Bivariate Statistics Exploration

Jacob Bruner

September 26, 2022

1 Introduction

In the real-world, statistics are often employed as objective measures of data. 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, or economists might draw on statistics to understand trends in consumer spending. Luckily for us, statistics often *are* helpful and insightful in disseminating patterns in data. But, again, it is all too easy to forget their short-comings.

In 1973, British mathematician Francis Anscombe formulated four datasets, each with 11 points, in order to demonstrate the pitfalls of many common statistical measures. In this paper, I will explore the ways in which these statistical measures can be misleading in interpreting the significance of, trends in, or validity of data using Anscombe's 'quartet' of points to highlight their oversimplification.

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

Figure 1: Computed Means of Dataset I-IV

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.

	$\bar{x}$	$\bar{y}$	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

Figure 2: Comparison of different Statistical Measures of Datasets I-IV

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 of the 'spread' or variation in a dataset. Similarly, ***'Peterson's Correlation Coefficient'*** (r) 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 ***covariance***, 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

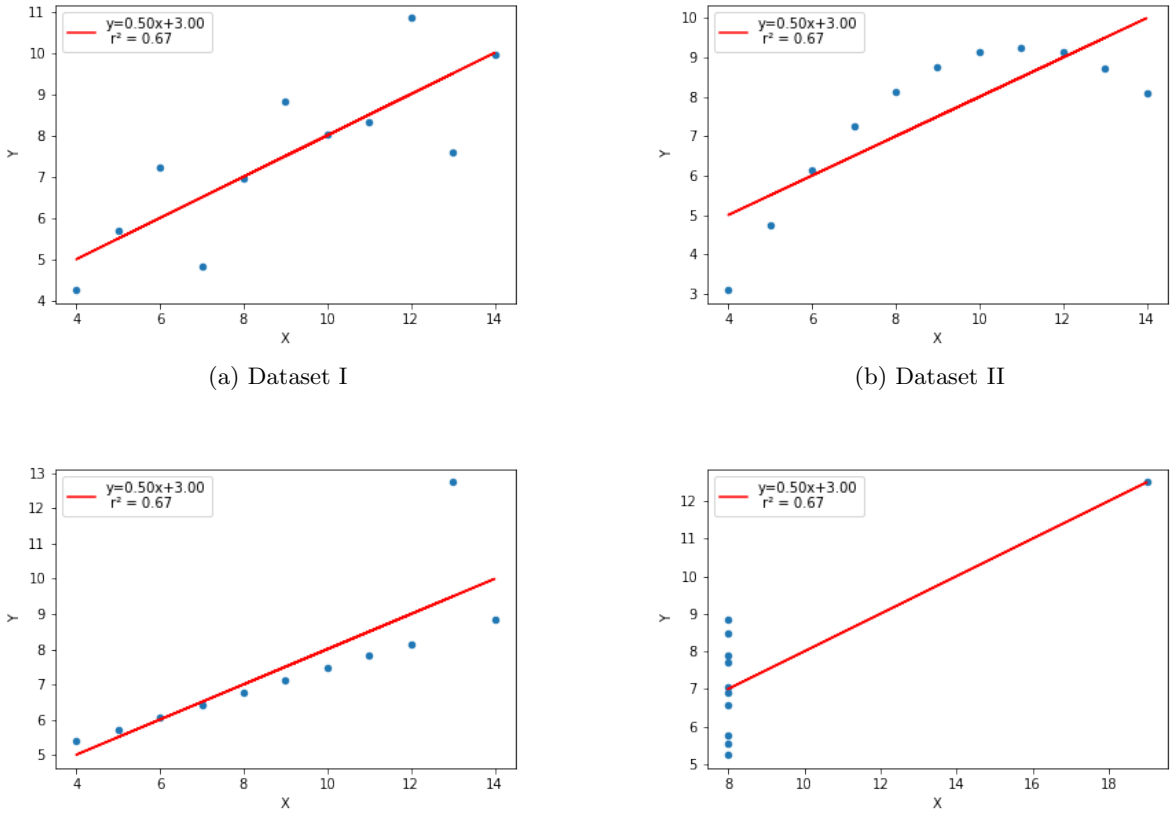
Using the python packages `matplotlib`, `scipy`, `numpy` and `pandas`, we can plot each dataset (as a dataframe contained in the array `dfs`). A linear regression is fit to the data using the `np.polyfit()` method with an exponent of 1, and the Pearson's coefficent is calculated using the `stats.pearsonr()` method. 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))
```

In so doing, we get the following interesting results (Figure 2):

3.3 Interpretation

Upon examination of the computed statistical measures (std, variance, 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 2, we see that each dataset is very distinct in its scatterplot pattern. In Figure 2a, its clear that the linear fit to the data indicates weak-moderate, positive correlation, with a healthy amount of deviation from the trendline. This graph's, the most straightforward of the four, statistical measures indicate helpful information about the data. For instance, the r^2 value of 0.67 reinforces the weak, positive spread outlined above. Similarly, the covariance of 5.5 and the variance $\sigma_x^2 = 11.0$, $\sigma_y^2 = 4.13$ reinforce the visual intuition of this dataset having a healthy amount of spread from the means $\bar{x}, \bar{y}$. As we transition to Figure 2b, we see a markedly different trend in the data. Visually, the scatterplot indicates an quadratic trend in x, indicating that the y values may vary as the negative square of x. Despite this, our linear trendlines (from a least squares regression) provides a line of 'best fit' equal to Figure 2a. Additionally, the other statistical measures of dataset II are exactly equivalent to that



(a) Dataset I

(b) Dataset II

(c) Dataset III

(d) Dataset IV

Figure 3: Plot of Datasets I-IV — X against Y with Lin. Reg.

of dataset I, despite the visual interpretation being different. For instance, one might interpret the r^2 value in Figure 2a as indicating the moderate (seemingly random) variation from the trendline, whereas in Figure 2b, we see the datapoints follow exactly the (not random) curve of a quadratic, despite still having the same r^2. If one were to perform a polynomial regression (perhaps taking the logarithm to linearize the polynomial exponent, then performing a least squares regression), the resulting trendcurve would likely be an extremely close fit to the data—a fact not illustrated by these statistical measures. In Figure 2c, we see a different outcome where, despite displaying a very linear trend, the computed statistics and trendline are greatly affected by a single outlier. Because of this, we see the trendline having the incorrect slope to match the (linear) trend in the rest of the data. Despite this outlier not being indicated in the computed statistical measures, viewing the graph demonstrates the highly linear relationship between x and y, which would be easily illustrated by removing the outlier or otherwise explaining it in the methodology. This fact highights the differnce between dataset III and dataset I, where, despite having the same statistical measures, Figure 2a displays almost uniformly random deviation from the line of best fit compared to Figure 2c which has a nicely behaved deviation given by the intersection of another trendline if the sole outlier were to be removed. In Figure 2d, we see how the