Data, Data Science, and Official Statistics

Yves Tillé
University of Neuchatel
Switzerland

Third workshop on methodologies for official statistics
December 2024, Rome

Unfortunate terminology

- “Statistics” comes from German *Statistik*
- Proposed by German economist Gottfried Achenwall (1719 - 1772).

Unfortunate terminology

- Body of knowledge that a statesman should possess.
- Knowledge needed to establish good governance.
- Now, scientific discipline that studies phenomena by collecting, processing, analysing, interpreting, modelling and forecasting data
- New terminology data science: vague and encompass statistics, information technology, computer science, database management and mathematics.

Statistics is a discipline that is not limited to “state‘” management.

Unfortunate terminology

- Data (1640): a fact a fact given or granted.
- Plural of *datum* that is the supine of the Latin verb *dare*, meaning "to give".
- In French : *données*, in Spanish, *datos*, in Italian *dati*.
- In German, *Datum* means date, *Daten* is data.
- Very bad term: Data is constructed following observation, experimentation or recovery from an administrative procedure (Wilkinson, 2005).

Unfortunate terminology

Becker (1952) pointed out that the word “data” therefore refers to observations that are taken for granted, when in fact they never are. By always using the supine of Latin verbs, we could have taken:

- *capta*, plural of the supine *captum* of the verb *capere* “to take”,
- *collecta*, plural of the supine *collectum* of the verb *colligere* “to collect”,
- *recepta*, plural of the supine *receptum recipere* “to receive”,
- *obtenta*, plural of the supine *obtentum* of the verb *obtinere* “to obtain”.

The statistician's main task

The statistician's main task is

- not only to process data,
- produce data,
- obtain data,
- correct data,
- try to deal with measurement errors.

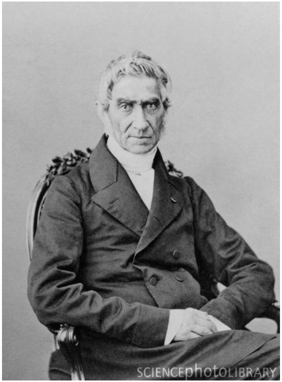
The data is not given.

The statistician's main task

- The word “data’ obliterates the heaviest and probably most important part of the statistician's job.
- This contributes to the myth that data on anything and everything is available at any time.
- This myth is more relevant than ever.
- With the development of the internet and the equally vague concept of big data, there is a tendency to think that all statistical information will be easily and directly available, which is a modern myth.

History

- 19th century, the predominant idea was that only exhaustive censuses could really produce reliable information (Quetelet, 1846).



History

Quetelet (1846, p. 293)

> *"La Place had proposed to substitute for the census of a large country, such as France, some special censuses in selected departments where this kind of operation might have more chances of success, and then to carefully determine the ratio of the population either at birth or at death. By means of these ratios of the births and deaths of all the other departments, figures which can be ascertained with sufficient accuracy, it is then easy to determine the population of the whole kingdom. This way of operating is very expeditious, but it supposes an invariable ratio passing from one department to another. [· · ·] This indirect method must be avoided as much as possible, although it may be useful in some cases, where the administration would have to proceed quickly; it can also be used with advantage as a means of control."*

History

Ken Brewer (2013)

> *"No, Laplace had not been the first person to use a ratio estimator, not even the first Frenchman (Stephan 1948). The Englishman John Graunt had used the ratio estimator in his estimation of the population of London (Graunt 1662). Well, perhaps he had not really used the ratio estimator (he probably hadn't used anything that would be recognized as a ratio estimator today, certainly not by a finicky survey statistician like me!), but he had admittedly used the Rule of Three."*

History

Argument used by Quetelet (1846, p. 293) to justify his position.

> *"To not obtain the faculty of verifying the documents that are collected is to fail in one of the principal rules of science. Statistics is valuable only by its accuracy; without this essential quality, it becomes null, dangerous even, since it leads to error."*

History

- Anders Nicolai Kiær (1896, 1899, 1903, 1905) proposed the use of partial data produced by survey sampling. (scandal!)

History

Georg von Mayr (Prussia) [$\cdots$] "It is especially dangerous to call for this system of representative investigations within an assembly of statisticians. It is understandable that for legislative or administrative purposes such limited enumeration may be useful - but then it must be remembered that it can never replace complete statistical observation. It is all the more necessary to support this point, that there is among us in these days a current among mathematicians who, in many directions, would rather calculate than observe. But we must remain firm and say: no calculation where observation can be done."

History

Guillaume Milliet (Switzerland). "I believe that it is not right to give a congressional voice to the representative method (which can only be an expedient) an importance that serious statistics will never recognize. No doubt, statistics made with this method, or, as I might call it, statistics, pars pro toto, has given us here and there interesting information; but its principle is so much in contradiction with the demands of the statistical method that as statisticians, we should not grant to imperfect things the same right of bourgeoisie, so to speak, that we accord to the ideal that scientifically we propose to reach."

History

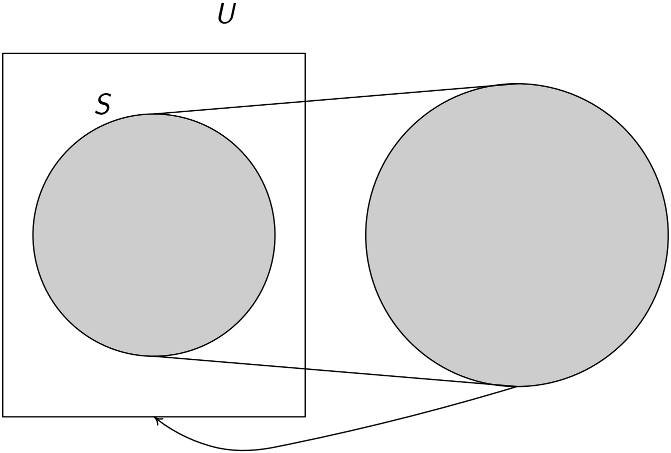
- It was only after 30 years that the idea of sampling was finally accepted by the International Statistical Institute (Jensen, 1926).
- With the development of computers, administrative files were created and began to be used as statistical sources.
- The accumulation of data on the Internet is intriguing. The volume is enormous.
- The question is whether this data can be put to good use, but real applications are extremely rare.

In official statistics, the main objective is the extrapolation to a population.

Data from the internet

- A wealth of information is available on the Internet.
- In general, this information cannot be extrapolated to the population as a whole.
- Problem with the definition of “unit of observations”.
- Notion like establishment or household are not simple.
- Extrapolation from the internet is often not possible.

Auxiliary Information	**Interest Information**
Auxiliary variables	**Interest variables**
X	Y
known or partially known	unknown



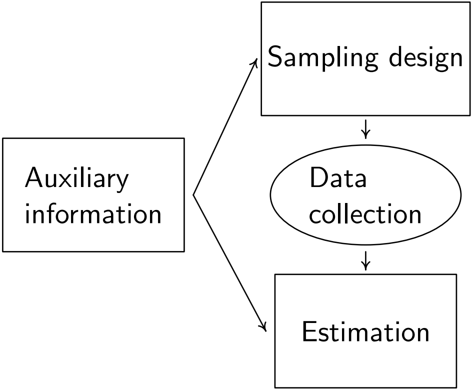
Extrapolation

Extrapolation

- Design-based or model-based approach or both.

$$\sum_{k\in S}\frac{y_k}{\pi_k} \text{ against } \sum_{k\in S} y_k + \sum_{k\in U\backslash S} \hat{y}_k.$$

- With design based approach, the inclusion probabilities must be known that implies that a register must be known on the population to select the sample.
- With the model based-approach, auxiliary variables on the population are needed.


Sampling design
Auxiliary information
Data collection
Estimation

To extrapolate, we need information from outside the sample.

A multitude of methods

- A multitude of methods: Weighting, mass imputation, modelling, prediction, propensity score methods, matching or calibration.
- All these methods aim to graft an extrapolation principle onto data that cannot be necessarily extrapolated a priori.
- For most methods, extrapolation can be formalised by a weighting system.
- It is up to the official statistics institutes to provide the framework for extrapolation, as the information needed to apply the principle of extrapolation cannot be produced by data without structure.

In official statistics, most methods aim to re-establish extrapolation.

Errors

- The fact that estimates contain errors can be more or less problematic.
- Errors do not contaminate all estimators in the same way.
- Suppose that two estimators of the same parameter at two different times $\widehat{\tau}_1$ and $\widehat{\tau}_2$ contain errors.
- The increase
$$100 \times \frac{\widehat{\tau}_2 - \widehat{\tau}_1}{\widehat{\tau}_1} \%$$
can be very accurate if $\widehat{\tau}_1$ and $\widehat{\tau}_2$ are strongly positively correlated.
- Application: growth of GDP, price index.

An error in the data is not necessarily a problem for an estimator.

The term “representative” means nothing

- Already used by Anders Nicolai Kiær in his 1896 article entitled “Observations and Experiments Concerning Representative Counts”.
- A “representative” sample would be a sort of reduced image of the population.
- Contradictory to Jerzy Neyman (1934) who showed almost a century ago that by over-representing the observations of the most dispersed categories, we can obtain much more precise estimates than with a supposedly representative sample.
- Generalised by Nedyalkova and Tillé (2008); Tillé and Wilhelm (2017)

The use of the word “representativeness” is an argument from authority, which shows above all that the person using this word doesn’t know much about sampling theory.

Use of unstructured partial data sets (for instance big data)

- Intensive research, for example: Yang and Kim (2020); Kim and Tam (2021); Kim (2022); Wu (2022); Valliant (2020); Jauslin and Tillé (2023); Chen et al. (2023); Wang et al. (2023); Beaumont et al. (2023).
- Often, the framework consists of grafting an extrapolation principle onto the unstructured data set, using another sample with an extrapolation principle.

Definition of the variables

- The variables are even more open to question.
- For example, the concept of “disposable income” has been codified to avoid ambiguity.
- There will always be room for interpretation.
- What’s more, a variable is measured at a particular date, in a particular territory, in a particular population.
- For these reasons, we know that there is never a perfect match, even between well-organised and up-to-date administrative files.

Administrative data

- Administrative files also contain errors.
- These errors are different from those encountered in official statistics.
- In official statistics, the data is produced for statistical purposes.
- They are often marred by non-response, but fairly standardised procedures make it possible to deal with this type of problem.
- Administrative data can have problems of quality and lack of updating, because they are produced for administrative purposes.
- If a variable is not really used from an administrative point of view, it may be completely neglected.
- For example, if we have a well-maintained housing register and a well-maintained population register, we can make many errors concerning the allocation of individuals to housing in the same building.
- The fact that some variables are of poor quality is not necessarily serious.
- For example, if we want to calculate the average size of households, we do not need to know which household each individual belongs to.

Calibrate the errors on the errors

- Variables with errors can also be useful for calibration.
- (Deville and Särndal, 1992). A weight w_k is calculated for each individual k in the sample S.
- If a variable y is only known in the sample, estimator of

$$Y = \sum_{k \in U} y_k \qquad \text{by} \qquad \widehat{Y}_C = \sum_{k \in S} w_k y_k.$$

- The weights are calibrated on p variables $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})^\top$

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{1}$$

- It does not matter whether the $\mathbf{x}_k$ variables contain errors, the only important thing is that they are correlated with y_k.
- Errors on $\mathbf{x}_k$ will not introduce bias on the calibrated estimator.

An error-filled variable can be used if it contains an extrapolation principle.

Errors must be calibrated to errors.

Massive non-organised data

- However, it is illusory to think that massive non-organised data, in streams or social networks, can be valued.
- These data do not meet any of the required qualities.
- It is not possible to determine the population from which the statistical units originate, nor the reference dates, and the statistical units cannot be matched with known statistical units.
- No extrapolation principle can be used to reconstruct estimators at population level.
- What's more, the Internet is not stable over time. Data evolves very quickly.
- Simply tracking the price of a product over time is far from straightforward.
- Despite all the fuss about big data in official statistics, applications remain anecdotal.

Machine learning

- Wikipedia (2024)'s Machine learning site:
  "Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions. Recently, artificial neural networks have been able to surpass many previous approaches in performance."
- In the article, we find: probability calculus, regression, cluster analysis, logistic regression, principal component analysis (PCA), fuzzy logic and neural networks.
- Some of these statistical methods are over a century old, such as the PCA proposed by Pearson (1901).
- Alongside this we see compression algorithms, numerical analysis methods and operational research.
- On reading this site, it seems that the definition is confused, that machine learning methods are not necessarily modern, and that a large proportion of machine learning methods are recycled classical statistical methods.

Machine learning

- A change in terminology is intended to make people believe that there has been a break in statistical methods, and we could even use the grandiloquent term “paradigm shift”.
- Methods concerning sparsity for high-dimensional data are old: like the Lasso penalization proposed by Tibshirani (1996) in the last century and the ridge regression was proposed by Hoerl and Kennard (1970) more than fifty years ago.
- Neural networks have also existed for several decades.
- Random forests were introduced by Breiman (2001) a quarter of a century ago.

Statistical learning is the new version of statistics in which not-so-new statistical methods have been recycled.

Inertia in the methods used

- Crisis of the *p*-value.
- Hypothesis testing remains a major argument in many fields where experimental statistics is applied.
- Repeated hypothesis testing and the HARKing methodology are still widely used (Kerr, 1998).
- HARKing (hypothesizing after the results are known).

Inertia in the methods used

- The *p*-value crisis does not seem to have had any real impact on usage.
- It is surprising to see how many scientific disciplines base their arguments essentially on hypothesis testing while there are fewer and fewer references to test theory in scientific journals on statistics.

Even if test theory seems outdated, it is still a master argument in science.

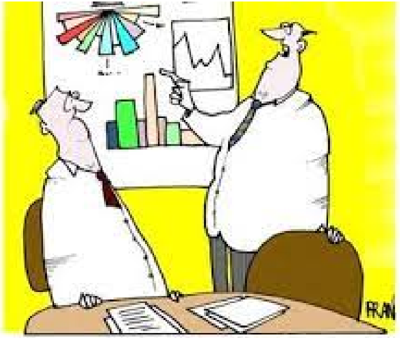
Use of forcasting methods

- Progress has clearly been made in forecasting methods.
- Certain new statistical methods (such as random forests) sometimes enable extremely accurate forecasts to be made, but it is difficult to really explain the relationships between the predicted variable and the auxiliary variables.
- In official statistics, the final objective is often not to make a forecast, but to extrapolate to a population.
- However, all high-performance forecasting methods can be useful.
- A prediction can be considered as a value of the variable of interest containing errors.
- Insofar as all the values of a population can be predicted, these predicted values provide a framework for extrapolation to the population.
- This idea has been the subject of numerous publications by, among others, Dagdoug et al. (2020a,b); Kim and Tam (2021); Yang and Kim (2020).

Any prediction can be useful even if the model can't be explained, provided that the predictions comply with an extrapolation principle.

Loss of confidence

- Loss of confidence in the statistical institutes.
- Response rates are falling in all surveys.
- Political pressure is mounting to reduce the statistical burden, especially for business surveys.
- It is difficult to ask again for information that has already been requested.



87% OF THE 56% WHO COMPLETED MORE THAN 23% OF THE SURVEY THOUGHT IT WAS A WASTE OF TIME

Loss of confidence

- In addition, concerns about data protection are becoming increasingly important.
- Statistical institutes are among the most closely scrutinised government bodies. This may seem odd, since statisticians are not interested in people or companies, but in parameters such as averages, totals and inequality indices.
- However, statistics are easy prey when it comes to data protection.
- It is easier to impose rules on a statistical institute than on a secret service or police force.
- Official statistics are therefore in a delicate position.
- There is an urgent need to make the most of all alternative sources to surveys and censuses.
- These difficulties are compounded by the fact that data seems to be available everywhere and all the time.

Fantasy of immediacy

- This fantasy of immediacy, of knowing everything about everything, this fantasy that everything can be found on the Internet, sometimes leads to the idea that official statistics can only be obtained from the Internet.
- This is obviously a myth. There are relatively few alternative sources that can be extrapolated to the populations of interest in official statistics.
- Added to this is the fact that government IT environments are often lagging behind or even obsolete.
- The fragility of systems in the face of hacking does not make the job any easier.

Challenge: administrative data

- The main challenge is to make the most of all administrative data.
- However, this data is often of poor quality because it has not been collected for statistical purposes.
- A major challenge is to harmonise data from very different sources.
- The battle is far from won.
- Administrative sources are sometimes delivered with considerable delay.
- The statistical units sometimes do not correspond, for example, between the statistical concept of the establishment and the taxable legal entity, or between the statistical concept of the household and the taxable family entity.
- Many countries, even in Europe, are unable to find out who lives where in the country, or how many citizens there are in a municipality where people actually live.
- Countries that do have registers sometimes find it difficult to use them, particularly when they are decentralised.

Reconciling sources

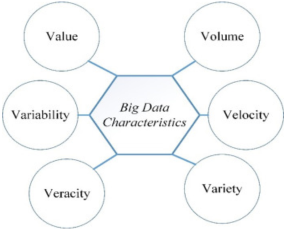
- Reconciling sources can sometimes be a headache, involving a considerable amount of work that has to be repeated every year.
- There is no such thing as immediate access to information.
- It is also difficult to recruit skilled workers in the data science fields of IT and statistics.
- Private companies often pay better for qualified people.
- States' inflexible pay scales do not allow high qualifications to be considered.

Outsourcing

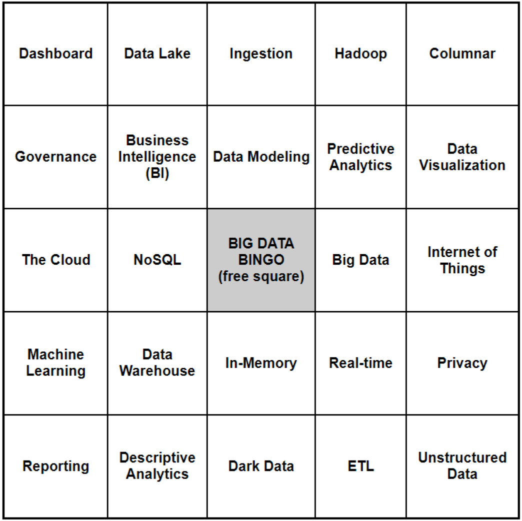
- Outsourcing may seem like an emergency solution, but outsourcing your core business is always a very bad idea.
- In highly technical fields, it is often the case that administrative and political leaders are incapable of understanding the work to be done and the real issues at stake.

Guruism

- In a media environment extolling the advances of data, there has been a significant development of improvised experts whose only skills sometimes lie in frenetic activity on social networks.
- This phenomenon of guruism generates an additional fog that makes it even more difficult to make informed decisions.
- Gurus are easy to recognise, however, because they generally give presentations with slides containing potatoes and arrows that contain only buzzwords and nothing technical or concrete.



Guruism



Dashboard	Data Lake	Ingestion	Hadoop	Columnar
Governance	Business Intelligence (BI)	Data Modeling	Predictive Analytics	Data Visualization
The Cloud	NoSQL	BIG DATA BINGO (free square)	Big Data	Internet of Things
Machine Learning	Data Warehouse	In-Memory	Real-time	Privacy
Reporting	Descriptive Analytics	Dark Data	ETL	Unstructured Data

Challenges 1: Best use of all the administrative data.

- Best use of all the administrative data.
- There is a lot of preliminary work to be done.
- Administrative databases are rarely perfect.
- If the data is not used for administrative purposes, it is rarely reliable.
- Variable definitions can vary considerably. A slightly different definition of what constitutes a household or a business can make merging files inextricable.
- The problem of temporality is also very important. Individuals, like businesses, move, die and are born. However, the main problem seems to me to be of a legal nature.
- The compartmentalisation of administrative services, the sometimes complicated relations between a State and its local authorities, and data protection legislation sometimes make simple administrative uses very complicated.
- Access sometimes requires legislative changes that can take a long time.

Challenges 2: Integration of sources.

- Integration of sources.
- Hundreds of publications have been devoted to this issue.
- However, there is no clearly defined methodology.
- Each case is unique.
- Methods as diverse as calibration, small area estimation, statistical matching and massive imputation can be used.
- Everything obviously depends on the data available and the existence of identifiers enabling matching.
- The use of data from the Internet will probably remain very anecdotal.
- However, new sources could be very valuable.
- In spatial statistics, for example, we now have aerial or satellite photos, sometimes of very high quality.
- This is reliable information that can be extrapolated to a population of interest.
- Making the most of this data will require statistical institutes to master new image analysis technologies.

Challenges 3: Fight against errors.

- Fight against errors.
- Correction, reweighting and imputation techniques are used to resolve problems of non-response.
- Calibration techniques can be used to reduce variance.
- Statisticians know how to deal with measurement errors, but each data set is unique.
- So there is no universal recipe in this area.
- In addition, new statistical methods of prediction need to be tested and integrated.
- They are likely to bring considerable improvements.

Challenges 4: Technological innovation.

- Technological innovation.
- Statistical methods are developing very rapidly.
- To cope with these innovations, the main problem is to recruit a skilled, high-quality workforce.
- I would therefore argue in favour of developing teaching and research in statistics.
- Scientific research is an approach that will make it possible to tackle questions that will arise in the future and that we do not yet know.
- It is therefore important to train students so that they are capable of continuing to learn, rather than focusing teaching on immediate skills.
- Like any business, statistical institutes need to develop research and scientific monitoring activities.

Outsourcing your core business means losing your soul.

Thank you.

References

Beaumont, J., Bosa, K., Brennan, A., Charlebois, J., and Chu, K. (2023). Handling nonprobability samples through inverse probability weighting with an application to statistics canada's crowdsourcing data. *Survey Methodology (accepted in 2023 and expected to appear in 2024)*.

Becker, H. (1952). Science, culture, and society. *Philosophy of Science*, 19(4):273–287.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brewer, K. R. W. (2013). Three controversies in the history of survey sampling. *Survey Methodology*, 39(2):249–262.

Chen, S., Woodruff, A. M., Campbell, J., Vesely, S., Xu, Z., and Snider, C. (2023). Combining probability and nonprobability samples by using multivariate mass imputation approaches with application to biomedical research. *Stats*, 6(2):617–625.

Dagdoug, M., Goga, C., and Haziza, D. (2020a). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. arXiv 2007.06298.

Dagdoug, M., Goga, C., and Haziza, D. (2020b). Model-assisted estimation through random forests in finite population sampling. arXiv 2002.09736.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jauslin, R. and Tillé, Y. (2023). An efficient approach for statistical matching of survey data through calibration, optimal transport and balanced sampling. *Journal of Statistical Planning and Inference*, 225:121–131.

Jensen, A. (1926). Report on the representative method in statistics. *Bulletin of the International Statistical Institute*, 22:359–380.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217.

Kiær, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9:176–183.

Kiær, A. N. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11:180–185.

Kiær, A. N. (1903). Sur les méthodes représentatives ou typologiques. *Bulletin de l'Institut International de Statistique*, 13:66–78.

Kiær, A. N. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14:119–134.

Kim, J. K. (2022). A gentle introduction to data integration in survey sampling. *The Survey Statistician*, 85:19—29.

Kim, J.-K. and Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2):382–401.

Nedyalkova, D. and Tillé, Y. (2008). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95:521–537.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Quetelet, A. (1846). *Lettres à S. A. R. le Duc régnant de Saxe-Cobourg et Gotha sur la théorie des probabilités appliquées aux sciences morales et politiques*. M. Hayez, Bruxelles.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tillé, Y. and Wilhelm, M. (2017). Probability sampling designs: Balancing and principles for choice of design. *Statistical Science*, 32(2):176–189.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Wang, Z., Yang, S., and Kim, J. K. (2023). Multiple bias-calibration for adjusting selection bias of non-probability samples using data integration. *arXiv preprint arXiv:2307.11651*.

Wikipedia (2024). Machine learning — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Machine_learning. [Online; accessed 30-May-2024].

Wilkinson, L. (2005). *The Grammar of Graphics (Statistics and Computing)*. Springer-Verlag, Berlin, Heidelberg.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2):283–311.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3:625–650.