an exponent of 1. This is accomplished like so:

```
for i in range(4):
  dfs[i].plot.scatter(x='x', y='y')
  reg = np.polyfit(dfs[i]['x'], dfs[i]['y'], 1)
  rval = stats.pearsonr(dfs[i]['x'], dfs[i]['y'])[0]
  plt.plot(dfs[i]['x'], np.polyval(reg, dfs[i]['x']), 'r', label='y={:.2f}x+{:.2f}\n r^2 =
      {:.2f}'.format(reg[0], reg[1], rval**2))
  plt.xlabel('x')
  plt.ylabel('y')
  plt.legend()
  plt.savefig('dataset{}.png'.format(i+1))
```

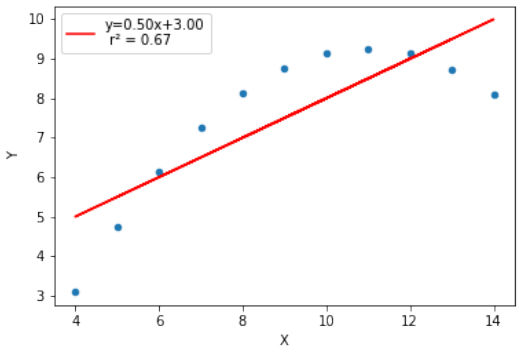
After doing so, we get the following interesting results:

3.3 Interpretation

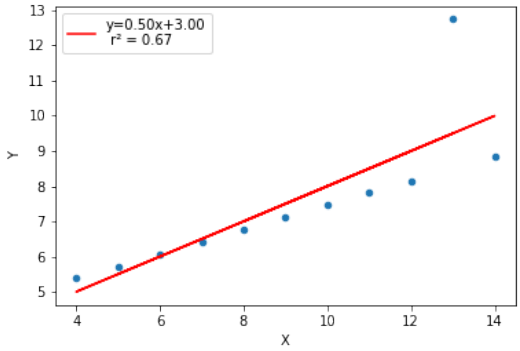
Upon examination of the computed statistical measures (std, varience, cov, etc.), the naive conclusion would be that each dataset contains extremely similar data and that similar conclusions could be drawn from each. For instance, the r^2 value for each, 0.67, indicates that there is a weak-to-moderate, positive correlation between the variables, and might imply that linear fits to each dataset would have a similar degree of inaccuracy. However, this notion is *categorically refuted* on even a cursory glance at the graphs. In Figure 1, we see that each dataset is very distinct in its scatterplot pattern. In Figure 1a, its clear that the linear fit to the data



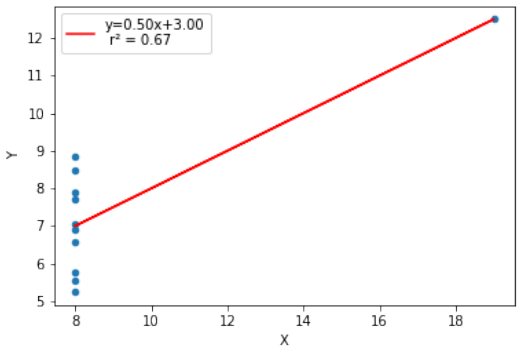
(a) Dataset I



(b) Dataset II



(c) Dataset III



(d) Dataset IV

Figure 1: Plot of Datasets I-IV — X against Y with Lin. Reg.