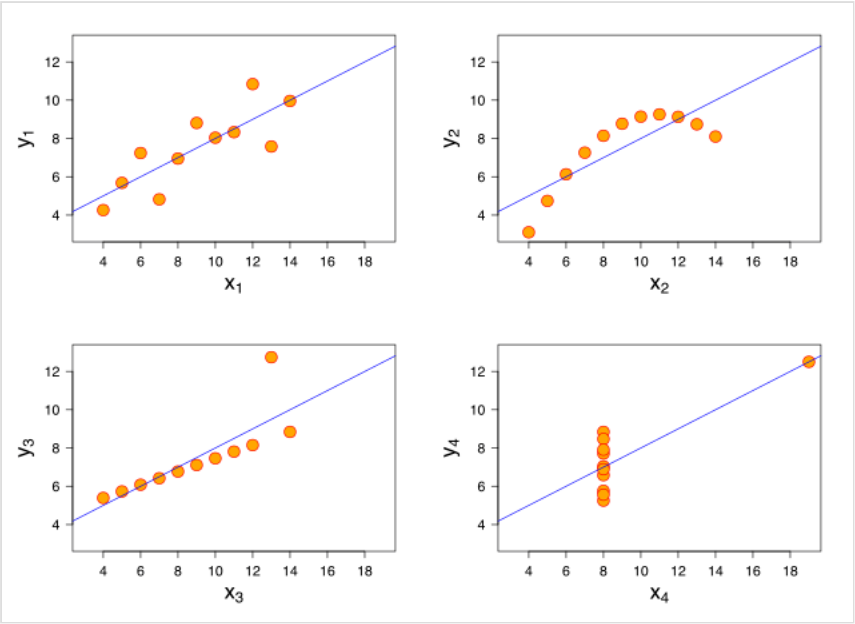




Anscombe's quartet

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".^[1]



The four datasets composing Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

Data

For all four datasets:

Property	Value	Accuracy
<u>Mean</u> of x	9	exact
Sample <u>variance</u> of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	± 0.003
<u>Correlation</u> between x and y	0.816	to 3 decimal places
<u>Linear regression</u> line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
<u>Coefficient of determination</u> of the linear regression: R^2	0.67	to 2 decimal places

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x .
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression

and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.^{[2][3][4][5][6]}

The datasets are as follows. The x values are the same for the first three datasets.^[1]

Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

It is not known how Anscombe created his datasets.^[7] Since its publication, several methods to generate similar datasets with identical statistics and dissimilar graphics have been developed.^{[7][8]} One of these, the *Datasaurus dozen*, consists of points tracing out the outline of a dinosaur, plus twelve other datasets that have the same summary statistics.^{[9][10][11]}

See also

- Datasaurus dozen
- Exploratory data analysis
- Goodness of fit
- Regression validation
- Simpson's paradox
- Statistical model validation

References

1. Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician*. **27** (1): 17–21. doi:10.1080/00031305.1973.10478966 (https://doi.org/10.1080%2F00031305.1973.10478966). JSTOR 2682899 (https://www.jstor.org/stable/2682899).
2. Elert, Glenn (2021). "Linear Regression" (http://physics.info/linear-regression/practice.shtml#4). *The Physics Hypertextbook*.
3. Janert, Philipp K. (2010). *Data Analysis with Open Source Tools* (https://archive.org/details/isbn_9780596802356/page/65). O'Reilly Media. pp. 65–66 (https://archive.org/details/isbn_9780596802356/page/65). ISBN 978-0-596-80235-6.
4. Chatterjee, Samprit; Hadi, Ali S. (2006). *Regression Analysis by Example*. John Wiley and Sons. p. 91. ISBN 0-471-74696-7.
5. Saville, David J.; Wood, Graham R. (1991). *Statistical Methods: The geometric approach*. Springer. p. 418. ISBN 0-387-97517-9.
6. Tufte, Edward R. (2001). *The Visual Display of Quantitative Information* (https://archive.org/details/visualdisplayofq00tufte) (2nd ed.). Cheshire, CT: Graphics Press. ISBN 0-9613921-4-2.
7. Chatterjee, Sangit; Firat, Aykut (2007). "Generating Data with Identical Statistics but Dissimilar Graphics: A follow up to the Anscombe dataset". *The American Statistician*. **61** (3): 248–254. doi:10.1198/000313007X220057 (https://doi.org/10.1198%2F000313007X220057). JSTOR 27643902 (https://www.jstor.org/stable/27643902). S2CID 121163371 (https://api.semanticscholar.org/CorpusID:121163371).
8. Matejka, Justin; Fitzmaurice, George (2017). "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing". *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. pp. 1290–1294. doi:10.1145/3025453.3025912 (https://doi.org/10.1145%2F3025453.3025912). ISBN 9781450346559. S2CID 9247543 (https://api.semanticscholar.org/CorpusID:9247543).
9. Matejka, Justin; Fitzmaurice, George (2017). "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing" (https://www.autodesk.com/research/publications/same-stats-different-graphs). *Autodesk Research*. Archived (https://web.archive.org/web/20201004003855/https://www.autodesk.com/research/publications/same-stats-different-graphs) from the original on 2020-10-04. Retrieved 2021-04-20.
10. Murray, Lori L.; Wilson, John G. (April 2021). "Generating data sets for teaching the importance of regression analysis" (https://onlinelibrary.wiley.com/doi/10.1111/dsji.12233). *Decision Sciences Journal of Innovative Education*. **19** (2): 157–166. doi:10.1111/dsji.12233 (https://doi.org/10.1111%2Fdsji.12233). ISSN 1540-4595 (https://search.worldcat.org/issn/1540-4595). S2CID 233609149 (https://api.semanticscholar.org/CorpusID:233609149).
11. Andrienko, Natalia; Andrienko, Gennady; Fuchs, Georg; Slingsby, Aidan; Turkay, Cagatay; Wrobel, Stefan (2020), "Visual Analytics for Investigating and Processing Data" (http://link.springer.com/10.1007/978-3-030-56146-8_5), *Visual Analytics for Data Scientists*, Cham: Springer International Publishing, pp. 151–180, doi:10.1007/978-3-030-56146-8_5 (https://doi.org/10.1007%2F978-3-030-56146-8_5), ISBN 978-3-030-56145-1, S2CID 226648414 (https://api.semanticscholar.org/CorpusID:226648414), retrieved 2021-04-20.

External links

- Department of Physics, University of Toronto (http://www.upscale.utoronto.ca/GeneralInterest/Harrison/Visualisation/Visualisation.html)
- Dynamic Applet (https://www.geogebra.org/m/tbwXxySn) made in GeoGebra showing the data & statistics and also allowing the points to be dragged (Set 5).

- [Animated examples from Autodesk \(https://www.autodeskresearch.com/publications/samestats\)](https://www.autodeskresearch.com/publications/samestats) called the "Datasaurus Dozen".
 - [Documentation \(https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/anscombe.html\)](https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/anscombe.html) for the datasets in [R](#).
-

Retrieved from "https://en.wikipedia.org/w/index.php?title=Anscombe%27s_quartet&oldid=1220464438"